UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICS
AND STATISTICS

REPORT 164

UNIVERSITÄT JYVÄSKYLÄ
INSTITUT FÜR MATHEMATIK
UND STATISTIK

BERICHT 164

STATISTICAL MODELLING OF SELECTIVE NON-PARTICIPATION IN HEALTH EXAMINATION SURVEYS

JUHO KOPRA

JYVÄSKYLÄ
2018

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICS
AND STATISTICS

REPORT 164

UNIVERSITÄT JYVÄSKYLÄ
INSTITUT FÜR MATHEMATIK
UND STATISTIK

BERICHT 164

STATISTICAL MODELLING OF SELECTIVE NON-PARTICIPATION IN HEALTH EXAMINATION SURVEYS

JUHO KOPRA

To be presented, with permission of the Faculty of Mathematics and Science
of the University of Jyväskylä, for public criticism in Auditorium H320
on February 24th, 2018, at 12 o'clock noon.

JYVÄSKYLÄ
2018

Editor: Pekka Koskela
Department of Mathematics and Statistics
P.O. Box 35 (MaD)
FI–40014 University of Jyväskylä
Finland

ISBN 978-951-39-7351-3 (print)
ISBN 978-951-39-7352-0 (pdf)
ISSN 1457-8905

Copyright © 2018, Juho Kopra
and University of Jyväskylä

University Printing House
Jyväskylä 2018

Abstract

Health examination surveys aim to collect reliable information on the health and risk factors of a population of interest. Missing data occur when some invitees do not participate the survey. If non-participation is associated with the variables to be studied, then the estimates based only on the participants cannot be generalised to the population of interest. In this case, the estimates have selection bias, which misleads the decision-makers.

The purpose of this thesis is to develop statistical methods to reduce the selection bias in the cross-sectional data using additional data sources. The data, which we use, comes from the National FINRISK Study, and we aim to estimate the prevalences of self-reported daily smoking and self-reported heavy alcohol consumption. The sources of additional information are follow-up data consisting of hospitalisations and causes of deaths, and questionnaire data collected from the non-participants of health examination by contacting them again, called re-contact data. Follow-up data give indirect information after the follow-up period about the health behaviour of non-participants during the health examination while the re-contact data give information similar to the health examination survey. This thesis presents methods for utilising these sources of additional information. Multiple imputation has been applied for the use of re-contact data, and Bayesian statistical modelling has been implemented for the use of follow-up data.

The thesis demonstrates that the use of additional data sources and these statistical methods leads to prevalence estimates for daily smoking and heavy alcohol consumption that are higher than those obtained from the participants only. Multiple imputation can be utilised for prevalence estimation if the re-contact data are available. Bayesian modelling is appropriate for the situation where re-contact data are not available but the follow-up data are and have follow-up period long enough to indicate about the differences between the participants and non-participants.

This thesis presents means for reducing the selection bias caused by non-participation. It is important to reduce the magnitude of the bias for obtaining more reliable information for example to support decision making. The statistical methods used in this thesis can also be applied to other fields of research than in the health studies.

Tiivistelmä

Terveystarkastustutkimusten tavoitteena on kerätä luotettavaa tietoa kohdepopulaation terveydentilasta ja riskitekijöistä. Terveystarkastustutkimuksissa puuttuvaa tietoa syntyy, kun osa tutkimukseen kutsutuista ei osallistu tutkimukseen, jolloin puhutaan poisjääneistä ja (osallistuja-) kadosta. Mikäli poisjäänti on yhteydessä tutkittaviin terveydellisiin tekijöihin, niin tutkimuksen osallistujilta lasketut tulokset eivät ole yleistettävissä alkuperäiseen kohdepopulaatioon. Tällöin sanotaan, että osallistujien estimaateissa on valikoitumisharhaa, joka vaikeuttaa päätöksentekoa.

Tämän väitöstutkimuksen tavoitteena on kehittää menetelmiä, joiden avulla voidaan pienentää valikoitumisharhaa poikkileikkausaineistossa lisätietoaineistoja käyttämällä. Käytössä on aineisto kansallisesta FINRISKI-tutkimuksesta ja kiinnostuksen kohteena on itseraportoidun päivittäisen tupakoinnin ja alkoholin suurkulutuksen vallitsevuus eli prevalenssi. Lisätiedon lähteinä ovat seuranta-aineistona sairaalakäynti- ja kuolinsyyaineistot sekä nk. uudelleenyhteydenottoaineisto, joka on kerätty terveystarkastuksesta poisjääneiltä ottamalla uudelleen yhteyttä terveystarkastuksen jälkeen. Seuranta-aineiston avulla voidaan seuranta-ajan jälkeen saada epäsuoraa informaatiota kohdepopulaation terveydentilasta ja elintavoista tutkimushetkellä, mutta uuden yhteydenoton kautta saadaan vastaavaa tietoa kuin varsinaisessa terveystarkastustutkimuksessa pian kyselyn jälkeen. Tämä väitöskirja esittää menetelmiä kumpaankin tilanteeseen. Uudelleenyhteydenottoaineistoa käytettäessä on sovellettu moni-imputointia ja seuranta-aineistoa käytettäessä bayeslaista tilastollista mallintamista.

Saaduista tutkimustuloksista nähdään, että lisätietoaineistoja ja käytettyjä tilastomenetelmiä hyödyntämällä saadaan korkeammat vallitsevuusestimaatit päivittäiselle tupakoinnille ja alkoholin suurkulutukselle kuin perustuen pelkästään osallistujilta saatuun aineistoon. Moni-imputointia voidaan käyttää apuna vallitsevuuden harhan pienentämisessä, mikäli uudelleenyhteydenotto on toteutettu. Bayeslainen mallintaminen soveltuu tilanteeseen, jossa uudelleenyhteydenottoaineistoa ei ole saatavilla, mutta seuranta-aineisto on ja seuranta-aika on tarpeeksi pitkä, jotta seuranta-aineistosta saadaan tietoa osallistujien ja poisjääneiden terveydentilasta ja terveyskäyttäytymisestä.

Tämä väitöskirja tarjoaa keinoja valikoituneen poisjäännin aiheuttaman harhan pienentämiseen. Harhan suuruuden pienentäminen on tärkeää luotettavamman tiedon saamiseksi esimerkiksi päätöksenteon tueksi. Työssä käytettyjä tilastomenetelmiä voidaan soveltaa myös muilla tieteenaloilla kuin terveystieteissä.

Acknowledgements

First, I am most grateful to my primary supervisor professor Juha Karvanen for giving me this opportunity to pursue a doctorship and to learn more about statistics.

Second, I would like to express my sincere gratitude to my secondary supervisor Dr Tommi Härkänen, whose supervising has been essential for the success of this thesis. His enthusiasm towards technical details of my work and particularly my programming codes has been important. I thank adjunct professor Hanna Tolonen for coordinating the Non-participation in Health Examination Studies -project (NoPaHES) and organising my visit as visiting researcher to the National Institute for Health and Welfare at Helsinki in December 2014. Hanna's epidemiological contribution to the articles of this thesis has been substantial. I also want to thank professor emeritus Antti Penttinen, who inspired me to study Bayesian methods. Antti has been my third supervisor.

I also want to express my gratitude to other co-authors Pekka Jousilahti, Kari Kuulasmaa, Pia Mäkelä and Jaakko Reinikainen for their work. It has been a pleasure to collaborate with you. I thank The Finnish Foundation for Alcohol Studies about their monetary support towards my thesis work. I thank The Academy of Finland for funding the NoPaHES -project, which has been a major source of funding for my thesis. I also appreciate European Survey Research Association's travel bursary for giving me an opportunity to participate in ESRA 2017 conference in Lisboa, Portugal. I thank The Department of Mathematics and Statistics for supporting my work and participation in the conferences in Turku (two times) and Helsinki. The staff of statistics and stochastics has helped me a lot by discussing and advising on teaching and administrative matters.

Lastly, I give my most profound gratitude to the love of my life, my wife Olga, and to my dearest parents who have given me all the care and support they can. Without you, this thesis would have never seen the daylight.

Kuopio, December 2017

Juho Kopra

List of original publications

I Kopra, J., Härkänen, T., Tolonen, H., Jousilahti, P., Kuulasmaa, K., Reinikainen, J. and Karvanen, J. (2017). Adjusting for selective non-participation with re-contact data in the FINRISK 2012 survey. Published online first in *Scandinavian Journal of Public Health*. doi: 10.1177/1403494817734774.

II Kopra J., Härkänen T., Tolonen H. and Karvanen J. (2015). Correcting for non-ignorable missingness in smoking trends. *Stat*, 4(1):1–14. A correction has been published as Kopra J., Härkänen T., Tolonen H. and Karvanen J. (2017). Correction: Correcting for non-ignorable missingness in smoking trends. *Stat*, 6(1):202–203.

III Kopra J., Karvanen J. and Härkänen T. (2017). Bayesian models for data missing not at random in health examination surveys. Published online first in *Statistical Modelling*. doi: 10.1177/1471082X17722605.

IV Kopra, J., Mäkelä, P., Tolonen H., Jousilahti, P. and Karvanen, J. (2018). Follow-up data improve the estimation of the prevalence of heavy alcohol consumption. Submitted.

The author of this thesis is the main contributor and writer of Articles I, II, III and IV. He has implemented the programming code needed for statistical analyses for papers II, III and IV, and has modified an earlier code for the statistical analysis of Article I. The co-authors have contributed to the planning of the research and writing of the articles.

Contents

Chapter 1

Introduction

Information about the health of a population is needed for decision-making purposes regarding health policy (Tolonen, 2013, p. 5). The World Health Organization (2015) lists various mortality rates, e.g. age-specific and disease-specific mortality rates, various risk factor indicators, e.g. total alcohol consumption and tobacco use, and indicators about coverage and functionality of health service as the key indicators of population health. Many of these health indicators are prevalences. The prevalence describes how widespread a disease or a risk factor is in the population (Rothman, 2012, p. 53). Prevalences are usually estimated using data from survey studies or register-based population studies (Rothman, 2012).

Health examination surveys (HESs) are population based surveys, which collect data using questionnaires and physical measurements (Tolonen, 2013). Physical measurements include measurements of weight, height, blood pressure etc. and collection of many biological samples. The HES data are useful in health policy decision-making as well as in planning and evaluation of prevention programs (Tolonen, 2013).

In an ideal situation, surveys provide diverse and reliable information about the *target population* (Thompson, 1997). In the best-case scenario, a random sample is drawn, the probability of belonging to the sample is known for each individual, and all the individuals of the sample are studied carefully. The results of this kind of survey can be generalised to the entire population of interest. This kind of sample is said to be *representative* with respect to target population.

Often, we cannot obtain data from all invitees, but only from persons who accept to participate. Persons who attend the physical health examination and reply to the questionnaire are called participants. The survey measurements are then collected from participants but cannot be collected from non-participants. Thus, the survey suffers from non-participation, which leads to missing data.

Selective non-participation refers to a situation, where non-participation (and participation) to the survey is associated with some variables measured in the survey.[1] This makes participants to overrepresent some of the population subgroups compared to the population of interest. Participation may be selective with respect to variables which are known for all persons invited to a survey, or with respect to variables which are to be measured in the survey. The former can be solved much easier than the latter one. If non-participation is present, a sample is representative with respect to the target population if non-participation

[1] Some studies use the term *selective non-response* (or *selective nonresponse*), and sometimes it is unclear whether the terms refer to decisive denial based on the topic of the study, or more generally as we see the term.

is not selective or if the non-participation can be adjusted with respect to the background variables and the non-participation is not associated with the variable of interest.

Not only can non-participation increase uncertainty of the estimates (or decreases precision) because of the smaller sample size but it also causes systematic error called *non-participation bias* or *selection bias* to the estimates. We say that, an estimator is unbiased if its expected value is the same as the true value in the population. The higher the non-participation, the more concerns it raises, creating larger potential bias (Nishimura et al., 2016).

Let us demonstrate how association between participation and the smoking may bias the results, see Figure 1.1. For the sake of example, let us assume that the true smoking prevalence in the population is known (usually it is not), and that is $0.2, 0.3, 0.4, 0.5,$ and 0.6 for five consecutive studies, respectively (black dashed line with filled circles). Let us also assume that the probability of participation, often called participation rate, is decreasing more rapidly for smokers than for non-smokers. For smokers, let the participation probability be $0.80, 0.65, 0.50, 0.35$ and 0.20 for each study from the first to the fifth (red line with triangles). For non-smokers, let the participation probability be $0.80, 0.75, 0.70, 0.65$ and 0.60 for each study from the first to the fifth (blue line with diamonds).

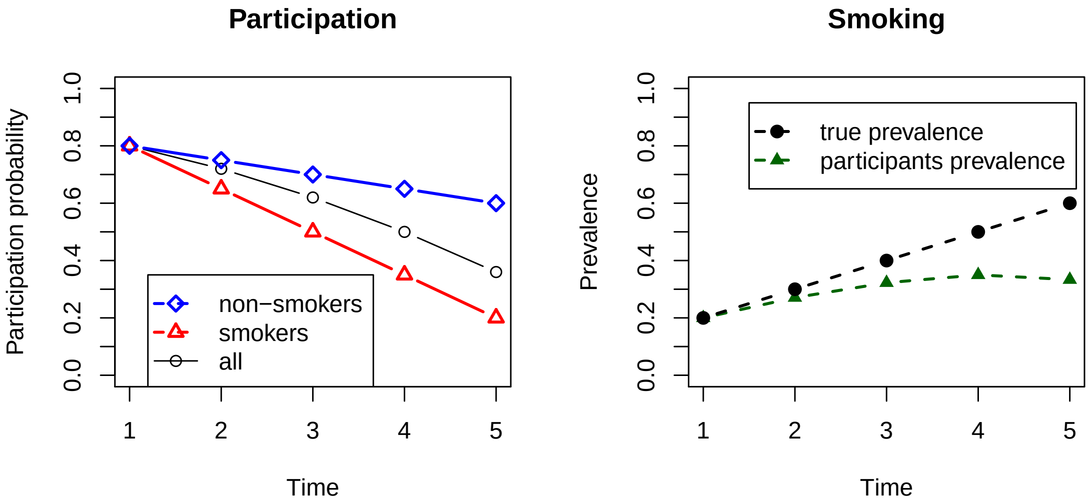

Figure 1.1: Artificial demonstration about how selective non-participation affects the estimation of smoking prevalence.

On the right panel of the Figure 1.1 we can see that the estimated smoking prevalence (dark green dashed line with filled triangles) is lower in the most time points than the true prevalence (black dashed line with filled circles). For the first study year, the estimation is unbiased because the participation probabilities for smokers and non-smokers were the same (left hand side). As the participation probabilities become more different between smokers and non-smokers, the participants' estimate becomes more distant to the true prevalence in the population. As the participation probabilities between the two groups are more different during the later years than in the beginning, there is also bias in the slope of smoking prevalence trends. Between the fourth and fifth time point, the prevalence estimated from participants appears to decline although the true trend increases. If the true trend had been decreasing, the estimated prevalence would also have been lower in that case, because of the selectivity mechanism.

This thesis utilises data from the National FINRISK Study (Borodulin et al., 2017), which is a series of cross-sectional surveys conducted by the National Institute for Health and Welfare. The FINRISK Study consist of consecutive surveys conducted once every five years in Finland during 1972–2012. The participation rate in these surveys has decreased noticeably. In 1972 the participation rate in the health examination was about 90% (Harald et al., 2007), while in 2012 it was only about 60% (Borodulin et al., 2013).

Previous studies have suggested that the FINRISK data are selective with respect to smoking and alcohol usage (Jousilahti et al., 2005; Tolonen et al., 2005; Karvanen et al., 2016). This creates a potential bias to the population estimates. These kinds of biased estimates cannot be generalized to the entire population. The biased results may critically misinform the decision-makers.

In this thesis, we develop statistical approaches capable of taking into account selective non-participation. We demonstrate that these approaches can reduce selection bias. We also compare the developed approaches with other commonly used methods, and apply the developed approaches for the FINRISK data to estimate the prevalences of daily smoking and heavy alcohol consumption.

In Article I, a multiple imputation (MI) approach is presented as a potential solution for the missing data problem. The paper utilises additional survey data called re-contact data collected among non-participants. The prevalence estimates for smoking and alcohol use are obtained. Articles II and III develop a different solution when re-contact data are not available, but follow-up data are. These two papers estimate the prevalence of daily smoking. Article IV applies the method of Article III to the heavy alcohol consumption. The obtained prevalence estimates in Article I–IV are higher than the estimates based on participants. Earlier results are based on participants only.

In Chapter 2, the data used in this thesis are presented. Chapter 3 describes the missing data problem. Chapters 4 and 5 describe the approaches of Article I and Articles II–IV, respectively. The Chapter 6 discusses the results and the implications of the work done in this thesis. Appendix A lists errata of included articles.

Chapter 2

Data from the National FINRISK Study

2.1 Survey data

The National FINRISK Study is a series of cross-sectional surveys or, more precisely, HESs conducted in Finland once in every five years (Borodulin et al., 2017). The purpose of the study is to investigate the health of adult population (25–74 -year-olds) of Finland. The first FINRISK Study survey took place in 1972 in North Karelia and Northern Savonia (Puska et al., 1973). At that time the study was called the North Karelia Project (Pohjois-Karjala -projekti in Finnish) (Puska et al., 1973). Later, new regions were added to the study (see Article II), and in 2012 the regions were North Karelia, Northern Savonia, Turku and Loimaa area, Helsinki and Vantaa area, and Oulu region. In 2007 Lapland was also part of the study (Peltonen et al., 2008). Recently, the name FINRISK has changed to FinTerveys and the design of the study has also changed. In this thesis, the data utilised are from the surveys conducted in 1972–2012 (Puska et al., 1973; Vartiainen et al., 1993; Korhonen et al., 1999; Laatikainen et al., 2003; Peltonen et al., 2008; Borodulin et al., 2013, 2017).

In The FINRISK Study, the participation rate has decreased during the period 1972–2012 (Harald et al., 2007; Borodulin et al., 2013). The participation variable can be defined as (1.) participation in the health examination, (2.) as a returning of a questionnaire, or (3.) as answering to the question regarding variable of interest. In Article I the first and the second definition are used, Articles II–IV use the third one. In 1972 the participation rate (1.) to a health examination was 88% while in 2012 it was only 58%, see Table 2.1. Similar phenomenon has also been observed in other countries (Galea and Tracy, 2007). Low participation potentially increases non-participation bias (Nishimura et al., 2016), which has also been briefly demonstrated in the example of Figure 1.1.

The National FINRISK Study data had in total $98,050$ invitees during 1972–2012, out of which $72,340$ persons participated in health examination, and $1,854$ persons returned only a study questionnaire. The yearly counts of invitees and participants are reported in Table 2.1. The number of invitees available in this thesis is slightly different to what is reported in Borodulin et al. (2017), because some parts of the FINRISK data were not available for us. The re-contact questionnaire data was collected in 2002, 2007 and 2012 in all areas. In 2002 and 2007 the questionnaire data without a health examination was collected in Lapland

(Peltonen et al., 2008).

Table 2.1: The number of invitees and participants of the FINRISK data. The questionnaire study of Lapland in 2007 ($n = 1260$) is not reported in the table.

	1972	1977	1982	1987	1992	1997	2002	2007	2012	All years
Invited	12,440	11,359	11,395	7,931	7,927	11,500	13,498	12,000	10,000	98,050
Participants	10,938	10,197	9,347	6,478	6,051	8,446	8,798	6,258	5,827	72,340
Non-participants	1,502	1,162	2,048	1,453	1,876	3,054	3,918	4,007	3,576	22,596
Only questionnaire	0	0	0	0	0	0	782	475	597	1,854

The survey design has varied over the years (Borodulin et al., 2017). The 1972 survey followed systematic sampling with respect to birth date and stratified sampling regarding area of recidence. In 1977 the survey applied simple random sampling stratified between areas. The third survey in 1982 utilised a stratified sampling between 10-years age-groups within the areas, and since 1987 the sampling design has been stratified sampling between 10-years age-groups within gender and areas.

The health examination of the FINRISK Study consists of for instance measurements of height and weight, systolic and diastolic blood pressure, pulse, circumference of waist and hip, and blood samples. Information of sample members' sociodemographic factors, the use of health services, diseases and symptoms, health behaviour (alcohol and tobacco use), nutrition habits and psychosocial factors are collected by questionnaire (Peltonen et al., 2008).

In 2002–2012 the FINRISK Study conducted a survey among non-participants. We call these surveys as re-contact surveys, as the non-participants of the initial survey has been contacted again, and subsample data of non-participants have been collected. For these kinds of surveys, there are three types of the invitees: *participants*, who participated initial survey, *re-contact respondents*, who where non-participants of the initial survey but returned the re-contact questionnaire, and *non-participants* who participated in the neither of the rounds.

2.2 Register data and record linkage

The survey data have been linked to multiple register data sets using the personal identification code. In Finland, the data obtained via register linkage is available for both participants and non-participants. The registers linked are Care Register for Health Care (HILMO) (National Institute for Health and Welfare, 2017), Cause of Death Register (Statistics Finland, 2014), and The Register of Completed Education and Degrees (Statistics Finland, 2017). The follow-up data is obtained from Care Register and Cause of Death Register. The HILMO register holds data about each person's hospital visits or hospitalisations with diagnoses (ICD-codes). The HILMO register holds hospitals visits since 1969. These additional data in this thesis contain information both on participants and non-participants, which is not possible in many countries.

2.3 Sources of auxiliary information in the FINRISK Study surveys

There are three sources of data for investigating the selectivity mechanism in this thesis. The first source is the data from survey frame, which typically hold demographic information, e.g. gender, age and location of living. We call these data as background data. Background data allow to investigate if participation is selective with respect to the background data, but the participation may still be selective with respect to the rest of the variables. Background data may be used to evaluate the differences between participants and non-participants with respect to age, for example, if it is available.

The second source is survey data consisting of data from the actual surveys and the surveys among non-participants. The variables of the survey data are called *survey variables* including the variables of interest. The variables of interest are daily smoking and heavy alcohol consumption, and we call them as *risk factors* as they increase the risk of diagnosed diseases. Survey data allow the comparison of survey participants and non-participants to some extent, but as only a sample of non-participants is available, the question if the re-contacts represent the non-participants of the actual survey remains. Only if we assume that the participants of the re-contact survey (aka. re-contact respondents) are a random sample of the non-participants, then we can compare participants' and re-contacts' estimates, and therefore evaluate the average health of these groups.

The third source is register-based follow-up data, which may be very useful in estimating the selectivity mechanism. For the FINRISK data, register data about disease diagnoses are available and can be used to evaluate the selectivity mechanism. Information about the relationship of the risk factors and disease diagnosis is central because the risk factors cannot be observed from the registers. Data about the disease diagnoses can be used to indicate whether there is a difference in the risk factors between participants and non-participants.

2.4 Notation

Let us define the notation for the data variables. Let M, T, X and Y stand for the participation indicator, disease outcome, background variables and risk factors, respectively. Let M take value $M = 1$ for participants and $M = 0$ for non-participants. The variables T stand for the age at the time of the first disease diagnosis as well as censoring status. The variables X consist of age at the time of a survey (baseline), gender and location of living. Regarding the work done in Article IV, the X also contains education. Notation $V = (X, Y, T)$ is used to refer to all data.

2.5 Differences between participants and non-participants

In the FINRISK Study, the differences in health between participants and non-participants have been observed. First, it is known from the survey frame that the non-participants are more often men than women (Jousilahti et al., 2005; Harald et al., 2007), and are younger persons than the participants (Peltonen et al., 2008; Borodulin et al., 2013). Second, surveys among non-participants have shown that the non-participants are more often smokers (Tolonen et al., 2005; Karvanen et al., 2016) and heavy alcohol consumers (Karvanen et al., 2016), have lower socioeconomic status than the participants (Harald et al., 2007), and are more often single than married (Tolonen et al., 2005). Third, studies utilising register-based data found out that non-participants have a higher total risk of death (Harald et al., 2007), and also a higher risk of smoking-related and alcohol-related death (Jousilahti et al., 2005).

Chapter 3

Missing data

Missing data in a survey are described by the terms item non-response or unit non-response (Little and Rubin, 2002, p. 6). In the unit non-response, a person does not participate in a study at all, and in the item non-response a person participates but does not provide an answer to a particular question. Both of these cause missing data which need to be taken into account in the analysis. The unit non-response yields missingness to both questionnaire based and physical measurement based data, while the item non-response makes only the values of particular variable(s) missing.

3.1 Types of missingness

Modelling of missing data (Little and Rubin, 2002) is guided by the nature of missingness. Traditionally, missing data have been classified into three categories, which are called missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Rubin, 1976). Seaman et al. (2013) presented a five class classification of missingness distinguishing between realized missingness and everywhere missingness. The classes are realized MCAR, everywhere MCAR, realized MAR, everywhere MAR and MNAR. We present types of missingness according to this classification.

Let us now denote the observation vector as V, and missingness vector as M. The missingness vector M indicates which values of V are observed and which missing. Let $\tilde{v}$ be the realized observation vector which may also contain missing values, and let $\tilde{m}$ be the realized missingness vector. Now, let $g_{\psi}(M = m|V = v)$ be a model for the missingness with parameters ψ. Let v and v^* be two values of the random variable V. Now, data are *realized MCAR* if for all ψ

$$g_{\psi}(\tilde{m}|v) = g_{\psi}(\tilde{m}|v^*) \text{ for all } v \text{ and } v^*. \tag{3.1}$$

Data are *everywhere MCAR* if for all ψ

$$g_{\psi}(m|v) = g_{\psi}(m|v^*) \text{ for all } m, v \text{ and } v^*. \tag{3.2}$$

In Example 1.1, the assumption of realized MCAR would mean that on the condition of the realized pattern of missingness the participation does not depend on smoking or other variables. The everywhere MCAR would mean the same for any pattern of missingness.

For the sake of the next two classes let $o(V, M)$ be observed data (random variable) and let $o(\tilde{v}, \tilde{m})$ be realized value of $o(V, M)$. The operator $o(\cdot, \cdot)$ picks the values of the first argument, which are observed according to the second argument. Now, data are *realized MAR* if for all ψ

$$g_\psi(\tilde{m}|v) = g_\psi(\tilde{m}|\tilde{v}) \text{ for all } v \text{ where } o(v,\tilde{m}) = o(\tilde{v},\tilde{m}). \tag{3.3}$$

Again, let v and v^* be two values of the random variable V. Now, data are *everywhere MAR* if for all ψ

$$g_\psi(m|v) = g_\psi(m|v^*) \text{ for all } m, v \text{ and } v^* \text{ where } o(v,m) = o(v^*,m). \tag{3.4}$$

The implications between the types of missingness are represented in Figure 3.1. The classes realized MCAR and realized MAR operate only on fixed (observed) missingness vector $\tilde{m}$, and classes everywhere MCAR and everywhere MAR operate on arbitrary missingness vector m. Realized MAR assumption is not as restrictive as realized MCAR and everywhere MCAR as it allows missingness to depend on the observed data but not on the missing data for fixed missingness vector $\tilde{m}$. In Example 1.1, data with MAR could have different participation probabilities for men and women or for different study years, but the participation cannot depend on missing data on smoking. Everywhere MAR assumption is similar, but it allows missingness vector m to be arbitrary.

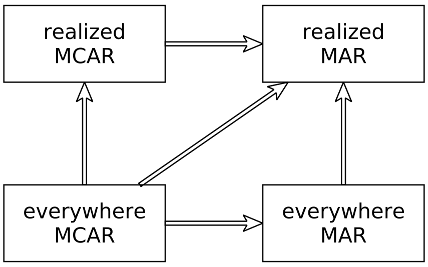

Figure 3.1: Implications between the definitions of missingness types in Equations (3.1)-(3.4).

If data do not belong to any of the previous four classes, data are MNAR. The MNAR data are difficult to analyse, because the missing data mechanism cannot be estimated without information on the actual values which are missing. In some cases this information can be obtained based on auxiliary data or previous measurements. Example 1.1 has MNAR data because participation is selective with respect to the variable of interest (smoking) which is missing for non-participants.

If the missing data mechanism is MNAR we say that it is *non-ignorable*, and otherwise the mechanism is said to be *ignorable*. The concepts ignorable or non-ignorable refer to whether or not the mechanism can be ignored in the likelihood of the model (Little and Rubin, 2002).

3.2 Statistical methods for handling of missing data

Statistical methods for analysing data with missing values can be divided into four non-exclusive classes: complete case methods, weighting methods, imputation methods, and model-based methods (Little and Rubin, 2002, p. 19).

The complete case analysis uses only the observations with all variables recorded. The observations with one or more missing values are excluded from the analysis. Complete case analysis is easy to use but it may result in seriously biased results if MCAR assumption does not hold (Little and Rubin, 2002, p. 41). In the Example 1.1. the data are MNAR but analysis utilises the MCAR assumption. The estimated smoking prevalence is biased in comparison to the true population prevalence.

A common analysis approach in survey sampling is to utilise inverse proportional weighting (IPW), where the weights are inverses of the sampling probabilities (Thompson, 1997). This is particularly used in stratified sampling where individuals belonging to different groups have different sampling probabilities. The weighting approach called post-stratification is often applied to survey data with non-participation (Brick and Kalton, 1996). If data are missing due to design, then the weights are known and otherwise they must be estimated from the data (Brick and Kalton, 1996; Little and Rubin, 2002, Chapter 3). For post-strafication, the missing data are due to non-participation, not due to design. Thus, the weights need to be estimated. Another method is propensity-score weighting (Rosenbaum and Rubin, 1983), which is a weighting based method for the situation where many variables explain missingness (Lunceford and Davidian, 2004). These approaches utilise implicitly the MAR assumption. The results are biased if data are MNAR (Brick and Kalton, 1996).

The multiple imputation (MI) method (Rubin, 1987) replaces the missing values with generated values. The artificial values may be obtained using other observations either from the same data set (hot-deck imputation) (Andridge and Little, 2010), different data set (cold-deck imputation), or be model-based.

The model-based methods utilize a model for observed data, which can be described using the likelihood function. First, a model for the data are built, then one must integrate with respect to the missing data. Bayesian modelling treats the model parameters and missing data the same way by integrating over them. For a frequentist analysis the EM algorithm (Dempster et al., 1977) or numeric integration is used. Both Bayesian modelling and model-based MI generate multiple artificial complete data sets. Calculating an average over an artificial data set gives an estimate of the prevalence of a risk factor. These estimates are combined into the final estimate.

