Adjusting for unmeasured confounding using validation data

Simplified two-stage calibration for survival and dichotomous outcomes

Hjellvik, Vidar; De Bruin, Marie L.; Samuelsen, Sven O.; Karlstad, Øystein; Andersen, Morten; Haukka, Jari; Vestergaard, Peter; de Vries, Frank; Furu, Kari

Published in:
Statistics in Medicine

DOI (link to publication from Publisher):
10.1002/sim.8131

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Adjusting for unmeasured confounding using validation data. 2-stage calibration simplified for dichotomous outcomes

Vidar Hjellvik[a]*, Marie L. De Bruin[b,c], Sven O. Samuelsen[d,a] Øystein Karlstad[a], Morten Andersen[e,f,g], Jari Haukka[h], Peter Vestergaard[i], Frank de Vries[b,j], Kari Furu[a]

In epidemiology one typically wants to estimate the risk of an outcome associated with an exposure after adjusting for confounders. Sometimes outcome and exposure and maybe some confounders are available in a large data set, whereas some important confounders are only available in a validation data set that is typically a subset of the main data set. A generally applicable method in this situation is the two stage calibration (TSC) method. We present a simplified easy-to-implement version of the TSC for the case where the validation data is a subset of the main data. We compared the simplified version to the standard TSC version for incidence rate ratios, odds-ratios, relative risks, and hazard ratios using simulated data and the simplified version performed better than our implementation of the standard version. The simplified version was also tested on real data and performed well.

Keywords: bias correction; epidemiology; two stage calibration; unmeasured confounding; validation data

[a] *Department of Chronic Diseases and Ageing, Norwegian Institute of Public Health, Oslo, Norway*
[b] *Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, The Netherlands*
[c] *Department of Pharmacy, Copenhagen Centre for Regulatory Science (CORS), University of Copenhagen, Denmark*
[d] *Department of Mathematics, University of Oslo, Norway*
[e] *Centre for Pharmacoepidemiology, Karolinska Institutet, Clinical Epidemiology Unit, Karolinska University Hospital Solna*
[f] *Department of Drug Design and Pharmacology, University of Copenhagen, Copenhagen, Denmark*
[g] *Research Unit of General Practice, University of Southern Denmark, Odense, Denmark*
[h] *Department of Public Health, University of Helsinki, Finland*
[i] *Departments of Clinical Medicine and Endocrinology, Aalborg University Hospital, Aalborg, Denmark*
[j] *Department of Clinical Pharmacy & Toxicology, Maastricht University Medical Centre+, Maastricht, The Netherlands*
* *Correspondence to: Vidar Hjellvik, Department of Chronic Diseases and Ageing, Norwegian Institute of Public Health, Oslo, Norway. E-mail: vidar.hjellvik@fhi.no*

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

1. Introduction

The CARING (CAncer Risk and INsulin analoGues) project [1] aimed to evaluate the risk of cancer associated with the use of human insulins and insulin analogues. Drug use data and cancer data were available for a large number of individuals, mainly from nationwide registers, whereas data on many possible confounders were available in subsamples (validation data), including participants in health surveys. In this case it is possible to use a calibration method applying the information on confounders in the subsample to the entire study population, but the number of generally applicable, robust, easy-to-implement methods of high quality has been limited. Several methods for dealing with unmeasured confounding in this setting have been described in the literature, but most of them require simplifications (e.g. a limited number of independent dichotomous confounders [2, 3]) and/or educated guesses (e.g. probability distributions assigned to bias parameters [3, 4]), or rely on assumptions that are not always fulfilled (e.g. surrogacy in the case of propensity score calibration (PSC)). The PSC method proposed by Stürmer et al in 2005 [5] allows for multiple confounders that may be correlated and follow any distribution (continuous, categorical, dichotomous). The propensity score (PS) of an individual is it's probability for being exposed, computed by logistic regression from the individual's covariates. PSC assumes that the outcome is a linear function of the gold standard PS (based on all of the covariates that are available in the validation data set) and the exposure, and that the error prone PS (based on the subset of the covariates that are available in the full data set) is independent of the outcome given the gold standard PS and the exposure (surrogacy) [5, appendix 2]. At least one of the confounders must be available in the full data set, and the validation data set must be representative of the full data set. PSC does not use outcome information from the validation data set.

The 2-stage calibration (TSC) method proposed by Lin and Chen [6, 7] in 2014 only relies on the assumption that the validation data is representative for the full data. The TSC method is more generally applicable and more robust than the PSC method, but more cumbersome to implement since two parameters (λ and ν — see equation 3) must be computed. In situations where $\lambda \approx \nu$, the loss by replacing the ratio λ/ν by 1 will be modest, and since estimation of these parameters can then be omitted, implementation of an approximate calibration method becomes straightforward. The assumption that $\lambda \approx \nu$ does not hold for continuous outcomes, but the results here indicate that it often holds for survival and dichotomous outcomes. This simplified TSC method is referred to as the TSCS method below.

The objective of this paper was to compare the performance of the TSCS method with the performance of the standard TSC method in different settings with dichotomous outcome and dichotomous exposure using simulated data, and to assess the performance of the TSCS method on real data. We present results for incidence rate ratios, odds-ratios, relative risks, and hazard ratios.

2. Simplified and standard 2-stage calibration

2.1. The standard 2-stage calibration estimator

Lin and Chen [6] present the TSC method in a logistic regression setting, using propensity scores based on incomplete and complete confounder information, respectively, as explanatory variables. They also indicate that the method can be extended to other situations (e.g. survival analyses and relative risk estimates), and can be used to adjust for confounders directly (not via propensity scores). We apply it to adjust four types of effect estimates for missing confounders directly: incidence rate ratios, hazard rate ratios, odds-ratios, and relative risks. Assume that we have a full data set where the variables $C_1, \ldots, C_p$ (referred to as "measured") are available, and that the additional variables $U_1, \ldots, U_q$ ("unmeasured") are available in a subset of the full data set (validation data set). Using the

variables directly (not propensity scores) as explanatory variables, consider the survival models

$$h_{\beta,C,U}(t) = h(t;\beta T + \beta_1 C_1 + \ldots + \beta_p C_p + \beta_{p+1} U_1 + \ldots + \beta_{p+q} U_q) \quad (1)$$

$$h_{\gamma,C}(t) = h(t;\gamma T + \gamma_1 C_1 + \ldots + \gamma_p C_p) \quad (2)$$

where t is time to outcome/censoring and T is exposure (0 or 1). For instance we could have the proportional hazards model $h(t;\eta) = h_0(t)\exp(\eta)$. Here, model 1 is the "true" model yielding ubiased estimates (assuming that all relevant variables are included), whereas model 2 yields biased estimates. Let $\bar{\beta}$ and $\bar{\gamma}$ be the estimates we would have obtained by fitting the two models to the full data set. Since the variables $U_1, \ldots, U_q$ are only available in the validation data set we can not obtain the estimate $\bar{\beta}$, but, according to the standard TCS method [6], a calibrated unbiased estimate $\tilde{\bar{\beta}}'$ can be calculated from the formula

$$\tilde{\bar{\beta}}' = \hat{\beta} - \frac{\lambda}{\nu}(\hat{\gamma} - \bar{\gamma}) \quad (3)$$

where $\hat{\beta}$ and $\hat{\gamma}$ are the estimates of β and γ obtained by fitting models 1 and 2 respectively, to the validation data set. Further, λ is the covariance between $\hat{\beta}$ and $(\hat{\gamma} - \bar{\gamma})$, and ν is the variance of $(\hat{\gamma} - \bar{\gamma})$. The variance of $\tilde{\bar{\beta}}'$ is given by

$$var(\tilde{\bar{\beta}}') = var(\hat{\beta}) - \frac{\lambda^2}{\nu}. \quad (4)$$

Equation 3 and 4 also hold if the survival models models 1 and 2 are replaced by generalized linear models.

