Research Article

Statistics in Medicine

Received XXXX

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

Comparisons of Risk Prediction Methods Using Nested Case-Control Data

Agus Salim[a*], Bénédicte Delcoigne[b], Krystyn Villaflores[a], Woon-Puay Koh[c], Jian-Min Yuan[d], Rob M. van Dam[e], Marie Reilly[b]

Using both simulated and real datasets, we compared two approaches for estimating absolute risk from nested case-control (NCC) data and demonstrated the feasibility of using the NCC design for estimating absolute risk. In contrast to previously published results, we successfully demonstrated not only that data from a matched NCC study can be used to unbiasedly estimate absolute risk, but also that matched studies give better statistical efficiency and classify subjects into more appropriate risk categories. Our result has implications for studies that aim to develop or validate risk prediction models. In addition to the traditional full cohort study and case-cohort study, researchers designing these studies now have the option of performing a NCC study with huge potential savings in cost and resources. Detailed explanations on how to obtain the absolute risk estimates under the proposed approach are given. Copyright © 0000 John Wiley & Sons, Ltd.

Keywords: absolute risk, cost efficiency, prediction models, prognosis, risk calculator, study design

1. Introduction

The main objective of a model such as the Adult Treatment Panel III risk calculator [1] is to predict the absolute risk of a disease. The cohort study design has traditionally been used to build or compare risk prediction models [2,3]. Recently, the case-cohort design has been demonstrated to estimate absolute risk unbiasedly [4-6]. An alternative to the case-cohort design is the nested case-control (NCC) design but until recently, the incidence density sampling of the NCC design has been considered as a relative weakness, because the control selection is tied to the particular outcome of interest. It was thus thought that selected controls could not be re-used to analyze other outcomes. However, recent methodological advances have enabled the re-use of controls [7-13]. The feasibility of a NCC design to estimate absolute risk was only established in limited scenarios with no matching [4,14] or when the matching variable is not a confounder [15]. Data from a multiphase NCC study has been used to estimate absolute risk [15]. However, details of the published methodology for this risk estimation are very brief and explanations on how it can be implemented using standard statistical software are lacking. In this article, using simulation studies and a real dataset, we compare two methods for estimating absolute risk

[a] *Mathematics and Statistics, La Trobe University, Bundoora VIC3086 Australia*
[b] *Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden*
[c] *Duke-NUS Graduate Medical School, Singapore*
[d] *Department of Epidemiology, University of Pittsburgh, USA* [e] *Saw Swee Hock School of Public Health, National University of Singapore*
[*] *Correspondence to: Department of Mathematics and Statistics, La Trobe University, Bundoora VIC3086 Australia*

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

from NCC data. We examine the performance of the two methods under a design with and without confounder matching, discuss the relative merits of each method and explain how these methods can be implemented using standard statistical software.

2. Statistical Methods

In a typical nested case-control study (NCC), the sole interest is in estimating the hazard ratios for exposures of interest. These estimates are traditionally obtained by performing conditional logistic regression where each case-control set is treated as a stratum. However, when absolute risk is also of interest, we need to estimate the baseline hazard function. Since the ratio of cases to controls in a NCC study is much higher than the ratio in the underlying population, estimating baseline hazard directly from NCC data will result in overestimation of absolute risk. The two methods we compare here differ in the way they estimate the hazard ratios and the baseline hazard function, although both implicitly assume a proportional hazard model,

$$\lambda_i(t) = \lambda_0(t)e^{\beta x_i + \gamma z_i},$$

where $\lambda_0(t)$ is the baseline hazard function, x_i denotes the main exposures of interest, z_i denotes the exposures of secondary interest or confounders and β and γ the log hazard ratios associated with x_i and z_i respectively.

Our interest lies in estimating the absolute risk of developing the outcome by time t,

$$F_i(t) = 1 - \exp(-\Lambda_0(t))^{\exp(x_i\beta + z_i\gamma)} \tag{1}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is the cumulative baseline hazard.

We compare two approaches for estimating absolute risk, the Langholz-Borgan (L-B) method [14] and the weighted method that combines a weighted partial likelihood approach [16] with weighted baseline hazard estimates [17]. The two methods are similar in that they both use sampling weights to adjust the (likelihood) contribution from controls, although the weight formulae are specific to each approach. The methods also differ in the way they handle subjects selected more than once as controls: with the L-B method, all such subjects are kept in their separate sets and analyzed as if they are distinct individuals; in the weighted approach, only unique individuals are pooled for analysis.