3.3 Approaches for the MNAR problem

In general, the estimation for MNAR problems requires some additional information or assumption(s). If the data are longitudinal, then missing data can be predicted based on a parametric model and other measurements on the same individual, even when data are MNAR (Little, 1995; Hedeker and Gibbons, 1997; Kenward and Molenberghs, 1999; Siddique and Belin, 2008). Additional information may origin from auxiliary data or informative priors (if utilising Bayesian methods). For a cross-sectional survey data, such as the FINRISK data, suitable sources of auxiliary information are surveys conducted among non-participants (re-contact survey data) and data obtained through registers. If the assumptions provide the use of additional information, those can be made regarding the parameters or the structure of the model, or functional form of the model. Sometimes assumptions are implemented utilising prior distributions (Bayesian methods). When it comes to the FINRISK data, we are convinced that the smoking and alcohol consumption habits could affect the participation, and that is why MNAR assumption cannot be excluded in advance (see Section 2.5).

Chapter 4

Multiple imputation

4.1 Overview

Single value imputation does not take the uncertainty (variation) into account, leading to underestimation of the variance of the estimator. This weakness is avoided in multiple imputation (MI) (Little and Rubin, 2002; van Buuren, 2012), which is a three-step procedure:

1. **Imputation:** Generate values to fill in missing data multiple times, e.g. $D = 5$ times. This step results in D complete data sets.

2. **Analysis:** Analyze each data set resulting D analysis results.

3. **Pooling:** Pool the analysis results to a single final result.

The idea of multiple imputation is based on Bayes' theorem (Little and Rubin, 2002), see Chapter 5. First, let data V consist of its observed data V_{obs} and missing data V_{mis}, $V = (V_{\text{obs}}, V_{\text{mis}})$, where $V_{\text{obs}} = o(V, M)$ with respect to notation used in Chapter 3. The aim of statistical analysis is to estimate a quantity θ, which has a posterior distribution $p(\theta|V_{\text{obs}})$ given the observed data. This can be written as

$$p(\theta|V_{\text{obs}}) = \int p(\theta|V_{\text{obs}}, V_{\text{mis}})p(V_{\text{mis}}|V_{\text{obs}})dV_{\text{mis}}, \tag{4.1}$$

where $p(\theta|V_{\text{obs}}, V_{\text{mis}})$ is the full data posterior, and predictive distribution $p(V_{\text{mis}}|V_{\text{obs}})$ is the conditional distribution of missing data given the observed data. Multiple imputation approximates the integral of Equation (4.1) by generating independent imputations $V_{\text{mis}}^{(d)} \sim p(V_{\text{mis}}|V_{\text{obs}})$ and then calculating the average over the imputations, which approximates the posterior

$$p(\theta|V_{\text{obs}}) \approx \frac{1}{D}\sum_{d=1}^{D} p(\theta|V_{\text{mis}}^{(d)}, V_{\text{obs}}). \tag{4.2}$$

The mean of θ can be approximated as

$$\overline{\theta} = E[\theta|V_{\text{obs}}] \approx \int \theta \frac{1}{D}\sum_{d=1}^{D} p(\theta|V_{\text{mis}}^{(d)}, V_{\text{obs}})d\theta \quad = \frac{1}{D}\sum_{d=1}^{D} E(\theta|V_{\text{mis}}^{(d)}, V_{\text{obs}}) = \frac{1}{D}\sum_{d=1}^{D} \hat{\theta}_d, \tag{4.3}$$

where $\hat{\theta}_d$ is the full data estimate for the dth imputed data set $(V_{\text{mis}}^{(d)}, V_{\text{obs}})$. Now, the variance can be estimated by

$$\begin{aligned} \text{Var}(\theta|V_{\text{obs}}) &= E[\text{Var}(\theta|V_{\text{mis}}, V_{\text{obs}})|V_{\text{obs}}] + \text{Var}[E(\theta|V_{\text{mis}}, V_{\text{obs}})|V_{\text{obs}}] && (4.4) \\ &\approx \frac{1}{D}\sum_{d=1}^{D}\text{Var}(\theta|V_{\text{mis}}^{(d)}, V_{\text{obs}}) + \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\theta}_d - \overline{\theta})^2, && (4.5) \end{aligned}$$

where $\text{Var}(\theta|V_{\text{mis}}^{(d)}, V_{\text{obs}})$ is the variance of θ calculated from imputed data set, and $\hat{\theta}_d$ and $\overline{\theta}$ were defined in Equation (4.3). Thus, the complete case methods can be used for each imputed data set.

In practice, it is important how the multiple imputation model for the distribution $p(V_{\text{mis}}|V_{\text{obs}})$ is built and estimated. As imputations are drawn from the imputation model, a problem occurs if explanatory variables have missing values. Multiple Imputations by Chained Equations (MICE) is an approach which solves this problem (van Buuren and Groothuis-Oudshoorn, 2011; van Buuren, 2012). At first step, MICE-algorithm samples initial imputations from the data. Then it utilises for each variable a fully conditional specification, which is a regression model defined by the user. Imputations for all variables with missing data are then generated at each iteration of the algorithm, and preferably 50 iterations are needed (van Buuren, 2012). The MICE approach does not guarantee the imputations to converge to a valid joint distribution, but this problem does not occur when variables with missing data are always missing together (Molenberghs and Kenward, 2007, p. 114). This holds in situations where we apply MICE.

4.2 Application of multiple imputation to the FINRISK data with re-contact survey

Article I analyses the FINRISK 2012 data containing re-contact survey data from the non-participants of the health examination. Such data are collected in two phases. First, a HES survey data are collected. Second phase collects a questionnaire data from the non-participants of the first survey. In this context, persons who participated the HES survey are called participants, persons who returned the re-contact questionnaire are called re-contact respondents, and persons who did not respond neither initial survey nor re-contact questionnaire are called non-participants. The variables of interest are daily smoking and heavy alcohol consumption. These kind of data provide useful additional information for analysing survey data with MNAR missingness.

For these kind of data multiple imputation need to be tailored to match the nature of the data. In its basic form, multiple imputation utilises MAR assumption, but the missingness are MNAR with respect to participation to the health examination. If re-contact data are available, it is possible to assume that the re-contact respondents represent the non-participants and that is why the data are MAR. In Article I we assumed that the re-contact respondents represent all non-participants when adjusted for background variables. Applying this assumption with the MICE algorithm requires using a model, which has different parameters for the participants and re-contact respondents. Multiple imputations for imputed variables are carried out using a regression model with fully conditional specification (van Buuren, 2012) and the rest of the variables are covariates in the MI model. The adjustment for the background variables can be made by using them as covariates in the MI model. If follow-up data are available, it can be used to investigate the validity of our modelling assumption. For the FINRISK 2012, retrospective follow-up data were not available. Instead, similar prospective data about hospital visits was utilised as a surrogate. In this case an association between the imputed variables and the number of hospital visits are utilised.

In Articles I and II the weighting has been applied to the imputed data. Because of the lack of detailed documentation for some of the older surveys, the weighting for the FINRISK data in Article II was not straightforward. The weights were based on population data obtained from Statistics Finland. For years 1972–1982 only approximate weights are available because the population data were neither available for the precise time of the sampling, nor available for municipalities which do not exist anymore.

Chapter 5

Bayesian modelling

5.1 Overview

Bayesian modelling is a paradigm, which utilises Bayes' theorem

$$p(\theta|V) = \frac{p(V|\theta)p(\theta)}{\int_{\theta'} p(V|\theta')p(\theta')d\theta'} \tag{5.1}$$

to update the prior probability distribution $p(\theta)$ with information of the data V via likelihood $p(V|\theta)$. The left hand side of the Equation (5.1) is called posterior (probability) distribution.

The prior probability distribution, or just prior distribution, must not be based on data which are to be analysed. The prior distribution may be based on previous data or expert opinion (O'Hagan et al., 2006). The prior distribution can be informative or uninformative. An informative prior holds subjective information about the parameter. Uninformative prior or vague prior reflects the absence of available information.

The task of Bayesian inference is to compute the posterior distribution $p(\theta|V)$ of the parameters. To complete this task, the integral in the denominator in the Bayes' theorem need to be calculated. When there are a lot of parameters, the integral is often too difficult to compute, and many methods generate samples of the posterior distribution using simulation. Markov chain Monte Carlo and importance sampling are often used algorithms for generating posterior samples (Robert and Casella, 2004).

Bayesian modelling is particularly useful, for instance, when multiple data sources are utilised (Spiegelhalter and Best, 2003), modelling requires a hierarchical structure (Gelman et al., 2013, Chapter 15), or when parameter uncertainty is important (Robert, 2007).

In the context of missing data, Bayesian modelling is useful because it allows setting up a joint model for observed data, missing data and parameters. Bayesian inference does not distinguish between missing data and parameters, being a reason why missing data are naturally handled in the model fitting process. As a consequence any separate imputation step is not needed.

Data augmentation (DA) (Tanner and Wong, 1987) is a Bayesian method of simulating posterior of θ when data V are partly missing. In DA the imputations and parameters are simulated iteratively from its full conditional probability distribution at each step t. First, start with initial values $\theta^{(0)}$ and repeat following procedure:

$$\begin{aligned}
&\text{Generate } V_{\text{mis}}^{(t+1)} \sim p(V_{\text{mis}}|V_{\text{obs}}, \theta^{(t)})\\
&\text{Generate } \theta^{(t+1)} \sim p(\theta|V_{\text{obs}}, V_{\text{mis}}^{(t+1)})
\end{aligned}$$

until the chains have converged. The convergence can be evaluated using Brooks-Gelman-Rubin -diagnostics (Brooks and Gelman, 1998). The algorithm is a special case of Gibbs' sampler (Gelfand et al., 1990), which samples each parameter from its full conditional at each timestep.

5.2 Application of Bayesian modelling in the analysis of survey data with follow-up

There are situations when follow-up data are useful for prevalence estimation of a risk factor under MNAR missingness. In Articles II, III and IV we consider the situation where follow-up data are available for both non-participants and participants of a survey study, and follow-up data are informative about the risk factor of interest. We first describe what data are used in each article, and then tell about the modelling.

Articles II and III focus on the prevalence of daily smoking and Article IV on the prevalence of heavy alcohol consumption. Article II uses 1972–1997 surveys, utilises follow-up of smoking-related diseases data from the date of health examination to the end of 2011. Article III utilises 1972–2007 surveys, follow-up data of the same diseases as in Article II from the date of health examination to the end of 2012. Neither of the Articles II or III have data from the education register. Article IV estimates the prevalence of heavy alcohol consumption in 1987–2007 surveys, uses follow-up data of diseases related to heavy alcohol usage starting from the survey data to the end of 2014. This article uses data from the education register.

The data about diseases and deaths can be utilised to estimate the prevalence of daily smoking although the survey data are MNAR. This is possible because daily smoking is an important predictor of diseases such as lung cancer and chronic obstructive pulmonary disease (COPD) (Doll and Hill, 1956; Wynder and Hoffmann, 1994; Cornfield et al., 2009). Also, observed lung cancer or COPD event indicates that the person is a daily smoker with high probability. Thus, information about disease events and deaths can be utilised to fill in missing values of an associated risk factor. The same applies for heavy alcohol consumption if follow-up data for alcohol-related diseases are available.

In order to take into account the advantages provided by these two linked data sets, a Bayesian model, consisting of three sub-models, is needed. The sub-models are:

1. participation model, in which participation in a survey $M = 1$ is explained by the risk factor Y and background variables X from the survey frame.

2. risk factor model to explain the variability of a risk factor Y using background variables X.

3. survival model, which describes the relationship between the risk factor Y and diseases obtained from follow-up data T adjusted for background data X.

These three sub-models are then utilised together as a Bayesian model for the data.

Based on Bayes' theorem, the missing data of smoking can be drawn from the full conditional distribution for the non-participants

$$y \sim p(Y = y|M = 0, X, T) \propto P(Y = y|X)P(M = 0|Y = y, X)P(T|Y = y, X).$$

As all the missing values in Y are for non-participants who have $M = 0$, the selectivity parameter of a participation model is not identifiable without the follow-up data. In Article III we found out that an informative prior to this parameter is needed, or otherwise the MCMC-chains in posterior computation may fail to converge. An alternative strategy is utilised in Article II, which assumes that M and Y are conditionally independent given background information X and follow-up data T. Thus data are MAR when follow-up data are available. This is not the same as having MAR assumption with respect to the original survey data (see Section 4.2). In the Article II the survival times were explicitly imputed, but in Articles III and IV the censoring mechanism was taken into account in the likelihood, and thus imputations for event times were not needed.

The Bayesian model was implemented using Just Another Gibbs Sample (JAGS) -software (Plummer, 2003), and R (R Core Team, 2014) and rjags -package (Plummer, 2015) was used to fit the model. The high number of missing values and highe autocorrelations between the model parameteres created a computational challenge for the model fitting. It took several days to fit the model for the FINRISK data.

Chapter 6

Discussion

This thesis deals with selective non-participation in HESs and the particular situation where survey data are MNAR. The problem was studied utilising the National FINRISK Study data with interest in estimating the prevalences of daily smoking and heavy alcohol consumption. It was known based on previous studies that participation in the FINRISK Study was associated with smoking and potentially associated with alcohol consumption. Studies from other countries report that both smoking (Christensen et al., 2015) and alcohol use (Zhao et al., 2009; Torvik et al., 2012; Gorman et al., 2014; Dawson et al., 2014; Boniface et al., 2017) are associated with participation.

We proposed two approaches to this problem. First, survey data collected among non-participants of health examination were used together with multiple imputation tailored for this problem (Article I). The proposed approach in the Article I utilises multiple imputation with an assumption that the probability distribution of risk factors for non-participants $p(Y|X, M = 0)$ can be estimated from the data of re-contacts. The prevalences of daily smoking and heavy alcohol consumption were observed to be 20–50 percentages or 2–4 percentage points higher than what was observed based on participants only. An alternative assumption that the re-contact respondents represent non-participants after adjustment for age and gender was considered and evaluated based on data about past hospitalisations. That assumption was supported by the data.

The supported assumption holds for the utilised data set, but it does not necessarily hold for other data. This is why it is important to evaluate the modelling assumptions if possible. If retrospective hospitalisation (follow-up) data are available, the use of such data is recommended for assumption evaluation. The use of past hospitalisations allows assumptions to be evaluated as soon as data are available. Currently, the re-contact data consist of only the questionnaire, which limits the use of proposed multiple imputation approach to those variables which are obtained from questionnaire.

Second, a HES data linked to follow-up data were utilised to estimate the smoking prevalence (Articles II and III) and the prevalence of heavy alcohol consumption (Article IV). The structure of the missingness and the model was described using a graphical representation (Karvanen, 2015). Two versions of a parametric Bayesian model for the data were developed. The first model assumed that the missingness mechanism is MAR if both survey data and follow-up data are used to predict missing values. More precisely, probability distribution of risk factors Y given background information X and follow-up data T does not depend on participation M, that is $p(Y|X, T, M = 0) = p(Y|X, T, M = 1)$. The second model called

Bayesian MNAR model (Articles III and IV) handles the situation where missingness mechanism is MNAR although the survey and follow-up data are used to predict the missing values. The modelling in these articles is based on assumption $p(T|X, Y, M) = p(T|X, Y)$, which is that the disease risk does not depend on participation given the background data X and risk factor Y. The estimation required an informative prior regarding the parameter describing the relationship between smoking (or alcohol use) and participation to allow the identifiability of the model. The posterior distribution was computed using Markov chain Monte Carlo methods. The identifiability of the Bayesian MNAR model (Article III) was demonstrated using simulation.

A major limitation of the Bayesian MNAR approach is that it requires follow-up data to be available for both participants and non-participants. Years or even decades of follow-up are needed to observe sufficient amount of smoking or alcohol related disease events to estimate the model parameters. The model needs to be carefully built to take into account all aspects of the data.

This thesis has also produced information about the differences between participants and non-participants. It was observed that young persons do not participate as often as older people and that participation has decreased in all age groups (Article II). Among the young male respondents the percentage of heavy alcohol consumers was estimated to be 15.9% (95% confidence interval: (12.5, 19.4)) which was higher than in other demographic groups (Article I). The cumulative hazards of smoking-related diseases were about six times higher for smoking men than non-smoking men, and about ten times higher for smoking women than non-smoking women (Article II). Non-participants have higher rates of smoking-based and alcohol-based diseases than participants among men and women (Articles III and IV).

The estimates of smoking prevalence and heavy alcohol consumption prevalence are key epidemiological results. For men, the (posterior) mean estimates of smoking prevalence in North Karelia and Northern Savonia were about 50% in 1972 and about 30% in 2007 according to Bayesian MNAR model (Article III). For women, corresponding mean estimates were 12–13% in 1972 and 17–20% in 2007. The 2012 estimates were 28.5% for men and 19.0% for women based on multiple imputation (Article I). A Bayesian modelling approach presented in Article II results in mean estimates of smoking prevalence to be 52% in 1972 and 32% in 1997 for men and 12% in 1972 and 18% in 1997 for women.

Both proposed approaches yield higher posterior expectations of prevalences and wider uncertainty estimates for smoking than the complete case analysis. For these data, the differences appear to be higher in 1977–1992 than in the rest of the years. It can be noticed that the Bayesian MNAR model finds only about one percentage point difference in 2007 for smoking prevalence, while Karvanen et al. (2016) estimates five percentage point difference using multiple imputation approach and re-contact data. Similarly, corresponding multiple imputation approach (Article I) founds a four percentage point difference for 2012. It can be speculated that this difference is because in 2007 only five years of follow-up were available, although it takes 20 years of daily smoking to develop a lung cancer or COPD.

Also, the proposed approaches provide wider credible and confidence intervals than the complete case analysis. Smaller uncertainty of the complete case analysis estimates are based on assumption that does not hold for the data. This being the case, we can say that the proposed approaches provide more realistic estimate of uncertainty than complete case analysis.

The prevalence estimates for heavy alcohol consumption are obtained from Bayesian MNAR model. For men, these prevalence estimates are about 10% in 1987 and about 13% in

2007. The highest estimate was 23% in 2002. For women, corresponding estimates are about 3% in 1987 and about 6% in 2007. The prevalence estimates of 2012 are 9.4% for men and 4.8% for women based on multiple imputation approach.

It appears that more attention should be paid to missing data in the analysis of the HES data. Our results suggest that earlier results based only on the participants' data are biased, underestimate the uncertainty, and that earlier results give overly positive image about the prevalences of daily smoking and heavy alcohol consumption. To provide more reliable estimates, the association between participation and the variables of interest need to be taken into account in the analysis. Bayesian modelling or multiple imputation approach can be used to analyse data with selective non-participation and MNAR missingness mechanism. Both of these approaches require additional data along with the HES data. For the future data collection, we would recommend to continue collecting re-contact data and to study potential modelling assumptions based on additional data set linked to the data. Potential topics for the future research are the improvements of computational efficiency for data with large number of missing values and the simultaneous use of re-contact and follow-up data.

Bibliography

Andridge, R. R. and Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64.

Boniface, S., Scholes, S., Shelton, N., and Connor, J. (2017). Assessment of non-response bias in estimates of alcohol consumption: applying the continuum of resistance model in a general population survey in england. *PloS one*, 12(1):e0170892.

Borodulin, K., Levälahti, E., Saarikoski, L., Lund, L., Juolevi, A., Grönholm, M., Jula, A., Laatikainen, T., Männistö, S., Peltonen, M., Salomaa, V., Sundvall, J., Taimi, M., Virtanen, S., and Vartiainen, E. (2013). Kansallinen FINRISKI 2012 -terveystutkimus - Osa 2: Tutkimuksen taulukkoliite. *Terveyden ja hyvinvoinnin laitos, Raportti 2013/22.* In Finnish. Accessed: 2017-12-11.

Borodulin, K., Tolonen, H., Jousilahti, P., Jula, A., Juolevi, A., Koskinen, S., Kuulasmaa, K., Laatikainen, T., Männistö, S., Peltonen, M., Perola, M., Puska, P., Salomaa, V., Sundvall, J., Virtanen, S. M., and Vartiainen, E. (2017). Cohort profile: The national FINRISK study. *International Journal of Epidemiology.*

Brick, J. M. and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(3):215–238.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Christensen, A. I., Ekholm, O., Gray, L., Glümer, C., and Juel, K. (2015). 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*, 110(9):1505–1512.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (2009). Smoking and lung cancer: recent evidence and a discussion of some questions. *International Journal of Epidemiology*, 38(5):1175–1191.

Dawson, D. A., Goldstein, R. B., Pickering, R. P., and Grant, B. F. (2014). Nonresponse bias in survey estimates of alcohol consumption and its association with harm. *Journal of Studies on Alcohol and Drugs*, 75(4):695–703.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38.

Doll, R. and Hill, A. B. (1956). Lung cancer and other causes of death in relation to smoking. *British Medical Journal*, 2(5001):1071–1081.

Galea, S. and Tracy, M. (2007). Participation rates in epidemiologic studies. *Annals of Epidemiology*, 17(9):643–653.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85(412):972–985.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian data analysis*. CRC press, Boca Raton, FL.

Gorman, E., Leyland, A. H., McCartney, G., White, I. R., Katikireddi, S. V., Rutherford, L., Graham, L., and Gray, L. (2014). Assessing the representativeness of population-sampled health surveys through linkage to administrative data on alcohol-related outcomes. *American Journal of Epidemiology*, 180(9):941–948.

Harald, K., Salomaa, V., Jousilahti, P., Koskinen, S., and Vartiainen, E. (2007). Non-participation and mortality in different socioeconomic groups: the FINRISK population surveys in 1972–92. *Journal of Epidemiology & Community Health*, 61(5):449–454.

Hedeker, D. and Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2(1):64.

Jousilahti, P., Salomaa, V., Kuulasmaa, K., Niemelä, M., and Vartiainen, E. (2005). 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 & Community Health*, 59(4):310–5.

Karvanen, J. (2015). Study design in causal models. *Scandinavian Journal of Statistics*, 42(2):361–377.

Karvanen, J., Tolonen, H., Härkänen, T., Jousilahti, P., and Kuulasmaa, K. (2016). Selection bias was reduced by recontacting nonparticipants. *Journal of Clinical Epidemiology*, 76:209–217.

Kenward, M. G. and Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, 8(1):51–83.

Korhonen, H. J., Jousilahti, P. J., Vartiainen, E., Juolevi, A., Sundvall, J., and Puska, P. (1999). FINRISKI 1997: kaupunkiraportti. Tutkimus kroonisten kansantautien riskitekijöistä, niihin liittyvistä elintavoista, oireista ja terveyspalvelujen käytöstä Helsingissä, Vantaalla, Joensuussa, Kuopiossa, Oulussa ja Turussa. Tutkimuksen toteutus ja perustaulukot. *Kansanterveyslaitoksen julkaisuja B: 4/1999.* In Finnish.

Laatikainen, T., Tapanainen, H., Alfthan, G., Salminen, I., Sundvall, J., Leiviskä, J., Harald, K., Jousilahti, P., Salomaa, V., and Vartiainen, E. (2003). FINRISKI 2002: Tutkimuksen toteutus ja tulokset 1. Perusraportti. *Helsinki: Publications of the National Public Health Institute.*

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data.* John Wiley & Sons, Second edition.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Molenberghs, G. and Kenward, M. (2007). *Missing data in clinical studies.* John Wiley & Sons.

National Institute for Health and Welfare (2017). Care register for health care. `http://www.thl.fi/en/web/thlfi-en/statistics/information-on-statistics/register-descriptions/care-register-for-health-care`. Accessed: 2017-12-11.

Nishimura, R., Wagner, J., and Elliott, M. (2016). Alternative indicators for the risk of non-response bias: A simulation study. *International Statistical Review*, 84(1):43–62.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities.* John Wiley & Sons.

Peltonen, M., Harald, K., Männistö, S., Saarikoski, L., Lund, L., Sundvall, J., Juolevi, A., Laatikainen, T., Aldén-Nieminen, H., Luoto, R., Jousilahti, P., Salomaa, V., Taimi, M., and Vartiainen, E. (2008). Kansallinen FINRISKI 2007 -terveystutkimus : Tutkimuksen toteutus ja tulokset: Taulukkoliite. *Kansanterveyslaitoksen julkaisuja B: 35/2008.* In Finnish. Accessed: 2017-12-11.

Plummer, M. (2003). JAGS: A program for analyses of Bayesian graphical models using Gibbs sampling. `https://www.r-project.org/conferences/DSC-2003/Proceedings/`. Online; Accessed 2017-12-14.

Plummer, M. (2015). rjags: Bayesian graphical models using MCMC. `https://CRAN.R-project.org/package=rjags`. Online; Accessed 2017-12-14.

Puska, P., Rimpelä, M., Sievers, K., Tuomilehto, J., Virtamo, J., Prunnila, T., and Karjalainen, Y. (1973). Pohjois-Karjala projektin peruskartoitus: toteutus ja perustaulukot. Kansanterveyden tutkimuskeskus (Pohjois-Karjala projekti). Kuopion korkeakoulun julkaisuja, kansanterveystiede B:1/1973. In Finnish.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation.* Springer Science & Business Media.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods.* Springer, New York, Second edition.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rothman, K. J. (2012). *Epidemiology: an introduction.* Oxford University Press, New York, Second edition.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics).* Wiley, New York.

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by "missing at random"? *Statistical Science*, 28(2):257–268.

Siddique, J. and Belin, T. R. (2008). Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data. *Computational Statistics & Data Analysis*, 53(2):405–415.

Spiegelhalter, D. J. and Best, N. G. (2003). Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine*, 22(23):3687–3709.

Statistics Finland (2014). Official Statistics of Finland (OSF): Causes of death [e-publication]. `http://tilastokeskus.fi/til/ksyyt/index_en.html`. Accessed: 2014-08-28.

Statistics Finland (2017). The register of completed education and degrees. `https://www.stat.fi/til/kou_en.html`. Accessed: 2017-12-11.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Thompson, M. (1997). *Theory of sample surveys.* Monographs on Statistics and Applied Probability 74. CRC Press, First edition.

Tolonen, H. (2013). EHES manual: Part A. Planning and preparation of the survey. Published online. `http://urn.fi/URN:ISBN:978-952-245-842-1`. National Institute for Health and Welfare. Helsinki, Finland.

Tolonen, H., Dobson, A., and Kulathinal, S. (2005). Effect of the trend estimates on the difference between survey respondents and non-respondents: results from 27 populations in the WHO MONICA Project. *European Journal of Epidemiology*, 20(11):887–98.

Torvik, F. A., Rognmo, K., and Tambs, K. (2012). Alcohol use and mental distress as predictors of non-response in a general population health survey: the hunt study. *Social Psychiatry and Psychiatric Epidemiology*, 47(5):805–816.

van Buuren, S. (2012). *Flexible imputation of missing data.* CRC press, Boca Raton, FL.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3).

Vartiainen, E., Jousilahti, P., Tamminen, M., Korhonen, H., Tuomilehto, J., Sundvall, J., Jauhiainen, M., and Puska, P. (1993). FINRISKI '92: Tutkimus kansanterveydellisistä riskitekijöistä, niihin liittyvistä elintavoista, oireista ja terveyspalvelujen käytöstä. tutkimuksen toteutus ja perustaulukot. Kansanterveyslaitoksen julkaisuja B 9/1993. Helsinki 1993.

World Health Organization (2015). Global reference list of 100 core health indicators. World Health Organization. Geneva, Switzerland. http://apps.who.int/iris/bitstream/10665/173589/1/WHO_HIS_HSI_2015.3_eng.pdf.

Wynder, E. L. and Hoffmann, D. (1994). Smoking and lung cancer: scientific challenges and opportunities. *Cancer Research*, 54(20):5284–5295.

Zhao, J., Stockwell, T., and MacDonald, S. (2009). Non–response bias in alcohol and drug population surveys. *Drug and Alcohol Review*, 28(6):648–657.

Appendix A: Corrections to articles

In Article II Section 3.1, p. 5 the first of the two equations should be:

$$T_i = (t_i, r_i) = \begin{cases} (t_i, 1), & \text{if an event is observed} \\ (t_i, 0), & \text{if an event is right censored,} \end{cases}$$

that is the value $r_i = 1$ should indicate observing and $r_i = 0$ is right censoring.

I

Adjusting for selective non-participation with re-contact data in the FINRISK 2012 survey.

Kopra, J., Härkänen, T., Tolonen, H., Jousilahti, P.,
Kuulasmaa, K., Reinikainen, J. and Karvanen, J.

Published online first in *Scandinavian Journal of Public Health.* (2017)
doi: 10.1177/1403494817734774.

Scandinavian Journal of Public Health, 1–9

ORIGINAL ARTICLE

Adjusting for selective non-participation with re-contact data in the FINRISK 2012 survey

JUHO KOPRA[1], TOMMI HÄRKÄNEN[2], HANNA TOLONEN[2], PEKKA JOUSILAHTI[2], KARI KUULASMAA[2], JAAKKO REINIKAINEN[2] & JUHA KARVANEN[1]

[1]Department of Mathematics and Statistics, University of Jyvaskyla, Finland, and [2]Department of Public Health Solutions, National Institute for Health and Welfare, Finland

Abstract

Aims: A common objective of epidemiological surveys is to provide population-level estimates of health indicators. Survey results tend to be biased under selective non-participation. One approach to bias reduction is to collect information about non-participants by contacting them again and asking them to fill in a questionnaire. This information is called re-contact data, and it allows to adjust the estimates for non-participation. *Methods:* We analyse data from the FINRISK 2012 survey, where re-contact data were collected. We assume that the respondents of the re-contact survey are similar to the remaining non-participants with respect to the health given their available background information. Validity of this assumption is evaluated based on the hospitalisation data obtained through record linkage of survey data to the administrative registers. Using this assumption and multiple imputation, we estimate the prevalences of daily smoking and heavy alcohol consumption and compare them to estimates obtained with a commonly used assumption that the participants represent the entire target group. *Results:* When adjusting for non-participation using re-contact data, higher prevalence estimates were observed compared to prevalence estimates based on participants only. Among men, the smoking prevalence estimate was 28.5% (23.2% for participants) and heavy alcohol consumption prevalence was 9.4% (6.8% for participants). Among women, smoking prevalence was 19% (16.5% for participants) and heavy alcohol consumption was 4.8% (3% for participants). ***Conclusions:* The utilisation of re-contact data is a useful method to adjust for non-participation bias on population estimates in epidemiological surveys.**

Key Words: *Selection bias, smoking, alcohol consumption, missing data*

Introduction

Health examination surveys (HES) are among the key data sources for the data-driven planning of national health policies. If the participants of the survey are a representative sample of the population of interest, then simple statistical estimates, such as sample averages, provide reliable support for decision-making. A major threat to the representativeness of survey data is selective non-participation. Under selective non-participation, survey participants do not represent the population of interest, which leads to bias in population-level health indicators. By health indicator, we mean a health-related population statistic, such as the prevalence of smokers. For example, if healthy people are more willing to participate in a survey than people with poor health, the health indicators give an overly positive impression of the health of the population. This makes the data misleading.

The sampling frame often provides background information, such as sex, age and region, on the sample members. This information reveals whether some demographic groups are under- or over-represented among the participants compared to non-participants.

