Journal of Clinical Epidemiology 108 (2019) 54−63

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Participation weighting based on sociodemographic register data improved external validity in a population-based cohort study

Carl Bonander[a,*], Anton Nilsson[b,c], Jonas Björk[b,d], Göran M.L. Bergström[e], Ulf Strömberg[a,f]

[a]Health Metrics Unit, Institute of Medicine, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[b]Division of Occupational and Environmental Medicine, Lund University, Lund, Sweden
[c]Centre for Economic Demography, Lund University, Lund, Sweden
[d]Clinical Studies Sweden, Forum South, Skåne University Hospital, Lund, Sweden
[e]Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, Sahlgrenska University Hospital, University of Gothenburg, Gothenburg, Sweden
[f]Research and Development, Region Halland, Halmstad, Sweden

Accepted 11 December 2018; Published online 16 December 2018

Abstract

Objective: To investigate whether inverse probability of participation weighting (IPPW) using register data on sociodemographic and disease history variables can improve external validity in a cohort study with selective participation.

Study Design and Setting: We fitted various IPPW models by logistic regression using register data for the participants ($n = 1,111$) and nonparticipants ($n = 1,132$) of a Swedish cohort study. For each of six diagnostic groups, we then estimated (1) weighted disease prevalence proportions and (2) weighted cross-sectional associations (odds ratios) between sociodemographic variables and disease prevalence. Using register data on the remaining individuals of the entire study population of men and women aged 50-64 years ($n = 22,259$), we addressed how the choice of variables used for IPPW influenced estimation errors.

Results: Disease prevalence proportions were generally underestimated in the absence of IPPW but became markedly closer to population values after IPPW using sociodemographic variables. We found limited evidence of selective participation bias in association estimates, but IPPW improved external validity when bias was present.

Conclusions: IPPW using sociodemographic register data can improve the external validity of disease prevalence estimates in cohort studies with selective participation. The performance of IPPW for association estimates merits further investigations in longitudinal settings and larger cohorts.

Keywords: Inverse probability weighting; External validity; Generalizability; Transportability; Nonresponse bias; Propensity score

1. Introduction

Improved methods for causal inference have attracted attention in a diverse number of fields over the last decades, including epidemiology [1]. Meanwhile, clinical trials often suffer from poor external validity [2,3], and the problem extends to epidemiological studies [4−8]. In both cases, nonrandom selection into the study sample, or selective participation, can introduce generalizability bias in descriptive public health measures, associations, and causal estimates [9,10]. Despite this, issues relating to external validity have received considerably less attention in epidemiologic research than those relating to internal validity.

In this article, we examine if inverse probability of participation weighting (IPPW) using external register data can improve external validity in cohort studies. The method, sometimes more generally referred to as inverse probability of sampling weighting [11], uses external data, for example, socioeconomic indicators, on the participants and nonparticipants of a study to estimate the propensity score for participation. Weighted analyses are then conducted using IPPW based on the inverse of the estimated propensity score, resulting in a weighted sample that is more similar to the target population (given that the invited individuals are a random subset of this population) [10,12]. The approach shares many traits with inverse probability of treatment weighting (IPTW) [13,14], the primary difference

Conflict of interest: The authors have no conflict of interest to declare.
Funding: This work was supported by grant 2017-00414 from the Swedish Research Council for Health, Working Life and Welfare (FORTE), Sweden.

* Corresponding author. Tel.: +4676-241 8269; fax: +4631-786 6783.
E-mail address: carl.bonander@gu.se (C. Bonander).

https://doi.org/10.1016/j.jclinepi.2018.12.011
0895-4356/© 2018 Elsevier Inc. All rights reserved.

What is new?

Key findings

- Our empirical validation of inverse probability of participation weighting (IPPW), which involved using register data to correct for selective participation in a Swedish cohort study, confirmed that the approach can decrease the effects of selective participation.
- By varying the covariate sets in the propensity score model and comparing the weighted estimates to known population parameters, we found that weighting on a set of sociodemographic variables reduced the estimation error of disease prevalence estimates to levels that were close to those obtained by unbiased random sampling.
- The errors were generally smaller for unweighted association estimates than for prevalence estimates, but IPPW decreased the error in at least one instance when selective participation bias appeared to be present.

What this adds to what was known?

- Previous methodological studies have demonstrated the validity of the method under certain modelling assumption, but this is the first study to assess the accuracy of the approach in a population-based cohort in register data with a true target population benchmark.

What is the implication and what should change now?

- Our results imply that IPPW models constructed using sociodemographic characteristics can improve the generalizability of population-based cohort studies suffering from selective participation.

being that the latter is used to minimize treatment selection. In particular, it shares a similar core assumption: that there is no selection into participation based on unobserved variables (conditional exchangeability; see references [11,15] for details). In most applications, it is impossible to know whether this assumption holds.

Previous research related to ours includes studies detailing the use of IPPW in randomized experiments [10,12,15,16], how to combine survey weighting and IPTW [17], and other methodological contributions [11,18–22], as well as applications to health surveys and cohort studies [9,23,24]. The survey literature also provides related methods aimed at estimating population totals, such as poststratification and calibration to correct for survey nonresponse [25,26]. However, this literature provides little practical guidance regarding which variables to include in the participation model when applied to cohort studies with selective participation. Is it sufficient to weight on demographic characteristics such as age and sex, or are indicators of socioeconomic status (SES) required? Can additional adjustments for disease history, if such data are available, further improve the estimates? Would the recommendations be the same for estimation of association measures as for estimation of disease prevalence proportions?

