Does low participation in cohort studies induce bias?

Additional material

Content:

A heuristic proof of the formula for the asymptotic standard error

We will here present a heuristic proof of the formula for the asymptotic variance of the difference between the estimate based on a random subpopulation and that based on the total population, i.e.

$$\mathrm{Var}_{as}\left(\hat{\theta}_{Sub}-\hat{\theta}_{Tot}\right)=\mathrm{Var}_{as}\left(\hat{\theta}_{Sub}\right)-\mathrm{Var}_{as}\left(\hat{\theta}_{Tot}\right) \quad (1)$$

First, we make a general observation. Let *Z* be given as the weighted average of two random variables *X* and *Y* using the inverse of their variances as weights:

$$Z=\left[\mathrm{Var}\left(X\right)^{-1}+\mathrm{Var}\left(Y\right)^{-1}\right]^{-1}\left[\mathrm{Var}\left(X\right)^{-1}X+\mathrm{Var}\left(Y\right)^{-1}Y\right].$$

If *X* and *Y* are uncorrelated then

$$\mathrm{Var}\left(X-Z\right)=\mathrm{Var}\left(X\right)-\mathrm{Var}\left(Z\right)$$

This surprising result follows as

$$\mathrm{Var}\left(Z\right)=\mathrm{Cov}\left(Z,X\right)=\left[\mathrm{Var}\left(X\right)^{-1}+\mathrm{Var}\left(Y\right)^{-1}\right]^{-1} \text{ and}$$

$$\mathrm{Var}\left(X-Z\right)=\mathrm{Var}\left(X\right)+\mathrm{Var}\left(Z\right)-2\,\mathrm{Cov}\left(Z,X\right)$$

Now consider two subpopulations 1 and 2 and let $\hat{\theta}_{Subi}$ denote the maximum likelihood estimate based on the data in the *i*th subpopulation. If we let

$$\tilde{\theta}=\left[\mathrm{Var}_{as}\left(\hat{\theta}_{Sub1}\right)^{-1}+\mathrm{Var}_{as}\left(\hat{\theta}_{Sub2}\right)^{-1}\right]^{-1}\left[\mathrm{Var}_{as}\left(\hat{\theta}_{Sub1}\right)^{-1}\hat{\theta}_{Sub1}+\mathrm{Var}_{as}\left(\hat{\theta}_{Sub2}\right)^{-1}\hat{\theta}_{Sub2}\right]$$

(2)

then it follows that $\mathrm{Var}_{as}\left(\hat{\theta}_{Sub}-\tilde{\theta}\right)=\mathrm{Var}_{as}\left(\hat{\theta}_{Sub}\right)-\mathrm{Var}_{as}\left(\tilde{\theta}\right)$.

Finally, we note that if the two subpopulations are a random partitioning of the total population, then $\tilde{\theta}$ is asymptotically equal to the maximum likelihood based on the total population $\hat{\theta}_{Tot}$.

It is important to note that equation 1 does not hold in general.

Description of the simulation study

The performance of the two methods for calculation of confidence intervals was assessed in a small simulation study covering scenarios identical or close to the actual data included in the adjusted analyses of IVF and preterm birth, and of smoking and SGA. The simulation set-up required specification of the distribution of the covariate pattern in the source population, the participation rate within each covariate pattern, and the dependence of the outcome on the covariates both among the participants and among the non-participants. For IVF and preterm birth the relevant model involved 96=2*2*3*2*2 different covariate patterns. In the simulations we used the observed distribution with the modification that patterns with less than 20 observations were assigned a small, but positive probability obtained by smoothing using a Poisson regression model. For each covariate pattern the participation rate was selected to be identical to the observed participation rate, except for rare patterns, where the participation rate was found by smoothing using a logistic regression model. For the associations between the covariates and the risk of preterm birth among the participants and among the non-participants we used logistic regression models of the same form as those used in the analysis of the actual data. In the simulation we considered three different scenarios reflecting different choices of the coefficients in the logistic model describing the relation between the covariates and preterm birth. In the first scenario the coefficients were identical to the estimates found in the actual data. In the two other scenarios we modified the adjusted odds ratios between IVF and preterm birth in order to consider other values of *ROR*. Since the source population is a mixture of participants and non-participants, the risk of preterm birth will not follow a logistic regression model. No closed form expressions for the coefficients in the

approximating logistic model exist, so these were found by calculating the expected probabilities and using these as weights in estimating the logistic regression model. For smoking and SGA the model involved two levels of exposure and 48=3*2*2*2 different covariate patterns. Here we also considered three scenarios established in a way similar to that used for IVF and preterm birth.