Correspondence: Juho Kopra, Department of Mathematics and Statistics, University of Jyvaskyla, PO Box 35 (MaD), FI-40014, Finland.
E-mail: juho.j.kopra@jyu.fi

Date received 4 January 2017; reviewed 9 August 2017; accepted 4 September 2017

© Author(s) 2017
Reprints and permissions: sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1403494817734774
journals.sagepub.com/home/sjp
SAGE

However, this information is insufficient to assess whether or not the non-participation is selective with respect to variables of interest.

Record linkage of HES studies with register-based data has shown that non-participants have higher alcohol consumption [1,2], and higher smoking and alcohol-related mortality [3], which indicates that participation is quite probably selective with respect to smoking and alcohol consumption. Non-participants also have a higher total mortality rate [4–6] than participants. Also, non-participants tend to be younger and less educated than participants [7,8]. They more often receive social welfare payments and have a higher unemployment rate [9].

Possible selectivity with respect to variables of interest (e.g. smoking and alcohol use) cannot be assessed using only data from the participants. In addition to the survey sample, we consider two additional sources of information: follow-up data and re-contact data.

Follow-up data are (time-to-event) data collected after the survey about diagnoses of diseases, with date details, date of death and causes of death. Re-contact data are data from a survey conducted among people who did not participate in the original survey.

Recently, adjustment methods using follow-up [10] and re-contact data [11] as an additional source of information have been proposed to reduce the selection bias. Kopra et al. [10] utilised a Bayesian survival model to impute the values of daily smoking using register-based follow-up data on chronic obstructive pulmonary disease and lung cancer. Karvanen et al. [11] used data on re-contact respondents' information and evaluated the modelling assumption using the five years of register-based follow-up data. A problem with these methods is that they can only be applied after the several years of follow-up after the survey has finished. In this paper, we use the same principal method as that presented in Karvanen et al. [11]. However, an important difference is that we evaluate the modelling assumption using data on past hospitalisations instead of follow-up data.

The aim of this paper is twofold. 1. To provide estimates for prevalences of self-reported heavy alcohol consumption and daily smoking adjusted for non-participation using data obtained through re-contact of non-participants. 2. To use register-based data on hospitalisation history for the evaluation of the assumption about the similarity of the health of re-contacted non-participants and the remaining non-participants.

Methods

Data

We use data from the National FINRISK 2012 survey, a HES among adults from five regions of Finland [12]. The survey was conducted in early 2012 with a total sample size of 10,000 invitees aged 25–74 years. The invitees were sampled from the national Population Information System using simple random sampling stratified by region, sex and 10-year age group. The respective Ethics Committee approved the survey when it was conducted. Written informed consent was obtained from survey participants.

Survey invitees received a letter of invitation with assigned time and place for an appointment (health examination), and a questionnaire to be filled in at home and returned at the health examination. One to two days before the appointment a reminder SMS message was sent to encourage the invitees to participate. If a person did not show up for an appointment, he or she was contacted to agree the time of a new appointment via a phone call or by a reminder postcard if the phone number was not known. Persons who participated in the health examination are referred to as participants.

After the original survey, a re-contact round was conducted. Persons who did not take part in a health examination received a re-contact letter in which they were asked to return the self-reported questionnaire using an envelope with pre-paid postage attached to the letter. This letter also contained a questionnaire which was identical to the previously sent questionnaire. Those individuals who returned the questionnaire after receiving the re-contact letter are called re-contact respondents. The time lag between the original survey and the re-contact round was 2–5 months. During that time some persons may have changed their smoking or alcohol-use habits, but we do not expect this to notably alter the results.

A total of 5827 invitees participated in the survey, yielding a 58.3% participation rate (i.e. having both the questionnaire and the health examinations completed). The re-contact round resulted in 597 returned questionnaires (14.3% of all non-participants), leaving 3576 non-participants without any self-reported information.

The data on background variables, sex, age and region, are available from the sampling frame for both participants and non-participants. Data on hospital visits and diagnoses (ICD codes) since 1969 are obtained for both participants and non-participants through record linkage to the Care Register for Health Care [13] using the unique personal identification code provided for every resident in Finland. We call these data 'hospitalisation history data'.

The survey sample is classified into three groups of people;

1. *Participants* who returned the questionnaire and participated in a health examination.

2. *Re-contact respondents* who did not participate in the survey after initial invitation, but did return the re-contact round questionnaire.
3. *Non-participants* who neither participated in health examination nor returned the re-contact round questionnaire.

The variables of primary interest are self-reported daily smoking and heavy alcohol consumption. Females who consumed more than 16 portions of alcohol per week and men who consumed more than 24 portions per week are defined as heavy alcohol users. One portion corresponds to 12 g of pure alcohol.

Modelling approach

In this paper, we fit two kinds of models: three alternative models to impute missing values in data and one model to evaluate the modelling assumptions. We apply R statistical software, version 3.3.1 [14] and the R package 'mice' [15] for multiple imputations, and R package 'pscl' for the evaluation of the modelling assumptions.

The alternative modelling assumptions that we consider here are as follows;

1. The participants represent the whole population of interest.
2. The participants represent the whole population of interest when adjusted for background variables.
3. The re-contact respondents represent all non-participants when adjusted for background variables.

Assumption (1) is a missing-completely-at-random (MCAR) assumption [16], leading to the complete-case analysis where data on participants are used to estimate the health indicators of non-participants. If assumption (1) holds, the non-participation is neither selective with respect to variables of interest nor background variables. It means that, for example, the average prevalence of smoking measured from the participants describes the smoking prevalence for the whole population even without adjusting for background variables. This assumption is made implicitly when estimates based on participants only are reported.

Assumption (2) is a missing-at-random (MAR) assumption that makes it possible to use data on participants to estimate the health of non-participants and, therefore, the health of the whole population provided that the background variables are collected. If assumption (2) holds, then non-participation is not selective with respect to variables of interest, but it may be selective with respect to the background variables. To estimate, for example, the prevalence of smoking for the non-participants, adjustment for the background variables is required.

The assumption (3) allows the use of data on re-contact respondents to estimate the health of non-participants, provided that background variables are collected for all invitees of a survey. This assumption can be interpreted as a version of the continuum of resistance model [17,18] where we adjust for background variables. Under assumption (3), the data are missing not at random (MNAR) with respect to HES participation, and MAR with respect to re-contact response.

Under assumption (3), participation may be selective with respect to variables of interest and background variables. However, the response to the re-contact questionnaire is not selective with respect to variables of interest, but it may be selective with respect to background information. If this assumption does not hold, the health indicators of the remaining non-participants cannot be estimated without bias unless some additional data are available.

Imputation models

We consider three different approaches imputing the health indicators. The approaches utilise either assumption (2) or (3). Assumption (1) is not used in imputation but is utilised if estimates based on data on participants only are considered to describe the health of the whole population. Our primary approach is called MI-MNAR, using multiple imputation (MI) with an assumption (3). In MI-MNAR the missing values for re-contact respondents and non-participants are imputed assuming that the parameters of the model are different for re-contact respondents and participants. Two alternative methods use assumption (2), and are called MI-MAR and MI-MAR-NR. In MI-MAR imputation the model parameters are the same in all groups. The MI-MAR-NR method uses no re-contact (NR) data at all, but is otherwise similar to MI-MAR.

For each imputed variable the multiple imputations are carried out using a regression model (fully conditional specification) [19]. The other variables are used as covariates in the imputation model. The imputed variables are daily smoking and heavy alcohol consumption, which are predicted by the following covariates: sex, age, region, education level, civil status, self-reported hypertension, self-reported high cholesterol and recency of blood pressure and cholesterol measurements. These variables are collected through the questionnaire and they are potential predictors for the lifestyle indicators and the participation. Covariates with missing data were imputed simultaneously with the main variables. The imputation models are specified as in Karvanen et al. [11]. In addition, we predict the number of hospitalisations for model-checking purposes based on the same covariates as for daily smoking and heavy alcohol consumption.

Evaluation of modelling assumptions

We evaluate the modelling assumptions (1)–(3) using the background variables and the hospitalisation history data. Assumption (1) is violated if there is an indication that either distribution of variables measured in the survey or distributions of background variables differ between participants and non-participants (including re-contact respondents). Assumption (2) does not hold if participants and non-participants (including re-contact respondents) differ with respect to their health indicators when conditioned on background variables. Assumptions (2) and (3) cannot be tested directly because there is no estimate of health indicator available for non-participants. Instead, they are evaluated by fitting a statistical regression model for the number of hospitalisations by each of the three groups and using the background variables as covariates.

We check if the hospitalisation rates differ between the participants, the re-contact respondents and the non-participants. A difference is interpreted as an evidence of differences in the health indicator distributions. If the hospitalisation rates for re-contact respondents can be assumed to be similar to non-participants' rates, then we can obtain information on the health of the non-participants from the re-contact data.

We utilise a zero-inflated negative binomial regression [20] as a model for the hospitalisation data to evaluate assumptions (2) and (3). The zero-inflated model consists of two parts: the excess zeros model; and the model for the counts. The count model utilises negative binomial distribution. The excess zero model describes the proportion of excess zeros (zero inflation) in addition to the zeros from the count model. Thus, a zero may occur from both of the models – the excess zero model or the count model.

We check the assumptions using full, five-year and one-year hospitalisation history data. The longer the history, the more hospitalisation events are expected. A high total number of events makes it easier to observe differences in the counts between the groups. However, as the health of individual changes over time, hospitalisation counts from a recent period may better describe the health at the time of the survey.

Results

The characteristics of the collected survey data are described in Table I. Among participants and re-contact respondents, there are slightly more women than men. Among non-participants, the opposite is true, which indicates that women are more eager to participate. The average age of non-participants is lower than that of both participants and re-contact respondents. The re-contact respondents seem to be less educated and more often single than the participants. For both men and women, there are more smokers among re-contact respondents than among participants. For men, the proportion of heavy alcohol consumption is 6.8% for both participants and re-contact respondents, but there is a lot of variation between the age groups. The proportion is much higher among the young re-contact respondents than among young participants. In the age group 25–34 years old, the proportion among re-contact respondents is exceptionally high (15.9%) compared to other age groups of re-contacts. Among the re-contact respondents of the age groups 55–65 and 65–74, the proportion drops below the rates of participants. For women, in all age groups, there is higher heavy alcohol consumption among re-contacts than for participants.

Re-contact respondents seem to be more often smokers and heavy alcohol users than participants, except for heavy alcohol consumption among men where the prevalence is the same for participants and re-contact respondents.

Table II shows the results for the assumption checking model. The risk of being hospitalised is higher for men than women, and the risk increases with age. A significant difference between participants and non-participants was observed for full, five-year and one-year hospitalisation histories, while no difference between re-contact respondents and non-participants was found for five-year and one-year histories. These findings indicate that assumption (2) does not hold, while assumption (3) is supported.

Table III presents the predicted hospitalisation counts per 1000 individuals for each length of hospitalisation history. The proposed method, the MI-MNAR, has predicted counts which match the best with the observed full cohort counts. This supports the assumption (3), which states that re-contacts represent the non-participants given their background variables. The match is more convincing for one-year and five-year histories than for full history. The hospitalisation counts for participants, re-contact respondents and non-participants can be compared with each other. It is interesting to see that hospitalisation counts per 1000 individuals are lower for female non-participants than for female re-contact respondents. For men, the contrary is true.

Table I. The averages and proportions, with their 95% confidence intervals, for background variables and health indicators.

	Participants	Re-contact respondents	Non-participants
N	5827	597	3576
Women, %	53.1 (51.5, 54.6)	53.3 (48.6, 58.0)	46.1 (44.0, 48.1)
Average age, years	49.7 (49.3, 50.0)	49.2 (48.1, 50.3)	44.9 (44.4, 45.3)
Age group 25–34, %	18.7 (17.5, 19.9)	21.3 (17.5, 25.2)	30.5 (28.6, 32.3)
Age group 35–44, %	18.0 (16.8, 19.2)	15.8 (12.4, 19.2)	21.7 (20.0, 23.4)
Age group 45–54, %	22.1 (20.9, 23.4)	22.8 (18.8, 26.7)	20.1 (18.4, 21.7)
Age group 55–64, %	23.7 (22.4, 25.0)	26.2 (22.1, 30.4)	17.7 (16.1, 19.3)
Age group 65–74, %	17.5 (16.3, 18.6)	13.9 (10.6, 17.1)	10.1 (8.9, 11.4)
Education			
High, %	37.6 (36.1, 39.1)	34.7 (30.3, 39.2)	–
Low, %	30.9 (29.5, 32.3)	34.8 (30.3, 39.3)	–
Civil status			
Married, %	52.4 (50.9, 54.0)	49.8 (45.1, 54.5)	–
Cohabiting, %	18.6 (17.4, 19.8)	17.1 (13.6, 20.7)	–
Single, %	15.4 (14.3, 16.5)	19.3 (15.6, 23.0)	–
Divorced, %	10.7 (9.7, 11.6)	11.4 (8.4, 14.4)	–
Widow, %	2.8 (2.3, 3.4)	2.5 (1.0, 3.9)	–
Daily smokers, men %	23.2 (21.9, 24.5)	28.5 (24.2, 32.7)	–
Age group 25–34, %	30.4 (29.0, 31.8)	26.1 (21.9, 30.2)	–
Age group 35–44, %	24.4 (23.1, 25.7)	36.3 (31.8, 40.9)	–
Age group 45–54, %	23.1 (21.9, 24.4)	24.2 (20.2, 28.2)	–
Age group 55–64, %	24.3 (23.0, 25.6)	31.4 (27.0, 35.8)	–
Age group 65–74, %	13.2 (12.2, 14.3)	23.3 (19.3, 27.2)	–
Daily smokers, women %	16.5 (15.3, 17.6)	19.7 (15.9, 23.4)	–
Age group 25–34, %	21.0 (19.8, 22.2)	16.3 (12.8, 19.8)	–
Age group 35–44, %	15.7 (14.6, 16.9)	24.0 (19.9, 28.0)	–
Age group 45–54, %	17.5 (16.3, 18.7)	19.6 (15.9, 23.4)	–
Age group 55–64, %	17.5 (16.3, 18.7)	28.1 (23.9, 32.3)	–
Age group 65–74, %	9.2 (8.3, 10.1)	4.7 (2.7, 6.7)	–
Heavy alcohol users, men %	6.8 (6.1, 7.6)	6.8 (4.4, 9.2)	–
Age group 25–34, %	5.9 (5.2, 6.7)	15.9 (12.5, 19.4)	–
Age group 35–44, %	5.3 (4.6, 6.0)	9.1 (6.4, 11.8)	–
Age group 45–54, %	9.6 (8.7, 10.5)	5.4 (3.2, 7.5)	–
Age group 55–64, %	7.2 (6.4, 8.0)	0.9 (0.0, 1.8)	–
Age group 65–74, %	5.3 (4.6, 6.0)	1.6 (0.4, 2.8)	–
Heavy alcohol users, women %	3.0 (2.5, 3.5)	5.0 (3.0, 7.1)	–
Age group 25–34, %	4.3 (3.6, 4.9)	6.8 (4.4, 9.1)	–
Age group 35–44, %	3.0 (2.5, 3.5)	6.4 (4.1, 8.7)	–
Age group 45–54, %	3.8 (3.2, 4.4)	5.1 (3.0, 7.2)	–
Age group 55–64, %	2.5 (2.0, 3.0)	4.5 (2.6, 6.5)	–
Age group 65–74, %	1.0 (0.7, 1.3)	1.5 (0.3, 2.6)	–

Table IV describes the prevalence of daily smoking and heavy alcohol consumption estimated with different imputation models. MI-MNAR imputation results show that the point estimate of the prevalence of daily smoking for men is 28.5%, which is 5.3 percentage points higher than what was measured from the participants only. For women, the corresponding imputed estimate is 19.0%, which is 2.5 percentage points higher than the estimate based on the participants only. For smoking, the estimates from participants only do not lie within the 95% confidence interval of MI-MNAR imputations for men, and for women they are barely within the confidence interval. The point estimates by MI-MAR-NR are, in all cases, lower than the point estimates of MI-MAR and MI-MNAR.

The prevalence of heavy alcohol consumption for men by the MI-MNAR method is 9.4%. This is much higher than one would expect based on the heavy alcohol consumption rates of participants (6.8%) and re-contact respondents (6.8%). The key factors in the imputation of heavy alcohol consumption are smoking, sex, age and region. Smoking strongly predicts heavy alcohol consumption in the estimated imputation model. Corresponding odds ratios for participants are 3.93 (2.87, 5.4) for men

Table II. Estimated parameters, with their 95% confidence intervals, from the zero-inflated negative binomial regression model for the number of hospital visits. The model was fitted using three periods of hospitalisation history data: full history, five-year history and one-year history. The reference levels for the categorical variables sex and region are men and North Karelia.

	Estimate (95% confidence interval)					
	Full history		Five years		One-year	
Count model						
Intercept	0.84	(0.71, 0.98)	–0.88	(–1.19, –0.57)	–1.79	(–2.49, –1.10)
Age: men (10 years)	0.18	(0.16, 0.20)	0.26	(0.21, 0.31)	0.27	(0.16, 0.39)
Age: women (10 years)	0.33	(0.30, 0.35)	0.22	(0.17, 0.28)	0.08	(–0.02, 0.18)
Sex (female)	–0.51	(–0.67, –0.34)	0.32	(–0.05, 0.69)	1.22	(0.35, 2.10)
Region: Northern Savonia	0.00	(–0.09, 0.09)	–0.03	(–0.20, 0.13)	0.04	(–0.23, 0.31)
Region: Turku and Loimaa	–0.16	(–0.25, –0.07)	–0.26	(–0.43, –0.09)	–0.26	(–0.55, 0.02)
Region: Helsinki and Vantaa	–0.30	(–0.37, –0.22)	–0.45	(–0.60, –0.31)	–0.46	(–0.70, –0.23)
Region: Oulu	0.03	(–0.05, 0.11)	–0.09	(–0.24, 0.06)	–0.16	(–0.41, 0.09)
Participant (yes)	–0.25	(–0.30, –0.21)	–0.60	(–0.71, –0.48)	–0.92	(–1.14, –0.70)
Re-contact respondent (yes)	–0.10	(–0.19, –0.01)	0.02	(–0.20, 0.24)	0.08	(–0.33, 0.50)
Zero model						
Intercept	22.19	(6.74, 37.63)	1.40	(0.53, 2.26)	1.73	(0.66, 2.80)
Age: men (10 years)	–9.23	(–15.32, –3.15)	–0.56	(–0.78, –0.34)	–0.31	(–0.55, –0.07)
Age: women (10 years)	–1.46	(–1.80, –1.11)	–0.44	(–0.65, –0.22)	–0.42	(–0.58, –0.26)
Sex (female)	–19.44	(–34.93, –3.96)	–0.46	(–1.78, 0.85)	0.99	(–0.35, 2.33)
Participant (yes)	–0.56	(–1.11, 0.01)	–1.59	(–2.66, –0.52)	–0.88	(–1.37, –0.40)
Re-contact respondent (yes)	–0.59	(–1.96, 0.78)	0.05	(–0.58, 0.68)	0.12	(–0.43, 0.67)

Table III. Hospitalisations per 1000 individuals by length of hospitalisation history, using: full history available, five-year history and one-year history. First four rows describe the actual data, and the next three show the results of multiple imputations. The results of multiple imputations are to be compared with the numbers from the full cohort.

	Estimate (95% confidence interval)		
	Full history	Five years	One year
Men:			
Full cohort	4305 (4126, 4484)	773 (718, 829)	182 (163, 201)
Participants only	3755 (3561, 3948)	647 (589, 705)	147 (127, 168)
Re-contact respondents	4072 (3394, 4751)	941 (671, 1212)	227 (136, 318)
Non-participants	5070 (4720, 5420)	919 (811, 1027)	223 (187, 259)
MI-MNAR	4050 (3725, 4374)	834 (689, 978)	188 (158, 217)
MI-MAR	3784 (3526, 4042)	667 (602, 731)	151 (130, 171)
MI-MAR-NR	3816 (3624, 4008)	675 (619, 732)	152 (133, 171)
Women:			
Full cohort	5445 (5237, 5653)	852 (783, 921)	180 (160, 200)
Participants only	5598 (5377, 5818)	799 (733, 865)	156 (138, 175)
Re-contact respondents	6514 (5581, 7446)	1073 (767, 1379)	269 (175, 363)
Non-participants	5179 (4733, 5625)	918 (760, 1076)	207 (161, 252)
MI-MNAR	5538 (5076, 5999)	929 (751, 1108)	222 (169, 275)
MI-MAR	5168 (4967, 5369)	767 (692, 842)	150 (132, 168)
MI-MAR-NR	5146 (4949, 5343)	760 (698, 821)	153 (133, 173)
Both:			
Full cohort	4880 (4742, 5018)	813 (769, 857)	181 (167, 195)
Participants only	4676 (4527, 4825)	723 (679, 767)	152 (138, 166)
Re-contact respondents	5293 (4696, 5890)	1007 (800, 1214)	248 (182, 313)
Non-participants	5124 (4845, 5403)	919 (826, 1012)	215 (186, 243)
MI-MNAR	4800 (4524, 5076)	882 (763, 1001)	205 (179, 231)
MI-MAR	4482 (4320, 4644)	717 (667, 767)	150 (137, 163)
MI-MAR-NR	4487 (4348, 4626)	718 (677, 759)	152 (139, 166)

MI: multiple imputation; MNAR: missing not at random; MAR: missing at random; NR: no re-contact.

and 4.1 (2.57, 6.56) for women. Further, it can be seen from Table I that heavy alcohol consumption is much more common among young re-contacts than among participants, and non-participation is much more common among young people than among others. These facts together explain why MI-MNAR

Table IV. Comparison of prevalence estimates of daily smoking and heavy alcohol consumption. The proposed method MI-MNAR is compared to alternative methods MI-MAR, MI-MAR-NR and estimates for the participants and re-contact respondents.

	Estimate (95% confidence interval)	
	Daily smokers (%)	Heavy alcohol users (%)
Men:		
Participants	23.2 (21.6, 24.8)	6.8 (5.9, 7.8)
Re-contact respondents	28.5 (22.9, 34.0)	6.8 (3.7, 9.9)
MI-MNAR	28.5 (25.9, 31.2)	9.4 (7.2, 11.6)
MI-MAR	24.8 (23.1, 26.5)	7.1 (5.7, 8.4)
MI-MAR-NR	23.7 (22.2, 25.1)	6.7 (5.7, 7.7)
Women:		
Participants	16.5 (15.2, 17.8)	3.0 (2.4, 3.6)
Re-contact respondents	19.7 (15.4, 24.0)	5.0 (2.6, 7.4)
MI-MNAR	19.0 (15.8, 22.2)	4.8 (3.4, 6.3)
MI-MAR	17.1 (15.6, 18.5)	3.2 (2.4, 3.9)
MI-MAR-NR	16.5 (15.0, 18.0)	3.1 (2.3, 3.9)
Both:		
Participants	19.6 (18.6, 20.6)	4.8 (4.2, 5.3)
Re-contact respondents	23.7 (20.3, 27.2)	5.9 (3.9, 7.8)
MI-MNAR	23.7 (21.5, 25.9)	7.1 (5.6, 8.6)
MI-MAR	20.9 (19.7, 22.0)	5.1 (4.4, 5.8)
MI-MAR-NR	20.1 (19.0, 21.1)	4.9 (4.3, 5.5)

MI: multiple imputation; MNAR: missing not at random; MAR: missing at random; NR: no re-contact.

leads to the high prevalence of heavy alcohol consumption in men.

Discussion

We studied the estimation of population-level health indicators from data that suffer from relatively high non-participation. The estimation utilised re-contact data; that is, data from the non-participants who answered to a survey questionnaire when contacted again, to adjust for non-participation. With data from FINRISK 2012, we estimated the prevalence of daily smoking and heavy alcohol consumption using the MI-MNAR approach. These estimates were compared with the estimates obtained using less elaborated MI-MAR and MI-MAR-NR approaches and with the straightforward inclusion of participants only.

These approaches relied on different assumptions. The MI-MNAR approach assumed that re-contact respondents represent all non-participants when adjusted for the background variables, while the MI-MAR approaches used a stronger assumption that participants represent the whole population when adjusted for the background variables. The inclusion of the participants only (complete-case analysis) used the strongest assumption that the participants represent the whole population.

The bias in the estimates depends on the validity of the assumptions. Many HES report that participants and non-participants differ on their health indicators, which violates the assumption of complete-case analysis. This is also the case for the FINRISK 2012 data, as the prevalence of daily smoking and heavy alcohol consumption for participants and re-contact respondents differ. We evaluated the other two assumptions using register-based history data about the hospitalisations of all people invited to the study. We checked if there were differences in numbers of hospitalisations between participants, re-contact respondents and the remaining non-participants when other variables were used as covariates.

We found out that if an individual had ever been hospitalised, the expected number of hospitalisations for re-contact respondents and non-participants were the same. In addition, we predicted the number of hospitalisations using three multiple imputation approaches. We observed that the predictions from the MI-MNAR approach matched best with the true number of hospitalisations. These findings support the assumption which is utilised by the MI-MNAR approach.

The evaluation of assumptions (2) and (3) was based on the idea that the number of hospitalisations is associated with the health status. If the number of hospitalisations differs between re-contact respondents and non-participants, then there is likely to be a difference in distributions of health indicators between the groups. Otherwise, the distributions are assumed to be the same. As we used the hospitalisations before the study, the symptoms are not caused by the health condition during the survey date but are associated with them.

This makes us think that the hospitalisations before the study are a less convincing source of evidence than prospective follow-up data that have been used earlier to evaluate the assumptions for FINRISK 2007 [11]. If the follow-up data are available, then we recommend using them [11]. Otherwise, we recommend using the proposed method instead of not checking the assumptions at all. Unlike prospective follow-up data, the hospitalisation history data are readily available shortly after the study. In principle, hospitalisation history data could be used directly in the imputation model such that instead of just evaluating the assumptions (1)–(3) the imputations would be predicted based on the hospitalisation history data. How to optimally do this and the benefit of doing it are questions to be further investigated.

The setup for FINRISK 2012 was similar to FINRISK 2007, which allows a comparison between the studies. Using the data from the participants only, the point estimates for smoking prevalence were 21.8% in 2007 [11] and 19.6% in 2012. Similarly, the prevalences of heavy alcohol consumption were estimated as 5.2% in 2007 and 4.8% in 2012. Thus, based on the participants only, there seems to be a positive development.

The results change if the MI-MNAR approach is used. Then, the estimated prevalence of daily smoking appears as 27.1% in 2007 [11] and 23.7% in 2012. Estimated prevalences of heavy alcohol consumption are 6.8% in 2007 and 7.1% in 2012. Thus, there seem to be notable differences in the prevalence estimates between the approaches. The MI-MNAR approach produces the widest confidence intervals in comparison to the MI-MAR, MI-MAR-NR and participants approach, all of which are based on unrealistic assumptions.

As noted by many authors [4,8,11,21–23], missing data caused by non-participation is a serious problem in HES. Our results support the idea that re-contact data can improve the reliability of the health indicators of non-participants and provide information about the selectivity.

Although the assumption for MI-MNAR holds for FINRISK 2012 data, it may not hold for other HES. For example, the LEIDEN 85-plus study [24] observed that the mortality risk of re-contact respondents was similar to that of elderly participants. In such a situation, re-contact data were not useful for bias reduction. As the populations of FINRISK and LEIDEN 85-plus differ a lot, the results are not directly comparable.

According to our knowledge, re-contact data have only occasionally been collected in HES. Our results suggest that re-contact data can provide information about the health indicators of non-participants and selectivity of non-participation. Therefore, we recommend that HES would collect re-contact questionnaire data and that the same self-reported questions would be asked for re-contacts and initial participants to allow comparison.

Obtaining representative estimates about sensitive health indicators associated with selective non-participation is important for data-driven decision-making in national health policy. Our work shows that re-contact data have the potential to help reduce the selection bias. When used together with hospitalisation register data, the assumptions for which the estimation of population-level health indicators is based on can be evaluated soon after the survey.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Academy of Finland (grant number 266251).

References

[1] Torvik FA, Rognmo K and Tambs K. Alcohol use and mental distress as predictors of non-response in a general population health survey: the HUNT study. *Soc Psychiatry Psychiatr Epidemiol* 2012;47(5):805–16.

[2] Gray L, McCartney G, White IR, 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.

[3] Christensen AI, Ekholm O, Gray L, et al. What is wrong with non-participants? Alcohol-, drug- and smoking related mortality and morbidity in a 12-year follow up study of participants and non-participants in the Danish Health and Morbidity Survey. *Addiction* 2015;110(9):1505–12.

[4] Jousilahti P, Salomaa V, Kuulasmaa K, et al. 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. *Epidemiol Community Health* 2005;59(4):310–15.

[5] Thygesen LC, Johansen C, Keiding N, et al. Effects of sample attrition in a longitudinal study of the association between alcohol intake and all-cause mortality. *Addiction* 2008;103(7):1149–59.

[6] Larsen SB, Dalton SO, Schüz J, et al. Mortality among participants and non-participants in a prospective cohort study. *Eur J Epidemiol* 2012;27(11):837–45.

[7] Søgaard AJ, Selmer R, Bjertness E, et al. The Oslo Health Study: the impact of self-selection in a large, population-based survey. *Int J Equity Health* 2004;3:3. 10.1186/1475-9276-3-3.

[8] Tolonen H, Dobson A and Kulathinal S; WHO MONICA Project. Effect on trend estimates of the difference between survey respondents and non-respondents: results from 27 populations in the WHO MONICA Project. *Eur J Epidemiol* 2005;20(11):887–98.

[9] Drivsholm T, Eplov LF, Davidsen M, et al. Representativeness in population-based studies: a detailed description of non-response in a Danish cohort study. *Scand J Public Health* 2006;34(6):623–31.

[10] Kopra J, Härkänen T, Tolonen H, et al. Correcting for non-ignorable missingness in smoking trends. *Stat* 2015;4(1):1–14.

[11] Karvanen J, Tolonen H, Härkänen T, et al. Selection bias was reduced by re-contacting nonparticipants. *J Clin Epidemiol* 2016;76(1):209–17.

[12] Borodulin K, Saarikoski L, Lund L, et al. *Kansallinen FINRISKI 2012 terveystutkimus Osa I: Tutkimuksen toteutus ja menetelmät*. Report, Osa I, THL, (2013).

[13] National Institute for Health and Welfare. Care Register for Health Care, https://www.thl.fi/en/web/thlfi-en/statistics/information-on-statistics/register-descriptions/care-register-for-health-care (accessed 13 October 2017).

[14] R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing,

Vienna, Austria, https://www.R-project.org/ (2015, accessed 13 October 2017).