2.1. NCC sampling and control selection

NCC studies use incidence density sampling for selecting controls. For each incident case, a number of controls is selected from the current riskset, which is defined as the set of all subjects still to experience onset, or for a study with confounder-matching, the set of all subjects with the same confounder values as the incident case who still have to experience onset. Since a control can appear in multiple risksets and the selection of controls for different incident cases is independent, controls can be selected more than once. Controls with longer follow-up will be available for selection for a larger number of cases, and so have higher overall probability of ever being selected. Hence, the probability of selection depend on variables available for all cohort members (time of entry, censoring times and censoring status) and in the case of confounder matching, it may also depend on the matching variables z assumed to be known for all cohort members. However, the probability of selection cannot depend on the expensive exposure x observed only for subjects selected into the study. This missing-at-random (MAR) assumption is required to ensure the validity of the inverse probability weighting (IPW) approach [7,11].

2.2. Langholz-Borgan method

The log hazard ratios are estimated by maximizing the partial likelihood,

$$L(\beta,\gamma)=\prod_i \frac{e^{x_i\beta+z_i\gamma}}{\sum_{j\in R_i} e^{x_j\beta+z_j\gamma}} \tag{2}$$

Given the log hazard ratio estimates $\hat{\beta}$ and $\hat{\gamma}$, the cumulative baseline hazard is estimated as

$$\hat{\Lambda}_0(t)=\sum_{i=1}^{n_c} \frac{I_{[t_i\le t]}}{\sum_{j\in R_i} w(t_i)\, e^{x_j\hat{\beta}+z_j\hat{\gamma}}}, \tag{3}$$

where $I(.)$ is the indicator function, n_c is the number of events (cases), t_i is the event time for the i^{th} case, R_i is the set containing the i^{th} case and the associated controls and $w(t_i)$ is the weight given by:

$$\frac{M_i}{m+1}, \tag{4}$$

where M_i is the number of individuals who are still at risk just prior to time t_i, and m is the required number of controls per case. Note that the weight is simply the inverse of the proportion of subjects being selected at time t_i; given that at time t_i we include $m+1$ subjects (m controls, 1 case) into the study out of M_i subjects, the proportion is simply $\frac{m+1}{M_i}$. Note also that the weight is time-dependent and a control selected multiple times may have different weights to reflect the time-dependent nature of the riskset size.

2.2.1. Estimation using R The log hazard ratios can be estimated using the `coxph` function in the `survival` package with x and z as covariates and a variable containing an identifier for each case-control set added as `strata`. The `coxph` function takes a `Surv` object as the dependent variable and for this approach, the survival time is specified as the time of onset for cases while for controls it is the time of onset for their matched case, with `event=0` for controls and 1 for cases.

Using the `coxph` object containing the estimated log hazard ratios, we then use the `basehaz` function with `centered=FALSE` option to estimate the (unweighted) baseline hazard. To obtain the weighted baseline hazard, we multiply the unweighted baseline hazard by the sampling weight in (4). The cumulative baseline hazard at a particular timepoint is then calculated by summing the weighted baseline hazard up to (and including) that timepoint.

2.2.2. Matching on confounders When performing a NCC study, there is a widely-accepted practice of selecting only controls that match the case on potential confounders. When *exact matching* is performed, we lose the ability to estimate hazard ratios associated with the confounders since for each case-control set $z_i=z_j$, so the terms involving γ will cancel out in equation (2). This inability to estimate γ affects the ability to estimate the absolute risk correctly. To overcome this, Ganna et al. used a $\hat{\gamma}$ value of zero in the absolute risk calculation [4] but this 'ad-hoc' approach is clearly not satisfactory and they found the absolute risk to be incorrectly estimated.

In the presence of *exact matching*, we propose calculating the absolute risk using stratum-specific baseline hazards where the strata are defined by unique values of the confounders. Our approach is motivated by the fact that we can write the survival function as

$$\begin{aligned} S_i(t) &= 1-F_i(t) \\ &= \left[\exp -\Lambda_0(t)^{e^{z_i\gamma}}\right]^{\exp(x_i\beta)} \\ &= \left[\exp -\Lambda_{0,Z=z_i}(t)\right]^{\exp(x_i\beta)} \end{aligned}$$

where $\Lambda_{0,Z=z_i}(t)$ is the baseline hazard for the stratum with $z = z_i$.