The Nordic countries are well known for their high-quality population and health care registers with individual-level data on the entire population, which can be linked to cohort members and nonparticipants through personal identification numbers [27]. Registers may therefore serve as a useful source for variables to include in the participation model. In this article, we use individual-level register data on a Swedish population–based cohort with documented selective participation [28], and on the entire study population, in an attempt to answer the questions posed previously. The data include a wide range of sociodemographic variables and disease history records on several diagnostic groups. These unique data allow us to assess the external validity of our estimates directly by comparing them to the true (register-based) prevalence proportions and associations observed in the target population. Hence, we can estimate if the conditional exchangeability assumption holds in various models. Specifically, we study systematic errors in (1) disease prevalence estimates and (2) cross-sectional association estimates (odds ratios [ORs]) between sociodemographic variables and disease prevalence, with regard to IPPW based on different sets of register variables.

2. Materials and methods

We used data from a pilot cohort study, pilot-SCAPIS (Swedish CArdioPulmonary Bio-Image Study), which was conducted in Gothenburg, Sweden, in 2012. One of the primary purposes of SCAPIS is to improve prediction of cardiopulmonary diseases by measuring rich phenotype data [29]. An extensive set of sociodemographic and disease history variables from Swedish registers has been linked to the cohort (see Section 2.1 for details). In this article, we focus only on the latter, as the analysis requires information on the true values in the study population (which is not available for cohort-specific data).

Recruitment to pilot-SCAPIS was based on a stratified random sample that had been drawn from six residential areas (three areas with high and three with low SES). The target population consisted of all residents in these areas between 50 and 64 years of age (N = 24,502). Invitations were sent to randomly selected residents within each area. The residents in the low-SES areas had a higher probability of being invited (12-13%) as compared to the residents in the high-SES areas (6-7%) (this sampling strategy was adopted to account for the expected lower response rates among low-SES populations). In total, 2,243 individuals were invited to participate in the cohort, 1,111 of which agreed to participate (50% response rate). A previous study found that participation in the cohort varied with regard to residential area, country of birth, civil status, education, occupational status, and disposable income, whereas disease history was less predictive of participation [28].

2.1. Data

We used an anonymous, individual-level data set for the entire target population ($N = 24{,}502$), constructed and matched to the cohort by Swedish register authorities (Statistics Sweden and the National Board of Health and Welfare). It contained indicators of invitation to and participation in pilot-SCAPIS, as well as register data on residential area, sociodemographic variables, and disease history. Data on country of birth, civil status, education, occupational status, income, sick leave, and retirement for 2011 were obtained from the Longitudinal integration database for health insurance and labor market studies (LISA) [30]. The disease history variables, which were obtained from the National Patient Register [31], included data on both inpatient and outpatient care during 2000-2011. We considered six diagnostic groups: (1) cardiovascular disease (CVD; ICD-10 codes I20-I25, I48, I50, I61-I64) or chest pain (R074); (2) cancer (C00-C97); (3) chronic obstructive pulmonary disease (COPD; J40-J47); (4) diabetes (E10-E14); (5) psychiatric disease (F20-F25, F28-F34, F38-F48, F60-F69); and (6) alcohol or substance abuse (F10-F19). We used this information to estimate disease prevalence, in this case defined as any occurrence of the disease requiring inpatient or outpatient care during the last 12 years before recruitment. Clearly, this is an imperfect measure of current disease status, but the same problem occurs both in the study population and in the sample, and our main concern here was that the information is comparable between the two.

2.2. Analysis

The analysis consisted of two main parts. First, we estimated IPPW using different sets of register variables. We then used these weights to estimate weighted disease prevalence proportions and compared these estimates to the true proportions in the study population. Next, we examined the impact of IPPW on the estimation errors for the cross-sectional association between a crucial socioeconomic indicator, educational attainment, and the disease prevalence proportions. This type of relationship is particularly interesting because education is often correlated with participation in cohorts [32,33], including pilot-SCAPIS [28] and is a known correlate of health-related behaviors and outcomes [34,35].

2.2.1. Participation modeling

We fitted a series of logistic regression models to estimate the propensity score for participation π_i in the cohort, iteratively varying the variable set to examine the performance of different options, using data from the invited sample ($n = 2{,}243$). We then predicted π_i to construct an inverse probability of participation weight $1/\pi_i$ for each individual i in the participant sample ($n = 1{,}111$) [15]. To correct for the stratified sampling scheme, we multiplied $1/\pi_i$ with sampling weights based on the inverse of the sampling probability within each area. In total, we fitted five different participation models, each accounting for the following:

1. sampling for invitation;
2. age and sex;
3. all individual-level sociodemographic variables (including age and sex, and sampling weights);
4. disease history variables (as well as age and sex, and sampling weights); and
5. all variables in (3) and (4).

The rationale behind each model is as follows. Model 1 served as a baseline to examine the extent of the estimation error when we only accounted for the stratified sampling scheme (recall that individuals from low-SES areas had a higher probability of being invited to participate). It reflects the expected impact of selective participation if simple random sampling had been used to invite individuals to the study, which we believe is of greater general interest than the specific sampling scheme used in pilot-SCAPIS. Model 2 accounts for age and sex, which is a potentially interesting case that addresses whether there might be a need for obtaining additional register data on nonparticipants. Models 3-5 include variables from different registers to address if register linkage is necessary for improving external validity or if one data source is sufficient.

When estimating model 4, we iteratively refit the participation model, always leaving out the disease history variable for which we were estimating prevalence or associations. We did this to mimic a scenario in which register data are used to correct for selective participation bias in study-specific variables.

2.2.2. Error analysis

Let y^{Ω} denote the true target population prevalence proportion for disease y, and $\widehat{y}_m$ denotes the estimated prevalence based on IPPW model m. For a particular y and m, the percentage error can then be defined aswhich we used to quantify estimation error as an estimate of bias. Here, an $\varepsilon_{y,m}$ that differs significantly from zero implies that the model may be biased. To determine statistical significance, we estimated 95% percentile bootstrap confidence intervals (CIs) using the *boot* package for R [36,37], re-estimating the entire procedure (including the participation models) within each of 5,000 bootstrap resamples to account for uncertainty in the weights.