[15] van Buuren S and Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011;45(3):1–67.

[16] Little RJA and Rubin DB. *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons, 2014.

[17] Lin IF and Schaeffer NC. Using survey participants to estimate the impact of nonparticipation. *Public Opin Q* 1995;59(2):236–58.

[18] Boniface S, Scholes S, Shelton N, et al. Assessment of non-response bias in estimates of alcohol consumption: applying the continuum of resistance model in a general population survey in England. *PloS One* 2017;12(1):e0170892.

[19] van Buuren S. *Flexible imputation of missing data*. Boca Raton, FL: CRC Press, 2012.

[20] Zeileis A, Kleiber C and Jackman S. Regression models for count data in R. *J Stat Softw* 2008;27(8):1–25.

[21] van Loon AJM, Tijhuis M, Picavet HSJ, et al. Survey non-response in the Netherlands: effects on prevalence estimates and associations. *Ann Epidemiol* 2004;13(2):105–10.

[22] Härkänen T, Karvanen J, Tolonen H, et al. Systematic handling of missing data in complex study designs–experiences from the Health 2000 and 2011 Surveys. *J Appl Stat* 2016;43(15):2772–90.

[23] Nummela O, Sulander T, Helakorpi S, et al. Register-based data indicated nonparticipation bias in a health study among aging people. *J Clin Epidemiol* 2011;64:1418–25.

[24] Bootsma-Van Der Wiel A, Van Exel E, De Craen AJM, et al. A high response is not essential to prevent selection bias: results from the Leiden 85-plus study. *J Clin Epidemiol* 2002;55(11):1119–25.

II

Correcting for non-ignorable missingness in smoking trends.

Kopra J., Härkänen T., Tolonen H. and Karvanen J.

Stat, 4(1):1–14. (2015)

A correction has been published as:

Stat, 6(1):202–203. (2017)

Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

Correcting for non-ignorable missingness in smoking trends

Juho Kopra[a,*], Tommi Härkänen[b], Hanna Tolonen[b] and Juha Karvanen[a]

Received 20 November 2014; Accepted 12 January 2015

Data missing not at random (MNAR) are a major challenge in survey sampling. We propose an approach based on registry data to deal with non-ignorable missingness in health examination surveys. The approach relies on follow-up data available from administrative registers several years after the survey. For illustration, we use data on smoking prevalence in Finnish National FINRISK study conducted in 1972–97. The data consist of measured survey information including missingness indicators, register-based background information and register-based time-to-disease survival data. The parameters of missingness mechanism are estimable with these data although the original survey data are MNAR. The underlying data generation process is modelled by a Bayesian model. The results indicate that the estimated smoking prevalence rates in Finland may be significantly affected by missing data. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: health examination survey; missing data; non-participation; registry data; smoking prevalence; survey sampling

1 Introduction

Participation rates in health examination surveys (HES) have been declining over the years in many countries. The declining participation rates inflict the estimation of health indicators in many ways. First, the low participation rates compromise the population representativeness of the sample because the participants and non-participants differ from each other. The non-participants are more often smokers (Shahar et al., 1996; Tolonen et al., 2005) and have higher risk of death (Jousilahti et al., 2005; Harald et al., 2007) compared with the participants. It has also been found that the non-participants tend to be men (van Loon et al., 2003; Sogaard et al., 2004), younger persons (Sogaard et al., 2004) and single (Shahar et al., 1996; Sogaard et al., 2004; Tolonen et al., 2005). Generally, the non-participants have been found to have lower socio-economic status (Jackson et al., 1996; van Loon et al., 2003; Drivsholm et al., 2006; Harald et al., 2007) and lower education (Shahar et al., 1996; Sogaard et al., 2004; Tolonen et al., 2005) than the participants. Second, the declining trends in participation rates may distort the trends of the estimated health indicators. Especially, if smokers, heavy alcohol users and obese are less eager to participate than they were decades ago, the trends of the health indicators may look more positive than they should.

In statistical terms, data from HES are missing not at random (MNAR), and consequently, the missingness mechanism cannot be ignored in the analysis (Little & Rubin, 2002). Although dealing with non-ignorable missingness is challenging in general, there are some methods for this. One of these is making functional assumptions for the joint

[a]Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, FI-40014, Finland
[b]National Institute for Health and Welfare, Helsinki, FI-00271, Finland
*Email: juho.j.kopra@jyu.fi

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

distribution of missing data and observed values (Little, 1993; Ekholm & Skinner, 1998). This is usually accompanied with a sensitivity analysis for evaluating the effect of assumed missingness mechanism (van Buuren et al., 1999). If study design is longitudinal, the modelling of non-ignorable missingness may be based on partially available repeated measurements (Ibrahim & Molenberghs, 2009). Recently, a subsample ignorable likelihood approach (Little & Zhang, 2011) was proposed for situations, where full data are available for some variables, while the other variables have missing data.

We propose an approach to correct for non-ignorable missingness in situations where follow-up data are available for both participants and non-participants. Finland is one of the few countries where follow-up data for the entire survey sample can be obtained through a record linkage to the administrative registers. Naturally, the follow-up data will not be available right after the survey but only many years later. Without further assumptions, the trends of health indicators can be therefore corrected only retrospectively.

As an illustration for our approach, we use the data from the National FINRISK studies (Laatikainen et al., 2003; Harald et al., 2007), which are one of the data sources used to evaluate public health in Finland. The data from the surveys carried out in 1972, 1977, 1982, 1987, 1992 and 1997 are included. The participation to the physical measurements have decreased from 95% in 1972 to 74% in 1997. Note that in the next section, we define participation differently. Under the decreasing participation, we estimate the prevalence of smoking utilizing the follow-up data available from the registers.

The relevant details of the FINRISK surveys are presented in Section 2. In Section 3, a Bayesian model is built for the analysis of non-ignorable missing data. Section 4 compares the trends for non-ignorable and ignorable approaches, and Section 5 concludes the paper.

2 FINRISK data and linked register data

The National FINRISK Study (earlier North Karelia Project) data arose from a setup where the original aim was to intervene to people of North Karelia via a health education campaign. Later, the data have been collected every five years to measure the risk factors of key diseases and to monitor public health. In addition to North Karelia, the neighbour province of Northern Savonia has been included in studies since the beginning. Later, Turku and Loimaa area, Helsinki and Vantaa area and Oulu province have joined the survey. The data from the surveys conducted in 1972–97 are used in this paper.

Sampling frame for the surveys has been the National Population Register. The survey design has changed over the study years (Table I), but at each study, the sampling has been stratified among the participating areas. In 1972, the sampling was systematic on birthdays, and people aged between 25 and 59 years were sampled. In 1977, the simple random samples was drawn from people aged between 30 and 64 years. In 1982, the survey was balanced between the 10-year age groups, and 25- to 64-year-old people were sampled. In years 1987, 1992 and 1997, the sampling design was balanced sampling between 10-year age groups within genders. In 1997, the eligible age was extended to 25–74 years in North Karelia and in Helsinki and Vantaa area.

The participation is defined as answering to the question about daily smoking. This definition leads to lower participation rates than reported elsewhere because some individuals participated otherwise but skipped the smoking questions. The participation seem to depend on age and gender but possibly also on smoking, which is to be investigated. The age dependency of the participation rate and its change over the period 1972–97 is shown in Figure 1. Smoking, together with other health indicators, was measured by using a multi-page questionnaire. Smoking questions classified each person either non-smokers, ex-smoker or current smoker. We model smoking using two classes, where the ex-smokers and non-smokers are considered as the same.

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

Table I. Description of sampling design and eligible cohorts and areas. Area codes are North Karelia, 2; Northern Savonia, 3; Turku and Loimaa, 4; Helsinki and Vantaa, 5; and Oulu province, 6. (* = Upper eligible age is 75 years for areas 2 and 5.) Marginal participation percentages are given for the studies. Two rightmost columns indicate the survey sample size and the corresponding count of observed lung cancer and COPD events for persons selected to the sample over the follow-up period.

Year	Cohort	Areas	Sampling design	Participation	Sample size	Events
1972	25–59	2,3	Systematic on birthdate, balanced between areas	86.0%	12,377	420
1977	30–64	2,3	Simple random sampling, balanced between areas	88.1%	11,319	373
1982	25–64	2,3,4	Balanced between 10-year age groups within areas	80.0%	11,332	281
1987	25–64	2,3,4	Balanced between 10-year age groups within areas within gender	79.9%	7,893	127
1992	25–64	2,3,4,5	Balanced between 10-year age groups within gender and areas	76.2%	7,895	97
1997	25–64*	2,3,4,5,6	Balanced between 10-year age groups within gender and areas	71.3%	11,423	140

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

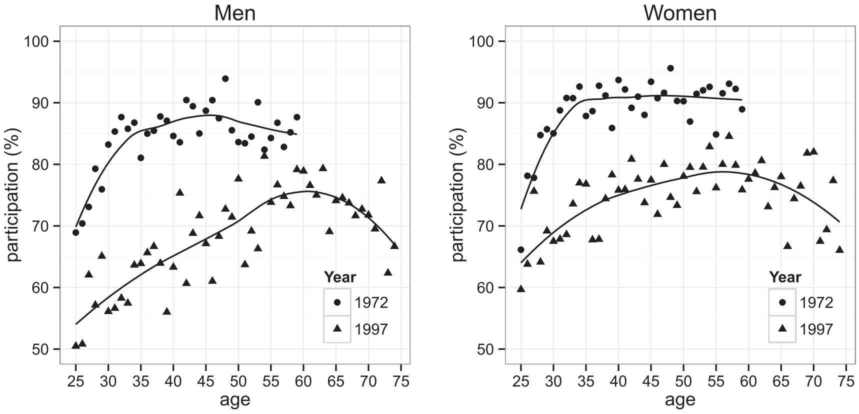

Figure 1. Participation rate as a function of age in 1972 and 1997. Each circle and triangle represents the observed proportion of participants within one-year age group over all study areas studied that year. The graph shows that the participation rates have decreased in all age groups for both men and women. The solid lines are calculated using locally weighted regression (Cleveland, 1979).

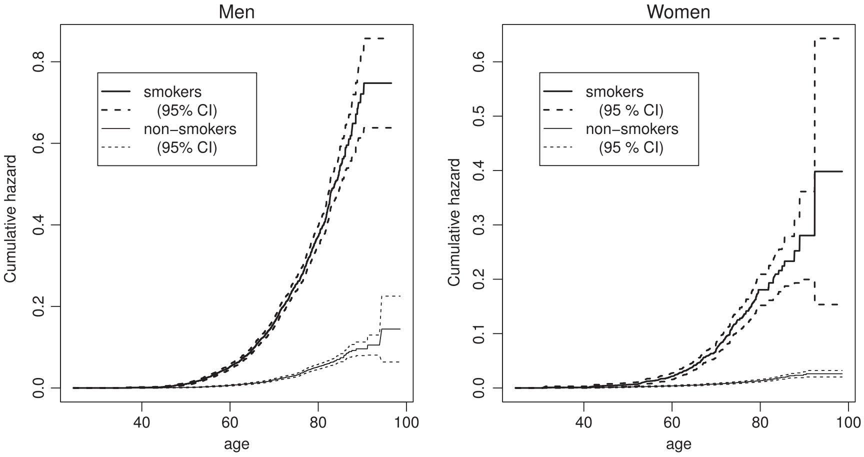

Figure 2. Cumulative hazard estimates with confidence intervals (CI) for smoking-based diseases of lung cancer and chronic obstructive pulmonary disease. Graphs are produced using the participant data only.

The sources of the follow-up data are Care Register for Health Care (HILMO) (National Institute for Health and Welfare, 2014) and the cause of death data (Statistics Finland, 2014). The follow-up data are linked to the survey data by personal identification number. The follow-up data contain the date and the cause of death and the cause of hospitalization. The diseases considered here are lung cancer (International Classification of Diseases (ICD) 10: C34, ICD9/ICD8: 162) and chronic obstructive pulmonary disease (COPD) (ICD10: J41-J44, ICD9/ICD8: 491-492) for which smoking is known to be the main risk factor (Doll & Hill, 1956; Wynder & Hoffmann, 1994; Mannino & Buist, 2007; Cornfield et al., 2009). The follow-up data are available for all persons (participants and non-participants) selected to the FINRISK samples. The effect of smoking to the onset of lung cancer and COPD for men and women is illustrated in Figure 2.

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

We denote our variables as follows. The smoking indicator variable is denoted as Y_i for person i. Background variables $X_i = (x_{1i}, x_{2i}, x_{3i}, x_{4i})$ for person i include the age at the beginning of the follow-up x_{1i}, area x_{2i} and gender x_{3i}, which origin from the registers. The variable x_{4i} is the study year.

The sample indicator $m_{1i} = 1$ indicates that person i has been chosen to a survey sample, and participation indicator $M_{2i} = 1$ indicates that he or she has participated to the survey. If $m_{1i} = 0$, then M_{2i} must also be 0 because people outside of the survey sample cannot take part. Variables X_i are observed for both the participants and non-participants, while Y_i is observed only from the participants.

The follow-up data consist of time-to-event-variable T_i and event indicator r_i, where T_i is the age at the diagnosis of the disease, i.e. the onset of lung cancer or COPD. Variable T_i is observed for the participants and non-participants. If a person has not been diagnosed until the end of follow-up period (31 December 2011), or if person dies for other causes, then the time-to-event-variable becomes right censored. In the case of right censoring, we know only that $T_i > c_i$ where c_i is person's age at censoring or age at death. The date of diagnosis can be the same as date of death, if person has not been diagnosed earlier, and lung cancer or COPD is the cause of death. If person recovers from lung disease and becomes repeatedly diagnosed, the time-to-event-variable holds the time of the earliest diagnosis.

3 Bayesian model for non-participation and smoking

3.1. Dependency structure and modelling assumptions

We present the structure of the model in Figure 3 using the concept of causal model with design (Karvanen, 2014). Figure 3 represents a causal model at the bottom where background variables $X_i = (x_{1i}, x_{2i}, x_{3i}, x_{4i})$ affect the probability of smoking $P(Y_i)$ and the risk of lung disease $P(T_i)$. In addition, smoking also has an effect on the risk of lung disease. These relations are described as arrows $X_i \rightarrow T_i, X_i \rightarrow Y_i$ and $Y_i \rightarrow T_i$ in the model graph. The causal relations of smoking and lung cancer (Doll & Hill, 1956; Wynder & Hoffmann, 1994; Cornfield et al., 2009) and smoking and COPD (Mannino & Buist, 2007) are known to exist. Also, it has been observed that the prevalence of smoking varies depending on the area, gender and age (Peltonen et al., 2008; Borodulin et al., 2013). Persons belonging to the sample have $m_{1i} = 1$ and are selected from population Ω, which in this case is the general Finnish population in geographically defined areas and age groups specified earlier. Sampling is based on the background register data, which is why we have $X_i \rightarrow m_{1i}$ in Figure 3. Participation, which is indicated by $M_{2i} = 1$, is affected by background variables ($X_i \rightarrow M_{2i}$) and smoking ($Y_i \rightarrow M_{2i}$). People may participate only if they are selected to the sample, which is indicated by the arrow $m_{1i} \rightarrow M_{2i}$ in the graph. If a person participates, he or she has $M_{2i} = 1$ and thus $Y_i^* = Y_i$. Otherwise, smoking indicator is missing $Y_i^* =$ NA. The background information as well as survival information T_i^* are collected for all persons in the sample. The follow-up variable T_i is a vector of two elements, the actual time variable t_i, either for the event-time or censoring time, and an indicator variable for censoring, denoted as r_i. The notation for this is

$$T_i = (t_i, r_i) = \begin{cases} (t_i, 0), & \text{if an event is observed} \\ (t_i, 1), & \text{if an event is right censored.} \end{cases}$$

The observed T_i^* is then defined as

$$T_i^* = \begin{cases} T_i & \text{if person i belongs to a sample: } m_{1i} = 1 \\ \text{NA}, & \text{if person i does not belong to a sample: } m_{1i} = 0. \end{cases}$$

The censoring due to deaths other than lung cancer or COPD is informative because smoking is a risk factor for many common causes of death. The usual way to deal with this kind of informative censoring is to define an additional endpoint for other deaths and use a competing risk model (Kalbfleisch & Prentice, 2002). However, this would create

Copyright © 2015 John Wiley & Sons, Ltd

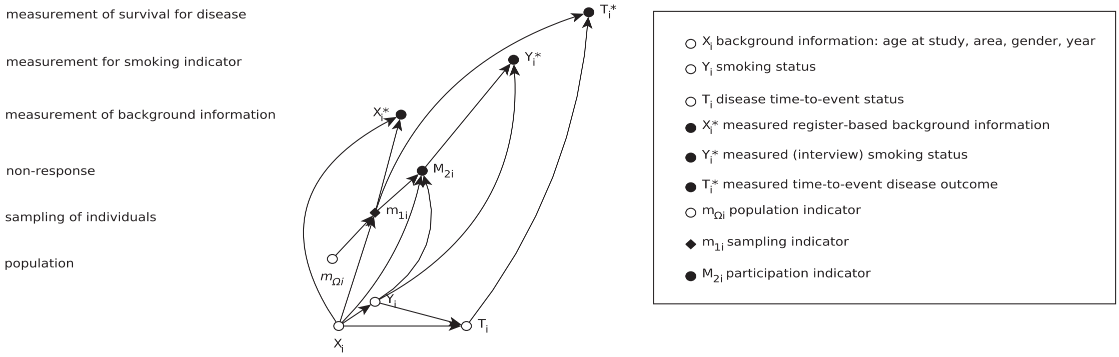
measurement of survival for disease
measurement for smoking indicator
measurement of background information
non-response
sampling of individuals
population
T_i*
Y_i*
X_i*
M_2i
m_1i
m_Ωi
Y_i
X_i
T_i
○ X_i background information: age at study, area, gender, year
○ Y_i smoking status
○ T_i disease time-to-event status
● X_i* measured register-based background information
● Y_i* measured (interview) smoking status
● T_i* measured time-to-event disease outcome
○ m_Ωi population indicator
◆ m_1i sampling indicator
● M_2i participation indicator

Figure 3. Illustration of variable dependencies and the data-collection process.

new problems because we would implicitly assume that all differences in the mortality between participants and non-participants are due to smoking. In reality, participants and non-participants differ also by many other risk factors, which work as confounders. Therefore, we have chosen to use only smoking-specific survival outcome in the analysis and to treat the censoring as non-informative. The implications to the results and alternative approaches are discussed in Section 5.

In Figure 3, the non-participation depends on smoking status Y_i, which means that the missingness mechanism is non-ignorable. In general, the non-ignorable missingness mechanism is not estimable from data. To overcome this issue, we use the follow-up data to make an additional assumption on the missingness mechanism.

We want to estimate the smoking prevalence for the whole sample, so we need to estimate the distributions

$$P(Y_i) = P(M_{2i} = 1)P(Y_i|M_{2i} = 1) + P(M_{2i} = 0)P(Y_i|M_{2i} = 0) \qquad i \in \Omega. \tag{1}$$

On the right-hand side of Equation (1), the probability of smoking for non-participants $P(Y_i|M_{2i} = 0)$ cannot be estimated using the observed data without making further assumptions. This may be written as

$$P(Y_i|M_{2i} = 0) = \int\int P(Y_i|T_i, X_i, M_{2i} = 0)P(T_i, X_i|M_{2i} = 0)dX_i dT_i, \qquad i \in \Omega$$

where $P(Y_i|T_i, X_i, M_{2i} = 0)$ is not estimable but $P(T_i, X_i|M_{2i} = 0)$ is estimable from observed data. We now assume that

$$P(Y_i|T_i, X_i, M_{2i} = 0) = P(Y_i|T_i, X_i, M_{2i} = 1), \qquad i \in \Omega \tag{2}$$

which means that, given the observations T_i and X_i, additional observation $M_{2i} = 1$ or $M_{2i} = 0$ does not give us any further understanding about the distribution of Y_i. Thus, for the rest of our paper, we restrict the models of interest to the cases for which Equation (2) holds. Now, the smoking prevalence (1) can be estimated if the probabilities $P(M_{2i} = 1)$, $P(Y_i|M_{2i} = 1)$, $P(M_{2i} = 0)$, $P(T_i, X_i|M_{2i} = 0)$ and $P(Y_i|T_i, X_i, M_{2i} = 1)$ can be estimated. The assumption (2) can be justified if the relation $Y_i \rightarrow T_i$ is strong; i.e. the early onset of lung cancer or COPD is a strong indicator of smoking. In practice, the model parameters for relations of X_i, Y_i and T_i are estimated using data from the participants only.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

3.2. Construction of posterior distribution

The model consists of two sub-models: a survival model for T_i^* and a logistic regression model for the smoking indicator Y_i^*. Next, the parametric forms for sub-models are considered.

Time-to-disease variable $T_i^*|m_{1i} = 1$ is assumed to follow Weibull distribution with a common shape parameter a and scale parameter b_i varying person by person. The distribution is left-truncated by the person's age $t_{0i} = x_{1i}$ at the beginning of follow-up. The likelihood contribution for observed disease cases can be written as

$$p(T = t_{1i}|a, b, r_i = 1, T > t_{0i}) = \frac{abt_{1i}{}^{a-1}\exp\left(-bt_{1i}{}^{a}\right)}{(1 - F(t_{0i}))} \qquad \text{for} \quad t_{1i} > t_{0i},$$

where $F(t)$ is cumulative distribution function for Weibull distribution. For censored cases $i : r_i = 0$, the likelihood contribution is the survival function

$$S(T > t_{1i}|a, b, r_i = 0, T > t_{0i}) = \exp\left(-b\left(t_{1i}{}^{a} - t_{0i}{}^{a}\right)\right) \qquad \text{for} \quad t_{1i} > t_{0i}.$$

Parameter b_i varies person by person based on the covariate measurements

$$\begin{aligned} \log(b_i) = \gamma_0 &+ \gamma_1 x_{3i} + \gamma_2 Y_i + \gamma_3 x_{3i} Y_i \\ &+ \gamma_{43} A_{3i} + \gamma_{44} A_{4i} + \gamma_{45} A_{5i} + \gamma_{46} A_{6i} \\ &+ \gamma_{53} x_{3i} A_{3i} + \gamma_{54} x_{3i} A_{4i} + \gamma_{55} x_{3i} A_{5i} + \gamma_{56} x_{3i} A_{6i} \\ &+ \gamma_{62} D_{2i} + \gamma_{63} D_{3i} + \gamma_{64} D_{4i} + \gamma_{65} D_{5i} + \gamma_{66} D_{6i} \\ &+ \gamma_{72} x_{3i} D_{2i} + \gamma_{73} x_{3i} D_{3i} + \gamma_{74} x_{3i} D_{4i} + \gamma_{75} x_{3i} D_{5i} + \gamma_{76} x_{3i} D_{6i}, \end{aligned} \tag{3}$$

where parameter γ_0 corresponds to lung disease risk of non-smoking men at baseline (year 1972, North Karelia), γ_1 indicates the difference of risks for non-smoking men and women, γ_2 indicates the effect of smoking for men at baseline and γ_3 describes how disease risk for smoking women is different from the risk of smoking men (at baseline). The $\gamma_{42}, \ldots, \gamma_{46}$ stand for how the other areas differ from the baseline area (North Karelia) for men. The coefficients $\gamma_{53}, \ldots, \gamma_{56}$ describe how the last-mentioned quantities differ between the women and men. The $\gamma_{62}, \ldots, \gamma_{66}$ are the differences of the study year to the baseline study (year 1972) for men, and $\gamma_{72}, \ldots, \gamma_{76}$ are the differences of women and men for that particular study year. In Equation (3), the variables $A_{2i}, \ldots, A_{6i}$ are indicators for the study area such that $A_{2i} = 1$ for the North Karelia (area 2), $A_{3i} = 1$ for the Northern Savonia (area 3), $A_{4i} = 1$ for Turku and Loimaa (area 4), $A_{5i} = 1$ for Helsinki and Vantaa (area 5) and $A_{6i} = 1$ for Oulu province (area 6). Similarly, $D_{1i}, \ldots, D_{6i}$ are indicators about the study year such that $D_{1i} = 1$ for 1972, $D_{2i} = 1$ for 1977, $D_{3i} = 1$ for 1982, $D_{4i} = 1$ for 1987, $D_{5i} = 1$ for 1992 and $D_{6i} = 1$ for 1997.

The smoking indicator is modelled also using logistic regression. The effects of gender x_{3i}, year of birth $x_{birth,i} = x_{4i} - x_{1i}$ and study year x_{4i} are included in the model. We assume that the smoking indicator is Bernoulli distributed

$$Y_i \sim \text{Bernoulli}(s_i)$$

with probability s_i such that

$$\text{logit}(s_i) = \alpha_{0,a,u,g} + \alpha_{1,a,u,g}(x_{birth,i} - 1930), \tag{4}$$

where $a = x_{2i}$ is area, $g = x_{3i}$ is gender and $u = x_{4i}$ is study year for person i. The coefficient $\alpha_{0,a,u,g}$ represents the intercept term for persons living in area a, of gender g, who were born in 1930 and were selected to the sample in year u. The year 1930 was chosen as a reference level because all the studies have some participants who were born in 1930. The coefficients $\alpha_{1,a,u,g}$ represents the impact of year of birth to the probability of smoking.

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

The information on the area (North Karelia or Northern Savonia) is missing for non-participants (2,664 in total) in 1972 and 1977. This missingness is due to accidentally lost data. These values are imported using multiple imputation with fixed probabilities $P(\text{area was Northern Savonia} \mid 1972) = 0.495$ and $P(\text{area was Northern Savonia} \mid 1977) = 0.493$. The imputation is not necessary for model fitting purposes but is needed for the comparison of the areawise smoking trends.

3.3. Model fitting and model diagnostics

The model was built and fitted using Just Another Gibbs Sampler (Plummer, 2003), which is a tool for Bayesian analysis (Gelman et al., 2013) of graphical models using Markov chain Monte Carlo (MCMC) (Robert & Casella, 2004). For all parameters, the prior distributions were set as normal distributions with zero mean and variance $\sigma^2 = 1,000$. Regarding the scale of the parameters, these priors are non-informative. Eight chains were run in parallel. Each of the chains had 200,000 iterations from which the first 40,000 were discarded as a burn-in. From the remaining 160,000 iterations, the values of each 250th iteration were stored to produce eight final thinned chains of 640 iterations. In total, we have $640 * 8 = 5,120$ realizations from posterior to use.

The MCMC convergence was monitored by Brooks–Gelman–Rubin convergence diagnostic (Brooks & Gelman, 1998). The diagnostics of all parameters were below 1.01 when values below 1.05 indicate convergence. One of the MCMC chains of the final model is visualized for two parameters in Figure 4. The Figure shows that the Weibull shape parameter is less well mixed than the other parameter. This is due to large autocorrelation caused by dependency on Weibull scale parameter. The better mixing on the smoking coefficient γ_2 is also visualized in the Figure. The majority of the parameters have good mixing. Posterior summaries of regression coefficients are given in Tables A.1 and A.2; see Appendix A.

The model diagnostics included a graphical comparison of the posterior predictive distribution against the observed values. The model was concluded to have a good fit to the data.

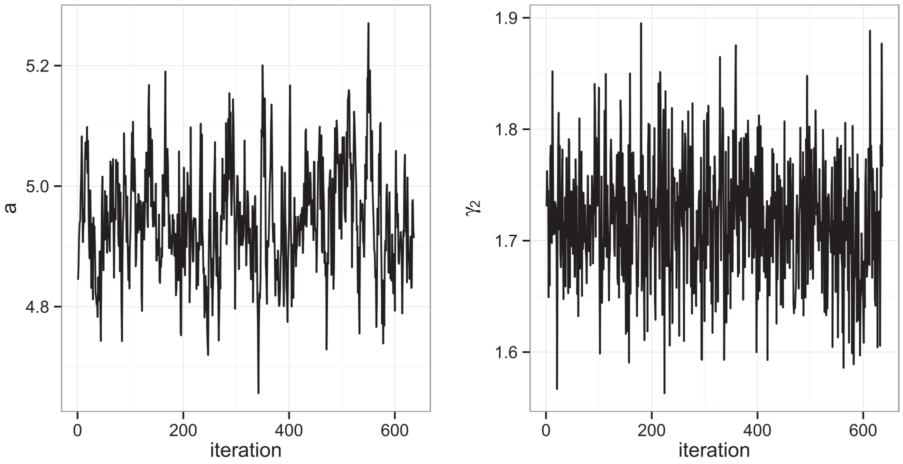

Figure 4. Chain plots of Markov chain Monte Carlo computation. Left: Weibull shape parameter *a*. Right: regression coefficient of smoking variable γ_2 of survival model.

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

4 Comparison of corrected and uncorrected smoking trends

To obtain knowledge about the smoking prevalence for the study populations, we apply data augmentation (Tanner & Wong, 1987) to impute the missing values of smoking for non-participants and take into account censoring of T_i. Because we apply Bayesian inference, the imputations are drawn from the posterior predictive distribution. First, the posterior samples of the regression coefficients are obtained using MCMC and participants data. Imputations of the smoking indicator for non-participants are drawn using the following procedure, which we implemented in R (R Core Team, 2014). The imputation depends on whether the event is observed or censored. If T_i is censored, then first event-time $\tilde{T}_i$ for T_i is generated using

$$\tilde{T}_i \sim P(\tilde{T}_i|X_i) = P(\tilde{T}_i|X_i, Y_i' = 1)P(Y_i' = 1|X_i) + P(\tilde{T}_i|X_i, Y_i' = 0)P(Y_i' = 0|X_i).$$

After that, use the imputed event-time $\tilde{T}_i$ to simulate $\tilde{Y}_i \sim P(\tilde{Y}_i|\tilde{T}_i, X_i)$. If T_i is observed, then simulate $\tilde{Y}_i \sim P(\tilde{Y}_i|T_i, X_i)$ straightforwardly based on the observed event-time.

After the imputation, the survey sampling design has to be taken into account. We may treat data with each imputation as a full dataset. To provide area-specific population-level estimates, we may then utilize inverse sampling probability weights (Lehtonen & Pahkinen, 2004). In addition to utilizing the sampling weights, the estimates were adjusted using WHO Scandinavian standardization weights (Ahmad et al., 2001) in order to make the smoking rates internationally comparable. As an outcome, we obtain area-specific trend estimates for both genders corresponding to each imputation. These trends can be considered as samples from the posterior distribution of the trends. The estimated model-based corrected trends are compared with the corresponding original trends in Figure 5 for North Karelia. The original or uncorrected trends were produced from the participant data only. The adjustment for sampling design and the WHO weights was the same as for the model-based trends.