2.2. The simplified 2-stage calibration estimator

Equation 3 is a special case of

$$\tilde{\bar{\beta}}'_\Theta = \hat{\beta} - \Theta(\hat{\gamma} - \bar{\gamma})$$

which is unbiased for all finite values of Θ (cf. Section 3 in [7]). The variance of $\tilde{\bar{\beta}}'_\Theta$ is $var(\tilde{\bar{\beta}}'_\Theta) = var(\hat{\beta}) + \Theta^2\nu - 2\Theta\lambda$ which is a quadratic function of Θ, with a minimum at $\Theta = \lambda/\nu$. Since the curve is flat near the minimum, the increase in variance caused by the assumption that $\Theta = 1$ is small if λ/ν is close to 1. On the other hand, estimating λ and ν adds uncertainty to equations 3 and 4, so there is a gain in precision by not estimating them.

In the simplified version of the TSC method it is assumed that $\lambda = \nu$, and equation 3 simplifies to

$$\tilde{\bar{\beta}} = \hat{\beta} - \hat{\gamma} + \bar{\gamma}. \quad (5)$$

The variance of $\tilde{\bar{\beta}}$ is given by (see Appendix A for details)

$$var(\tilde{\bar{\beta}}) = var(\hat{\beta}) - var(\hat{\gamma}) + var(\bar{\gamma}). \quad (6)$$

Equation 6 is only valid if $\lambda = \nu$, but the error is small if $\lambda \approx \nu$. We have checked the validity of this approximation for dichotomous outcomes by simulations.

3. Method evaluation using simulated data

3.1. Estimation of λ and ν

To compare the performance of $\tilde{\bar{\beta}}'$ and $\tilde{\bar{\beta}}$, a total of 150 000 full data sets of censored survival data with $N = 128\,000$, 32 000, or 8000 were simulated (50 000 data sets for each N). 1500 parameter sets were randomly generated according

to Table 1, and for each parameter set 100 full data sets with at least one exposed individual with outcome were generated according to the description in Sections 3.3-3.5. For each full data set, 5 validation data sets were generated, containing random selections of respectively 1%, 5%, 10%, 20%, and 40% of the individuals in the full data set (the probability of being selected, though, depended on the exposure and outcome status through the parameters $P_{T,vd}$ and $P_{D,vd}$ in Table 1). The 100 data sets generated from the same parameter set are referred to as a group. For each validation data set size, only the groups where all of the 100 validation data sets had at least one exposed individuals with outcome were used.

For each full data set, $\bar{\beta}$ and $\bar{\gamma}$ were obtained from models 1 and 2 or their glm-equivalents (since this was simulated data we actually had access to the "unmeasured" covariates in the full data sets), and from each validation data set, $\hat{\beta}$ and $\hat{\gamma}$ were obtained, as well as the calibrated estimates $\tilde{\bar{\beta}}'$ and $\tilde{\bar{\beta}}$. All simulations and analyses were done in R [8]. To compute $\tilde{\bar{\beta}}'$, estimates of ν and λ in equation 3 are needed. We were not able to implement the method described in [6] in R, and used the following approach: For each group of 100 full data sets that were simulated with identical parameters, and the respective validation data sets, ν was estimated as the empirical variance of the 100 differences $\hat{\gamma} - \bar{\gamma}$. Correspondingly, λ was estimated as the empirical covariance between $\hat{\beta}$ and $\hat{\gamma} - \bar{\gamma}$. That is, for each group j of 100 data sets with identical parameters, $\nu_{j,\%_{\text{vd}}}$ and $\lambda_{j,\%_{\text{vd}}}$ were estimated for $\%_{\text{vd}} = 1$, 5, 10, 20, 40, as

$$\nu_{j,\%_{\text{vd}}} = \frac{1}{99}\sum_{i=1}^{100}(\hat{\gamma}_{i,j,\%_{\text{vd}}} - \bar{\gamma}_{i,j} - \mu_{1,j,\%_{\text{vd}}})^2$$

$$\lambda_{j,\%_{\text{vd}}} = \frac{1}{99}\sum_{i=1}^{100}(\hat{\gamma}_{i,j,\%_{\text{vd}}} - \bar{\gamma}_{i,j} - \mu_{1,j,\%_{\text{vd}}})(\hat{\beta}_{i,j,\%_{\text{vd}}} - \mu_{2,j,\%_{\text{vd}}})$$

where

$$\mu_{1,j,\%_{\text{vd}}} = \frac{1}{100}\sum_{i=1}^{100}(\hat{\gamma}_{i,j,\%_{\text{vd}}} - \bar{\gamma}_{i,j,\%_{\text{vd}}}) \qquad \text{and} \qquad \mu_{2,j,\%_{\text{vd}}} = \frac{1}{100}\sum_{i=1}^{100}\hat{\beta}_{i,j,\%_{\text{vd}}}.$$

Four type of effect estimates were obtained by fitting four different models to the simulated data - two time dependent and two time independent: incidence rate ratios (Poisson regression with time as offset); hazard rate ratios (Cox regression); odds-ratios (logistic regression); and relative risks (log-binomial regression).

3.2. Performance of the simplified TSC

To check the assumption that $\lambda = \nu$ and/or the validity of the above approach to estimate λ and ν, boxplots of the 1500 estimates of λ/ν were generated for each of the five validation data set sizes. Figure 1 a) shows the result for incidence rate ratios. The median ratio was very close to 1, independently of $\%_{\text{vd}}$, but its variance increased with increasing $\%_{\text{vd}}$ (Figure 1a). (The variances of λ and ν, however, both decreased with increasing $\%_{\text{vd}}$; data not shown). The numbers at the top of the panels indicate how many groups of 100 the estimates are based on (groups where at least one exposed individual had the outcome in all validation data sets).

To check the presicion and accuracy of $\tilde{\bar{\beta}}'$ and $\tilde{\bar{\beta}}$, boxplots of the differences $\tilde{\bar{\beta}}' - \bar{\beta}$ and $\tilde{\bar{\beta}} - \bar{\beta}$ were generated, where $\bar{\beta}$ is the fully adjusted estimate from the full data set (Figure 1b). The bias of both $\tilde{\bar{\beta}}'$ and $\tilde{\bar{\beta}}$ was on average small, but $\tilde{\bar{\beta}}'$ was more precise, in particular for large $\%_{\text{vd}}$'s.

The performance of the variance estimates in equations 4 and 6 was assessed by boxplots of the estimates themselves, and by comparing the variance estimates with the empirical variance of $\tilde{\bar{\beta}}'$ and $\tilde{\bar{\beta}}$ in the groups with identical parameters, i.e. with

$$s'^2_{j,\%_{\text{vd}}} = \frac{1}{99}\sum_{i=1}^{100}(\tilde{\bar{\beta}}'_{i,j,\%_{\text{vd}}} - \mu'_{j,\%_{\text{vd}}})^2 \qquad \text{and} \qquad s^2_{j,\%_{\text{vd}}} = \frac{1}{99}\sum_{i=1}^{100}(\tilde{\bar{\beta}}_{i,j,\%_{\text{vd}}} - \mu_{j,\%_{\text{vd}}})^2 \tag{7}$$

where $j = 1, \ldots, 1500$ is group number,

$$\mu'_{j,\%_{\text{vd}}} = \frac{1}{100}\sum_{i=1}^{100} \tilde{\tilde{\beta}}'_{i,j,\%_{\text{vd}}} \qquad \text{and} \qquad \mu_{j,\%_{\text{vd}}} = \frac{1}{100}\sum_{i=1}^{100} \tilde{\tilde{\beta}}_{i,j,\%_{\text{vd}}}.$$

The variance of $\tilde{\tilde{\beta}}'$, as estimated by equation 4, was in many cases < 0, in particular for small validation data sets (Figure 1c). The variances of $\tilde{\tilde{\beta}}$, as estimated by equation 6, was > 0 in all cases with $\%_{\text{vd}} \geq 10$, and in all cases with $\%_{\text{vd}} = 5$ and $N > 8000$. The distribution of $\widehat{var}(\tilde{\tilde{\beta}})$ increased with increasing $\%_{\text{vd}}$ (Figure 1c). This is because the average number of exposed individuals with outcome in the full dataset decreases with increasing $\%_{\text{vd}}$ due to the demand that at least one exposed individual in all validation data sets in a group of 100 should have the outcome. The distribution of $\widehat{var}(\tilde{\tilde{\beta}}')$ gradually approached that of $\widehat{var}(\tilde{\tilde{\beta}})$ as $\%_{\text{vd}}$ increased, getting quite close for $\%_{\text{vd}} = 40$. (Figure 1c).