To evaluate the overall performance of each model, we averaged the absolute percentage error, $|\varepsilon_{y,m}|$, over all disease history variables (we did not use $\varepsilon_{y,m}$ to this end, as negative and positive errors can cancel each other out). In this case, $|\varepsilon_{y,m}|$ can be interpreted as a measure of accuracy, which is the sum of bias and precision. The precision part varies depending on, for example, sample size and population variance, and is thus case specific. As such, there is no natural null hypothesis for accuracy (in contrast to bias). To solve this problem, we constructed an empirical null by drawing 10,000 random samples of size $n = 1{,}111$ (the number of participants in pilot-SCAPIS) to obtain a stable estimate of the expected accuracy in an unbiased scenario with perfect participation. This null can also be interpreted as the expected, case-specific precision under zero mean bias, given the sample size and variables under study, which served as a benchmark to compare with the overall performance of the IPPW models.

Following the same procedure as for the disease prevalence estimates, we then examined how the bias and accuracy of association estimates were affected by IPPW by studying the association (measured in log-odds) between low educational attainment (primary education vs. secondary or university education) and each disease variable. We estimated unadjusted ORs, as our primary concern was generalizability, not internal validity.

3. Results

3.1. Summary of register data

Data on most variables deviated more strongly from the target population values among the cohort participants than in the originally invited (stratified random) sample (Table 1). Such deviations can be seen for both socioeconomic characteristics and disease prevalence estimates, which were generally underestimated when based on the cohort participants alone. Using sampling weights to account for the stratified sampling scheme alleviated the problem in the originally invited sample, as expected. However, the errors remained large among the participants.

3.2. Errors in estimated prevalence proportions

We present the percentage errors (estimated bias) in disease prevalence estimates from different participation models, along with 95% bootstrap CIs, in Fig. 1. The sampling weight model (model 1) underestimated the prevalence of alcohol and substance abuse, diabetes and psychiatric disease, and overestimated the prevalence of cancer. The point estimates of CVD and COPD were also lower than in the target population, but they did not differ significantly from the population proportions.

IPPW weighting using age and sex (model 2) did not bring about further improvement than model 1, which might be expected, given the marginal differences in age and sex between the participant sample and the population observed in Table 1. However, after weighting on sociodemographic variables (model 3), the estimates that differed significantly in model 1 were generally closer to the population counterparts, and their CIs now contained the true values. The disease history model (model 4) exhibited a similar performance as models 1 and 2 and using sociodemographic variables and disease history variables together in the full model (model 5) only marginally affected the estimated bias as compared to model 3.

3.3. Errors in association estimates

We present unadjusted prevalence ORs for the association between low educational attainment and the six disease history variables in Table 2, along with group-specific prevalence estimates. Before turning to the error analyses, we first show the estimates based on the sampling weight model (model 1) in comparison to estimates obtained from model 3.

The ORs in the target population imply that the odds of disease are higher among individuals with low educational attainment, except for cancer, where the odds are lower (Table 2). We note that the sample ORs were overestimated for most disease variables. However, the only instance in which the CI did not contain the true population OR (in model 1) was the estimate for alcohol or substance abuse. The corresponding OR with IPPW (model 3) differed less pronouncedly. The group-specific prevalence estimates changed quite meaningfully in some instances (especially for psychiatric disease), but their relative difference remained roughly the same (Fig. 2). We note that while the point estimates generally changed, and were closer to the population ORs on average (see section 3.4), the overall conclusions regarding the direction and approximate size of the association from a study estimating these models would have remained the same with or without IPPW. Results for income, country of birth, and occupational status give rise to similar conclusions as for education (Appendix Table A1).

The error analyses, which also included the other IPPW models, confirmed that the error (in the log-OR) was largest for alcohol or substance abuse (Fig. 3). This estimate differed significantly from the population value in IPPW models 1, 2, and 4, whereas the estimated bias was smaller in models 3 and 5.

3.4. Average model performance

We present the mean absolute percentage errors in prevalence proportions and association estimates, averaged over all studied diseases, in Fig. 4. As can be seen there, the CIs for models 3 and 5 covered the random sampling benchmark, whereas the other models did not. This is true for both prevalence proportions and association estimates, although the accuracy gain was more apparent for the prevalence estimates. Using prevalence ratios as an alternative association measure yielded almost identical results, but the average deviation in accuracy from the random sampling benchmark was smaller (nonsignificant in all models) on the prevalence difference scale (Appendix Figure A1). We also present the results in terms of absolute errors instead of percentage errors in Appendix Figures A2-A4 (the conclusions are identical to the main results).

Table 1. Prevalence in the target population on all register-based variables included in the study compared to estimates based on the two (invited, participant) samples from the pilot-SCAPIS, with and without sampling weights