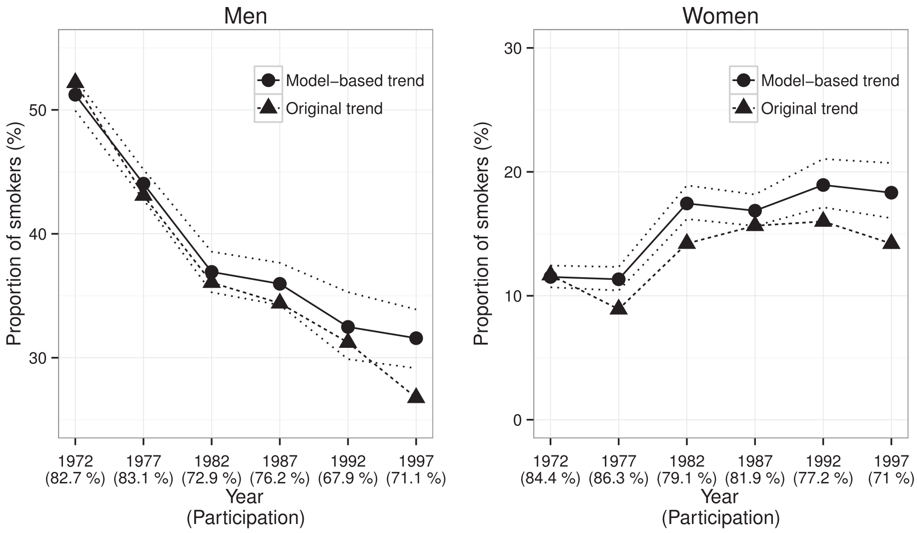

Figure 5. Model-based trend and original trend for men (left) and women (right) in North Karelia province. North Karelia was chosen because of the most visible change in the trends among the areas. Two dotted lines represent 95% credible interval (CI) of the posterior distribution for corrected trends. Both the model-based and the original trend use WHO Scandinavian standardization weights.

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

Table II. Observed and model-based smoking proportions for the study in 1997 adjusted using WHO Scandinavian standardization weights. The two rightmost columns describe the 95% credible intervals of model-based trends. Participant smoking is the same as "original trend" in Figure 5.

Gender	Area	Participant smoking (%)	Model-based total smoking (%)	95% Credible interval	
Men	North Karelia	26.8	31.6	29.2	33.9
Men	Northern Savonia	30.7	31.8	29.5	34.0
Men	Turku and Loimaa	32.4	33.7	31.1	36.2
Men	Helsinki and Vantaa	26.1	32.7	30.1	35.6
Men	Oulu province	30.1	32.3	29.5	35.2
Women	North Karelia	14.2	18.3	16.3	20.7
Women	Northern Savonia	17.0	19.1	17.1	21.1
Women	Turku and Loimaa	20.5	23.6	21.3	26.0
Women	Helsinki and Vantaa	22.6	27.7	25.3	30.4
Women	Oulu province	19.1	22.2	20.0	24.3

In Figure 5, the difference between the trends increases as the participation rate decreases. In addition, it seems that the difference of the trends in most time-points is larger for women than for men. On the other hand, the largest difference in the corrected and non-corrected prevalence estimates is 6.6 percentage points (relative difference of 25%), which is observed for men in Helsinki and Vantaa in 1997. The comparison of the model-based and original smoking prevalence trends for the study year 1997 is presented in Table II.

5 Discussion

We have proposed an approach to overcome the challenges with non-ignorable missing data in epidemiological studies and have applied it to estimate the population trends of smoking in Finland in 1972–97. The approach uses follow-up data to obtain information on risk factors missing at baseline. Thanks to the administrative registers in Finland, the follow-up data are available also for non-participants. Smoking has been selected as the risk factor of interest because it is a strong risk factor of lung cancer and COPD and potentially has an effect to the decision on the HES participation.

We evaluated the proportion of smokers combining the available information from both the participants and non-participants for the FINRISK study. Our results indicate that the levels of smoking prevalence is affected when the information provided by lung cancer and COPD time-to-event data is accounted to provide an estimate for the smoking of non-participants.

In general, statistical modelling under the non-ignorable missingness requires external information on the missingness mechanism. It can be argued that the inclusion of follow-up data provides the information needed. The situation can be formally described using causal models with design and then modelled by a Bayesian model. The idea of utilizing existing causal knowledge to fix non-ignorable missingness is not restricted to survival models.

The approach is limited by the availability of follow-up data. It takes years or decades until the follow-up data on lung cancer and COPD can be used to model the missing data mechanism. It is unclear to what extent the approach can be applied in other countries because register-based baseline and follow-up data sets are not usually available for non-participants. Although the approach may not be directly applicable in a study, the results from other similar studies, where the approach has been applied, may provide a starting point for the prior setting and the sensitivity analyses.

Censoring was treated as non-informative, which may cause some bias to the estimates. As smoking is a risk factor for many common causes of death, an individual censored due to other deaths is more likely to be a smoker than an individual censored due to the end of the follow-up. It is therefore expected that the actual proportions of smokers

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

could be even higher than the corrected proportions reported here. Improved estimation would require a competing risk approach with a comprehensive set of risk factors and a number of disease-specific endpoints. This is left as future work.

Inclusion of information about smoking as a time-dependent process would yield more realistic expressions of smoking in different age groups. The effect of smoking years could be then considered as a covariate for the lung diseases. With the current model, it is assumed that observed lung disease diagnosis, e.g. at age 50 years is equally strong indication about smoking, no matter if the person is diagnosed five or 25 years after the survey. In reality, individuals may have started or stopped smoking after the survey was conducted.

The presented approach may be utilized with data arising in forthcoming FINRISK surveys. In addition, the model could be used to give recommendations on the sample size and the stratification.

Our work reminds that data with MNAR situation may be changed to missing at random using additional assumption and external information. This allows us to provide estimates that describe the whole population instead of the restricted sample of survey participants.

Appendix A: Regression coefficients

Table A.1. Posterior summaries of the estimated parameters for the smoking model, reduced to the parameters of North Karelia.

Description of related variable	Parameter	Mean	SD	2.5%	97.5%
Men born at 1930 in 1972	$\alpha_{0,1972,1,2}$	0.086	0.043	0.003	0.168
Men born at 1930 in 1977	$\alpha_{0,1977,1,2}$	−0.294	0.047	−0.384	−0.203
Men born at 1930 in 1982	$\alpha_{0,1982,1,2}$	−0.672	0.069	−0.806	−0.538
Men born at 1930 in 1987	$\alpha_{0,1987,1,2}$	−0.888	0.084	−1.052	−0.724
Men born at 1930 in 1992	$\alpha_{0,1992,1,2}$	−1.092	0.159	−1.409	−0.779
Men born at 1930 in 1997	$\alpha_{0,1997,1,2}$	−1.449	0.107	−1.661	−1.244
Women born at 1930 in 1972	$\alpha_{0,1972,2,2}$	−2.106	0.070	−2.242	−1.971
Women born at 1930 in 1977	$\alpha_{0,1977,2,2}$	−2.452	0.086	−2.625	−2.287
Women born at 1930 in 1982	$\alpha_{0,1982,2,2}$	−2.361	0.111	−2.583	−2.153
Women born at 1930 in 1987	$\alpha_{0,1987,2,2}$	−2.412	0.128	−2.670	−2.164
Women born at 1930 in 1992	$\alpha_{0,1992,2,2}$	−2.599	0.219	−3.039	−2.183
Women born at 1930 in 1997	$\alpha_{0,1997,2,2}$	−2.833	0.204	−3.239	−2.449
Difference of year of birth to 1930 (men in 1972)	$\alpha_{1,1972,1,2}$	0.001	0.004	−0.007	0.009
Difference of year of birth to 1930 (men in 1977)	$\alpha_{1,1977,1,2}$	0.008	0.005	−0.0004	0.017
Difference of year of birth to 1930 (men in 1982)	$\alpha_{1,1982,1,2}$	0.013	0.005	0.003	0.022
Difference of year of birth to 1930 (men in 1987)	$\alpha_{1,1987,1,2}$	0.019	0.005	0.009	0.028
Difference of year of birth to 1930 (men in 1992)	$\alpha_{1,1992,1,2}$	0.017	0.007	0.002	0.032
Difference of year of birth to 1930 (men in 1997)	$\alpha_{1,1997,1,2}$	0.029	0.005	0.019	0.038
Difference of year of birth to 1930 (women in 1972)	$\alpha_{1,1972,2,2}$	0.043	0.007	0.030	0.056
Difference of year of birth to 1930 (women in 1977)	$\alpha_{1,1977,2,2}$	0.050	0.008	0.034	0.066
Difference of year of birth to 1930 (women in 1982)	$\alpha_{1,1982,2,2}$	0.057	0.007	0.044	0.070
Difference of year of birth to 1930 (women in 1987)	$\alpha_{1,1987,2,2}$	0.049	0.006	0.037	0.062
Difference of year of birth to 1930 (women in 1992)	$\alpha_{1,1992,2,2}$	0.049	0.009	0.031	0.066
Difference of year of birth to 1930 (women in 1997)	$\alpha_{1,1997,2,2}$	0.049	0.007	0.035	0.064

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

Table A.2. Posterior summaries of the estimated parameters for the survival model (includes all parameters).

Description of related variable	Parameter	Mean	SD	2.5%	97.5%
Weibull shape parameter	a	4.257	0.111	4.041	4.475
Intercept (men)	γ_0	−21.848	0.501	−22.817	−20.859
Gender (women)	γ_1	−1.352	0.164	−1.665	−1.031
Smoking	γ_2	1.772	0.061	1.653	1.893
Interaction of smoking and gender	γ_3	0.559	0.116	0.328	0.786
Northern Savonia	γ_{43}	0.070	0.054	−0.036	0.175
Turku and Loimaa	γ_{44}	−0.298	0.085	−0.466	−0.134
Helsinki and Vantaa	γ_{45}	−0.389	0.131	−0.652	−0.139
Oulu province	γ_{46}	−1.290	0.300	−1.931	−0.743
Interaction of Northern Savonia and women	γ_{53}	−0.274	0.131	−0.528	−0.023
Interaction of Turku and Loimaa and women	γ_{54}	0.654	0.161	0.337	0.970
Interaction of Helsinki and Vantaa and women	γ_{55}	0.509	0.241	0.037	0.963
Interaction of Oulu province and women	γ_{56}	1.072	0.477	0.104	2.006
Year 1977	γ_{62}	−0.242	0.074	−0.382	−0.094
Year 1982	γ_{63}	0.017	0.074	−0.125	0.164
Year 1987	γ_{64}	−0.090	0.105	−0.295	0.117
Year 1992	γ_{65}	−0.185	0.126	−0.433	0.062
Year 1997	γ_{66}	0.134	0.107	−0.075	0.345
Interaction of women and year 1977	γ_{72}	0.269	0.162	−0.042	0.582
Interaction of women and year 1982	γ_{73}	−0.240	0.174	−0.581	0.092
Interaction of women and year 1987	γ_{74}	0.182	0.212	−0.234	0.600
Interaction of women and year 1992	γ_{75}	0.506	0.225	0.058	0.950
Interaction of women and year 1997	γ_{76}	0.343	0.212	−0.078	0.753

Acknowledgments

The work was supported by the Academy of Finland (grant number 266251).

References

Ahmad, OB, Boschi-Pinto, C, Lopez, AD, Murray, CJ, Lozano, R & Inoue, M (2001), 'Age standardization of rates: a new WHO standard', *GPE Discussion Paper Series*. http://www.who.int/healthinfo/paper31.pdf.

Borodulin, K, Levälahti, E, Saarikoski, L, Lund, L, Juolevi, A, Grönholm, M, Jula, A, Laatikainen, T, Männistö, S, Peltonen, M, Salomaa, V, Sundvall, J, Taimi, M, Virtanen, S & Vartiainen, E (2013), *Kansallinen FIN-RISKI 2012—terveystutkimus—Osa 2: tutkimuksen taulukkoliite*. http://urn.fi/URN:ISBN:978-952-302-054-2, In Finnish. Accessed: 2014-09-12.

Brooks, SP & Gelman, A (1998), 'General methods for monitoring convergence of iterative simulations', *Journal of Computational and Graphical Statistics*, **7**(4), 434–455.

Cleveland, WS (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association*, **74**(368), 829–836.

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

Cornfield, J, Haenszel, W, Hammond, EC, Lilienfeld, AM, Shimkin, MB & Wynder, EL (2009), 'Smoking and lung cancer: recent evidence and a discussion of some questions', *International Journal of Epidemiology*, **38**(5), 1175–1191.

Doll, R & Hill, AB (1956), 'Lung cancer and other causes of death in relation to smoking', *British Medical Journal*, **2**(5001), 1071–1081.

Drivsholm, T, Eplov, LF, Davidsen, M, Jorgensen, T, Ibsen, H, Hollnagel, H & Borch-Johnsen, K (2006), 'Representativeness in population-based studies: a detailed description of non-response in a Danish cohort study', *Scandinavian Journal of Public Health*, **34**(6), 623–631.

Ekholm, A & Skinner, C (1998), 'The Muscatine children's obesity data reanalysed using pattern mixture models', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**(2), 251–263.

Gelman, A, Carlin, JB, Stern, HS, Dunson, DB, Vehtari, A & Rubin, DB (2013), *Bayesian Data Analysis*, *CRC press*, Boca Raton, FL.

Harald, K, Salomaa, V, Jousilahti, P, Koskinen, S & Vartiainen, E (2007), 'Non-participation and mortality in different socioeconomic groups: the FINRISK population surveys in 1972–92', *Journal of Epidemiology and Community Health*, **61**(5), 449–454.

Ibrahim, JG & Molenberghs, G (2009), 'Missing data methods in longitudinal studies: a review', *TEST*, **18**(1), 1–43.

Jackson, R, Chambless, LE, Yang, K, Byrne, T, Watson, R, Folsom, A, Shahar, E & Kalsbeek, W (1996), 'Differences between respondents and nonrespondents in a multicenter community-based study vary by gender and ethnicity', *Journal of Clinical Epidemiology*, **49**(12), 1441–1446.

Jousilahti, P, Salomaa, V, Kuulasmaa, K, Niemelä, M & Vartiainen, E (2005), '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 & Community Health*, **59**(4), 310–315.

Kalbfleisch, JD & Prentice, RL (2002), *The Statistical Analysis of Failure Time Data*, Second, *John Wiley & Sons*, Hoboken.

Karvanen, J (2014), 'Study design in causal models', *Scandinavian Journal of Statistics*, **DOI: 10.1111/sjos.12110**.

Laatikainen, T, Tapanainen, H, Alfthan, G, Salminen, I, Sundvall, J, Leiviskä, J, Harald, K, Jousilahti, P, Salomaa, V & Vartiainen, E (2003), *FINRISKI 2002: tutkimuksen toteutus ja tulokset 1. Perusraportti.* https://www.julkari.fi/bitstream/handle/10024/78519/2003b7-1.pdf, Online; Accessed 2013-11-21.

Lehtonen, R & Pahkinen, E (2004), *Practical Methods for Design and Analysis of Complex Surveys*, *Wiley*, Chichester, England.

Little, RJ (1993), 'Pattern-mixture models for multivariate incomplete data', *Journal of the American Statistical Association*, **88**(421), 125–134.

Little, RJ & Rubin, DB (2002), *Statistical Analysis with Missing Data*, *Wiley*, Hoboken, New Jersey.

Little, RJ & Zhang, N (2011), 'Subsample ignorable likelihood for regression analysis with missing data', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **60**(4), 591–605.

Mannino, DM & Buist, AS (2007), 'Global burden of COPD: risk factors, prevalence, and future trends', *The Lancet*, **370**(9589), 765–773.

National Institute for Health and Welfare (2014), *Care register for health care*. http://www.thl.fi/en/web/thlfi-en/statistics/information-on-statistics/register-descriptions/care-register-for-health-care, Accessed: 2014-08-28.

Copyright © 2015 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.73

Peltonen, M, Harald, K, Männistö, S, Saarikoski, L, Lund, L, Sundvall, J, Juolevi, Anne, Laatikainen, T, Aldén-Nieminen, H, Luoto, R, Jousilahti, P, Salomaa, V, Taimi, M & Vartiainen, E (2008), *Kansallinen FINRISKI 2007—terveystutkimus: tutkimuksen toteutus ja tulokset: taulukkoliite*. http://urn.fi/URN:NBN:fi-fe201204193299, In Finnish. Accessed: 2014-09-12.

Plummer, M (2003), 'JAGS: A program for analyses of Bayesian graphical models using Gibbs sampling', Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) in Hornik, K, Leisch, F & Zeileis, A (eds), TU Wien, Vienna, Austria. ISSN 1609-395X, http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, *R Foundation for Statistical Computing*, Vienna, Austria.

Robert, CP & Casella, G (2004), *Monte Carlo Statistical Methods*, Second, *Springer*, New York.

Shahar, E, Folsom, AR & Jackson, R (1996), 'The effect of nonresponse on prevalence estimates for a referent population: insights from a population-based cohort study', *Annals of Epidemiology*, **6**(6), 498–506.

Sogaard, AJ, Selmer, R, Bjertness, E & Thelle, D (2004), 'The Oslo health study: the impact of self-selection in a large, population-based survey', *International Journal of Equity in Health*, **3**(3), DOI 10.1186/1475-9276-3-3.

Statistics Finland (2014), *Official statistics of Finland (OSF): causes of death [e-publication]*. http://tilastokeskus.fi/til/ksyyt/index_en.html, Accessed: 2014-08-28.

Tanner, MA & Wong, WH (1987), 'The calculation of posterior distributions by data augmentation', *Journal of the American Statistical Association*, **82**(398), 528–540.

Tolonen, H, Dobson, A & Kulathinal, S (2005), 'Effect of the trend estimates on the difference between survey respondents and non-respondents: results from 27 populations in the WHO MONICA project', *European Journal of Epidemiology*, **20**(11), 887–898.

van Buuren, S, Boshuizen, HC & Knook, DL (1999), 'Multiple imputation of missing blood pressure covariates in survival analysis', *Statistics in Medicine*, **18**(6), 681–694.

van Loon, AJ, Tijhuis, M, Picavet, HS, Surtees, PG & Ormel, J (2003), 'Survey non-response in the Netherlands: effects on prevalence estimates and associations', *Annals of Epidemiology*, **13**(2), 105–110.

Wynder, EL & Hoffmann, D (1994), 'Smoking and lung cancer: scientific challenges and opportunities', *Cancer Research*, **54**(20), 5284–5295.

Copyright © 2015 John Wiley & Sons, Ltd

Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.146

Correction: Correcting for non-ignorable missingness in smoking trends

Juho Kopra[a]*, Tommi Härkänen[b], Hanna Tolonen[b] and Juha Karvanen[a]

Received 21 April 2017; Accepted 21 April 2017

There is a mistake in the real data example in the article Kopra, J, Härkänen, T, Tolonen, H, & Karvanen, J (2015) "Correcting for non-ignorable missingness in smoking trends", *Stat*, **4**(1), 1–14. In Figure 5, the trends labelled as "Original trend" are incorrect due to misspecified weighting. The mistake appears also in the third column of Table 2 in the original article. The corrected figure and table are presented below. The correction changes some of our conclusions. The differences in the trends for men are small. Especially, the large difference observed for men in Helsinki and Vantaa in 1997 seems to disappear. For women, the differences between the trends are slightly smaller than in the original article but still notable.

Table 2. Observed and model-based smoking proportions for the study in 1997 adjusted using the World Health Organization Scandinavian standardization weights. The two rightmost columns describe the 95% credible intervals of model-based trends. Participant smoking is the same as "Original trend" in Figure 5.

Gender	Area	Participant smoking (%)	Model-based total smoking (%)	95% credible interval	
Men	North Karelia	32.4	31.6	29.2	33.9
Men	Northern Savonia	31.5	31.8	29.5	34.0
Men	Turku and Loimaa	33.4	33.7	31.1	36.2
Men	Helsinki and Vantaa	32.0	32.7	30.1	35.6
Men	Oulu province	30.7	32.3	29.5	35.2
Women	Northern Karelia	16.9	18.3	16.3	20.7
Women	North Savonia	17.2	19.1	17.1	21.1
Women	Turku and Loimaa	21.1	23.6	21.3	26.0
Women	Helsinki and Vantaa	26.3	27.7	25.3	30.4
Women	Oulu province	19.7	22.2	20.0	24.2

[a]Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, FI-40014, Finland
[b]National Institute for Health and Welfare, Helsinki, FI-00271, Finland
*Email: juho.j.kopra@jyu.fi

Copyright © 2017 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.146

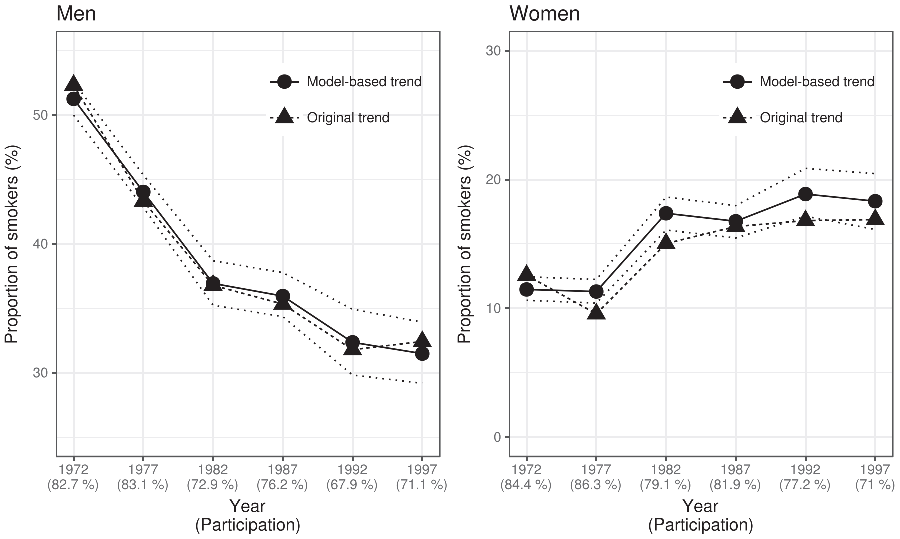

Figure 5. Model-based trend and original trend for men (left) and women (right) in North Karelia province. Two dotted lines represent 95% credible interval of the posterior distribution for corrected trends. Both the model-based and the original trend use the World Health Organization Scandinavian standardization weights.

Copyright © 2017 John Wiley & Sons, Ltd

III

Bayesian models for data missing not at random in health examination surveys.

Kopra J., Karvanen J. and Härkänen T.

Published online first in *Statistical Modelling.* (2017)
doi: 10.1177/1471082X17722605.

Statistical Modelling 2017; **xx(xx)**: 1–16

Bayesian models for data missing not at random in health examination surveys

Juho Kopra[1], Juha Karvanen[1] and Tommi Härkänen[2]
[1]Department of Mathematics and Statistics, University of Jyvaskyla, Jyväskylä, Finland.
[2]National Institute for Health and Welfare, Helsinki, Finland.

Abstract: In epidemiological surveys, data missing not at random (MNAR) due to survey nonresponse may potentially lead to a bias in the risk factor estimates. We propose an approach based on Bayesian data augmentation and survival modelling to reduce the nonresponse bias. The approach requires additional information based on follow-up data. We present a case study of smoking prevalence using FINRISK data collected between 1972 and 2007 with a follow-up to the end of 2012 and compare it to other commonly applied missing at random (MAR) imputation approaches. A simulation experiment is carried out to study the validity of the approaches. Our approach appears to reduce the nonresponse bias substantially, whereas MAR imputation was not successful in bias reduction.

Key words: Bayesian estimation, data augmentation, follow-up data, health examination surveys, multiple imputation, survival analysis

Received October 2016; **revised** May 2017; **accepted** June 2017

1 Introduction

Population level estimates of risk factors are of major interest in epidemiology. Data on risk factors such as blood pressure, cholesterol level, body mass index, alcohol consumption and daily smoking are often collected in health examination surveys (HES). In an HES, the data on risk factors are gathered usually via both questionnaires and physical measurements. The trends of population-level risk factors are monitored, and they are valuable input for policy decisions.

Missing data by unit nonresponse occurs in an HES as invitees neither participate to physical measurements nor return a survey questionnaire. The decision about participation have been found to depend on the risk factors, such as smoking (Shahar et al., 1996), either directly or via a common cause such as health awareness. This may be deduced from the fact that the non-participants have a higher risk of death (Jousilahti et al., 2005; Harald et al., 2007; Karvanen et al., 2016). This dependence causes missing data to be classified as missing not at random (MNAR; Rubin, 1976). Because the data are MNAR, the population-level risk factors calculated from the participants' data are biased, and they usually give an overly healthy view of

Address for correspondence: Juho Kopra, Department of Mathematics and Statistics, P.O. Box 35 (MaD), FI-40014 University of Jyväskylä, Finland.
E-mail: juho.j.kopra@jyu.fi

© 2017 SAGE Publications 10.1177/1471082X17722605

the population. Biased estimates of risk factor prevalence may seriously misinform decision-makers. Instead of analysing only the participants data, the posterior distributions of risk factor levels of a whole sample, including non-participants, should be estimated. This requires external information and modelling assumptions.

In this article, we demonstrate how follow-up data on endpoints associated with the risk factor of the interest provides external information that allows us to reduce the bias caused by selective non-participation. We propose a Bayesian method for the estimation of risk factor prevalence and the missing data mechanism, when the data are MNAR.

Our datasets origin from the FINRISK studies, which are national HES providing information about the health of Finns. We improve and extend an earlier work (Kopra et al., 2015) on the estimation of smoking prevalence from FINRISK data. The key improvements are: a fully Bayesian model is used, the survival model is more flexible and informative prior is utilized instead of assumption of conditional independence (Kopra et al., 2015; equation (2)). Differently from Kopra et al. (2015), the study years 2002 and 2007 are included in the modelling.

Next section describes the data of the FINRISK studies and follow-up. Section 3 presents the Bayesian model and the priors that we apply to smoking prevalence estimation. Section 4 explains model fitting, and Section 5 provides a simulation study on the proposed approach. We evaluate alternative methods in Section 6. In Section 7, we apply our approach to real data from the FINRISK studies and provide smoking prevalence estimates for both men and women. Section 8 discusses the results and methods presented.

2 Data description

Our HES data contain eight FINRISK studies conducted in selected geographical areas of Finland once in every five years in 1972–2007 (Laatikainen et al., 2003; Harald et al., 2007). In each study year, persons were selected to the FINRISK studies in a random sampling stratified by region, gender and 10-year age group. Our data are restricted to the two regions (Northern Savonia and North Karelia) that have been included in all eight studies. In total, the data contain 52 325 persons including 9 928 persons with missing smoking indicator.

Each person selected to the study received a letter of invitation, in which he or she was asked to fill in a survey questionnaire and participate to physical measurements in the local survey site. If the person participated, the filled questionnaire was collected and the physical measures were taken. If the person did not participate, then risk factors are missing, but background variables, study year, age, gender and region are known from the sampling frame. Table 1 shows that the participation rates have dramatically decreased from 1972 to 2007. It can be also seen that women have participated more actively than men in all study years. We also know that person's age affects participation (Kopra et al., 2015).

Our HES data were linked together with follow-up data of all participants and non-participants. The follow-up data contains the exact dates and diagnoses (ICD

Table 1 Participation rates (%) and size of survey sample (*n*) by gender, region and year. The participation rates of 1972 and 1977 are approximated (*) as the region information of non-participants is missing for these years

		North Karelia		Northern Savonia	
Year		Men	Women	Men	Women
1972	%	84.3*	88.5*	88.4*	91.3*
	n	2 641	2 607	3 574	3 555
1977	%	85.7*	89.0*	90.1*	93.0*
	n	2 323	2 382	3 223	3 391
1982	%	76.1	83.2	80.8	86.0
	n	2 007	2 019	1 810	1 566
1987	%	78.8	85.3	80.3	86.3
	n	1 971	1 976	979	988
1992	%	68.2	80.8	75.9	83.8
	n	984	993	982	990
1997	%	72.1	75.3	70.8	79.8
	n	1 052	1 020	990	997
2002	%	66.5	76.2	66.2	78.2
	n	1 021	1 011	1 000	1 000
2007	%	63.0	71.9	61.1	70.4
	n	811	825	817	820
Total	%	77.5	83.5	81.5	87.1
	n	12 810	12 833	13 375	13 307

codes) of hospitalizations and deaths. In Finland, this kind of follow-up data can be collected from administrative registers for both participants and non-participants. The follow-up period started at the time of study for each person and ended on 31st December 2012 for all FINRISK study years. Thus, the length of the follow-up period varies by study years.

It is well known that smoking is a key risk factor for lung cancer and chronic obstructive pulmonary disease (COPD) (Doll and Hill, 1956; Mannino and Buist, 2007). Thus, we use lung cancer and COPD events together as an endpoint. Table 2 shows that non-participants have a higher rate of disease events than participants.

Table 2 The total count of observed lung cancer and COPD events, events per 1 000 follow-up person years and participation rate by region and gender

Region	Gender	Participant	Events	Events/1 000 years	Participation (%)
N. Karelia	Men	Yes	387	1.75	77.4
N. Karelia	Men	No	166	3.14	
N. Savonia	Men	Yes	479	1.85	81.6
N. Savonia	Men	No	129	2.85	
N. Karelia	Women	Yes	75	0.28	83.6
N. Karelia	Women	No	43	1.02	
N. Savonia	Women	Yes	62	0.21	87.0
N. Savonia	Women	No	33	0.94	

We limit in our analysis the age range to 25–64 years old and select the subset of healthy persons with respect to our endpoint variables. The two exceptions are 1972 and 1977 studies, which have age ranges of 25–59 and 30–64 years old, respectively.

3 Bayesian model

The modelling is based on the idea that although it is impossible to directly observe the smoking status of non-participants, we can obtain information on smoking indirectly from the follow-up data. More precisely, the modelling uses the observed incidence differences of the smoking-based diseases between participants and non-participants, which allows us to adjust the estimates of smoking prevalence. Full Bayesian approach is applied, and model fitting is executed using Markov chain Monte Carlo (MCMC) methods (Robert and Casella, 2004).

3.1 Notation for the data

We introduce our model using causal models with design (Karvanen, 2015) and make a difference between measurements and underlying causal variables. The model is presented in Figure 1. For each person $i = 1, \dots, N$ invited to the survey, we denote participation indicator by M_i, which takes the value $M_i = 1$, if person i participated, and value $M_i = 0$ otherwise. Value $M_i = 0$ indicates missing risk factor data. The indicator of self-reported daily smoking is denoted by Y_i and the corresponding measurement by Y_i^*. Variable Y_i takes value 1, if a person is a daily smoker, and 0 otherwise. Class $Y_i = 0$ includes earlier smokers who quitted. The value of Y_i^* is known for the participants, then $Y_i^* = Y_i$, but missing for the non-participants. We denote by vector $\boldsymbol{X}_i^*$, the variables age a_i, gender g_i, region r_i and study year s_i in the background data observed for all sample members. The values $g_i = 0$ stand for men, and $g_i = 1$ for women. The North Karelia region is denoted by $r_i = 0$ and Northern Savonia by $r_i = 1$.