In R, fitting stratum-level baseline hazards is achieved by adding the matching variable as `strata` instead of as an ordinary covariate. Note also that estimating stratum-specific baseline hazards for z does not require the proportional hazards assumption, so this approach should also work well when the effect of a matching variable cannot be modelled using the PH model.

The `basehaz` function with `centered=FALSE` option is used to estimate the (unweighted) baseline hazard. The weighted baseline hazard is obtained by multiplying by the inverse of the stratum-level sampling weights (see Section 2.2.3) and the stratum-level cumulative baseline hazard is calculated as the sum of all weighted baseline hazards for that stratum, up to (and including) the particular timepoint.

2.2.3. Sampling weight for matched study By making cases and controls more similar, matching generally creates selection bias. Even when the matching variable is not a confounder, statistical adjustment is needed for valid inference [18]. Under the L-B approach, this adjustment is reflected in the weight, $w(t_i)$. In a matched design, the riskset will only consist of those subjects still at risk who have the same matching covariate values as the incident case. At time t_i, the size of this riskset is $M_i^{z=z_i}$, where the superscript reflects the fact that the riskset contains only those subjects still at risk whose z values are the same as the incident case. The stratum-specific weight is then calculated simply as $w(t_i) = \frac{M_i^{z=z_i}}{m+1}$.

2.3. Weighted method

We note that in the L-B method, subjects that were selected more than once as controls, or selected controls who later became cases, are kept in their separate sets and analysed as if they are distinct individuals. Samuelsen [16] proposed an alternative method in which the matching is broken and unique individuals are pooled for analysis, keeping only one record for controls that appeared in multiple sets. If a selected control later becomes a case, the case record is kept for analysis. Denoting the pool of individuals for analysis as Ω, the log hazard ratios are estimated by maximizing the weighted partial likelihood,

$$L(\beta, \gamma) = \prod_i \frac{e^{\boldsymbol{x_i\beta + z_i\gamma}}}{\sum_{j\in\Omega_i} w_j \, e^{\boldsymbol{x_j\beta + z_j\gamma}}} \tag{5}$$

where Ω_i is the subset of individuals in the pool that were still at risk at time t_i, the ith onset time. Note that now the term involving γ will no longer disappear from the likelihood since subjects in the pool do not necessarily have the same values of z.

Given the log hazard ratio estimates, the cumulative baseline hazard can be estimated using the method suggested by Cai and Zheng [17],

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n_c} \frac{I_{[t_i \le t]}}{\sum_{j\in\Omega_i} w_j \, e^{\boldsymbol{x_j\hat{\beta} + z_j\hat{\gamma}}}}. \tag{6}$$

The weights, w_j, are calculated as the inverse of the probability of individual j being included in the NCC study. For a study without confounder-matching, this probability of inclusion (assuming that all cases are selected) is given by,

$$p_j = 1 - \prod_{i, s_j \le t_i \le e_j} \left[1 - \frac{m}{M_i - 1}\right] \left[1 - Y_j(t_i)\right], \tag{7}$$

where $Y_j(t_i)$ is the indicator of whether individual j has become a case by time t_i and the product is taken across all onset times when individual j is still in the cohort (between start time s_j and the end time e_j), and thus available to be selected either as a control or a case. The end time e_j is equal to the onset time for cases and for controls, it is their censoring time. Hence, the product is simply the probability of never being included in the study and the form of the product reflects

the fact that under incidence density sampling, the sampling at different times is independent and the probability of not being excluded at time t_i is $1 - \frac{m}{M_i - 1}$ if j is still to experience the event, i.e. $Y_j(t_i) = 0$, and is 0 if j is already a case, i.e $Y_j(t_i) = 1$. For a subject j that eventually becomes a case, regardless of how many times he/she was eligible to be selected as a control, the probability of inclusion will be 1 because the last product term in equation (7) at t_j is zero when the individual became a case. For a control that never became a case (hence $Y_j(t_i) = 0, \forall t_i$), the probability of inclusion can be simplified as

$$p_j = 1 - \prod_{i \in \Omega, s_j \leq t_i \leq e_j} \left[1 - \frac{m}{M_i - 1}\right], \tag{8}$$