	Target (N = 24,502)	Invited (n = 2243)				Participants (n = 1111)			
		Unweighted		Sampling weights		Unweighted		Sampling weights	
Characteristics	%	%	Error[a]	%	Error[a]	%	Error[a]	%	Error[a]
Sociodemographic variables									
Age, years									
50–54	33.9	32.4	−4	31.5	7	29.5	−13	29.3	−14
55–59	31.8	32.9	3	33.9	7	34.6	9	35.1	10
60–64	34.3	34.7	1	34.6	1	35.9	5	35.6	4
Female	48.5	50.3	4	50.5	4	50.1	3	52	4
Country of birth									
Sweden	66.1	56.0	−15	65.8	0	68.8	4	77	16
Nordic	5.4	7.4	36	6.1	13	7.7	42	6.1	13
EU	5.6	6.7	19	5.4	4	5.1	−8	4	−29
Non-EU	7	9.6	37	7.1	1	6.6	−6	4.5	−36
Outside Europe	16	20.4	28	15.6	3	11.8	−26	8.5	−47
Living single	39.3	44.9	14	41	4	37.4	−5	33.9	−14
Low-SES area	45.7	63.2	38	45.7	0	49.7	9	32.6	−29
Education									
Primary/missing	22.3	26.1	17	22.1	−1	17.4	−22	14.5	−35
Secondary	46.1	45.1	−2	44.2	−4	46.7	1	45.2	−2
University	31.6	28.8	−9	33.7	7	35.9	14	43	28
Occupational status									
Employed	70.6	64.6	−8	71.2	1	77.8	10	82.5	17
Unemployed	10	11.0	10	10	0	8.8	−12	7.8	−22
Sick leave	5.4	7.2	34	5.3	−2	3.2	−42	2.1	−61
Retired	14	17.2	23	13.6	−3	10.2	−27	7.6	−46
Disposable income									
Q1	24.9	31.3	26	25.7	3	18.7	−25	14.8	−41
Q2	24.9	28.0	13	25.5	2	27.6	11	24.4	−2
Q3	24.9	23.0	−8	24.8	0	28.4	14	29.1	17
Q4	24.9	17.7	−29	23.9	−4	25.3	02	31.6	27
Disease history									
Cardiovascular disease[b]	14.1	14.69	4	13.1	−7	13.7	−3	12.3	−13
Cancer	7.6	8.13	7	8.7	14	9.4	23	9.6	26
Chronic obstructive pulmonary disease	3.3	3.71	12	3.2	−3	3.2	−2	2.7	−18
Diabetes	4.4	5.40	23	4.4	0	3.5	−20	3	−32
Psychiatric disease	9.7	10.81	11	9.2	−5	7.7	−20	6.4	−34
Alcohol or substance abuse	3.8	4.29	13	3.7	−3	2.3	−38	2	−47
Average absolute percentage error									
All variables			16		4		16		24
Sociodemographic			17		3		14		23
Disease history			12		5		18		28

[a] Percentage error compared to the true population value.
[b] Including chest pain.

4. Discussion

Our results imply that IPPW using sociodemographic register data can improve the external validity of disease prevalence estimates. Study participation rates are generally higher in high-SES populations [38], and selective participation bias can occur if these factors are related to the study variables of interest. If this is the case for a

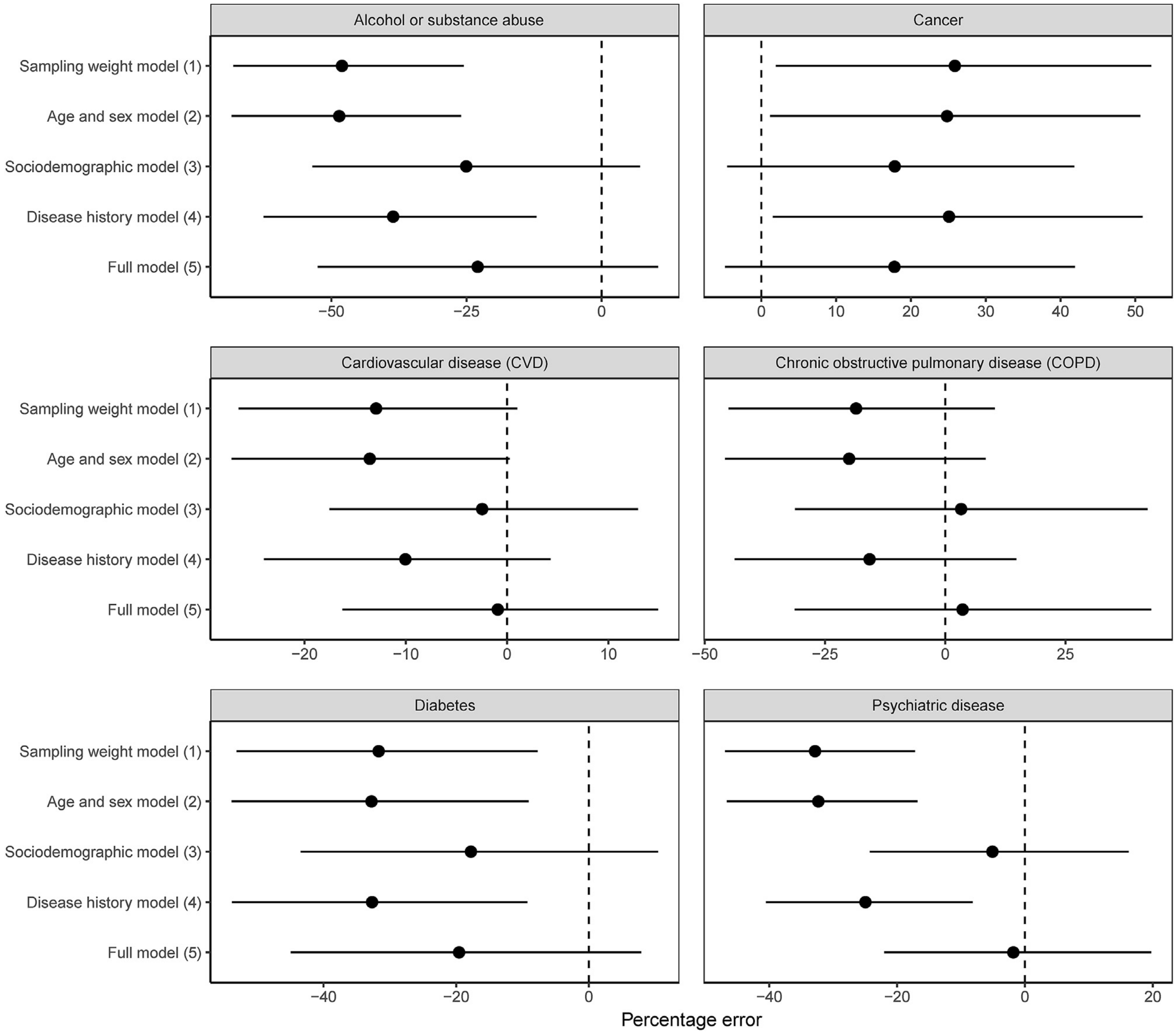

Fig. 1. Percentage error in the estimated target population prevalence, by disease group and participation model, with 95% percentile bootstrap confidence intervals (5,000 resamples).