We denote T_i as the age at the day of diagnosis, which may also be the age at the time of death, if a person dies without previous lung cancer or COPD diagnoses and the death is caused by either of the two diseases. If the person has not been diagnosed, the corresponding measurement T_i^* is missing, and T_i is right-censored. Variable $T_{\text{cens},i}$ is the age of the person i in the end of the year 2012, which is the end of our follow-up period, or the age of death for the person who has died before the end of the year 2012. The variable $T_{\text{obs},i}$ is the minimum of T_i and $T_{\text{cens},i}$, so $T_{\text{obs},i} = \min(T_i, T_{\text{cens},i})$, and $T^*_{\text{cens},i} = T_{\text{cens},i}$ and $T^*_{\text{obs},i} = T_{\text{obs},i}$.

3.2 Submodels

The joint model for data from an HES linked with follow-up consists of three submodels:

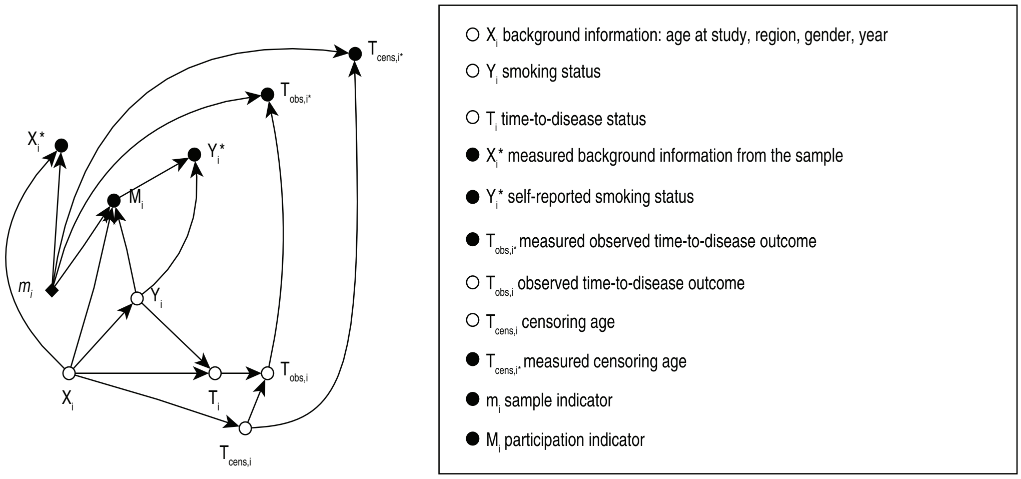
T_{cens,i^*}
T_{obs,i^*}
X_i^*
Y_i^*
M_i
m_i
Y_i
X_i
T_i
$T_{obs,i}$
$T_{cens,i}$
○ X_i background information: age at study, region, gender, year
○ Y_i smoking status
○ T_i time-to-disease status
● X_i* measured background information from the sample
● Y_i* self-reported smoking status
● T_{obs,i^*} measured observed time-to-disease outcome
○ $T_{obs,i}$ observed time-to-disease outcome
○ $T_{cens,i}$ censoring age
● T_{cens,i^*} measured censoring age
● m_i sample indicator
● M_i participation indicator

Figure 1 A graph representing the model and the dependencies between the variables of HES data and the follow-up data. Direct causal effects are represented as arrows. Measurement variables are denoted with asterisk, for example, X_i^*, and are presented as filled circles. The causal variables do not have asterisk symbol (e.g., X_i), and they are drawn unfilled to indicate that they are not observed directly but via measurement variables. The measurement variables always have one participation indicator (m_i or M_i) and one causal variable as their parent. The graph tells that X_i^*, $T_{\text{obs},i}^*$ and $T_{\text{obs},i}^*$ are collected for each member of the sample, while Y_i^* is measured only for participants and is missing for the non-participants

1. a participation model in which participation is explained by daily smoking and background variables (arrow $\boldsymbol{X}_i \rightarrow M_i$ in Figure 1),
2. a risk factor model for daily smoking given the background variables (arrow $\boldsymbol{X}_i \rightarrow Y_i$) and
3. a survival model for the follow-up data given the daily smoking and background variables (arrows $Y_i \rightarrow T_i$ and $\boldsymbol{X}_i \rightarrow T_i$).

These three submodels together form a joint model for the data, which we call Bayesian MNAR model, see Figure 1. The arrows $\boldsymbol{X}_i \rightarrow M_i$ and $Y_i \rightarrow M_i$ correspond to the participation submodel that can be written as $P(M_i = 1|\boldsymbol{X}_i, M_i)$. The arrow $\boldsymbol{X}_i \rightarrow Y_i$ corresponds to the risk factor submodel (distribution $P(Y_i|\boldsymbol{X}_i)$), and the arrows $Y_i \rightarrow T_i$ and $\boldsymbol{X}_i \rightarrow T_i$ correspond to the survival model (distribution $P(T_i|Y_i, \boldsymbol{X}_i)$). All the submodels are fitted together because each of them contains the indicator of smoking, which has missing values to be imputed.

3.3 Participation model

First, our model for participation indicator M_i is

$$\text{logit}(P(M_i = 1|\boldsymbol{X}_i, Y_i)) = \alpha_{0[g_i,s_i]} + \eta_{[g_i,s_i]}Y_i + \alpha_{1[g_i,Y_i]}(a_i - 45) + \alpha_2 r_i, \tag{3.1}$$

where g_i, s_i, a_i and r_i are part of $\boldsymbol{X}_i$, and they stand for gender, study year, age and region, respectively. Variable Y_i stands for smoking. The roles of model parameters $\alpha_{0[g_i,s_i]}$, $\eta_{[g_i,s_i]}$, $\alpha_{1[g_i,Y_i]}$ and α_2 are explained further. Parameter $\alpha_{0[g_i,s_i]}$ is a regression coefficient (intercept) which varies over the levels of g_i and s_i, $i = 1, \ldots, N$. The variable g_i is binary and s_i has eight possible values, which create the total of 16 intercept parameters. The parameters $\eta_{[g_i,s_i]}$ are gender-specific regression coefficients modelling how daily smoking affects participation in each year. We also take into account how the age of person affects participation; the gender-specific coefficients $\alpha_{1[g_i,0]}$ and $\alpha_{1[g_i,1]}$ model how age affects participation for non-smokers and smokers, respectively. The parameter α_2 describes the differences in participation between the two regions.

3.4 Risk factor model

Next, we need to model smoking indicator Y_i by background variables $\boldsymbol{X}_i = (g_i, s_i, a_i, r_i)$. We use a logistic regression model

$$\operatorname{logit}(P(Y_i = 1|\boldsymbol{X}_i)) = \beta_{0[g_i,r_i,s_i]} + (s_i - a_i - 1\,938)\beta_{1[g_i,r_i,s_i]}, \tag{3.2}$$

where coefficients β_0 and β_1 vary between groups defined by combinations of gender g_i, region r_i and study year s_i, similarly as in (3.1). The year of birth $s_i - a_i$ for person i is centred at its rounded population average 1 938 in the model.

3.5 Survival model

To define a survival model for $P(T_i|\boldsymbol{X}_i, Y_i)$, a counting process notation is used. Let $N_i(t)$ stand for the count of disease diagnoses up to age t for person i. Let $dN_i(t)$ be the increment of the counting process over one-year time interval $(t, t+1)$, and let t take discrete values $25, 26, \ldots, 100$. Now, we model $dN_i(t)$ with a piecewise constant hazard model assuming that for each one-year time period, the gender-specific hazard $h_{0,g}(t)$ remains constant ($g = 0, 1$ stands for the gender). The model for follow-up data is

$$dN_i(t) \sim \text{Poisson}(\lambda_i(t)) \tag{3.3}$$

$$\lambda_i(t) = \begin{cases} \exp(\gamma_1 Y_i)\, h_{0,0}(t), & \text{given that } T_i \geq t \text{ and } g = 0 \\ \exp(\gamma_2 Y_i)\, h_{0,1}(t), & \text{given that } T_i \geq t \text{ and } g = 1 \\ 0, & T_i < t, \end{cases} \tag{3.4}$$

where γ_1 and γ_2 model how smoking increases the hazard for men and women, respectively.

3.6 Prior distributions

For the participation model, we use informative prior distribution for the difference between smokers and non-smokers. An informative prior for $\eta_{[g_i,s_i]}$ is derived as follows. We consider a 45-year-old non-smoker, who participates with probability $p = 0.7$, and elicit the corresponding prior probability for a smoker, who is otherwise similar. We consider that there is a 15% chance that the participation prior probability p is less than 0.50, about 30% chance for less than 0.60 and 50% chance for less than 0.70. These considerations together with an assumption on a logistic distribution for $\eta_{[g_i,s_i]}$ lead to prior distribution

$$\eta_{[g_i,s_i]} \sim \text{Logistic}(\mu = 0, s = 2.05^{-1}),$$

which makes the prior distribution of p to have expected value $E(p) = 0.676$ and 95% credible interval $[0.281, 0.933]$. Here, logistic distribution density function is

$$f_{\text{logistic}}(x|\mu, s) = \frac{e^{(x-\mu)/s}}{s(1 + e^{(x-\mu)/s})^2},$$

for $x, \mu \in \mathbb{R}$ and scale parameter $s > 0$.

The prior distributions for participation model coefficients $\alpha_{0[g_i,s_i]}$ and $\alpha_{1[g_i,Y_i]}$ and risk factor model parameters $\beta_{0[g_i,r_i,s_i]}$ and $\beta_{1[g_i,r_i,s_i]}$ are normal distributions with mean $\mu = 0$ and variance $\sigma^2 = 1\,000$ (uninformative priors).

Survival model parameters γ_1 and γ_2 are also a priori normally distributed with $\mu = 0$ and $\sigma^2 = 1\,000$. Our prior distribution for baseline hazard $h_{0,g}(t)$ is monotonically increasing with age

$$\begin{aligned} &h_{0,g}(25) \sim \text{Uniform}(0, 20) \\ &h_{0,g}(t) \sim \text{Uniform}(h_{0,g}(t-1), 20), \qquad \text{where } t = 26, 27, \ldots, 100, \end{aligned}$$

where g stands for gender, 0 for men and 1 for women. This means that model assumes that risk of smoking-based diseases only increases with age. This assumption seems to be in agreement with our data.

4 Model fitting

As the number of model parameters (316) and missing values (9 928) is large, there are over 10 000 variables to sample at each iteration of the MCMC model fitting process. This creates a computational challenge for Bayesian model fitting. The Markov chains typically require thousands of iterations or more to obtain satisfactory convergence, which requires a lot of computing time.

To impute the missing values for smoking indicators, the data augmentation was applied (Tanner and Wong, 1987). A Bayesian MNAR model described in Figure 1 and Sections 3.3, 3.4 and 3.5 was used. The augmented data for smoking indicator

Y_i are drawn from fully conditional distribution $P(Y_i|M_i = 0, \boldsymbol{X}_i, T_i)$, given its parent nodes ($\boldsymbol{X}_i$) and child nodes (M_i and T_i).

We used Just Another Gibbs Sampler software (JAGS; Plummer, 2003), R (R Core Team, 2016) and `rjags` package (Plummer, 2015) to fit the model. Seven parallel MCMC chains were used. Each chain had 9 000 burn-in iterations, 45 900 actual iterations with thinning interval 75, which makes a total of 612 iterations per chain to be recorded. The time consumed for this model fitting using parallel chains was about 107 hours, which makes 7 seconds per each iteration and less than 0.7 milliseconds per each parameter for one iteration. The high absolute number of missing values (9 928) explains the long running time. Because of high autocorrelations in the chains, we decided to use a long thinning interval and many iterations to reduce the autocorrelations in the saved iterations of the MCMC and larger sample of the posterior for the final trends. The convergence of the chains was examined both visually and using Brooks–Gelman $\hat{R}$-diagnostics (Brooks and Gelman, 1998). All the $\hat{R}$ test statistics for model coefficients were below 1.01 which indicates convergence.

5 Simulation study

We carried out a simulation experiment to demonstrate the performance of the model. The simulated data allow us to compare the performance of the estimated prevalence of smoking with true prevalence, which is known with the simulated data but not with the HES data. The actual values of the background variables from the FINRISK study were used together with parameter estimates from a preliminary analysis. Thus, the simulation experiment had conditions similar to the real data, for example, the smoking prevalence had a decreasing trend for men and an increasing trend for women. For both genders, the participation was selective and the participation rate had a decreasing trend.

The simulation was implemented using R language, and the data were simulated from the Bayesian MNAR model presented in Section 3. Because of the computational burden of model fitting, the model was fitted into a single simulated dataset.

The model fitting for the Bayesian MNAR model with simulated data was implemented as described in Section 4. All the $\hat{R}$-diagnostics were below 1.01 which indicates convergence. We inspected the posterior correlations between the model parameters and found strong correlations (≥ 0.9 or ≤ -0.9) between some of the parameters. In the risk factor model, the strongest correlations were observed between the parameters β_0 and β_1. The median of these correlations was -0.434 and the range was $[-0.905, 0.661]$. Conditioning with M_i, likely causes these posterior correlations. On the participation model, the strongest posterior correlations were found between the $\alpha_{0[g_i,s_i]}$ and $\eta_{[g_i,s_i]}$. For men, those eight correlations ranged between -0.972 and -0.899, and for women between -0.860 and -0.665. In the survival model, strong positive correlations occurred particularly between the hazards of consecutive years, which is natural to this type of models. The highest correlations for men were 0.920 and for women 0.952.

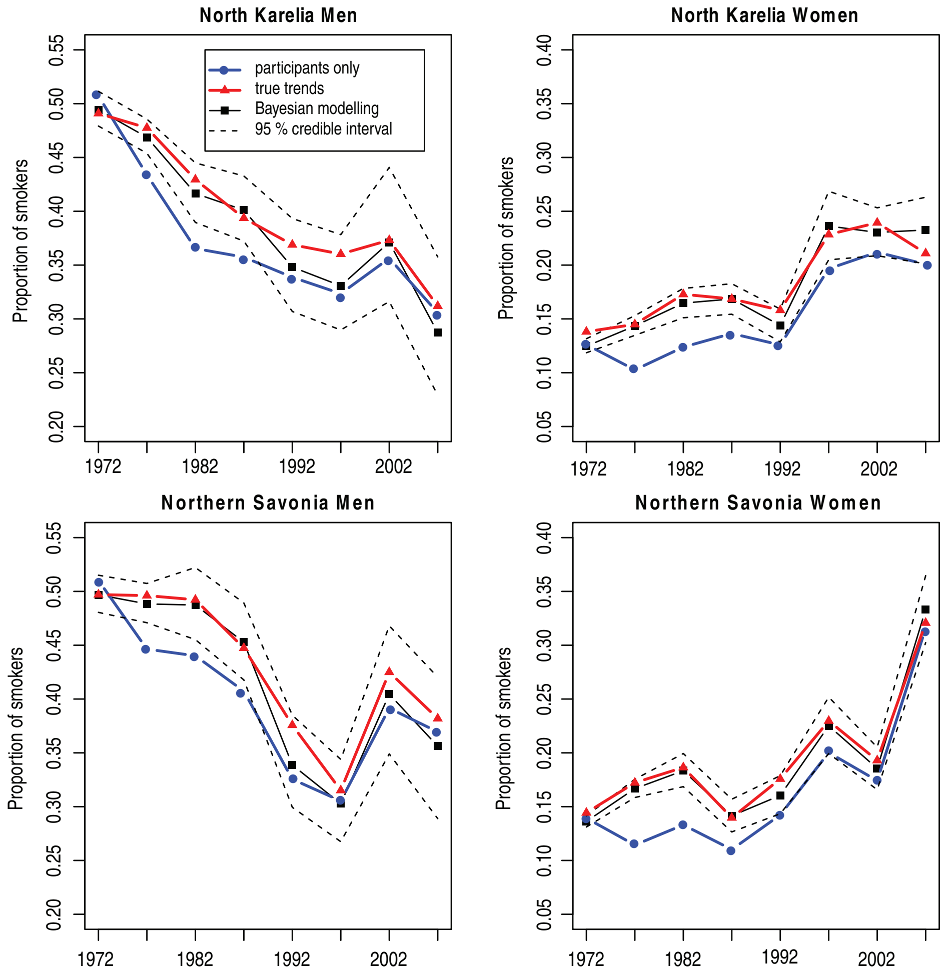

Figure 2 Trends for the simulation experiment. The red line with triangles is the true proportion of smokers used in the simulation. The blue line with circles is estimated from participants only and the black line with squares is model-based posterior mean with 95% credible intervals (dashed black line) calculated from the simulated data

It can be seen from Figure 2 that with an exception of women in 1972, the true prevalence is located inside the credible interval and there is no indication of systematic bias in the posterior mean. In contrast, the prevalence calculated using only the participants systematically underestimates the true prevalence. As the same family of models was used both to simulate the data and to fit the parameters, the Bayesian MNAR approach is expected to perform very well. However, the experiment demonstrates that the model can be estimated from the data.

6 Alternative methods

In addition to the Bayesian MNAR approach, we considered two alternative modelling approaches and the complete case analysis. Both modelling approaches utilize the missing at random (MAR) assumption, and the complete case analysis uses data on participants only. In terms of Figure 1, the MAR assumption omits the arrow $Y_i \rightarrow M_i$, which means that participation is not selective with respect to smoking.

First, the Bayesian MAR approach differs from the Bayesian MNAR approach such that the entire survival model (3.4) is omitted, and the regression coefficient $\eta_{[g_i,s_i]}$ is fixed to zero in participation model (3.1).

We also used the frequentist MAR model which was implemented using the `mice` package in R (van Buuren and Groothuis-Oudshoorn, 2011). The missing smoking indicator was imputed using a logistic regression model that had full interactions between year of birth, gender, region and year, and full interactions between gender, event indicator and age at the event/censoring. The year of birth and the age at event/censoring were used as linear covariates and the other variables were categorical.

We fitted these alternative approaches to data simulated from the MNAR model which is selective with respect to smoking. The trend estimates are presented in Appendix in Table A2. We calculated root mean square errors (RMSE) for the Bayesian MNAR model and each of the alternative approaches. The RMSE was calculated using formula

$$\text{RMSE}(y, y_{\text{true}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_{\text{true},i})^2},$$

where y is the estimated smoking prevalence (%) and y_{true} is the true smoking prevalence from simulation. The RMSE was calculated over the regions, the genders and the study years, which gives one RMSE value for each approach.

The Bayesian MNAR approach had the smallest RMSE, 1.65. The Bayesian MAR and the frequentist MAR methods have very similar RMSE with each other, 3.34 and 3.37, respectively. These two methods were slightly more accurate than the complete case approach with RMSE 3.48.

7 Application to FINRISK data

The trends of daily smoking for the FINRISK data estimated using the Bayesian MNAR model are reported in Figure 3 and compared to participant trends, which are often reported in HES. The difference between the smoking prevalence estimates of the complete case (participants) and the Bayesian MNAR approaches is the highest for the study years 1977, 1982 and 1987 (Figure 3 and Table A1 in Appendix).

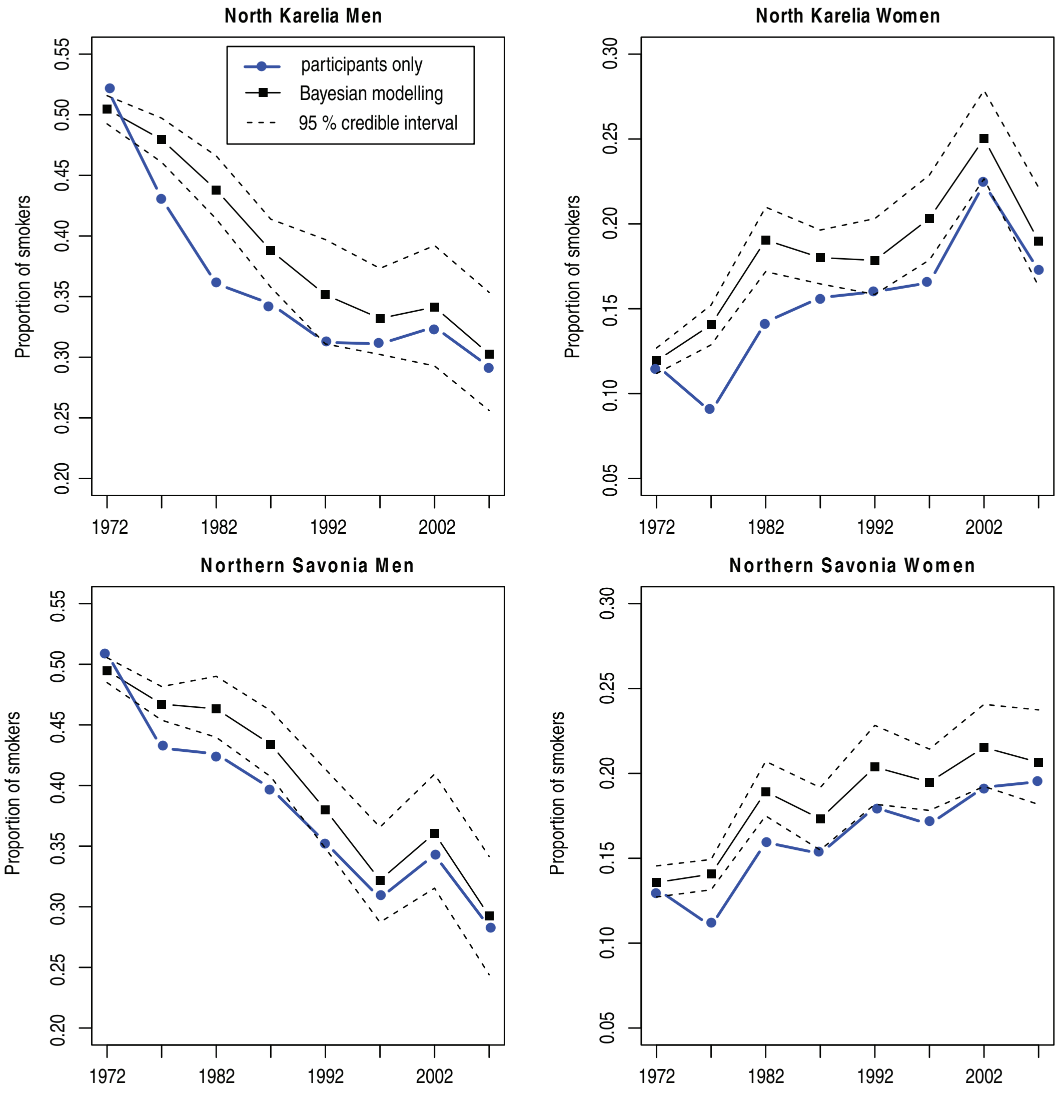

Figure 3 Participant trends (blue line with circles) and model-based posterior trends (black line with squares) with 95% credible intervals (dashed black line) for the FINRISK data. For numeric presentation of trends, seeTable A2 in Appendix

Starting from 1992, the complete case trends are within the 95% credible interval of the Bayesian MNAR model, but they are systematically below the posterior mean. The proportion of missing data is higher for the later study years than the earlier ones, which makes the credible intervals wider.

The values of the region variable r_i for the study years 1972 and 1977 were missing for non-participants. We used a single imputation with fixed probabilities $P(r_i = 1|s_i = 1972) = 0.495$ and $P(r_i = 1|s_i = 1977) = 0.493$ as in Kopra et al. (2015). We decided to use single imputation a prior to model fitting because the regions seemed to be rather similar with respect to smoking prevalence and the participation rates were high.

We executed sensitivity analysis to find out if model can be fitted with even less informative prior. We tried prior distributions for η with s parameter set to $(2.05/2)^{-1}$, which corresponds to doubling the prior variance. We found out that the Markov chains do not converge. More precisely, convergence problems were found with η-parameters for which the $\hat{R}$s ranged between 1.097 and 1.828 after 45 900 iterations and 9 000 burn-in. Thus, it appears that vague prior distributions are not applicable.

8 Discussion

We have proposed the Bayesian MNAR modelling approach to reduce the non-participation bias and applied the approach to the FINRISK studies. In a simulation experiment, we compared our approach to the Bayesian MAR approach, to the complete case analysis, and the frequentist MAR imputation. The latter two were easier to use than the Bayesian MNAR approach, but did not substantially reduce the non-participation bias. The proposed approach appears to reduce the non-participation bias.

Trends by the Bayesian and the frequentist MAR approaches are essentially the same as participants' trends. Thus, the MAR approaches do not reduce the bias of risk factor levels by much. Although there may be ways to improve the imputation model (White and Royston, 2009), the MAR imputation do not account for selective non-participation.

The information about non-participants' risk factor levels comes from the survival data. The risk factor of interest must be a strong predictor of the survival outcome. The estimation of survival model parameters requires that a sufficient number of events have been recorded. This implies that the length of follow-up must be long enough, and if the event is rarely observed, the number of persons must be high. The information obtained from the survival model may be insufficient by itself and need to be supported by an informative prior on the selection mechanism.

In many countries, it is not technically or legally possible to link the HES data of non-participants to follow-up data. The requirement on the availability of survival data for both participants and non-participants is a major limitation for the proposed approach.

The Bayesian model fitting is often a computational challenge, particularly when the amount of missing data is large. The memory management and computation time were issues in our case study due to a large amount of missing data. In our first attempts, we tried to save all the imputations, which filled the RAM memory of the computer (16 GB) quite rapidly. We later realized that it is possible to calculate sufficient (summary) statistics, for example, count of smokers and non-smokers by gender, year and region, and store only them. We also reduced the time required per one iteration by coarsening the continuous covariates (the age) of the survival model and summing over discrete or discretized covariates, and by modelling the number of events in each risk group using Poisson distribution.

Another challenge was the posterior correlations caused by the model structure and non-random missingness. One possibility to alleviate this problem could be to develop a custom MCMC algorithm with blocked updating of the parameters with high posterior correlations (Haario et al., 2001). However, this often requires custom programming and is, therefore, much more laborious than an application based on JAGS, which we have applied. Thus, if one needs to use data with more missing data than in our case study or multiple variables with MNAR missingness, we recommend using some specialized MCMC algorithms or possibly the iterated importance sampling algorithm (Celeux et al., 2006).

Additional fully observed covariates can be added into the model without excessive increase in computational burden. If more variables with missing data are imputed, the model fitting is slowed down proportional to increase of absolute amount of missing values. This is because at each MCMC iteration all the missing values need to be imputed.

Selective non-participation in HES is an important problem that may have implications to the decisions on health policy. Our solution is not simple to implement, but the reduction of selection bias makes it worth of the effort.