Leaving out the negative variance estimates, $\widehat{var}(\tilde{\tilde{\beta}})$ was on average slightly smaller than the empirical variance of $\tilde{\tilde{\beta}}$ in particular for small validation data sets and small N. For $\%_{\text{vd}} \geq 10$ the bias was small. For $\%_{\text{vd}} = 10$ the median ratio $\widehat{var}(\tilde{\tilde{\beta}}')/s^2$ was 0.90, 0.93, and 0.94 for $N = 8000$, 32 000 and 128 000, respectively (Figure 1d). In the leftmost column of Figure 1, the variance estimates were multiplied by the ad-hoc factor $k = \exp(\{1 - \log_2(N/1000)/13\}/\%_{\text{vd}})$, which resulted in practically unbiased variance estimates except in the extreme case of $N = 8000$ and $\%_{\text{vd}} = 1$. The median ratio for $\%_{\text{vd}} = 10$ was now 0.97, 0.99, and 0.98 for $N = 8000$, 32 000 and 128 000, respectively.

Results corresponding to the last row of Figure 1 for relative risks, odds-ratios, and hazard-ratios are shown in Supplementary Figure 1, and the TSCS-method seemed to perform best in the odds-ratio setting in the sense that the underestimation of the variance for small validation data sets was smaller for odds-ratios than for the other effect estimates.

As an alternative to estimating λ and ν from groups of 100 data sets, we also tried a bootstrap approach where λ and ν were estimated for each data set separately by generating 100 bootstrap replicas of the data set. The details are given in the Supplementary material, and results from the bootstrap approach are presented in Supplementary Figure 2. The bootstrap approach yielded similar results as the "groups-of-100-approach", and could be an alternative to the method described in the appendix of [6] for applying the standard TSC method to a given data set. The bootstrap approach also yielded negative variance estimates, and they most often occured in data sets where $\hat{\lambda}$ was large, which makes sense since $var(\tilde{\tilde{\beta}}') \approx var(\hat{\beta}) - \lambda$ if $\lambda \approx \nu$ in equation 4. $\hat{\lambda}$ was strongly negatively correlated with the logarithm of the number of exposed cases. Whether negative variance estimates would occur if λ and ν were estimated using the methods in the appendix of [6], we do not know.

A third way to estimate λ and ν is via the dfbeta-function in R. In general this approach yielded $\lambda \approx \nu$, but the estimates of λ and ν were larger than those obtained by the bootstrap, resulting in more negative variance estimates when plugged into equation 4. We don't present results from the dfbeta-method, but the method is implemented in an R-script provided in the Supplementary material for applying the TCS and TSCS mehods. A simulated example data set is also available as Supplementary material.

3.3. Data generation, step 1. Simulation of p measured and q unmeasured covariates

Simulations with $2 \leq p + q \leq 4$ covariates were performed, of which $p \geq 1$ were measured and $q \geq 1$ were unmeasured. Measured and unmeasured covariates are denoted by $C_1, \ldots, C_p$ and $U_1, \ldots, U_q$, respectively. When it is irrelevant which covariates are measured and which are not, the notation $Z_1, \ldots, Z_{p+q}$ is used. The covariates were generated as follows: First the `mvrnorm` function in R was used to create an N x $(p + q)$ matrix with variables

$Z'_1, \ldots, Z'_{p+q}$ from a multinormal distribution with mean 0.5 and covariance matrix equal to a sub-matrix of

$$\Sigma_{HS} = 0.15^2 \times \begin{bmatrix} 1 & \rho_{1,2} = .440 & \rho_{1,3} = .533 & \rho_{1,4} = .201 & \rho_{1,5} = -.005 & \rho_{1,6} = .069 & \rho_{1,7} = .245 \\ & 1 & \rho_{2,3} = .355 & \rho_{2,4} = .207 & \rho_{2,5} = .094 & \rho_{2,6} = .114 & \rho_{2,7} = .410 \\ & & 1 & \rho_{3,4} = .303 & \rho_{3,5} = .211 & \rho_{3,6} = .011 & \rho_{3,7} = .273 \\ & & & 1 & \rho_{4,5} = .045 & \rho_{4.6} = -.254 & \rho_{4,7} = .382 \\ & & & & 1 & \rho_{5,6} = -.058 & \rho_{5,7} = .141 \\ & & & & & 1 & \rho_{6,7} = -.426 \\ & & & & & & 1 \end{bmatrix}$$

where the $\rho_{i,j}$'s are correlations measured in real data from Norwegian health surveys (subscripts i and j represents from 1 to 7: age, total cholesterol, systolic blood pressure, BMI, resting heart rate, HDL cholesterol, and log(triglycerides)). In the simulations, each of the $p+q$ simulated covariates were randomly assigned the role of one of the seven health survey variables.

The `rank` function was then applied on the columns of the matrix generated by `mvrnorm` to create variables uniformly distributed on [0,1] as $Z_j = \texttt{rank}(Z'_j)/N$, where $Z'_j, j = 1, \ldots, p+q$ are the Gaussian distributed variables, and N is the sample size. The uniform distribution was chosen to obtain more individuals with low or high exposure and/or outcome probabilities than a Gaussian distribution would yield.

After generation of the covariates, they were assigned the role of measured or unmeasured, "positive" or "negative", and factorized according to Table 1 (the relevant parameters are p, p^-, q^-, is.factor, and levels). See also Supplementary Figure 3 for a definition of positive and negative confounding.

3.4. Data generation, step 2: Simulation of exposure T

The association between covariates and exposure probability was simulated as exponential or logistic. First, the exposure probability was calculated for each individual as

$$P(T_i = 1) = \exp\{\log(\alpha_0) + \log(\alpha_{C1})C_{1,i} + \cdots + \log(\alpha_{Cp})C_{p,i} + \log(\alpha_{U1})U_{1,i} + \cdots + \log(\alpha_{Uq})U_{q,i}\} \quad (8)$$

in the exponential case, or, in the logistic case, as

$$P(T_i = 1) = \alpha_0 \, f(\alpha_{C1}, C_{1,i}) \times \cdots \times \, f(\alpha_{Cp}, C_{p,i}) \times f(\alpha_{U1}, U_{1,i}) \times \cdots \times \, f(\alpha_{Uq}, U_{q,i}) \quad (9)$$

where

$$f(\alpha, Z) = \left\{1 + \alpha e^{10(Z-0.5)}\right\} \left\{1 + e^{10(Z-0.5)}\right\}^{-1}.$$

Here, α_0 is the base exposure risk for an individual i with $Z_{ji} = 0 \; \forall \, j$, and α_j, $j > 0$ is the relative risk for exposure associated with $Z_{ji} = 1$ as compared to $Z_{ji} = 0$. Z_{ji}, $i = 1, \ldots, N$ are for each $j, j = 1, \ldots, p+q$, uniformly distributed on [0,1].

The base risk α_0 was chosen so that a mean exposure risk $\bar{\alpha} \in \{.01, .02, .04, .08, .16\}$ was obtained. This was achieved by first generating exposure probabilities with $\alpha_0 = .01$ and then multiplying the obtained individual exposure probabilities $P(T_i = 1)$ by $\bar{\alpha} / \frac{1}{N} \sum_{i=1}^{N} P(T_i = 1)$. If the resulting set of α's yielded exposure probabilities $P(T_i = 1) > 1$ for an individual i , α_0 was adjusted so that $P(T_i = 1) \leq 1 \; \forall \, i$. Finally, each individual i was assigned a random number $R_{i,T}$ between 0 and 1, and if $R_{i,T} < P(T_i = 1)$, individual i was considered exposed.