For each scenario we considered source population sizes of 25,000 and 50,000. Each combination of scenario and sample size was simulated 5,000 times. The results of the simulations were summarized by the observed coverage probability of a nominal 95% interval of *ROR*, i.e. by computing how often the true *ROR* was contained in the interval estimated $ROR \cdot \exp(\pm 1.96 \cdot \text{se})$. All analyses and simulations were made using STATA version 8.2.

Interpretation of the estimates and the confidence intervals

We will here make some comments on the interpretation of the estimates and their confidence intervals. As an example we will consider the proportion of primiparae. In the data used in the study we had 45.7 % primiparae in the source population and 50.4% among the participants in the DNBC. This resulted in an observed ratio of 1.103. That is, among the pregnancies in North Jutland County and in the Aarhus Municipality, we found a 10.3% overrepresentation of primiparae among the participants in the DNBC. The confidence interval (1.089-1.117) presented in the paper reflects the uncertainty of this estimate when considering similar cohort studies in a similar setting. This is also the usual interpretation of the confidence interval. Another relevant question is: How large is the over- or underrepresentation of primiparae in the entire DNBC? Since we believe that the pregnancies in the North Jutland County and in the Aarhus Municipality are a representative sample of the eligible population of the DNBC, the 1.10 is a valid estimate of this ratio. The eligible population of the DNBC is finite (approximately 310,000 pregnancies) and we therefore have a finite population problem. An approximate confidence interval in such a situation can be obtained by multiplying the standard error with the finite population adjustment factor $\sqrt{1-\text{sample fraction}} = \sqrt{1-0.16} = 0.92$. This gives the confidence interval (1.115-1.090).

The observed adjusted ROR for BMI and stillbirth was 0.97 with the infinite sample space confidence interval (0.48-1.96) (Table 2). The confidence interval for ROR in the entire DNBC would be slightly narrower and becomes (0.51-1.85).

Additional table. Participation rates and crude relative odds ratios for 3 exposure-outcome associations*

	Participation rates			Crude relative odds ratio		
					(95% CI)	
	Outcome			ROR[†]	Equation[‡]	Bootstrap[§]
IVF and preterm birth	Preterm	Term	Total			
No treatment[‖]	30%	32%	32%	1.00		
IVF	34%	38%	38%	0.94	(0.68 - 1.28)	(0.67 - 1.31)
Total	30%	32%	32%			
Smoking and SGA	SGA	not SGA	Total			
Non-smoker[‖]	29%	33%	33%	1.00		
0-10 cig/day	23%	27%	27%	0.98	(0.79 - 1.21)	(0.80 - 1.20)
>10 cig/day	24%	22%	22%	1.24	(0.95 - 1.63)	(0.96 - 1.61)
Total	26%	32%	32%			
BMI and stillbirth	Stillbirth	Live birth	Total			
<18.5	-	27%	27%			
18.5-24.9[‖]	42%	33%	33%	1.00		
25.0-29.9	39%	31%	31%	1.01	(0.58 - 1.73)	(0.57 - 1.79)
30+	33%	28%	28%	0.94	(0.47 - 1.85)	(0.45 - 1.93)
Total	39%	32%	32%	-		

* Based on the populations displayed in Table 2. Associations are: In vitro fertilization (IVF) and preterm birth, smoking and birth of a small-for-gestational-age infant (SGA), and body mass index and antepartum stillbirth.

[†] ROR, relative odds ratio; CI, confidence interval; ref., reference.

[‡] Computed using equation 1, see text for details.

[§] Based on a non-parametric bootstrap sample of size 200.

[‖] Reference category.

Participation was in general non-differential, i.e. the dependence on the outcome category was consistent across exposure categories, the only exception being heavy smokers where also the highest bias was observed. Here the reverse pattern relative to the reference category was seen with a higher participation rate in births with SGA than in births without SGA.