Acknowledgements

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

APPENDIX

Table A1 The trends and 95 % credible intervals of simulation experiment and alternative approaches

		Northern Karelia		North Savonia	
Year	Method	Men	Women	Men	Women
1972	Bayes+MAR	50.9 (50.2, 51.7)	12.2 (11.8, 12.7)	50.3 (49.6, 51.0)	14.2 (13.7, 14.6)
1972	Bayes+MNAR	49.5 (47.8, 51.2)	12.4 (11.7, 13.2)	48.8 (47.3, 50.5)	14.7 (14.0, 15.5)
1972	Complete case	50.6 (48.5, 52.7)	12.2 (10.8, 13.5)	50.0 (48.2, 51.7)	14.2 (13.0, 15.4)
1972	mice	51.0 (49.0, 53.0)	12.2 (10.8, 13.6)	50.4 (48.6, 52.1)	14.2 (13.0, 15.5)
1972	True	49.1	12.9	48.9	15.3
1977	Bayes+MAR	42.6 (41.8, 43.4)	8.8 (8.5, 9.2)	42.9 (42.2, 43.7)	11.9 (11.5, 12.3)
1977	Bayes+MNAR	46.9 (45.7, 48.3)	12.2 (11.2, 13.2)	47.3 (45.8, 48.8)	16.6 (15.7, 17.7)
1977	Complete case	42.6 (40.5, 44.8)	8.9 (7.7, 10.1)	43.0 (41.1, 44.8)	11.9 (10.8, 13.1)
1977	mice	42.8 (40.7, 45.0)	8.8 (7.4, 10.3)	43.1 (41.3, 44.9)	11.9 (10.6, 13.2)
1977	True	47	13.2	47.5	17.5
1982	Bayes+MAR	38.4 (37.2, 39.5)	14.5 (13.9, 15.2)	42.0 (40.6, 43.3)	13.1 (12.3, 13.9)
1982	Bayes+MNAR	44.0 (41.4, 46.5)	18.3 (16.8, 19.9)	46.7 (43.5, 50.2)	17.7 (16.0, 19.4)
1982	Complete case	38.2 (35.8, 40.6)	14.5 (12.8, 16.2)	41.7 (39.0, 44.4)	13.0 (11.2, 15.0)
1982	mice	38.6 (36.2, 41.1)	14.4 (12.7, 16.0)	42.0 (39.5, 44.6)	13.0 (11.1, 15.0)
1982	True	45	19.6	47.7	19.7
1987	Bayes+MAR	35.1 (34.0, 36.4)	15.5 (14.8, 16.2)	40.1 (38.3, 42.0)	13.0 (12.0, 14.1)
1987	Bayes+MNAR	38.9 (36.2, 41.1)	16.3 (15.2, 17.4)	44.7 (41.8, 47.3)	13.9 (12.6, 15.3)
1987	Complete case	34.8 (32.4, 37.2)	15.4 (13.7, 17.2)	39.8 (36.2, 43.4)	12.9 (10.7, 15.3)
1987	mice	35.0 (32.5, 37.5)	15.5 (13.8, 17.2)	40.5 (36.9, 44.2)	13.3 (10.9, 15.6)
1987	True	39	17.4	44.1	15.5
1992	Bayes+MAR	35.0 (32.9, 37.1)	12.7 (11.6, 13.9)	37.7 (35.7, 39.6)	15.0 (13.9, 16.2)
1992	Bayes+MNAR	32.6 (28.9, 37.2)	15.6 (13.6, 17.8)	36.0 (32.2, 40.1)	16.7 (14.9, 18.9)
1992	Complete case	33.8 (30.3, 37.4)	12.6 (10.3, 15.0)	36.8 (33.2, 40.4)	14.8 (12.4, 17.4)
1992	mice	35.2 (31.4, 39.0)	12.7 (10.3, 15.0)	38.0 (34.3, 41.8)	15.2 (12.7, 17.8)
1992	True	38	16	39.8	17.4
1997	Bayes+MAR	35.2 (33.1, 37.4)	19.4 (18.1, 20.8)	33.4 (31.7, 35.3)	18.4 (17.2, 19.8)
1997	Bayes+MNAR	34.3 (29.8, 40.0)	19.9 (17.8, 22.0)	34.1 (29.7, 38.9)	19.1 (16.7, 22.0)
1997	Complete case	34.0 (30.5, 37.7)	18.9 (16.2, 21.7)	32.6 (29.3, 36.0)	17.9 (15.3, 20.6)
1997	mice	35.4 (31.5, 39.2)	19.5 (16.8, 22.1)	33.8 (30.3, 37.4)	18.3 (15.9, 20.8)
1997	True	34.9	21.8	33.5	21.2
2002	Bayes+MAR	44.2 (42.0, 46.3)	17.9 (16.5, 19.3)	35.2 (33.1, 37.2)	22.2 (20.7, 23.7)
2002	Bayes+MNAR	44.2 (38.3, 50.2)	17.2 (15.1, 19.6)	36.2 (30.7, 42.9)	22.8 (20.4, 25.5)
2002	Complete case	42.6 (38.8, 46.3)	17.2 (14.6, 20.0)	33.6 (30.1, 37.2)	22.0 (19.1, 25.0)
2002	mice	44.2 (40.5, 48.0)	18.1 (15.3, 20.8)	35.2 (31.3, 39.1)	22.1 (19.5, 24.8)
2002	True	44.3	19.1	35.3	25
2007	Bayes+MAR	41.8 (39.2, 44.2)	29.6 (27.7, 31.5)	35.2 (32.7, 37.7)	22.4 (20.6, 24.2)
2007	Bayes+MNAR	39.7 (33.0, 46.4)	28.1 (24.7, 31.5)	33.6 (27.7, 40.5)	21.2 (18.2, 24.5)
2007	Complete case	39.9 (35.7, 44.2)	28.9 (25.3, 32.7)	33.3 (29.3, 37.5)	21.7 (18.5, 25.1)
2007	mice	42.1 (37.8, 46.3)	29.4 (26.1, 32.7)	35.0 (30.6, 39.3)	22.5 (19.3, 25.8)
2007	True	37.9	28.8	32.3	23.5

Table A2 Participant trends and model-based posterior trends with 95% credible intervals for the FINRISK data

		Northern Karelia		North Savonia	
Year	Method	Men	Women	Men	Women
1972	Bayes	50.3 (49.0, 51.9)	11.9 (11.0, 12.9)	49.7 (48.6, 50.8)	13.5 (12.7, 14.5)
1972	Complete case	52.2 (50.3, 54.5)	11.7 (11.0, 13.9)	50.9 (49.4, 53.0)	13.2 (12.0, 14.6)
1977	Bayes	48.5 (46.5, 51.0)	13.2 (12.0, 14.5)	46.4 (45.1, 47.9)	14.4 (13.4, 15.4)
1977	Complete case	43.1 (41.1, 45.4)	8.9 (7.5, 10.4)	43.1 (41.4, 45.0)	11.0 (9.8, 12.3)
1982	Bayes	43.9 (41.3, 46.5)	18.9 (17.0, 21.1)	46.5 (44.0, 49.0)	18.8 (17.2, 20.6)
1982	Complete case	36.1 (34.0, 39.1)	14.2 (12.6, 15.9)	42.6 (40.6, 45.6)	15.9 (14.4, 18.2)
1987	Bayes	39.0 (36.4, 42.5)	17.8 (16.3, 19.6)	43.6 (40.7, 46.7)	17.2 (15.5, 19.2)
1987	Complete case	34.4 (33.7, 38.9)	15.7 (13.3, 16.6)	39.8 (36.8, 43.6)	15.3 (13.0, 17.7)
1992	Bayes	35.5 (31.2, 39.8)	17.6 (15.6, 20.0)	38.3 (35.2, 41.7)	20.1 (18.0, 22.1)
1992	Complete case	31.2 (28.3, 36.1)	16.0 (14.2, 19.2)	35.3 (32.2, 39.7)	18.0 (15.8, 20.7)
1997	Bayes	33.5 (29.8, 37.5)	20.2 (17.9, 22.9)	32.5 (28.6, 36.8)	19.5 (17.8, 21.5)
1997	Complete case	31.1 (28.3, 35.4)	16.5 (14.4, 19.6)	30.7 (27.7, 35.5)	17.0 (14.7, 19.6)
2002	Bayes	34.5 (30.4, 39.2)	24.8 (22.2, 27.6)	36.5 (32.4, 41.1)	21.3 (19.0, 24.1)
2002	Complete case	32.5 (28.4, 35.8)	22.4 (19.9, 25.4)	34.3 (31.2, 38.6)	19.2 (17.0, 22.1)
2007	Bayes	30.4 (25.5, 35.5)	18.6 (15.9, 21.7)	29.3 (24.4, 35.5)	20.4 (17.6, 23.5)
2007	Complete case	29.2 (26.0, 34.0)	17.3 (14.5, 20.5)	28.3 (25.7, 34.1)	19.5 (16.9, 23.0)

References

Brooks SP and Gelman A (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–55.

Celeux G, Marin J-M and Robert CP (2006) Iterated importance sampling in missing data problems. *Computational Statistics & Data Analysis*, **50**, 3386–3404.

Doll R and Hill AB (1956) Lung cancer and other causes of death in relation to smoking. *British Medical Journal*, **2**, 1071–81.

Haario H, Saksman E and Tamminen J (2001) An adaptive metropolis algorithm. *Bernoulli*, **7**, 223–42. URL https://projecteuclid.org/euclid.bj/1080222083 (last accessed 14 August 2017).

Harald K, Salomaa V, Jousilahti P, Koskinen S and Vartiainen E (2007) Non-participation and mortality in different socioeconomic groups: The FINRISK population surveys in 1972–92. *Journal of Epidemiology & Community Health*, **61**, 449–54.

Jousilahti P, Salomaa V, Kuulasmaa K, Niemelä M and Vartiainen E (2005) 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 & Community Health*, **59**, 310–15.

Karvanen J (2015) Study design in causal models. *Scandinavian Journal of Statistics*, **42**, 361–77.

Karvanen J, Tolonen H, Härkänen T, Jousilahti P and Kuulasmaa K (2016) Selection bias was reduced by recontacting nonparticipants. *Journal of Clinical Epidemiology*, **76**, 209–17.

Kopra J, Härkänen T, Tolonen H and Karvanen J (2015) Correcting for non-ignorable missingness in smoking trends. *Stat*, **4**, 1–14.

Laatikainen T, Tapanainen H, Alfthan G, Salminen I, Sundvall J, Leiviskä J, Harald K, Jousilahti P, Salomaa V and Vartiainen

E (2003) *FINRISKI 2002: Tutkimuksen toteutus ja tulokset*. Helsinki, Finland: National Public Health Institute.

Mannino DM and Buist AS (2007) Global burden of COPD: Risk factors, prevalence, and future trends. *The Lancet*, **370**, 765–73.

Plummer M (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vol. 124, p. 125). Wien, Austria: Technische Universit at Wien.

Plummer M (2015) *rjags: Bayesian graphical models using MCMC*. R package version 3–15. URL https://CRAN.R-project.org/package=rjags (last accessed 14 August 2017).

R Foundation for Statistical Computing (2016) *R: A language and environment for statistical computing*. Vienna, Austria. URL https://www.R-project.org/ (last accessed 14 August 2017).

Robert CP and Casella G (2004) *Monte Carlo statistical methods*. New York, NY: Springer.

Rubin DB (1976) Inference and missing data. *Biometrika*, **63**, 581–92.

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

Tanner MA and Wong WH (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–40.

van Buuren S (2012) *Flexible imputation of missing data*. Boca Raton, FL: CRC press.

van Buuren S and Groothuis-Oudshoorn K (2011) Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, **45**, 1–67. URL http://www.jstatsoft.org/v45/i03/ (last accessed 14 August 2017).

White IR and Royston P (2009) Imputing missing covariate values for the Cox model. *Statistics in Medicine*, **28**, 1982–98.

IV

Follow-up data improve the estimation of the prevalence of heavy alcohol consumption.

Kopra, J., Mäkelä, P., Tolonen H., Jousilahti, P. and Karvanen, J.

Submitted.

Follow-up data improve the estimation of the prevalence of heavy alcohol consumption

January 24, 2018

Juho Kopra[1], Pia Mäkelä[2], Hanna Tolonen[2], Pekka Jousilahti[2] and Juha Karvanen[1]

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

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

Corresponding author:
Juho Kopra
Department of Mathematics and Statistics
P.O. Box 35 (MaD)
FI-40014 University of Jyväskylä
Finland
Email: juho.j.kopra@jyu.fi
Telephone: +358408053455
Fax: –

Abstract

Aims: We aim to adjust for potential non-participation bias in the prevalence of heavy alcohol consumption.

Methods: Population survey data from Finnish health examination surveys conducted in 1987-2007 were linked to the administrative registers for mortality and morbidity follow-up until end of 2014. Utilising these data, available for both participants and non-participants, we model the association between heavy alcohol consumption and alcohol-related disease diagnoses.

Results: Our results show that the estimated prevalence of heavy alcohol consumption is on average 1.5 times higher for men and 1.8 times higher for women than what was obtained from participants only (complete case analysis). The magnitude of the difference in the mean estimates by year varies from 0 to 9 percentage points for men and from 0 to 2 percentage points for women.

Conclusion: The proposed approach improves the prevalence estimation but requires follow-up data on non-participants and Bayesian modelling.

Keywords: survey, Bayesian analysis, register linkage, non-response bias, data missing not at random

Introduction

Reliable information about the prevalence of heavy alcohol consumption is important because alcohol-related health problems and undesired social consequences (Klingemann and Gmel, 2001) cause significant costs in many countries (Rehm et al., 2009). Prevalence estimates can be obtained through health surveys, but the low participation rates (Galea and Tracy, 2007) imperil the reliability of the results. If the participation is selective with respect to alcohol consumption, the estimates of alcohol use suffer from non-participation bias, which hinders their usability for decision-making. If non-participants have worse health than participants, the bias usually leads to an overly positive image of the health of the population.

Empirical evidence suggests that participation is often selective concerning alcohol consumption. Studies from Canada (Zhao et al., 2009), England (Boniface et al., 2017), Finland (Karvanen et al., 2016; Kopra et al., 2017b), Norway (Torvik et al., 2012), Scotland (Gorman et al., 2014) Sweden (Romelsjö, 1989), and the United States (Dawson et al., 2014) conclude that non-participants drink more alcohol than participants. Follow-up studies have shown that non-participants tend to have a higher risk of alcohol-related diseases (Romelsjö, 1989; Jousilahti et al., 2005; Gorman et al., 2014; Christensen et al., 2015; Karvanen et al., 2016), and increased risk of hospitalisations and death (Jousilahti et al., 2005; Christensen et al., 2015; Karvanen et al., 2016), which indicates that non-participants tend to use more alcohol than participants. A study from Netherlands (Lahaut et al., 2002) found that non-participants are more often abstainers than participants, which is not directly interpretable as a conflicting result because many social factors may be associated with abstaining. An older study from Sweden did not find an indication of selective participation (Halldin, 1985).

In addition to selective non-participation, bias may be introduced by imperfect coverage of the target population by the survey sampling frame, and by questionnaire design. First, if some individuals of target population cannot be invited to a survey, the sample does not represent the population of interest and the estimates will be biased. Mäkelä and Huhtanen (2010) observed that in Finland, persons who cannot be invited to a survey due to missing home address have about four times higher risk for alcohol-related deaths. This caused a small bias in the population estimates. Second, Livingston and Callinan (2015) claim that quantity-frequency design of the alcohol use questions underestimates alcohol consumption by one-third compared to asking about drinking with a within-location beverage-specific design. Gmel (2000) reported that alcohol as a subject of survey study does not have an impact on participation in comparison to similar questionnaire without alcohol-related questions.

Studies from the United States (Dawson et al., 2014) and Finland (Mäkelä, 2003) have shown that the non-participation bias cannot be adjusted using just weights depending on basic demographic variables. Some studies adjust for selective non-participation utilising continuum of resistance model (Zhao et al., 2009; Meiklejohn et al., 2012; Boniface et al., 2017) but there are also other methods (Karvanen et al., 2016; Kopra et al., 2017b).

In Finnish health surveys participation has been decreasing, while reported alcohol consumption has been mainly increasing from 1960s to 2007 (Mäkelä et al., 2012). Jousilahti et al. (2005) report that in Finland non-participants have a higher risk of alcohol-specific diseases and death (Jousilahti et al., 2005), which is why we expect the estimates of heavy alcohol consumption to be biased. The difference in the disease risk could be explained by heavier alcohol consumption among non-participants. From previous studies (Harald et al., 2007; Hirvonen, 2017), we know that participation in the FINRISK Study is affected by age, gender, area, and education (Reinikainen et al., 2017).

We aim to adjust for selective non-participation for heavy alcohol consumption and to estimate the prevalence of heavy alcohol consumption with reduced bias. We present a Bayesian solution that is based on mortality and morbidity follow-up data.

Methods

Data

We used data from the National FINRISK Study, which is a series of cross-sectional health examination surveys (Borodulin et al. , 2017) conducted in Finland every fifth year since 1972. We analyzed data for the years 1987–2007. Years 1972–1982 as well as 2012 were excluded because the questions of alcohol consumption were too different from the questions in 1987–2007.

In 1987 and 1992 studies the questions were essentially the same. In 1997 the study questions regarding the usage of cider or mild wine (alc. vol under 5%) were added. Otherwise the study remained the same as earlier. In 2002, question regarding comsumption of red wine and other wines were separated from each other. Also in 20020, the total alcohol consumption was no more base don several alcohol beverage-specific questions but participants were adviced to calculate their number of alcohol portions (standard drinks) consumed and sign asuitable class from quantity-frequency table. In 2007, the alcohol questions were the same as in 2002, but the instructions for the calculation of daily alcohol consumption were improved.

The surveys provide data on 25–74 -year-old adults from six regions of Finland. We restrict the data to people aged 25–65 -years since oldest age group 65–74 years old was not available in all areas until 2007. The survey consists of questionnaires, and a health examination carried out at a local study clinic. The sample was drawn from the National Population Register and was stratified by region, gender and 10-year age-group. In total, there are $44,317$ invitees including $31,567$ participants. The survey data contains self-reported alcohol consumption and background variables, age, gender, region and study year for the whole sample. The survey utilised a beverage-specific quantity-frequency questions on alcohol use in the first three surveys and graduate frequency measure in the latter two surveys. The questions related to alcohol consumption are provided in Appendix B. The study questionnaires in 1987, 1992 and 1997 asked alcohol usage one type of alcohol beverages at a time: beer, spirits/vodka, long drink or cider (in 1997), wines and mild wines (in 1992 and 1997). In 2002 and 2007 the questionnaire was different, and individuals reported their alcohol consumption for all beverage types in one question.

From these questions, a number of standard drinks consumed per week during a previous 12 months was calculated. One standard drink equals 12 grams of pure alcohol which is equivalent to, e.g, one bottle of beer (33cl, 4.7 volume percent of alcohol). Based on the number of standard drinks, the (self-reported) total amount of 100% alcohol consumed in the previous 12 months was calculated.

Since our main interest was to estimate the prevalence of heavy alcohol users, we classified participants as heavy alcohol consumers and others (non-heavy alcohol consumers) as follows. The persons who reported consuming on average at least 24 standard drinks per week for men or at least 16 standard drinks per week for women during the one-year period before the examination were considered heavy alcohol consumers.

The survey data were linked to three registers: The Register of Completed Education and Degrees (Statistics Finland, 2016), Care Register for Health Care (National Institute for Health and Welfare , 2017) and Cause of Death Register (Statistics Finland, 2017) using personal identification code. The register-based data were available for both participants and non-participants. The level of education is categorised according to the International Standard Classification of Education (ISCED, 2011). We classified education into three levels: 1) high level (tertiary education, ISCED levels 5-8), 2) middle level (secondary education, ISCED levels 3-4) and 3) low level (primary education or less or unknown, ISCED levels 0-2). The Care Register gives data about the hospital visits with dates and ICD-codes for both participants and non-participants. From the Causes of Death Register, we obtain data about dates and ICD-codes of the cause of death.

Follow-up data contains the time-to-event (age) and ICD code of the first alcohol-related disease diagnosis or death. The ICD-codes we considered to be alcohol related are listed in Table 1. The follow-up begins from the survey and ends at the end of 2014. Persons who have neither alcohol-related disease diagnosis nor alcohol-related cause of death are censored at the end of the follow-up. Deaths not related to alcohol are treated as censorings.

Table 1: The ICD-codes interpreted as alcohol-related events.

ICD-9:

291	Alcohol-induced mental disorders
303	Alcohol dependence syndrome
357.5	Alcoholic polyneuropathy
425.5	Alcoholic cardiomyopathy
535.3	Alcoholic gastritis
571.0	Alcoholic fatty liver
571.1	Acute alcoholic liver disease
571.2	Alcoholic cirrhosis of liver
571.3	Alcoholic liver damage, unspecified
577.0D-F	Alcoholic disease of the pancreas, acute
577.1C-D	Alcoholic disease of the pancreas, chronic
980.0	Toxic effect ethyl alcohol
980.2	Toxic effect of isopropyl alcohol
980.8	Toxic effect of other specified alcohols
980.9	Toxic effect of other unspecified alcohol
E851	Accidental poisoning by alcohol

ICD-10:

F10	Mental and behavioural disorders due to use of alcohol
G31.2	Degeneration of nervous system due to alcohol
G62.1	Alcoholic polyneuropathy
G72.1	Alcoholic myopathy
I42.6	Alcoholic cardiomyopathy
K29.2	Alcoholic gastritis
K70	Alcoholic liver disease
K85.2	Alcohol-induced acute pancreatitis
K86.0	Alcohol-induced chronic pancreatitis
T51	Toxic effect of alcohol
X45	Accidental poisoning or other exposure to alcohol
Y15	Poisoning by and exposure to alcohol, undetermined intent

Complete case analysis

The complete case analysis (e.g. mean estimate from the participants) assumes that participation is not selective concerning alcohol consumption. Violations of this assumption lead to bias. We compared the results of complete case analysis to a Bayesian approach, which relies on more realistic assumptions and allows for selective non-participation concerning heavy alcohol use.

Modelling approach

We applied a Bayesian approach introduced in (Kopra et al., 2017a) to estimate the prevalence of heavy alcohol consumption. The Bayesian model consists of three sub-models which are fitted simultaneously. The sub-models are:

1. Participation model,
2. Risk factor model, and
3. Survival model.

The mathematical formulas for the models are given in Appendix A.

The participation model describes which variables affect participation. Participation is defined as a binary indicator (0 or 1) for the availability information on alcohol consumption. This model is a logistic regression model with linear covariates for study year and age, and categorical variables for the region (4 levels), education (3 levels) and the alcohol consumption (binary). The model also takes into account the possible interactions of gender and study year, gender and alcohol consumption, and study year and alcohol consumption.

The risk factor model describes how alcohol consumption (heavy or non-heavy) varies by background variables. By background variables we mean age, gender, region, study year and education. The model is a logistic regression model with interactions for the year of birth with gender, region, study year and education.

The survival model describes the relationship between alcohol consumption and alcohol-related diseases. All disease events are combined and modelled as one survival outcome. The survival model is a piecewise constant hazard model with one-year baseline hazard period terms. The model assumes monotonically increasing baseline hazard, which is accomplished using prior specification. In addition to baseline hazard, alcohol consumption is used as a regressor. Both baseline hazard terms and the regression coefficient are gender-specific. The model assumes that the disease risk of non-participants must be between the risks of heavy alcohol consumer participants and other participants. This follows from our reasoning that if the risk of non-participants were the same as the risk of heavy alcohol consumers, we would expect all of them be heavy alcohol consumers. Similarly, if the risk of non-participants equaled to the risk of non-heavy alcohol consumers, we would expect none of them to be heavy alcohol consumers.

Prior distributions

We used weakly informative prior distributions that reflect the existing knowledge but have variances large enough to allow for surprises. This approach is recommended in textbooks on Bayesian statistics (Gelman et al., 2014), and there exists guidelines for elicitation of prior distributions (O'Hagan et al. , 2006). The participation model needed an informative prior for the effect of heavy alcohol consumption on the participation, i.e., for the strength of selectivity mechanism. Some degree of subjectivity cannot be avoided in the prior specification. To define a weakly informative prior, we took a 45-year-old non-heavy alcohol consumer who participates with probability 0.7 as a reference and considered the prior probability for a heavy alcohol consumer who is otherwise similar. We elicit that there is 25% chance that person participates with probability p lower than 0.5 ($P(p \leq 0.5) = 0.25$), 35% chance for $p \leq 0.6$, and 50% chance for $p \leq 0.7$. The functional form of the prior distribution was chosen to be logistic distribution. These elicitations lead to logistics prior distribution with expected value zero and variance 1/2.05.

In the survival model, we applied monotonically increasing baseline hazards separately for men and women. The prior distribution for the first hazard term (25–26-year-olds) was a uniform distribution with a range from 0 to 20. From second hazard term (26–27-year-olds) to the last hazard term (99–100-years-old), each term had a uniform distribution with the lower limit being the value of previous baseline hazard term and upper limit 20.

All the remaining model parameters had normally distributed priors with zero mean and variance 1000. The prior distributions are presented using mathematical notation in Table 4 of Appendix A.

Imputations and model fitting

Alcohol consumption was missing for the non-participants. These missing values (heavy or non-heavy) were imputed simultaneously with Bayesian model fitting using data augmentation (Tanner and Wong, 1987). The model was fitted using Markov chain Monte Carlo (MCMC) (Robert and Casella, 2004) and implemented with Just Another Gibbs Sampler (JAGS) -software (Plummer, 2003) and R software (R Core Team, 2017) with `rjags` -package (Plummer, 2015). The convergence of MCMC chains were investigated using Brooks-Gelman $\hat{R}$ diagnostics (Brooks and Gelman, 1998), and all the $\hat{R}$s were below 1.01 which indicates convergence. The model fitting utilised computational resources of IT Center for Science Ltd (CSC).

Results

Descriptive statistics

In Table 2, we present the descriptive statistics on age, education and gender. These variables are examined by study year, and comparisons can be made between participants and non-participant as well as between heavy alcohol consumers and other alcohol consumers.

The average age of the non-participants was lower than the average age of the participants. Over the years, the average age appears not to have changed much for the non-participants, but it has slightly increased for the participants. Among participants, the average age of heavy alcohol consumers has increased more rapidly than for non-heavy alcohol consumers. The average age of heavy alcohol consumers was 41.7 in 1987 (44.4 for non-heavy) and it has increased between each study being 47.3 for heavy alcohol consumers and 45.5 for non-heavy alcohol consumers in 2007. The average age of non-heavy alcohol consumers has also increased between the studies, except between the 1997 and 2002 when it decreased by 0.2 years.

The level of education has increased for both participants and non-participants during the study period. The non-participants tend to have low education more often than participants, and participants tend to have high education more often than non-participants. In 1987, there were a higher proportion of highly educated participants among heavy alcohol consumers than among non-heavy alcohol consumers. In 2007, the situation was opposite; the proportion of highly educated persons is higher for non-heavy alcohol consumers than for the heavy alcohol consumers. The proportion of women among participants has slightly increased from 52.0% to 53.4% during 1987–2007. Among non-heavy alcohol consumers, the proportion is higher: 52.7%–55.3%. Women are a minority among heavy alcohol consumers. There were 15.9% women among heavy alcohol consumers in 1987, and the proportion has notably increased being 27.8% in 2007.

The proportion of women was higher among the participants than among non-participants. The proportion of women among participating heavy alcohol users has been rapidly increasing over the years, while the corresponding proportion had not increased by much among non-heavy alcohol consumers.

The number of invitees, the participation rate and the number of events for both participant and non-participant men and women are presented in Table 3. During the study period, the proportion of heavy alcohol consumers

Table 2: Description of background information by study year for non-participants, participants, and heavy and non-heavy alcohol consumers among participants.

		Participants		
Year	Non-participants	All	Heavy alcohol consumers	Moderate alcohol consumers
Average age:				
1987	42.5	44.4	41.7	44.4
1992	41.8	44.7	44.5	44.7
1997	42.8	45.0	45.0	45.0
2002	42.3	44.9	45.7	44.8
2007	41.9	45.6	47.3	45.5
High education (%):				
1987	13.6	18.5	23.0	18.4
1992	18.4	26.6	30.2	26.5
1997	24.7	29.9	31.4	29.8
2002	25.6	35.7	32.7	35.8
2007	27.7	38.6	34.3	38.8
Middle education (%):				
1987	30.6	31.8	31.0	31.8
1992	35.6	34.9	34.4	34.9
1997	37.7	38.1	37.6	38.1
2002	41.8	40.6	43.5	40.4
2007	45.3	44.2	45.5	44.1
Low education (%):				
1987	55.8	49.7	46.0	49.8
1992	46.1	38.5	35.4	38.6
1997	37.6	32.1	31.0	32.1
2002	32.6	23.8	23.9	23.8
2007	27.0	17.2	20.2	17.1
Women (%):				
1987	42.3	52.0	15.9	52.7
1992	40.5	53.0	21.2	54.1
1997	43.2	52.9	22.8	54.3
2002	41.6	53.6	29.0	54.9
2007	42.7	53.9	27.8	55.3

Table 3: Number of invitees, the participation rate, the prevalence of heavy alcohol consumption based on participants and Bayesian modelling (posterior mean), and the number of alcohol-related incident events (per 1000 follow-up years) for the non-participant and the participant men and women.

Year	Invited N	Participation rate	Prevalence for participants	Posterior mean	Alcohol-related incident events (per 1000 follow-up years) Participants	Non-participants
Men						
1987	3910	79.5%	5.0%	9.6%	202 (2.8)	91 (5.2)
1992	3888	73.3%	7.7%	15.0%	168 (2.9)	128 (6.5)
1997	4034	70.0%	9.2%	9.7%	150 (3.2)	103 (5.4)
2002	3955	66.5%	14.4%	22.9%	118 (3.6)	62 (3.8)
2007	3202	61.8%	11.0%	12.8%	47 (3.1)	35 (3.7)
Women						
1987	3961	85.1%	0.7%	2.4%	52 (0.6)	29 (2.0)
1992	3951	81.0%	2.5%	4.4%	46 (0.7)	23 (1.5)
1997	4031	75.8%	3.7%	5.3%	25 (0.5)	37 (2.3)
2002	4019	75.4%	5.7%	5.9%	34 (0.9)	13 (1.0)
2007	3278	71.3%	4.0%	5.8%	12 (0.7)	10 (1.3)

has increased for both men and women among participants, and simultaneously the participation rate has decreased.

The probabilities for not having alcohol-related disease diagnosis up to the given age for men and women are presented by Kaplan-Meier survival plots (Figure 1). The top row shows that the non-participants were more likely to have alcohol-related diagnoses than participants. The lower row shows that the risk for non-participants lies between the risks of heavy and non-heavy alcohol consumers, which is a requirement for the utilised Bayesian model. The number of persons with a disease diagnosed in each group is reported next to the survival curve in the Figure 1.

Adjusted prevalences of heavy alcohol consumption

Figure 2 presents the trends of the prevalence of heavy alcohol consumption, based on complete case analysis and the Bayesian modelling. It can be seen that the mean estimates of the Bayesian approach lie above the estimates of the complete case analysis. The numeric values are presented in Table 3.

To compare the prevalence estimates based on participants only, and the

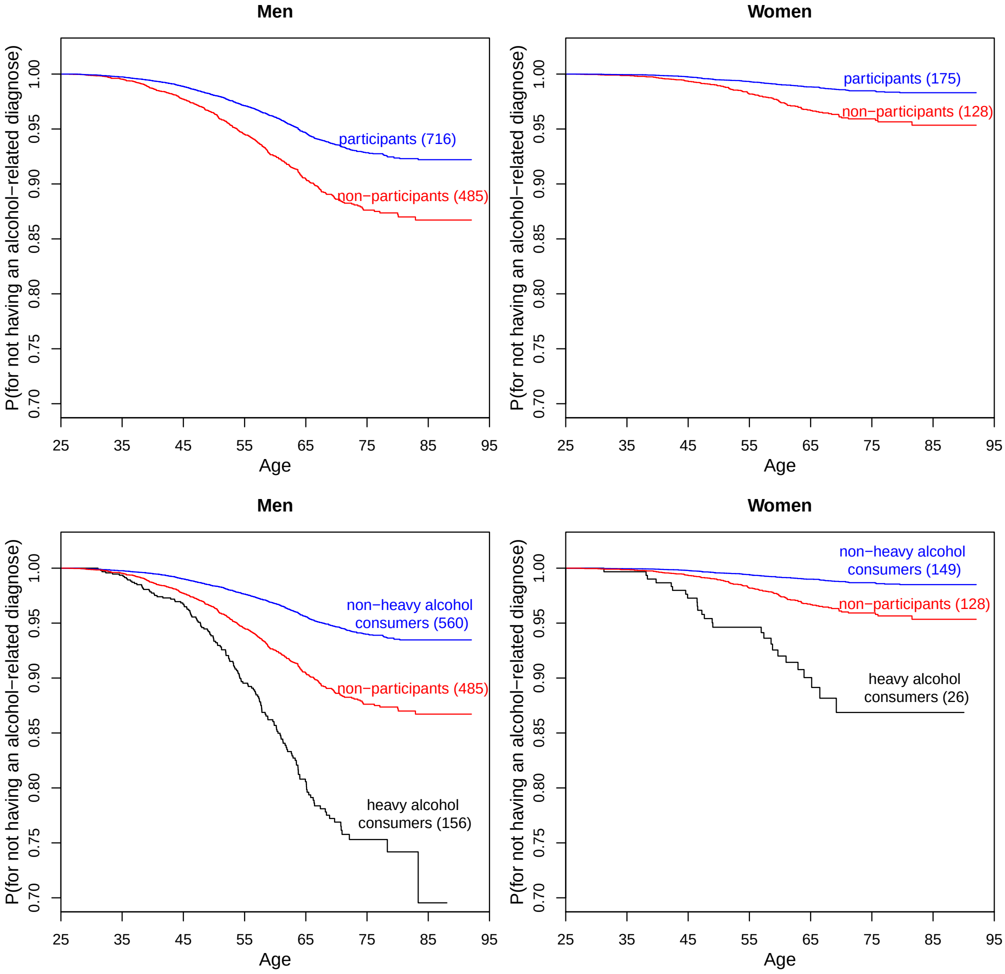

Figure 1: Kaplan-Meier survival plots for men and women comparing the probabilities of not having alcohol-related diagnoses among participants and non-participants (upper panels) and for heavy, non-heavy alcohol consumers and non-participants (lower panels). The number of persons with a disease diagnosed in each group is reported within parenthesis.

posterior estimate for the prevalence of entire survey, absolute and relative differences can be calculated. For men, the absolute difference of the yearly prevalence estimates for 1987–2007 are 4.6, 7.3, 0.5, 8.6, and 1.8 percentage points calculated from Table 3, respectively. Those lead to average difference of 4.6 percentage points. The corresponding relative differences for men are 1.93 (i.e. almost a two-fold difference), 1.95, 1.06, 1.6 and 1.17, respectively, and average relative difference is 1.5. For women, the corresponding values are yearly absolute differences; 1.7, 1.9, 1.6, 0.3 and 1.9, respectively, leading to average absolute difference of 1.5 percentage points. The yearly relative differences are 3.39, 1.77, 1.42, 1.04 and 1.47, respectively, leading to average relative difference of 1.8, see Table 3.