particular cohort, the approach studied here may be useful for improving the generalizability of study results. However, before generalizing our results to other settings, some unique features of the studied cohort are worth noting. For instance, weighting only on age and sex (model 2) did not improve generalizability to the target population. This is not surprising, given that these factors appear to have had a minimal impact on participation in the studied cohort. This question may therefore be worth investigating in other cohorts, perhaps with greater age spans. Another result that may be specific to our studied age group is that the cancer prevalence proportion was overestimated in the participant sample, in contrast to the other diseases history variables considered. This result is probably due to the fact that three common cancer diseases (with relatively high incidences also in the age span up to 64 years) are more frequently diagnosed in groups with high SES: cutaneous malignant melanoma and breast and prostate cancers (because mammography screening and prostate-specific antigen testing are more common in high-SES groups) [39–41].

Although our results concerning association measures also suggest that IPPW using sociodemographic register data can correct for selective participation (when present), the overall conclusions are not as obvious as for the prevalence estimates. We only found one instance that required a correction to begin with but our general impression is that these analyses may have been hampered by low statistical power to detect heterogeneity in association estimates between different population groups. It is important to keep in mind that the effects of IPPW on bias may vary between, and within, populations. Given that the participation model

Table 2. Simple logistic regression results for the association (prevalence odds ratios) between low educational attainment and disease history in six diagnostic groups, comparing inverse probability of participation weighted estimates to OR in the target population

	Target			Sampling weight model (1)			Sociodemographic model (3)		
	Prevalence (%)		Association	Prevalence (%)		Association	Prevalence (%)		Association
Disease history	**Low educ.**	**Mid/high educ.**	**OR**	**Low educ.**	**Mid/high educ.**	**OR (95% CI)**	**Low educ.**	**Mid/high educ.**	**OR (95% CI)**
CVD[b]	18.0	12.9	1.48	18.5	11.2	1.81 (1.14, 2.76)	18.5	12.3	1.61 (1.02, 2.45)
Cancer	6.2	8.0	0.76	7.3	9.9	0.72 (0.38, 1.20)	7.6	9.4	0.82 (0.43, 1.39)
COPD	4.8	2.9	1.69	6.5	2.1	3.26 (1.46, 6.71)	7.7	2.2	3.78 (1.63, 7.86)
Diabetes	7.1	3.6	2.05	6.6	2.4	2.95 (1.28, 5.92)	6.5	2.7	2.47 (1.08, 4.93)
Psychiatric disease	13.0	8.8	1.54	11.2	5.7	2.06 (1.16, 3.37)	14.9	7.6	2.12 (1.18, 3.49)
Alcohol or substance abuse	6.4	3.0	2.23	7.4	1.0	7.58[a] (3.11, 18.79)	6.7	1.7	4.08 (1.66, 10.36)

Abbreviations: CI, confidence interval; CVD, cardiovascular disease; COPD, chronic obstructive pulmonary disease.

[a] 95% percentile bootstrap confidence interval does not contain the target population odds ratio.

[b] Including chest pain.