3.5. Data generation, step 3: Simulation of time to outcome D

Time t (years) to outcome for individual i was simulated as exponentially distributed with $f(t) = \eta e^{-\eta t}$, where

$$\eta = \exp\{\xi_{0,D} - \log(\xi_T)T_i - \log(\xi_{C1})C_{1,i} - \ldots - \log(\xi_{Cp})C_{p,i} - \log(\xi_{U1})U_{1,i} - \ldots - \log(\xi_{Uq})U_{q,i}\} \quad (10)$$

where $\xi_{0,D}$ determines the baseline hazard, which was chosen so that a randomly determined mean outcome risk $\bar{\xi}$ was (approximately) obtained. This was achieved by setting

$$\xi_{0,D} = 1.5 + N^{-1} \sum_{i=1}^{N} \{\log(\xi_T) + \log(\xi_1) Z_{1,i} + \ldots + \log(\xi_{p+q}) Z_{p+q,i}\} - \log(\bar{\xi}).$$

Similarly, time to censoring $t_{S,i}$ was simulated independently of $Z_1, \ldots, Z_{p+q}$, as exponentially distributed with parameter

$$\eta = \exp(\xi_{0,S}).$$

Follow-up time was set to 5 years. If $t_{D,i} < \min(t_{S,i}, 5)$, individual i was considered as having the outcome at time $t_{D,i}$, otherwise as censored at time $\min(t_{S,i}, 5)$. With $\xi_{0,D} = 4$ and all of the other ξ's of equation 10 being equal to one, the probability of surviving 5 years was about 90%, and $\xi_{0,S}$ was chosen in the range from 5 to 6, corresponding to censoring probabilities of about 3% and 1%, respectively, at low outcome probabilities.

4. Method evaluation using real data

4.1. Motivation

Although the simulated data sets represent a wide range of different scenarios, real data may have properties that were not taken into account in the simulations, e.g. regarding the functional form of the dependencies between covariates and exposure/outcome. For example, the simulated exposure probability increased exponentially or logistically with increasing values of the covariates, and the relationship between each covariate and exposure/outcome was simulated independently of the correlations between the covariates. Also, the validation data set is not always a random sample of the full data set in the real world. To check the robustness of the simplified TSC method on real data, including situations where the validation data are not representative of the full data in all respects, the methods were applied to a data set containing information from the Norwegian Prescription Database (NorPD), the Cancer Registry of Norway, Statistics Norway, and Norwegian health surveys conducted in the period 1975-2003.

4.2. Data

The study population consisted of health survey participants from three different health surveys: The counties study in 1975-1988 ($N \approx 43\,000$), the 40 year programme in 1985-1999 ($N \approx 371\,000$), and the CONOR study in 1994-2003 ($N \approx 106\,000$). The numbers refer to individuals alive and 40-70 years old in 2004. The health survey data were obtained from the Norwegian Institute of Public Health, and data on exposure and outcome for the participants were obtained by linkage to the NorPD, the Cancer Registry, and Statistics Norway.

The NorPD [9] contains data on all drugs dispensed from pharmacies to Norwegian Citizens since 1/1-2004, including date of dispensing and type of drug according to the Anatomical Therapeutic Chemical (ATC) classification system [10]. The Cancer Registry of Norway was established in 1951, and contains data on all cancer diagnoses for all Norwegian Citizens. From Statistics Norway we have obtained data on marital status and gross income from a nationwide census in 2001, as well as date of death. The data were linked using the encrypted personal identification number that all Norwegian citizens are given at birth or immigration.

Eight combinations of outcome and exposure were tested. The outcome was i) death or ii) incident cancer (any cancer type), and the exposure was i) prevalent use of oral antidiabetics (ATC group A10B) in 2004 or ii) prevalent use of insulins (A10A) in 2004 or iii) incident use of oral antidiabetics after 2004 or iv) incident use of insulins

after 2004. Follow-up was from 1/1-2005 to 1/1-2011. Prevalent drug users were treated as exposed from 1/1-2005. Incident drug users were treated as unexposed until the first prescription of an exposure drug. The analyses were performed separately for men and women.

The following covariates were included: Age (birth year), year of health survey participation, BMI, log(triglycerides), systolic blood pressure, smoking, marital status in 2001, and gross income in 2001. The first five were continuous variables. Smoking was categorical with 5 levels: never-smoker, ex-smoker with < 5000 or ≥ 5000 pack-years, or current smoker with < 5000 or ≥ 5000 pack-years (number of pack-years was computed as number of years smoked times number of cigarettes per day divided by 20). Marital status was dichotomous (married/cohabitating or not). Income was categorical with 6 levels: 0, 1-100, 101-200, 201-300, 301-400, > 400 thousand NOK per year.

Since the purpose of this exercise was to test the method on real data rather than to estimate the risk or cancer or death associated with use of antidiabetic drugs, the motivation behind the choice of covariates was to have both continuous and categorical variables represented, with various degree of confounding potential. For the same reason, a detailed description of the data is not given.

To reduce the relative importance of age as a confounder, only individuals aged 40-70 years in 2004 were included. Individuals with missing values on any covariate were excluded. When the outcome was cancer, individuals with a cancer diagnosis before 2005 were excluded. When the exposure was incident drug use, individuals who were prescribed the relevant drug before 2005 were excluded. The exact number of individuals in the full data set for a given combination of exposure and outcome depended on the number of baseline exclusions (cf Table 2). Before the exclusions the full data set included 248 415 men and 271 726 women.

4.3. Methods

For each combination of exposure, outcome, and sex, 500 validation data sets each consisting of a 10% random sample of the full data set was constructed. Confidence intervals of the corrected estimates were calculated as follows:

1. Draw n_V individuals without replacement from the full data set 500 times, thus creating $m = 500$ validation data sets.
2. For $i = 1, \ldots, m$,
   - compute $\hat{\beta}_i^*$ and $\hat{\gamma}_i^*$ from validation data set i.
   - compute $\tilde{\bar{\beta}}_i^* = \hat{\beta}_i^* - \hat{\gamma}_i^* + \bar{\gamma}$.
   - compute $var(\tilde{\bar{\beta}}_i^*) = var(\hat{\beta}_i^*) - var(\hat{\gamma}_i^*) + var(\bar{\gamma})$.
3. Estimate $\tilde{\bar{\beta}}$ by the median of the $\tilde{\bar{\beta}}_i^*$'s
4. Estimate $\widehat{var}(\tilde{\bar{\beta}})$ by the median of the $var(\tilde{\bar{\beta}}_i^*)$'s.
5. Estimate the 95% confidence interval of $\tilde{\bar{\beta}}$ by $\tilde{\bar{\beta}} \pm 1.96\sqrt{\widehat{var}(\tilde{\bar{\beta}})}$

Age is a strong confounder for the association between drug use and cancer/death. To test the effect of non-representative validation data sets with respect to age, some validation data sets were drawn from i) the youngest half of the full data, ii) the oldest half of the full data, and iii) the median age (54 years) $\pm$ 2 years (narrow age range).

First, a 'gold standard' $\bar{\beta}$ estimate was obtained from the full data set using a model including all of the covariates. Then, one or more of the covariates were taken to be unmeasured, i.e. only available in the validation data. The error-prone $\bar{\gamma}$ was estimated based on the remaining covariates, and $\hat{\beta}$ and $\hat{\gamma}$ calculated from the validation data set were used to compute the corrected $\tilde{\bar{\beta}}$ according to equation 5. The corrected $\tilde{\bar{\beta}}$ should be close to $\bar{\beta}$ if the calibrations worked well.