For men, the mean estimates based on Bayesian model vary year by year, but the credible intervals do not exclude the possibility of a monotonically increasing trend from 1987 to 2002. The complete case estimates are outside of the 90% credible interval of Bayesian trends in 1987, 1992, and 2002.

The credible intervals are narrower for women than for men. For women, the complete case prevalence estimates are outside of the 90% credible intervals of Bayesian trends in 1987 and 1992, and are within the credible interval in 1997, 2002 and 2007.

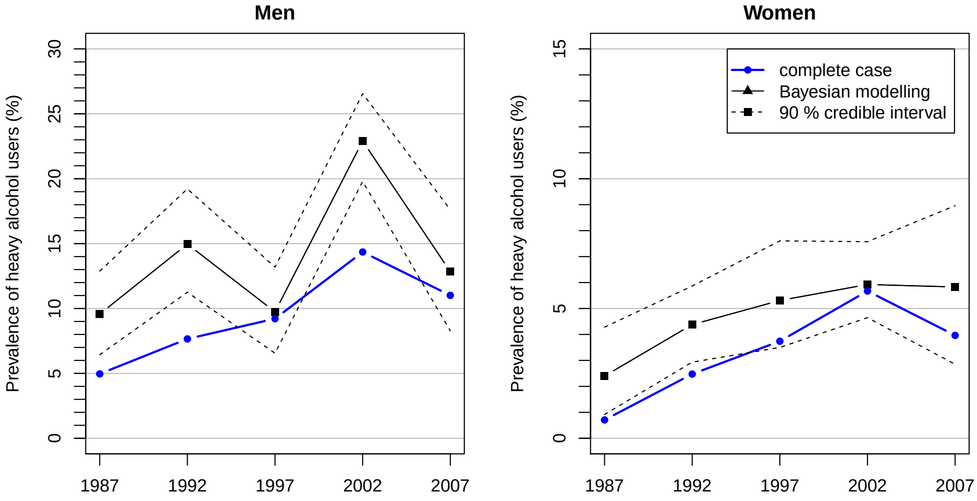

Figure 2: Comparison of prevalence estimates of complete case analysis and Bayesian multiple imputation adjusted for education. Note that the scales of the vertical axis for men and women are different from each other.

Discussion

There is evidence that non-participation in a survey asking about alcohol consumption is selective with respect to heavy alcohol consumption in Finland and in many other countries. We studied the prevalence of heavy alcohol consumption based on data from the National FINRISK Study, which suffer from selective non-participation. In FINRISK data, the average self-reported alcohol consumption for men was equal to 5.9 liters and for women 1.9 liters of pure 100% alcohol per year. For comparison, the national consumption statistics by National Institute for Health and Welfare (2016) show that the average yearly consumption of 100% alcohol for persons at least 15 years old was in the range of 10–13 liters per person during 1987–2007. Thus, in FINRISK data the self-reported consumption is about 60–70% lower what has been reported in the national consumption statistics (which were not used in our modelling in any way). Although many reasons can partly explain the differences between the consumption statistics and self-reported data, e.g. questionnaire design and imperfect matching of survey frame with the target population, the differences between non-participants and participants in the follow-up data summarized in Figure 1 suggest that selection bias is present.

We observed differences in alcohol-related events for participants and non-participants. Non-participants had significantly increased risk for alcohol-related disease or death compared to participants, and men had a higher risk than women. This phenomenon has also been observed for other data, see (Romelsjö, 1989), (Gorman et al., 2014) and (Christensen et al., 2015).

When participation is selective with respect to variables to be studied, which is the case for alcohol use, the estimates from complete case analysis are affected by non-participation bias and the real level of uncertainty is hidden, e.g. confidence intervals are not wide enough when complete case analysis is used. Mäkelä (2003) and Dawson et al. (2014) demonstrated that this kind of bias cannot be reduced for alcohol data with demographic information.

We compared the estimates obtained by a complete case analysis to estimates obtained by adjusting for non-participation with a full Bayesian modelling approach. The Bayesian approach gave a higher estimate of heavy alcohol consumption than the complete case analysis. Our approach reduced the bias and made the uncertainty visible. We estimated that the magnitude of bias is 0–9 percentage points for men and 0–2 percentage points for women in the FINRISK data. The Bayesian mean estimate was on average 1.5 times higher for men and 1.8 times higher for women compared to participants.

The use of our approach requires follow-up data and background variables for the entire invited sample (including non-participants), follow-up time long enough to observe alcohol-related disease events and Bayesian modelling. The first requirement cannot be fulfilled in many countries because of lack of register data or legal restrictions for data linkage. The second requirement means that the prevalence estimates will be available only several years after the survey. This requirement may be relaxed if there exist earlier surveys that can be assumed to share the same model parameters with the current survey. The third requirement is the easiest to fulfill because it only calls for statistical expertise that is widely available.

To conclude, the prevalence of heavy alcohol consumption based on survey participants only appears to be biased downward for both men and women. The magnitude of observed absolute bias was larger for men than women. The proposed non-participation adjustment approach is useful in context of alcohol research when follow-up data on non-participants are available, and the modelling requirements are met. The follow-up data can be used to improve the estimation of the prevalence of heavy alcohol consumption.

Declaration of conflicting interest

The Authors declare that there is no conflict of interest.

Funding

This work was supported by the Finnish Foundation for Alcohol Studies and Academy of Finland [grant numbers 266251 and 311877].

References

Boniface S, Scholes S, Shelton N, Connor J. (2017) Assessment of non-response bias in estimates of alcohol consumption: applying the continuum of resistance model in a general population survey in England. PloS one, 12, e0170892.

Borodulin K, Tolonen H, Jousilahti P et al. (2017) Cohort Profile: The National FINRISK Study. Int J Epidemiol doi: 10.1093/ije/dyx239 http://dx.doi.org/10.1093/ije/dyx239 (28 November 2017, date last accessed).

Brooks SP, Gelman A. (1998) General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 7: 434-455.

Christensen AI, Ekholm O, Gray L, Glümer C, Juel K. (2015) What is wrong with non-participants? Alcohol-, drug- and smoking related mortality and morbidity in a 12-year follow up study of participants and non-participants in the Danish Health and Morbidity Survey. Addiction 110: 1505-12.

Dawson DA, Goldstein RB, Pickering RP, Grant BF. (2014) Nonresponse bias in survey estimates of alcohol consumption and its association with harm. J Stud Alcohol Drugs 75: 695-703.

Galea S, Tracy M. (2007) Participation rates in epidemiologic studies. Ann Epidemiol 17: 643-653.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. (2014) Bayesian Data Analysis, Third edition. Boca Raton, FL, USA: Chapman & Hall/CRC.

Gmel G (2000) The effect of mode of data collection and of nonresponse on reported alcohol consumption: a splitsample study in Switzerland. Addiction 95: 123-134.

Gorman E, Leyland AH, McCartney G et al. (2014) Assessing the representativeness of population-sampled health surveys through linkage to administrative data on alcohol-related outcomes. Am J Epidemiol 180: 941-948.

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

Halldin J. (1985) Alcohol consumption and alcoholism in an urban population in central Sweden. Acta Psychiat Scand 71: 128-140.

Harald K, Salomaa V, Jousilahti P, Koskinen S, Vartiainen E. (2007) Non-participation and mortality in different socioeconomic groups: the FINRISK population surveys in 1972-92. J Epidemiol Commun H 61: 449-454.

Hirvonen E. (2017) Puuttuvuuden mallintaminen FINRISKI -tutkimuksessa (in Finnish). Masters thesis, University of Jyväskylä, `http://urn.fi/URN:NBN:fi:jyu-201706192945`. (19 October 2017, date last accessed).

International Standard Classification of Education: ISCED 2011. UIS, Montreal, Quebec.

Jousilahti P, Salomaa V, Kuulasmaa K, Niemelä M, Vartiainen E. (2005) 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. J Epidemiol Commun H 59: 310-315.

Karvanen J, Tolonen H, Härkänen T, Jousilahti P, Kuulasmaa K. (2016) Selection bias was reduced by re-contacting nonparticipants. J Clin Epidemiol 76: 209-217.

Klingemann, H, Gmel, G (Eds.) (2001) Mapping the social consequences of alcohol consumption. Dordrecht, The Netherlands, Kluwer Academic Publishers.

Kopra J, Härkänen T, Tolonen H, Karvanen J. (2015) Correcting for non-ignorable missingness in smoking trends. Stat, 4: 1-14.

Kopra J, Karvanen J, Härkänen T. (2017a) Bayesian models for data missing not at random in health examination surveys. Stat Model, Advance online publication, doi:10.1177/1471082X17722605.

Kopra, J, Härkänen, T, Tolonen, H, Jousilahti, P, Kuulasmaa, K, Reinikainen, J, Karvanen, J. (2017b) Adjusting for selective non-participation with re-contact data in the FINRISK 2012 survey. Scand J Public Healt, Advance online publication, doi:https://doi.org/10.1177/1403494817734774.

Lahaut VM, Jansen HA, Van de Mheen D, Garretsen HF. (2002) Adjusting for selective nonparticipation with recontact and hospitalisation history data. Alcohol Alcoholism 37: 256-260.

Livingston M, Callinan S. (2015) Underreporting in alcohol surveys: whose drinking is underestimated? J Stud Alcohol Drugs 76: 158-164.

Meiklejohn J, Connor J, Kypri K (2012) The effect of low survey response rates on estimates of alcohol consumption in a general population survey. PLoS One 7: e35527.

National Institute for Health and Welfare *Care Register for Health Care*. Helsinki, Finland. `http://www.thl.fi/en/web/thlfi-en/statistics/information-on-statistics/register-descriptions/care-register-for-health-care`. (26 October 2017, date last accessed).

National Institute for Health and Welfare: *Alcoholic Beverage Consumption 2016 [e-publication]*. Helsinki, Finland. `https://www.thl.fi/fi/tilastot/tilastot-aiheittain/paihteet-ja-riippuvuudet/alkoholi/alkoholijuomien-kulutus`. (14 September 2017, date last accessed).

Mäkelä P. (2003) Impact of correcting for nonresponse by weighting on estimates of alcohol consumption. J Stud Alcohol 64: 589-596.

Mäkelä P, Huhtanen P. (2010) The effect of survey sampling frame on coverage: The level of and changes in alcoholrelated mortality in Finland as a test case. Addiction 105: 1935-1941.

Mäkelä P, Tigerstedt C, Mustonen H. (2012) The Finnish drinking culture: change and continuity in the past 40 years. Drug Alcohol Rev 31: 831-840.

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T. (2006) Uncertain Judgements: Eliciting Experts' Probabilities. Chichester, England, John Wiley & Sons.

Plummer M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.* In Proceedings of the 3rd international workshop on distributed statistical computing. **124**, p. 125. Wien, Austria: Technische Universit at Wien.

Plummer M. (2015). *rjags: Bayesian Graphical Models using MCMC.* R package version 3-15. `https://CRAN.R-project.org/package=rjags`. (8 November 2017, date last accessed).

R Foundation for Statistical Computing (2017) *R: A Language and Environment for Statistical Computing.* Vienna, Austria. `https://www.R-project.org/`. (8 November 2017, date last accessed).

Robert CP, Casella G. (2004) Monte Carlo Statistical Methods. Vienna, Austria, Springer Texts in Statistics.

Rehm J, Mathers C, Popova S, Thavorncharoensap M, Teerawattananon Y, Patra J. (2009) Global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders. Lancet, 373: 2223-2233.

Reinikainen J, Tolonen H, Borodulin K et al. (2017) Participation rates by educational levels have diverged during 25 years in Finnish health examination surveys. Eur J Public Health, Advance online publication, doi:10.1093/eurpub/ckx151.

Romelsjö A. (1989) The relationship between alcohol consumption and social status in Stockholm. Has the social pattern of alcohol consumption changed?. Int J Epidemiol 18: 842-851.

Official Statistics of Finland (OSF): *Causes of death [e-publication].* Helsinki, Finland. `http://tilastokeskus.fi/til/ksyyt/index_en.html`. (26 October 2017, date last accessed).

Statistics Finland *The Register of Completed Education and Degrees.* Helsinki, Finland. `https://www.stat.fi/til/kou_en.html`. (8 November 2017, date last accessed).

Tanner M, Wong W. (1987) *The calculation of posterior distributions by data augmentation.* J Am Stat Assoc 82: 528-540.

Torvik FA, Rognmo K, Tambs K. (2012) Alcohol use and mental distress as predictors of non-response in a general population health survey: the HUNT study. Soc Psych Psych Epid 47: 805-816.

Vartiainen E, Laatikainen T, Peltonen M et al. (2009) Thirty-five-year trends in cardiovascular risk factors in Finland. Int J Epidemiol 39: 504-518.

Zhao J, Stockwell TIM, MacDonald S. (2009) Nonresponse bias in alcohol and drug population surveys. Drug Alcohol Rev 28: 648-657.

Appendix A

The statistical model is based on work presented in (Kopra et al., 2017a), which utilises similar model to estimate the prevalence of smoking.

Notation for the data

For each individual $i = 1, \ldots, N$ invited to a survey, we denote M_i being indicator of participation ($M_i = 1$ for participants and $M_i = 0$ for non-participants). Background information X_i consists of both survey frame and education information and is available for both participants and non-participants. The survey frame has variables gender g_i, region r_i, age a_i and study year s_i. The education is denoted by e_i, and values 1, 2 and 3 corresponds to high, middle and low education, respectively. Thus $X_i = (g_i, r_i, a_i, s_i, e_i)$. The heavy alcohol consumption Y_i is a binary variable such that heavy alcohol consumers have $Y_i = 1$ and non-heavy alcohol consumers have $Y_i = 0$. The variable T_i is the age at the first diagnosis of any of the alcohol-related diseases. The T_i is right censored and left-truncated at the age when the person entered the study.

Participation model

The participation model

$$\begin{aligned} \operatorname{logit}(P(M_i = 1 | X_i, Y_i)) = {} & \alpha_{0[g_i,s_i]} + \alpha_{1[g_i,s_i,e_i]} + \eta_{[g_i,s_i]} Y_i \\ & + \alpha_{2[g_i,Y_i]}(a_i - 45) + \alpha_{3[r_i]}, \end{aligned} \tag{1}$$

is a logistic regression model with following parameters. First, parameter $\alpha_{0[g_i,s_i]}$ is a constant where notation $[g_i, s_i]$ indicates that there are independent α_0 parameters for all levels of gender g_i and study year s_i. Second, parameter $\alpha_{1[g_i,s_i,e_i]}$ is the regression coefficient for education levels. For the lowest education level this parameter is forced to be 0. The parameter $\eta_{[g_i,s_i]}$ describes how heavy alcohol consumption affects participation. For this parameter, we need an informative prior. The parameter $\alpha_{2[g_i,Y_i]}$ describes how age at study affect participation. Finally, $\alpha_{3[r_i]}$ is a term for the region. For one of the regions, this parameter is forced to be 0. We selected a model that included important factors affecting participation while ensuring the convergence of the MCMC chains in Bayesian inference.

Risk factor model

The model for risk factor (heavy alcohol consumption) is

$$\text{logit}(P(Y_i = 1|X_i)) = \beta_{0[g_i,r_i,s_i,e_i]} + (s_i - a_i - 1938)\beta_{1[g_i,r_i,s_i,e_i]}. \tag{2}$$

The risk factor model is stratified by gender g_i, region r_i, study year s_i and education e_i using similar notation as in (1). The parameter $\beta_{0[g_i,r_i,s_i,e_i]}$ is constant for persons born in 1938. The parameter $\beta_{1[g_i,r_i,s_i,e_i]}$ determines how the heavy alcohol consumption prevalence changes with the year of birth.

Survival model

Let $dN_i(t)$ be the number of new events (increment) for the individual i at the time t. The increment follows a Poisson distribution with intensity parameters $\lambda_i(t)$. The intensity $\lambda_i(t)$ is modelled independently for both genders consisting of one-year period piecewise-constant baseline hazard terms $h_{0,0}(t)$ for men and $h_{0,1}(t)$ for women, and heavy alcohol consumption term $\exp(\gamma_1 Y_i)$ and $\exp(\gamma_2 Y_i)$ indicating the effect of heavy alcohol consumption for men and women, respectively

$$\begin{aligned} dN_i(t) &\sim \text{Poisson}(\lambda_i(t)) \\ \lambda_i(t) &= \begin{cases} \exp(\gamma_1 Y_i)\, h_{0,0}(t), & \text{given that } T_i \geq t \text{ and } g = 0 \\ \exp(\gamma_2 Y_i)\, h_{0,1}(t), & \text{given that } T_i \geq t \text{ and } g = 1 \\ 0, & T_i < t. \end{cases} \end{aligned}$$

Prior distributions

The prior distributions are specified in Table 4. The prior distributions for piecewise constant hazard terms $h_{0,0}(t)$ and $h_{0,1}(t)$ are specified such that the hazard becomes increasing function with respect to t.

Table 4: Prior distributions.

Notation	Distribution	Interpretation
Participation model		
η	$\text{Logistic}(0, \tau = 2.05)$	How heavy alcohol consumption affects the participation.
$\alpha_0, \alpha_1, \alpha_2, \alpha_3$	$\text{N}(0, 1000^{-1})$	Other parameters.
Risk factor model		
β_0, β_1	$\text{N}(0, 1000^{-1})$	Other parameters.
Survival model		
$h_{0,0}(25)$	$\text{Unif}(0, 20)$	Hazard for men at age 25–26.
$h_{0,1}(25)$	$\text{Unif}(0, 20)$	Hazard for women at age 25–26.
$h_{0,0}(t), t = 26, 27, \ldots$	$\text{Unif}(h_{0,0}(t-1), 20)$	Hazard for men at age t.
$h_{0,1}(t), t = 26, 27, \ldots$	$\text{Unif}(h_{0,1}(t-1), 20)$	Hazard for women at age t.
γ_1	$\text{N}(0, 1000^{-1})$	How heavy alcohol consumption affects hazard for men.
γ_2	$\text{N}(0, 1000^{-1})$	How heavy alcohol consumption affects hazard for women.

Appendix B

The study questions in 1987

CONSUMPTION OF ALCOHOL

1. **Do you use any alcoholic drinks, even occasionally (f. ex. beer, wine or spirits)?**

   1 yes
   2 no, but I have not quitted completely
   3 no, because I quit using alcohol ...... years ago
   4 I have never used alcohol

   If you have quitted alcohol use, please specify, why did you quit?

	no	yes
For health reasons	1	2
For economic reasons	1	2
For other reasons	1	2

2. **Have you during the past year (last 12 months) had any alcohol (beer, wine or spirits)?**

   1 yes
   2 no (for your part, the questions are completed)

3. **How often do you usually drink beer (III or IV A)?**

   1 daily
   2 a few times a week
   3 about once a week
   4 few times a month
   5 about once a month
   6 about once in a few months
   7 3 - 4 times a year
   8 twice a year
   9 once a year or more seldom

0 never

4. **How much do you usually drink beer at a time?**

   1 less than one bottle
   2 1 bottle
   3 2 bottles
   4 3 bottles
   5 4 - 5 bottles
   6 6 - 9 bottles
   7 10 - 14 bottles
   8 15 bottles or more
   9 I do not drink beer

5. **How often do you usually drink wine (light or strong, also home made)?**

   1 daily
   2 a few times a week
   3 about once a week
   4 a few times a month
   5 about once a month
   6 about once in a few months
   7 3 - 4 times a year
   8 twice a year
   9 once a year or more seldom
   0 never

6. **How much do you usually drink wine at a time?**

   1 half a glass
   2 one glass
   3 two glasses
   4 about half a big bottle
   5 a little less than one big bottle
   6 about one big bottle
   7 from one to two big bottles

8 more than two big bottles
9 I do not drink wine

7. **How often do you usually drink spirits?**

   1 daily
   2 a few times a week
   3 about once a week
   4 a few times a month
   5 about once a month
   6 about once in a few months
   7 3 - 4 times a year
   8 twice a year
   9 once a year or more seldom
   0 never

8. **How much do you usually drink spirits at a time?**

   1 less than one restaurant measure (less than 4 cl)
   2 one restaurant measure (about 4 cl)
   3 two restaurant measures (about 8 cl)
   4 3 - 4 restaurant measures
   5 5 - 6 restaurant measures (about quarter liter)
   6 7 - 10 restaurant measures
   7 about a half liter bottle
   6 more than a half liter bottle
   7 I do not drink spirits

9. **How often have you during the last 12 months had so much beer, wine or spirits that you have felt intoxicated?**

   1 a few times a week or more often
   2 about once a week
   3 a few times a month
   4 about once a month
   5 about once in two months
   6 4 - 5 times a year
   7 2 - 3 times a year
   8 once a year
   9 not even once

The changes in questions from 1987 to 1992

The questions 1, 6 and 7 have with changes in text. We have highlighted the removed text with strikeout font (e.g. ~~removed~~) and added text with italic font (e.g. *added*). The changes are in comparison with the previous survey.

1. **Do you use any alcoholic drinks, even occasionally (f. ex. beer, wine or spirits)?**

   1 ~~yes~~ *yes, at least once a month*
   2 ~~no, but I have not quitted completely~~ *yes, less than once a month*
   3 no, because I quit using alcohol ...... years ago
   4 I have never used alcohol

   ~~If you have quitted alcohol use, please specify, why did you quit?~~

	~~no~~	~~yes~~
~~For health reasons~~	~~1~~	~~2~~
~~For economic reasons~~	~~1~~	~~2~~
~~For other reasons~~	~~1~~	~~2~~

6. **How much do you usually drink wine at a time?**

   1 half a glass
   2 one glass
   3 two glasses
   4 ~~about one small bottle~~ *about half a big bottle*
   5 a little less than one big bottle
   6 about one big bottle
   7 from one to two big bottles
   8 more than two big bottles
   9 I do not drink wine

7. **How much do you usually drink spirits at a time?**

   1 less than one restaurant measure (less than 4 cl)
   2 one restaurant measure (about 4 cl)
   3 two restaurant measures ~~(about 8 cl)~~
   4 3 - 4 restaurant measures
   5 5 - 6 restaurant measures ~~(about quarter liter)~~

6 7 - 10 restaurant measures
7 about a half liter bottle
6 more than a half liter bottle
7 I do not drink spirits

The changes in questions from 1992 to 1997

The questions 4 and 6 have with changes in text. We have highlighted the removed text with strikeout font (e.g. ~~removed~~) and added text with italic font (e.g. *added*). The changes are in comparison with the previous survey.

4. **How much do you usually drink beer at a time?** ***(1 bottle = 1/3 liters.)***

   1 less than one bottle
   2 1 bottle
   3 2 bottle
   4 3 bottles
   5 4 - 5 bottles
   6 6 - 9 bottles
   7 10 - 14 bottles
   8 15 bottles or more
   9 I do not drink beer

6. **How much do you usually drink wine at a time?**

   1 half a glass
   2 one glass *(1 glass = c. 12 cl)*
   3 two glasses
   4 about half a ~~big~~ bottle *(1 bottle = 0,75 l)*
   5 a little less than one ~~big~~ bottle
   6 about one ~~big~~ bottle
   7 from one to two ~~big~~ bottles
   8 more than two ~~big~~ bottles
   9 I do not drink wine

The changes in questions from 1997 to 2002

In 2002 the questions 3–8 have been replaced with a new question number 3.

3. **How often did you drink the following amounts in one day during the last 12 months?** Instruction: Start answering from the first row. Mark (x) the most suitable 'How often?' alternative. Then continue row at a time down in the same manner. Please mark only one alternative per row.

1 dose = bottle (1/3 liter) beer (class III)
or a glass (12 cl) of light wine
or a glass (8 cl) of strong wine
or a glass (4 cl) of spirits or other strong liquor

Bottle (0.33 liter) beer (class IV), Gin Long Drink or strong cider	=	1.25 doses
Large bottle (0.5 liter) beer (class III)	=	1.5 doses
Large bottle (0.5 liter) beer (class IV)	=	2 doses
Bottle (0.75 liter) wine	=	7 doses
Bottle (0.75 liter) strong wine	=	10 doses
Bottle (0.5 liter) strong alcohol (e.g. Koskenkorva)	=	12 doses

Doses per day	Never	Once a month or more seldom	2-3 times a month	About once a week	2-3 times a week	4-5 times a week	6-7 times a week
15 or more	□	□	□	□	□	□	□
13-14	□	□	□	□	□	□	□
11-12	□	□	□	□	□	□	□
9-10	□	□	□	□	□	□	□
7-8	□	□	□	□	□	□	□
5-6	□	□	□	□	□	□	□
3-4	□	□	□	□	□	□	□
1-2	□	□	□	□	□	□	□

The changes in questions from 2002 to 2007

In 2007 the new question number 4 has been updated with a small change in the instructions and a change in the categories of consumed doses per day.

4. **How often did you drink the following amounts in one day during the last 12 months?**
Instruction: Start answering from the first row. Mark ~~(x)~~ the most suitable 'How often?' alternative. Then continue row at a time down

in the same manner. Please mark only one alternative per row.

1 dose = bottle (1/3 liter) beer (class III)
or a glass (12 cl) of light wine
or a glass (8 cl) of strong wine
or a glass (4 cl) of spirits or other strong liquor

Bottle (0.33 liter) beer (class IV), Gin Long Drink or strong cider	=	1.25 doses
Large bottle (0.5 liter) beer (class III)	=	1.5 doses
Large bottle (0.5 liter) beer (class IV)	=	2 doses
Bottle (0.75 liter) wine	=	7 doses
Bottle (0.75 liter) strong wine	=	10 doses
Bottle (0.5 liter) strong alcohol (e.g. Koskenkorva)	=	12 doses

Doses per day	At least 4 times a week	2-3 times a week	About once a week	1-2 times a month	3-10 times a year	1-2 times a year	Never
18 or more	1	2	3	4	5	6	7
13-17	1	2	3	4	5	6	7
8-12	1	2	3	4	5	6	7
5-7	1	2	3	4	5	6	7
3-4	1	2	3	4	5	6	7
1-2	1	2	3	4	5	6	7

REP. UNIV. JYVÄSKYLÄ DEPT. MATH. STAT.

131. SEPPÄLÄ, HEIKKI, Interpolation spaces with parameter functions and L_2-approximations of stochastic integrals. (18 pp.) 2011
132. TUHOLA-KUJANPÄÄ, ANNA, On superharmonic functions and applications to Riccati type equations. (17 pp.) 2012
133. JIANG, RENJIN, Optimal regularity of solutions to Poisson equations on metric measure spaces and an application. (13 pp.) 2012
134. TÖRMÄKANGAS, TIMO, Simulation study on the properties of quantitative trait model estimation in twin study design of normally distributed and discrete event-time phenotype variables. (417 pp.) 2012
135. ZHANG, GUO, Liouville theorems for stationary flows of generalized Newtonian fluids. (14 pp.) 2012
136. RAJALA, TUOMAS, Use of secondary structures in the analysis of spatial point patterns. (27 pp.) 2012
137. LAUKKARINEN, EIJA, On Malliavin calculus and approximation of stochastic integrals for Lévy processes. (21 pp.) 2012
138. GUO, CHANGYU, Generalized quasidisks and the associated John domains. (17 pp.) 2013
139. ÄKKINEN, TUOMO, Mappings of finite distortion: Radial limits and boundary behavior. (14 pp.) 2014
140. ILMAVIRTA, JOONAS, On the broken ray transform. (37 pp.) 2014
141. MIETTINEN, JARI, On statistical properties of blind source separation methods based on joint diagonalization. (37 pp.) 2014
142. TENGVALL, VILLE, Mappings of finite distortion: Mappings in the Sobolev space $W^{1,n-1}$ with integrable inner distortion. (22 pp.) 2014
143. BENEDICT, SITA, Hardy-Orlicz spaces of quasiconformal mappings and conformal densities. (16 pp.) 2014
144. OJALA, TUOMO, Thin and fat sets: Geometry of doubling measures in metric spaces. (19 pp.) 2014
145. KARAK, NIJJWAL, Applications of chaining, Poincaré and pointwise decay of measures. (14 pp.) 2014
146. JYLHÄ, HEIKKI, On generalizations of Evans and Gangbo's approximation method and L^∞ transport. (20 pp.) 2014
147. KAURANEN, AAPO, Space-filling, energy and moduli of continuity. (16 pp.) 2015
148. YLINEN, JUHA, Decoupling on the Wiener space and variational estimates for BSDEs. (45 pp.) 2015
149. KIRSILÄ, VILLE, Mappings of finite distortion on generalized manifolds. (14 pp.) 2015
150. XIANG, CHANG-LIN, Asymptotic behaviors of solutions to quasilinear elliptic equations with Hardy potential. (20 pp.) 2015
151. ROSSI, EINO, Local structure of fractal sets: tangents and dimension. (16 pp.) 2015
152. HELSKE, JOUNI, Prediction and interpolation of time series by state space models. (28 pp.) 2015
153. REINIKAINEN, JAAKKO, Efficient design and modeling strategies for follow-up studies with time-varying covariates. (36 pp.) 2015
154. NUUTINEN, JUHO, Maximal operators and capacities in metric spaces. (22 pp.) 2016
155. BRANDER, TOMMI, Calderón's problem for p-Laplace type equations. (21 pp.) 2016
156. ÄRJE, JOHANNA, Improving statistical classification methods and ecological status assesment for river macroinvertebrates. (30 pp.) 2016
157. HELSKE, SATU, Statistical analysis of life sequence data. (40 pp.) 2016
158. RUOSTEENOJA, EERO, Regularity properties of tug-of-war games and normalized equations. (16 pp.) 2017
159. ZHANG, YI, Planar Sobolev extension domains. (13 pp.) 2017
160. YLITALO, ANNA-KAISA, Statistical inference for eye movement sequences using spatial and spatio-temporal point processes. (45 pp.) 2017
161. LINDQVIST, PETER, Notes on the p-Laplace equation. (106 pp.) 2017
162. LEHTONEN, JERE, Injectivity results for the geodesic ray transform. (15 pp.) 2017
163. NICOLUSSI GOLO, SEBASTIANO, Topics in the geometry of non-riemannian lie groups. (14-pp.) 2017

ISBN 978-951-39-7351-3 (print)
ISBN 978-951-39-7352-0 (pdf)
ISSN 1457-8905