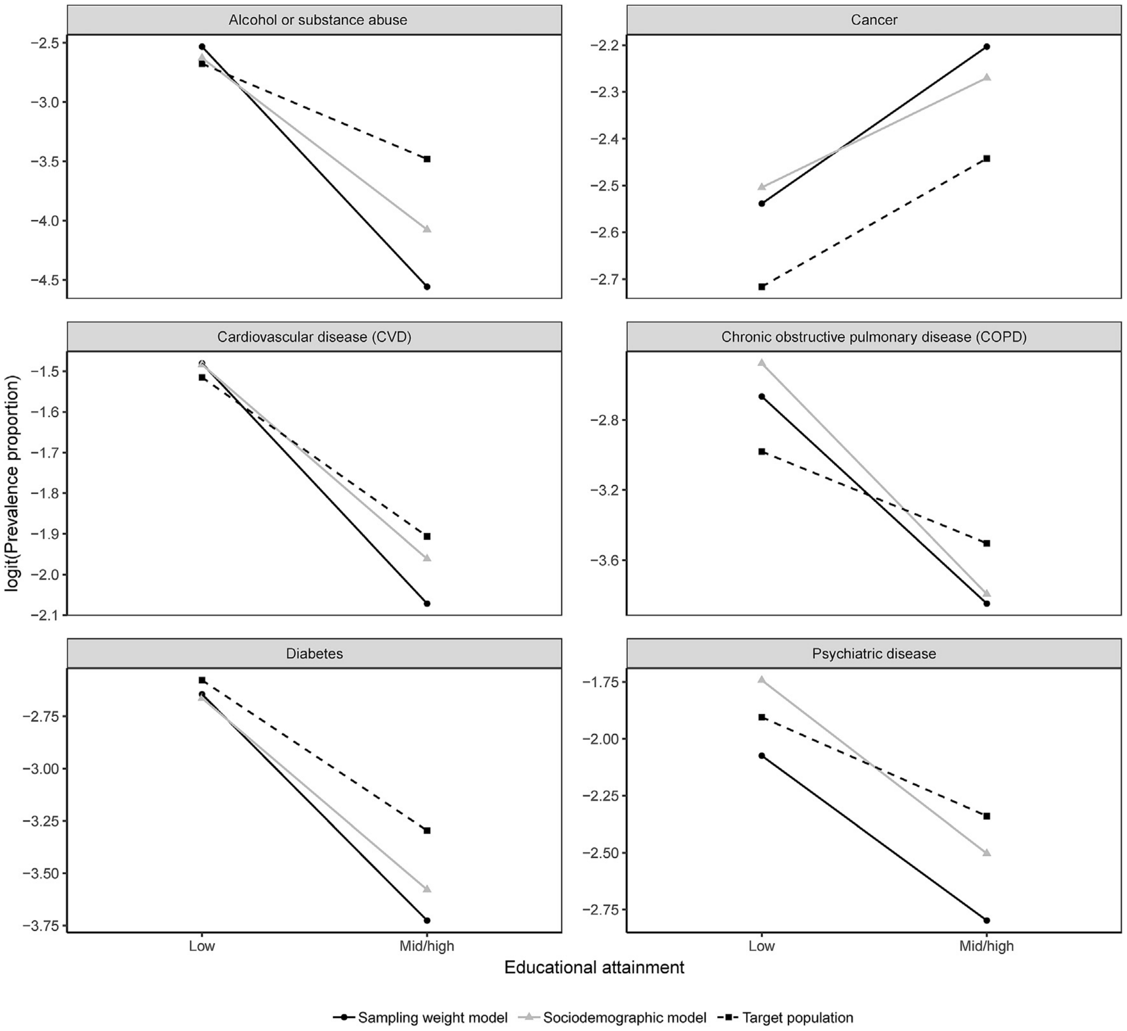

Fig. 2. Prevalence proportions of six disease history variables (in log-odds) by educational attainment, comparing the sampling weight and sociodemographic models to the values obtained from the target population data.

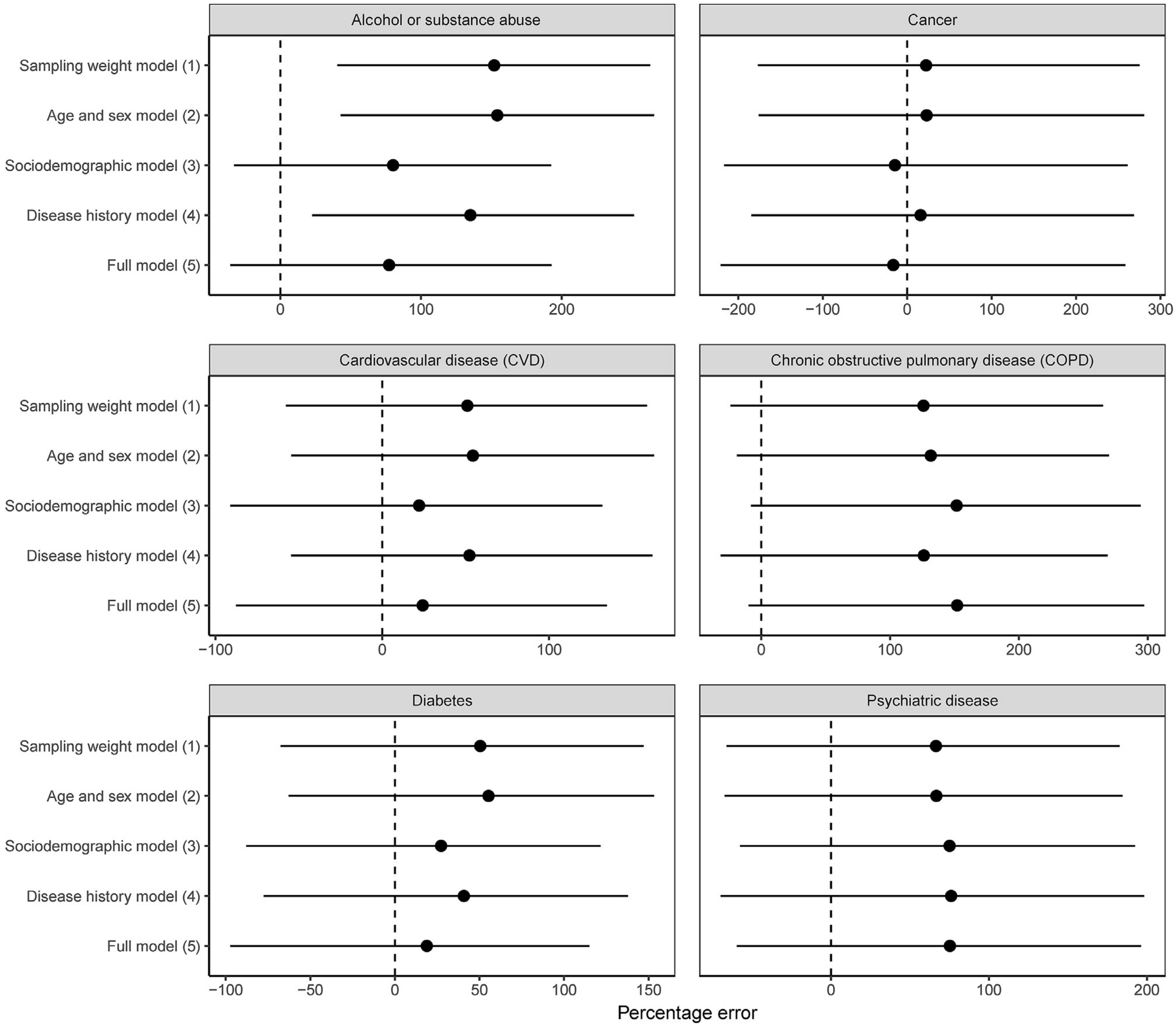

Fig. 3. Percentage error from the target population log-odds ratio for low educational attainment, by disease outcome and participation model, with 95% percentile bootstrap confidence intervals (5,000 resamples).

is correctly specified, IPPW should reduce bias in settings with large variations in exposure effect size across the variables that influence participation (i.e., effect modification by these variables) [17]. The results obtained here could simply indicate that there is little heterogeneity in the measured associations due to the factors affecting participation in pilot-SCAPIS. It is therefore important to address the selection model carefully on a case-by-case basis (directed acyclic graphs may be useful as guidance in such applications [28,42]).