4.4. Results

Figure 2 shows the results of the method applied to real data with prevalent insulin use in 2004 as exposure and death as outcome. Those who used insulin in 2004 were twice as likely to die before 2011 as those who did not use insulin after adjusting for all of the covariates mentioned above ($\exp(\bar{\beta}) = \exp(0.66) = 1.93$). This is unlikely to be caused by insulin itself, but rather a consequence of insulin users having diabetes and other comorbidities, i.e. insulin use is a proxy of diabetes. Thus, the 'gold standard' estimate here is biased. However, the bias in the 'error prone' estimates from the full data set (triangles) increases when more covariates are left out, whereas the corrected estimates (crosses) are quite close to the original 'gold standard' estimate (bullet). That is also true when none of the covariates are adjusted for, although the uncertainty in $\bar{\tilde{\beta}}$ is larger than in $\bar{\beta}$. The confidence intervals based on equation 6 (method 2 above) are quite close to the gold standard interval, whereas those computed by method 1 are somewhat wider.

Summary statistics for all of the 16 combinations of exposure, outcome and sex are given in Table 2 (row 11 corresponds to Figure 2). The performance of the TSCS method when age is adjusted for and the validation data set is age-restricted is summarized in Table 3. The validation data sets drawn from the oldest half of the full data still yield good calibrated estimates, whereas the young and the narrow age group validation data sets tend to yield over-calibrated estimates. This is probably because both exposure and outcome are more common in the oldest population.

A graphical representation of the results of Table 3 is given in Figure 3 together with corresponding results when no covariates are adjusted for in $\bar{\gamma}$, and when age, BMI, triglycerides and smoking are adjusted for (the NONE-row and the A,B,T,P-row, respectively, in Figure 2). When no covariates are adjusted for, the simplified TSC method performs well when the validation data are representative for the full data, otherwise it tends to under-calibrate (top panel of Figure 3). When age, BMI, triglycerides, and smoking are adjusted for, the method tends to over-calibrate when the validation data set has a young or narrow age range and the outcome is death ($\bar{\tilde{\beta}} - \bar{\beta} < 0$, bottom panel of Figure 3). When the outcome is cancer, there is very little residual confounding left (compared to the gold standard estimate). Still, $\bar{\tilde{\beta}}$ is closer to $\bar{\beta}$ than is $\bar{\gamma}$ in all of the eight exposure/sex combinations when the validation data is not age restricted.

The interquartile range is in general smallest for the validation data sets with old people, which is logical since both outcome and exposure are more prevalent in this group.

5. Method application: drug use and lung cancer

In the CARING project, an active-comparator analysis was performed, comparing users of insulin glargin to users of human insulin. Several additional confounders were adjusted for [1], and no increased risk associated with use of insulin glargin or detemir was found. Therefore we do not go further along that path here. In 2016, Pottegård et al [11] published a systematic screening approach for identifying associations between prescribed drugs and cancer risk, using nationwide Danish health registries. At the second ATC-level the strongest associations were found between ATC code R03 (drugs for obtructive airway diseases) and various subtypes of lung cancer after adjusting for age, sex, education and morbidity (Charlson Comorbidity Index (CCI)) (Table 3 in [11]). Clearly, these odds-ratios are confounded by smoking since smokers tend to use R03-drugs and to have lung cancer more often than non-smokers, but there is no nationwide register with data on smoking status for the entire population. However, the Norwegian health surveys include data on smoking habits on a subsample of the population, and we were able to approximately replicate the Danish study on Norwegian data with additional adjustments for smoking using the TSCS method with the health surveys as validation data. That is, we did a nested case-control study using prescription and cancer registry data from January 1, 2004 and 1953, respectively, through 2014. Cases were individuals with an incident

lung cancer diagnosed between January 1, 2009 and December 31, 2014, at an age of 18-85 years, and with no cancer diagnoses (except non-melanoma skin cancer) before the incident lung cancer diagnosis. The index date was defined as the 15th in the month of diagnosis (the exact day of diagnosis was not available). Lung cancer was defined as ICD-10 C33-34 (we could not distinguish between C33 and C34 in the data), and the subtypes were defined by histology codes as given in Appendix B. For each case, 100 random controls, matched on birth year and sex, were selected from all Norwegian citizens that were cancer free at the case's index date. Exposure was the number of prescription fills for R03-drugs from January 1, 2004 to one year before index date. Odds-ratios for the outcomes associated with exposure was estimated by comparing individuals with 8 fills or more with individulas with 0-1 fills. Odds-ratio estimates from the full data were adjusted for education (5 levels) and comorbidity (CCI based on data from the Norwegian Patient Registry from 2008 (individual data is not available before 2008 for this register)). The CCI was computed using ICD-10 codes from Table 1 in [13] and weights from [12]. The odds-ratios from the full data set were calibrated by using smoking data (pack-years as described in Section 4.2) from the health surveys.

The odds-ratio from the full Norwegian data set was significantly higher then the corresponding odds-ratio from the Danish data for three of the five lung cancer subtypes (Table 4). The reduction in the risk estimates achieved by calibration ranged from 41% and 61% (a reduction in OR from 2 to 1, from 3 to 2, and from 4 to 3 corresponds to a risk reduction of 100%, 50% and 33.3%, respectively), but all of the calibrated odds-ratios were still significantly larger than 1. Most of the effect can probably be explained by residual confounding. The smoking data are mainly from the period 1985-2000, and some of those who reported current smoking in the health surveys probably quitted smoking before 2009. On the other hand, for those who continued smoking after the health survey participation, the number of pack-years at index date was higher than the number at the survey date (the latter is adjusted for in Table 4). However, assuming that those who reported current smoking in the health surveys continued smoking until 2009, adjusting for number of pack-years per 2009 had hardly no effect on the estimates. Neither had additional adjustments for other potential confounders (BMI, triglycerides, total cholesterol, physical activity and year of health survey participation).

Whereas Pottegård et al used 10 controls per case, we used 100. This was because the odds-ratios were rather dependent on the selection of controls with 10 controls per case. For example, the odds-ratios for squamous cell carcinoma (line 2 in Table 4) varied from 3.22 to 3.49 for 10 different random selections of 10 controls per case. The controls used for Table 4 were not matched to cases on health survey participation. Doing such matching only changed the calibrated odds-ratios in the five c-lines in Table 4 with -0.03, -0.02, 0.00, -0.02 and 0 from top to bottom, but for the "Other" and "Adenocarcinoma" categories there were not enough controls participating in the health survey for some of the youngest health survey participating cancer cases.

6. Discussion

In this study, a simplified version of the 2-stage calibration method presented by Lin and Chen [6] in 2014 has been suggested for regression models with dichotomous outcomes. The simplified version (TSCS) is very easy to apply, and it performed better than our implementation of the standard method (TSC) on simulated data. It also performed well on real data. The TSCS method assumes that $\lambda = \nu$, but if $\lambda \approx \nu$ the error in assuming that they are equal is small. We have estimated λ and ν for dichotomous outcomes by generating groups of 100 data set with identical parameters, and by bootstrapping individual data sets. Both approaches gave estimates of λ/ν close to 1. The error caused by violation of the assumption that $\lambda = \nu$ is illustrated in Supplementary Figure 4.

The (simplified) TSC method assume that the *associations* between exposure, outcome, and confounders are the same in the validation data as in the full data, but the validation data need not necessarily be representative of the full data with respect to the *distribution* of exposure, outcome, and confounders. This was illustrated on a real

data set by letting the validation data have an older, younger, or narrower age distribution than the full data. In the case where no confounders were available in the full data, the TSCS method undercorrected, but if age was available in the full data, the method performed fairly well if the validation data consisted of people from the older half of the full data. When the validation data had a young or narrow age distribution, the TSCS method tended to overcorrect.

For a continous outcome, the assumption that $\lambda = \nu$ does not hold. See Supplementary Figure 5 for an example.

Neither does the assumption hold in the "errors-in-variables" situation where the "unmeasured" confounders are not completely missing in the full data, but are measured with less precision than in the validation data set (Supplementary Figure 6). Still, the TSCS yielded more presice estimates than the TSC in our simulations (Supplementary Figure 6 f), and the variance estimates of the two methods performed similarly (Supplementary Figure 6 g and h). In some of our simulated data sets the validation set cases were oversampled and in this situation the $\hat{\beta}$ may be biased when using the same cohort estimating technique. It is interesting to note that the calibration technique still worked well. This is perhaps a reflection of the $\hat{\gamma}$ also having a similar bias compared to $\bar{\gamma}$ and the correction term then appear to compensate also for the bias.

Strengths and limitations. The TSCS method can handle complex confounder scenarios (many correlated continuous or categorical unmeasured confounders), and it is easy to apply. In the current settings with dichotomous outcomes, the TSCS estimate was precise and accurate, actually more precise than our implementation of the standard TSC estimate since estimating λ and ν in the latter adds uncertainty. However, the assumption that $\lambda = \nu$ doesn't hold with a continuous outcome variable, or in the "errors-in-variables" situation.

In summary, the TSCS method appears to be a precise and accurate easy-to-use calibration method for regression models with a dichotomous outcome where the unmeasured covariates are completely missing in the full data set.

Acknowledgements

The research leading to the results of this study has received funding from the European Communitys Seventh Framework Programme (FP-7) under grant agreement number 282 526, the CARING project. The funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report.

Disclaimer

The study has used data from the Cancer Registry of Norway and the Norwegian Patient Registry. The interpretation and reporting of these data are the sole responsibility of the authors, and no endorsement by the Cancer Registry of Norway nor the Norwegian Patient Registry is intended nor should be inferred.

Appendix A. Variance of the simplified TSC estimator.

In the standard TCS method [6], the variance of the corrected estimate is calculated from the formula

$$var(\tilde{\hat{\beta}}) = var(\hat{\beta}) - \lambda^2/\nu.$$

Setting $\lambda = \nu$, this simplifies to

$$\begin{aligned} var(\tilde{\hat{\beta}}) &= var(\hat{\beta}) - \nu \\ &= var(\hat{\beta}) - var(\hat{\gamma} - \bar{\gamma}) \\ &= var(\hat{\beta}) - \{var(\hat{\gamma}) + var(\bar{\gamma}) - 2cov(\hat{\gamma}, \bar{\gamma})\}. \end{aligned} \tag{A1}$$

If the validation data set is not a subset of the full data set, the covariance term in equation A1 is zero, but normally this is not the case. Let the full data set consist of N independent observations $\{X_1, \ldots, X_n, X_{n+1}, \ldots, X_N\}$ where $\{X_1, \ldots, X_n\}$ is the validation data set. Here, $X_i = \{D_i, T_i, C_{1,i}, ..., C_{p,i}\}$ is a vector of exposure, outcome, and measured confounders for individual i. Clearly, $\bar{\gamma}$ is a function of $\{X_1, \ldots, X_N\}$. Now, assume that $\bar{\gamma} = \bar{\gamma}_w$ where $\bar{\gamma}_w$ is a weighted average of two estimates from two subsets of the data set:

$$\bar{\gamma}_w = f(X_1, \ldots, X_n, X_{n+1}, \ldots, X_N) = \frac{w_1 f(X_1, \ldots, X_n) + w_2 f(X_{n+1}, \ldots, X_N)}{w_1 + w_2} \doteq \frac{w_1\hat{\gamma} + w_2\breve{\gamma}}{w_1 + w_2}, \tag{A2}$$

where $f(X_{n+1}, \ldots, X_N) \doteq \breve{\gamma}$ is the estimate corresponding to $\hat{\gamma}$ based on the observations that are *not* in the validation data set and $f(X_1, \ldots, X_n) = \hat{\gamma}$. Then, if $\bar{\gamma} = \bar{\gamma}_w$, the covariance term of equation A1 is

$$cov(\hat{\gamma}, \bar{\gamma}) = cov\left(\hat{\gamma}, \frac{w_1\hat{\gamma} + w_2\breve{\gamma}}{w_1 + w_2}\right) = \frac{w_1}{w_1 + w_2} var(\hat{\gamma}) + \frac{w_2}{w_1 + w_2} cov(\hat{\gamma}, \breve{\gamma})$$

where $cov(\hat{\gamma}, \breve{\gamma}) = 0$ since the X's are independent. With the weights

$$w_1 = var(\hat{\gamma})^{-1} \doteq \hat{\sigma}^{-2} \text{ and } w_2 = var(\breve{\gamma})^{-1} \doteq \breve{\sigma}^{-2}$$

we have

$$cov(\hat{\gamma}, \bar{\gamma}) = \frac{1}{w_1 + w_2}$$

and

$$var(\bar{\gamma}) = var\left\{\frac{\breve{\sigma}^2}{\hat{\sigma}^2 + \breve{\sigma}^2}\hat{\gamma} + \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \breve{\sigma}^2}\breve{\gamma}\right\} = \frac{\breve{\sigma}^4}{(\hat{\sigma}^2 + \breve{\sigma}^2)^2}\hat{\sigma}^2 + \frac{\hat{\sigma}^4}{(\hat{\sigma}^2 + \breve{\sigma}^2)^2}\breve{\sigma}^2 = \frac{\hat{\sigma}^2\breve{\sigma}^2}{\hat{\sigma}^2 + \breve{\sigma}^2} = \frac{1}{w_1 + w_2}.$$

Thus,

$$cov(\hat{\gamma}, \bar{\gamma}) = var(\bar{\gamma})$$

and equation A1 simplifies to

$$var(\tilde{\hat{\beta}}) = var(\hat{\beta}) - var(\hat{\gamma}) + var(\bar{\gamma}).$$

To check the assumption that $\bar{\gamma} = \bar{\gamma}_w$ we estimated $\hat{\gamma}$ and $\breve{\gamma}$ from each of 77 500 data sets with $N = 128\,000$ and plugged the estimates into equation A2. For w_1 and w_2 we used the inverse of the squared standard errors of the estimates. The resulting estimates $\bar{\gamma}_w$ were very close to $\bar{\gamma}$ when the probability of being selected to the validation data set was independent of either exposure or outcome (see Supplementary Figure 7). When the probability for being selected to the validation data set was doubled for those with an outcome (the diseased) and halved for the exposed, or vise versa, $\bar{\gamma}_w - \bar{\gamma} \approx 0.2$ for $\%_{\text{vd}} = 40$. When the selection probability was doubled or halved for both exposed and diseased, $\bar{\gamma}_w - \bar{\gamma} \approx -0.2$. For smaller validation data sets the difference $\bar{\gamma}_w - \bar{\gamma}$ decreased proportionally with $\%_{\text{vd}}$ (≈ 0.1 for $\%_{\text{vd}} = 20$ and 0.05 for $\%_{\text{vd}} = 10$, etc.

Appendix B. Definition of lung cancer subtypes

Definition of subtypes of lung cancer (ICD-10 C33-34) by histology codes:

Lung cancer subtype	Histology codes
Adenocarcinoma	81403 82303 82503 82523 82603 83103 83233 84803 85003
Squamous cell carcinoma	80703 80713 80833
Small cell carcinoma	80413 80443
Other (non-small cell)	80463
Carcinoid	82403 82463 82493
Large cell carcinoma	80123
Other	All remaining codes

References

1. But A, De Bruin ML, Bazelier MT, Hjellvik V, Andersen M, Auvinen A, Starup-Linde J, Schmidt MK, Furu K, de Vries F, Karlstad Ø, Ekstrøm N, Haukka J. Cancer Risk among insulin users: comparing analogues with human insulin in the CARING five-country cohort study. *Diabetologia* 2017; 60(9), 1691–1703.
2. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Safety* 2006; 15, 291–303.
3. Lash TL, Schmidt M, Jensen AØ, Engebjerg MC. Methods to apply probabilistic bias analysis to summary estimates of association. *Pharmacoepidemiol Drug Safety* 2010; 19, 638–644.
4. Lash TL, Fox MP, Fink AK. *Applying quantitative bias analysis to epidemiological data*. New York: Springer; 2010. 192 p.
5. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting Effect Estimates for Unmeasured Confounding with Validation Data using Propensity Score Calibration. *Am J Epidemiol* 2005; 162, 279–289.
6. Lin HW, Chen YH. Adjustment for missing confounders in studies based on observational databases: 2-stage calibration combining propensity scores from primary and validation data. *Am J Epidemiol* 2014; 180, 308–317.
7. Chen YH, Chen H. A unified approach to regression analysis under double-sampling designs. *J R Statist Soc* 2000; 60, 449–460.
8. R Core Team 2016-2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
9. Furu K, Wettermark B, Andersen M, Martikainen JE, Almarsdottir AB, Srensen HT. The Nordic Countries as a cohort for pharmacoepidemiological research. *Basic Clin Pharmacol Toxicol* 2010; 106, 86–94.
10. WHO Collaborating Centre for Drug Statistics Methodology. ATC/DDD Index 2017. Available from www.whocc.no/atc_ddd_index/, accessed 10 December 2018.
11. Pottegård A, Friis S, dePont Christensen R, Habel LA, Gagne JJ, Hallas J. Identification of Associations Between Prescribed Medications and Cancer: A Nationwide Screening Study. *EBioMedicine* 2016; 7, 73–79.
12. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation. *J Chron Dis* 1987; 5, 373–383.
13. Thygesen SK, Christiansen CF, Christensen S, Lash TL and Srensen HT. The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of Patients. *BMC Medical Research Methodology* 2011; 11:83.