Our study estimated (period) prevalence of various diseases, based on register data collected at baseline, and the results are not necessarily generalizable to longitudinal contexts. It will be of interest to address external validity when analyzing longitudinal follow-up data of the cohort, considering associations with incidence and mortality rates. Mortality rates often differ between participants and nonparticipants in cohort studies [4–8], and to which extent the IPPW method can improve the external validity of risk estimates for mortality requires additional research.

Finally, we also point out that the sampling scheme for recruiting cohort members may influence selective participation bias. For example, in a situation where invitation is done randomly, instead of applying oversampling in low-SES residential areas as in pilot-SCAPIS, unweighted estimates may show worse external validity because oversampling populations with lower participation rates can partly counter the error (as seen in Table 1). In this case, using sampling weights on their own may do more harm than good because they effectively remove this correction. This illustrates the importance of considering how the

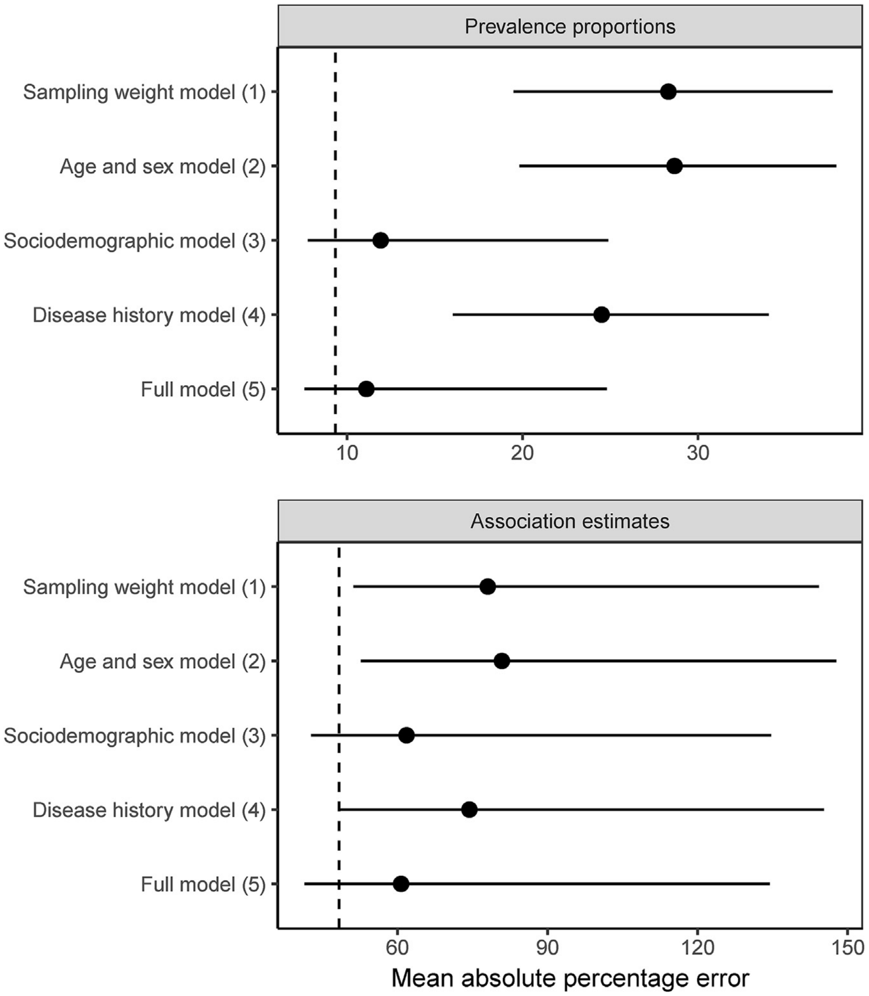

Fig. 4. Mean absolute percentage error for all six studied disease history variables by estimate type and participation model with 95% percentile bootstrap confidence intervals (5,000 resamples), compared to a random sampling benchmark without selective participation (vertical line).

sampling method influences selective participation bias in cohort studies.

In conclusion, the IPPW method shows promise in correcting for estimation errors caused by selective participation (when present). Our study suggests that access to sociodemographic register data may be sufficient to improve external validity, but the benefits of the method should be investigated further using clinical exposures and prospective outcomes, as well as in larger cohorts, which is part of our planned work.

Acknowledgments

This article was finalized during the last author's sabbatical at the School of Public Health, Imperial College London, which was supported by the Wenner-Gren Foundations. SCAPIS is supported by the Swedish Heart and Lung Foundation, the Knut and Alice Wallenberg Foundation, the Swedish Research Council and VINNOVA. The SCAPIS pilot study also received funding from the Sahlgrenska Academy at Gothenburg University and strategic grants from ALF, Sweden/LUA in Western Sweden. The authors are grateful to all the participants in this study. A special thanks all test personnel at the SCAPIS test center in Gothenburg.

Authors' contribution: C.B. contributed to conceptualization, methodology, formal analysis, data curation, writing of original draft, review and editing, visualization, and software. A.N. contributed to methodology and review and editing. J.B. contributed to methodology, funding acquisition, data curation, resources, and review and editing. G.B. contributed to resources, project administration, and supervision. U.S. contributed to funding acquisition, data curation, resources, review and editing, supervision, and project administration.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.12.011.

References

[1] Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health 2006;60:578−86.

[2] Elting LS, Cooksley C, Bekele BN, Frumovitz M, Avritscher EBC, Sun C, et al. Generalizability of cancer clinical trial results: prognostic differences between participants and nonparticipants. Cancer 2006;106:2452−8.

[3] Evans A, Kalra L. Are the results of randomized controlled trials on anticoagulation in patients with atrial fibrillation generalizable to clinical practice? Arch Intern Med 2001;161:1443−7.