Tables

Table 1. Parameters used in the simulations. For each 100th simulated data set the parameters in the first column were randomly assigned values among those in the second column, with probabilities indicated in the third column. "Uniform" means that each value is equally probable. All parameters were simulated independent of each other.

parameter(s)	value(s)	probability	remark
$p+q$	2, 3, 4	uniform	number of covariates
p	1,..., p+q-1	uniform	no of measured covariates
ξ_T	0.5, 1, 1.5, 3	uniform	see equation 10
α_{C1}, α_{U1},ξ_{C1},ξ_{U1}	1, 2, 4	uniform	see equations 8 and 10
$\xi_{Cj},\ j>1$	$k_j \times \xi_{C1}$, $k_j \in [0, 0.1, \ldots, 1.0]$	uniform	a)
p^-	0, ..., p	0.6, $0.4/p$, ..., $0.4/p$	b)
q^-	0, ..., q	0.6, $0.4/q$, ..., $0.4/q$	b)
$\bar{\alpha}$, $\bar{\xi}$ ($N = 8000$)	0.04, 0.08, 0.16	uniform	c)
$\bar{\alpha}$, $\bar{\xi}$ ($N = 32\,000$)	0.02, 0.04, 0.08, 0.16	uniform	c)
$\bar{\alpha}$, $\bar{\xi}$ ($N = 128\,000$)	0.01, 0.02, 0.04, 0.08, 0.16	uniform	c)
$\xi_{0,S}$	$5 \le \beta_{0,S} \le 6$	uniform	d)
$P_{D,vd}$, $P_{T,vd}$	0.5, 1, 2	0.25, 0.5, 0.25	e)
is.factor	FALSE, TRUE	0.8, 0.2	f)
levels	0,2,3,4,5,6	0.5, 0.1, 0.1, 0.1, 0.1, 0.1	g)
covariance matrix	Σ_{HS}		h)
functional form	exponential		

a) If $p > 1$ then η_{Mj}, $j = 2 \ldots, p$ were generated as follows: $p-1$ numbers $v_1, \ldots, v_{p-1}$ were drawn with replacement from the set $[0, 0.1, 0.2, \ldots, 1.0]$, and sorted in decrasing order, resulting in factors $v_{(1)} \ge, \ldots, \ge v_{(p-1)}$. Then η_{M2} and, if $p = 3$, η_{M3}, were calculated as $\eta_{M2} = v_{(1)}\eta_{M1}$ and $\eta_{M3} = v_{(2)}\eta_{M1}$. The parameters η_{Uj}, β_{Mj}, β_{Uj}, $j > 1$ were generated equivalently.
b) p^- and q^- denote the number of 'negative' measured and unmeasured confounders, respectively, i.e. the number of covariates that were simulated to reduce the outcome risk ($\beta_j < 1$). The probablilty of having zero negative measured confounders was 0.6, and the probability of having k negative measured confounders, $1 \le k \le p$ was $0.4/p$. Same for unmeasured negative confounders. If, for example, $q = 4$ and $q^- = 2$, then $\beta_{U1} = 1/\beta'_{U1}$ and $\beta_{U2} = 1/\beta'_{U2}$ where β'_{U1} and β'_{U2} are the β's generated as under remark c).
c) $\bar{\alpha}$ and $\bar{\xi}$ denote mean probabilites of exposure and outcome, respectively, averaged over all individuals in the generated data set. These parameters determined the base risks parameters α_0 and $\xi_{0,D}$ in equations 8, 9 and 10
d) $\beta_{0,S}$ determines the probability of being censored. $\beta_{0,S} = 5$ and $\beta_{0,S} = 6$ imply censoring probabilities of about 3% and 1% respectively at low outcome probabilities.
e) $P_{D,vd}$ denotes the probability of being selected for the validation data set for an individual with the outcome compared to one without. Equivalently, $P_{T,vd}$ denotes the probability of being selected for an exposed individual compared to a non-exposed individual. For an exposed individual with the outcome, the probability is $P_{D,vd}P_{T,vd}$ times the probability for an non-exposed individual without the outcome. Restrictions: If $\bar{\eta} = 0.01$ and $P_{T,vd} = 0.5$ or $\bar{\eta} = 0.32$ and $P_{T,vd} = 2$, $P_{T,vd}$ was changed to 1. If $\bar{\beta} = 0.01$ and $P_{D,vd} = 0.5$ or $\bar{\beta} = 0.32$ and $P_{D,vd} = 2$, $P_{D,vd}$ was changed to 1.
f) The probability that at least one of the covariates is a categorical variable is 0.2.
g) If is.factor = TRUE, then the probability for each covariate of being continuous is 0.5, and of being a categorical variable with k categories, $k = 2, \ldots, 6$, is 0.1.
h) Σ_{HS} is defined in section Section 3.3.

Table 2. Summary characteristics of data from Norwegian health surveys, restricted to participants aged 40-70 years in 2004. N =Number of subjects, $\%_T$ =percentage exposed, $\%_D$ =percentage with outcome, $N_{T\&D}$ =number of exposed with outcome, $\bar{\beta}$ =gold standard estimate (adjusted for age, BMI, log(triglycerides), smoking, marital status, income, systolic blood pressure, year of survey).

No.	Outcome	Exposure	Users	Sex	N	$\%_T$	$\%_D$	$N_{T\&D}$	$\bar{\beta}$
1	Cancer	Insulin	New	Men	230 843	0.80	3.80	52	-0.11
2	Cancer	Insulin	New	Women	250 296	0.50	3.61	31	0.05
3	Cancer	Insulin	Prev	Men	233 936	1.32	3.82	156	0.04
4	Cancer	Insulin	Prev	Women	252 552	0.89	3.62	115	0.15
5	Cancer	Oral AD	New	Men	227 864	3.53	3.78	198	-0.19
6	Cancer	Oral AD	New	Women	248 294	2.31	3.60	132	-0.06
7	Cancer	Oral AD	Prev	Men	233 936	2.60	3.82	329	0.02
8	Cancer	Oral AD	Prev	Women	252 552	1.69	3.62	205	0.02
9	Death	Insulin	New	Men	238 360	0.88	2.47	146	1.00
10	Death	Insulin	New	Women	264 176	0.55	1.63	85	1.43
11	Death	Insulin	Prev	Men	241 600	1.34	2.55	258	0.66
12	Death	Insulin	Prev	Women	266 596	0.91	1.68	160	0.86
13	Death	Oral AD	New	Men	235 243	3.62	2.45	172	-0.10
14	Death	Oral AD	New	Women	262 029	2.37	1.63	106	0.14
15	Death	Oral AD	Prev	Men	241 600	2.63	2.55	391	0.31
16	Death	Oral AD	Prev	Women	266 596	1.71	1.68	197	0.35

Table 3. Errors in crude ($\bar{\gamma}$) and corrected ($\tilde{\bar{\beta}}$) estimates for the data sets in Table 2. In each line, $\tilde{\bar{\beta}}$ is the median of 500 estimates from 500 validation data sets randomly selected from the (age-restricted) full data set. The crude estimates are corrected for age only.