[4] Walker M, Shaper AG, Cook DG. Non-participation and mortality in a prospective study of cardiovascular disease. J Epidemiol Community Health 1987;41:295−9.

[5] Mattila VM, Parkkari J, Rimpelä A. Adolescent survey non-response and later risk of death. A prospective cohort study of 78,609 persons with 11-year follow-up. BMC Public Health 2007;7:87.

[6] Hara M, Sasaki S, Sobue T, Yamamoto S, Tsugane S. Comparison of cause-specific mortality between respondents and nonrespondents in a population-based prospective study: ten-year follow-up of JPHC Study Cohort I. Japan Public Health Center. J Clin Epidemiol 2002;55:150−6.

[7] Heilbrun LK, Nomura A, Stemmermann GN. The effects of non-response in a prospective study of cancer: 15-year follow-up. Int J Epidemiol 1991;20:328−38.

[8] Ferrie JE, Kivimäki M, Singh-Manoux A, Shortt A, Martikainen P, Head J, et al. Non-response to baseline, non-response to follow-up and mortality in the Whitehall II cohort. Int J Epidemiol 2009;38:831−7.

[9] Etter J-F, Perneger TV. Analysis of non-response bias in a mailed health survey. J Clin Epidemiol 1997;50:1123−8.

[10] Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. J R Stat Soc Ser A Stat Soc 2011;174:369−86.

[11] Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. Epidemiology 2017;28:553−61.

[12] Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. Prev Sci 2015;16:475−85.

[13] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41−55.

[14] Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550−60.

[15] Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. Am J Epidemiol 2010;172:107−15.

[16] Miratrix LW, Sekhon JS, Theodoridis AG, Campos LF. Worth weighting? How to think about and use weights in survey experiments. Polit Anal 2018;26:275−91.

[17] Dugoff EH, Schuler M, Stuart EA. Generalizing observational study results: applying propensity score methods to complex surveys. Health Serv Res 2014;49:284−303.

[18] Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika 2009;96:723−34.

[19] Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. Stat Sci 2014;29:579−95.

[20] Wooldridge JM. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. Port Econ J 2002;1:117−39.

[21] Dahabreh IJ, Robertson SE, Stuart EA, Hernan MA. Transporting inferences from a randomized trial to a new target population 2018: arXiv:1805.00550v1 (preprint). Available at https://arxiv.org/abs/1805.00550. Accessed January 2, 2019.

[22] Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol 2017;186:1010−4.

[23] Corry NH, Williams CS, Battaglia M, McMaster HS, Stander VA. Assessing and adjusting for non-response in the millennium cohort family study. BMC Med Res Methodol 2017;17:16.

[24] Korkeila K, Suominen S, Ahvenainen J, Ojanlatva A, Rautava P, Helenius H, et al. Non-response and related factors in a nation-wide health survey. Eur J Epidemiol 2001;17:991−9.

[25] Deville J-C, Särndal C-E. Calibration estimators in survey sampling. J Am Stat Assoc 1992;87:376−82.

[26] Deville J-C, Särndal C-E, Sautory O. Generalized raking procedures in survey sampling. J Am Stat Assoc 1993;88:1013−20.

[27] Maret-Ouda J, Tao W, Wahlin K, Lagergren J. Nordic registry-based cohort studies: possibilities and pitfalls when combining Nordic registry data. Scand J Public Health 2017;45:14−9.

[28] Björk J, Strömberg U, Rosengren A, Toren K, Fagerberg B, Grimby-Ekman A, et al. Predicting participation in the population-based Swedish cardiopulmonary bio-image study (SCAPIS) using register data. Scand J Public Health 2017;45:45−9.

[29] Bergström G, Berglund G, Blomberg A, Brandberg J, Engström G, Engvall J, et al. The Swedish CArdioPulmonary BioImage Study: objectives and design. J Intern Med 2015;278:645−59.

[30] Statistics Sweden. Longitudinal integration database for health insurance and labour market studies (LISA) n.d. Available at: http://www.scb.se/en/services/guidance-for-researchers-and-universities/vilka-mikrodata-finns/longitudinella-register/longitudinal-integration-database-for-health-insurance-and-labour-market-studies-lisa/. Accessed August 15, 2018.

[31] Ludvigsson JF, Andersson E, Ekbom A, Feychting M, Kim J-L, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. BMC Public Health 2011;11:450.

[32] Strandhagen E, Berg C, Lissner L, Nunez L, Rosengren A, Torén K, et al. Selection bias in a population survey with registry linkage: potential effect on socioeconomic gradient in cardiovascular risk. Eur J Epidemiol 2010;25:163−72.

[33] Howe LD, Tilling K, Galobardes B, Lawlor DA. Loss to follow-up in cohort studies. Epidemiology 2013;24:1−9.

[34] Lleras-Muney A. The relationship between education and adult mortality in the United States. Rev Econ Stud 2005;72:189−221.

[35] Furnée CA, Groot W, van den Brink HM. The health effects of education: a meta-analysis. Eur J Public Health 2008;18:417−21.

[36] R Core Team. R. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.

[37] Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions 2017: R package version 1.3-20.

[38] Galea S, Tracy M. Participation rates in epidemiologic studies. Ann Epidemiol 2007;17:643−53.

[39] Jiang AJ, Rambhatla PV, Eide MJ. Socioeconomic and lifestyle factors and melanoma: a systematic review. Br J Dermatol 2015;172:885−915.

[40] Lundqvist A, Andersson E, Ahlberg I, Nilbert M, Gerdtham U. Socioeconomic inequalities in breast cancer incidence and mortality in Europe—a systematic review and meta-analysis. Eur J Public Health 2016;26:804−13.

[41] Kilpeläinen TP, Talala K, Raitanen J, Taari K, Kujala P, Tammela TLJ, et al. Prostate cancer and socioeconomic status in the Finnish randomized study of screening for prostate cancer. Am J Epidemiol 2016;184:720−31.

[42] Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15:615−25.