Validation data age distribution:						Full	Young	Old	Narrow
No.	Outcome	Exposure	Users	Sex	$\bar{\gamma} - \bar{\beta}$	$\tilde{\bar{\beta}} - \bar{\beta}$			
1	Cancer	Insulin	New	Men	0.07	0.001	-0.011	-0.009	-0.013
2	Cancer	Insulin	New	Women	0.09	0.002	0.008	-0.020	-0.042
3	Cancer	Insulin	Prev	Men	0.03	0.002	-0.023	0.001	-0.022
4	Cancer	Insulin	Prev	Women	0.07	0.005	0.035	-0.033	0.008
5	Cancer	Oral AD	New	Men	0.04	0.002	-0.022	-0.011	-0.033
6	Cancer	Oral AD	New	Women	0.08	-0.003	-0.013	-0.012	-0.027
7	Cancer	Oral AD	Prev	Men	0.05	0.004	-0.055	0.006	-0.051
8	Cancer	Oral AD	Prev	Women	0.10	0.004	0.008	-0.022	-0.037
9	Death	Insulin	New	Men	0.26	0.006	-0.107	0.024	-0.045
10	Death	Insulin	New	Women	0.20	0.007	-0.106	0.011	-0.146
11	Death	Insulin	Prev	Men	0.31	0.005	-0.044	0.016	-0.023
12	Death	Insulin	Prev	Women	0.31	0.003	-0.007	0.006	-0.036
13	Death	Oral AD	New	Men	0.15	0.000	-0.166	0.028	-0.067
14	Death	Oral AD	New	Women	0.16	0.001	-0.122	0.007	-0.103
15	Death	Oral AD	Prev	Men	0.30	0.000	-0.157	0.027	-0.049
16	Death	Oral AD	Prev	Women	0.30	0.006	-0.074	0.019	-0.107

Prepared using simauth.cls

Table 4. Odds-ratios for five types of lung cancer associated with use of drugs for obstructive airway diseases (ATC-code R03). Adjusted for education and comorbidity in the full population (nationwide registers) and calibrated with smoking data from health surveys.

Lung cancer type	Exposed: Cases/odds	Unexposed: Cases/odds	OR (95% CI)	note	Risk red.
	1824/ 0.254	5947/ 0.081	2.61 (2.45-2.78)	a)	
Squamous cell carcinoma	710/0.0294	1589/0.0074	3.35 (3.06-3.68)	b)	
	156/0.0283	377/0.0067	2.12 (1.92-2.34)	c)	52%
	147/ 0.231	563/ 0.084	2.43 (1.96-3.00)	a)	
Carcinoid	56/0.0169	261/0.0084	1.91 (1.41-2.58)	b)	
	10/0.0132	62/0.0088	1.54 (1.12-2.12)	c)	41%
	680/ 0.213	2843/ 0.087	2.08 (1.89-2.29)	a)	
Other (non-small cell)	225/0.0228	750/0.0083	2.37 (2.02-2.76)	b)	
	30/0.0129	166/0.0068	1.54 (1.30-1.83)	c)	61%
	618/ 0.209	2685/ 0.087	2.03 (1.83-2.24)	a)	
Other	726/0.0304	1455/0.0073	3.53 (3.21-3.87)	b)	
	134/0.0268	340/0.0072	2.29 (2.07-2.54)	c)	49%
	2100/ 0.176	11087/ 0.091	1.63 (1.55-1.72)	a)	
Adenocarcinoma	955/0.0205	3652/0.0085	2.15 (2.00-2.32)	b)	
	229/0.0193	928/0.0075	1.51 (1.40-1.64)	c)	56%

a) Numbers copied/computed from Table 3 in [11]. Odds-ratios are computed from nationwide Danish registries, adjusted for education and comorbidity. Number of cases and odds are given for the exposed and the unexposed.
b) Odds-ratios computed from nationwide Norwegian registries, adjusted for education and comorbidity.
c) Odds-ratios from b) calibrated by applying smoking information (pack-years) from Norwegian health surveys. The number of cases and the odds are for the health survey population only.
Last column: Risk reduction is defined as (1-(1-ORc)/(1-ORb))*100, where ORb and ORc are the odds-ratios in rows b) and c).

Figure captions

Figure 1. Comparisons of the simplified and the standard TSC method based on simulated full data sets with $N = 8000$, 32 000, and 128 000 observations. Five validation data sets were generated for each full data set, as random samples of 1%, 5%, 10%, 20%, and 40%, respectively, of the full data set. For each N, 500 groups of 100 full data sets with identical parameters were simulated. For each validation data set size only groups where all validation data sets had at least one exposed individual who experienced the outcome were included. The number of included groups are given at the top of each panel in column a). Each boxplot shows the median, the lower and upper quartiles, and the 5% and 95% quantiles of the relevant variables. Column a) The ratio λ/ν in equation 3. Each of the λ's and ν's was calculated from a group of 100 data sets with identical parameters (cf. Table 1). Column b) The distribution of the differences between $\tilde{\bar{\beta}}'$ and $\bar{\beta}$ (white boxes) and between $\tilde{\bar{\beta}}$ and $\bar{\beta}$ (grey boxes). Column c) The distribution of the estimates $\widehat{var}(\tilde{\bar{\beta}}')$ (white boxes) and $\widehat{var}(\tilde{\bar{\beta}})$ (grey boxes), as calculated from equation 4 and equation 6, respectively. Column d) The distribution of the ratios $\widehat{var}(\tilde{\bar{\beta}}')/s^2$ (white boxes) $\widehat{var}(\tilde{\bar{\beta}})/s^2$ (grey boxes), where s^2 is the empirical variance of the 100 estimates of $\tilde{\bar{\beta}}'$ and $\tilde{\bar{\beta}}$ with identical parameters, given by equation 7. Only validation data sets with positive variance estimates are included. Column e) The distribution of the ratio $k \times \widehat{var}(\tilde{\bar{\beta}})/s^2$, where $k = \exp(\{1 - \log_2(0.001N)/13\}/\%_{\text{vd}})$ and s^2 is the empirical variance of the 100 estimates of $\tilde{\bar{\beta}}$ with identical parameters, given by equation 7. Only validation data sets with positive variance estimates are included.

Figure 2. Results from application to real data. Exposure is insulin use in 2004, and outcome is death. Black bullet (top): 'Gold standard' estimate $\bar{\beta}$ from the full data set adjusted for all of the covariates, i.e. A=age, B=BMI, T=log(triglycerides), P=packyears (smoking), M=marital status, I=income, S=systolic blood pressure, Y=year of survey. Triangles: Error prone estimates $\bar{\gamma}$ from the full data set, with only the covariates indicated to the left adjusted for. Crosses: Corrected estimates $\tilde{\bar{\beta}}$ from equation 5, with all covariates available in the validation data set. Each cross represent the median of 500 randomly sampled 10% validation data sets. 95% confidence intervals are given for all estimates. For the corrected estimates they are computed from the 2.5% and 97.5% quantiles of the 500 variance estimates from equation 6. See Section 4.3 for details. For the gold standard and crude estimates they are computed from the standard errors of the effect estimates. The dotted and dashed vertical lines indicate the gold standard estimate with 95% confidence interval.

Figure 3. Results from application to real data. The middle panel is a graphical representation of Table 3. The lines in Table 3 are in the figure ordered after increasing $\bar{\gamma}$, and the numbers on the x-axis corresponds to the No.-column in Table 3. For each combination of exposure, outcome, and sex, $\tilde{\bar{\beta}}$ was computed from 500 randomly sampled 10% validation data sets. The upper panel shows $\bar{\gamma} - \bar{\beta}$ and $\tilde{\bar{\beta}} - \bar{\beta}$ when no covariates are adjusted for in the crude estimates, and in the lower panel Age, BMI, smoking (packyears) and log(triglycerides) are adjusted for. Circles and black lines indicate the median and the interquartile range of the 500 differences $\tilde{\bar{\beta}} - \bar{\beta}$, respectively.