2.3.1. Sampling weight for matched study For a study with confounder-matching, the probability of inclusion needs to be modified to reflect the fact that subjects are only eligible to be selected as controls if they have the same confounder values as the incident case. Salim et al. [8,9] proposed that the weight be calculated using the following modified formula to reflect the sampling process,

$$p_j = 1 - \prod_{i, s_j \leq t_i \leq e_j} \left[1 - \frac{m}{M_i^{z=z_i} - 1} I(z_i = z_j)\right] [1 - Y_j(t_i)], \tag{9}$$

where $I(z_i = z_j)$ is the indicator that individual j and the current incident case (i) have the same z values and $M_i^{z=z_i}$ is the number of subjects at risk at t_i with the values of matching factors $z = z_i$. Note that now the product term, which reflects the probability of being excluded (not selected) at time t_i, will be equal to 1 whenever the incident case does not have the same confounder value as individual j, reflecting the fact that j is not eligible for selection.

2.3.2. Estimation using R The log hazard ratios can be estimated using the `coxph` function in the `survival` package with x and z as covariates and using the variable containing the sampling weight in the `weight` argument. The standard errors of the estimates are calculated using the robust variance formula, achieved by specifying option `robust=TRUE` in the `coxph` function. The survival time in the `Surv` object is specified as time of onset for cases and time of the latest sampling for controls, i.e., time of onset for the latest matched case.

Using the `coxph` object containing the estimated log hazard ratios, we then use the `basehaz` function with `centered=FALSE` option to estimate the cumulative baseline hazard in equation (6). Note that, unlike in the L-B approach, since the sampling weight is already taken into account in the `coxph` object, there is no need to further multiply the hazards with sampling weights.

3. Simulation Studies

We performed simulations by generating data from a proportional hazards model, with the baseline hazard assumed to have a Weibull form. We first generated values for the following variables: gender and age (variables in $\boldsymbol{z_i}$), cholesterol, HDL, SBP, smoking status and antihypertensive treatment status (variables in $\boldsymbol{x_i}$) for 50,000 cohort members. These values (including those for categorical variables) were first generated using a multivariate normal distribution with the means and covariance matrix given in Supplementary Tables 1 and 2, as observed from a NCC study of coronary heart disease conducted within the Singapore Chinese Health Study (SCHS) [19]. The simulated values for age were then rounded to the nearest integer and for the variables gender, smoking status and antihypertensive treatment status, the simulated values were further categorized into binary values, so that the observed means were equal to the population means from the SCHS.

Given the simulated values of the risk factors above, we generated a time of onset for the disease, assuming the PH model with the following log hazard ratio parameters estimated from the SCHS data: $\gamma_{Age} = 0.05, \gamma_{Gender} = 0.47, \beta_{Chol} = 0.01, \beta_{HDL} = -0.02, \beta_{SBP} = 0.01, \beta_{Antihyp} = 0.29, \beta_{Smoke} = 0.54$). From the SCHS data, we also found that the estimates of baseline hazard can be well-approximated by a Weibull function, $\lambda_0(t) = \lambda\alpha t^{\alpha-1}$ where λ and α were set equal to their estimates from the SCHS data, namely 6.23×10^{-6} and 1.011, respectively.

We generated 500 realisations, each of size N = 50,000, with random censoring times generated from an exponential distribution with a rate of 0.05, giving an average length of follow-up of 20 years. Individuals with a censoring time occurring before the time of onset are censored, and a maximum censoring time was set to be t = 25 years, resulting in approximately 96% of cohort members being censored. For simplicity, we assumed that all individuals enter the study at time t = 0 so that the time of onset is measured as time (years) since the start of the study. Note that since their regression coefficients are non-zero, age and gender are associated with the onset time. Furthermore, age and gender are correlated to the other risk factors through the covariance matrix in Supplementary Table 2. Thus, age and gender can be regarded as confounders of the association of the other factors with the disease.

Hence, for each simulated cohort, we have the following information for every cohort member: time of entry or start time ($s_j = 0$ for all subjects), censoring/onset time (e_j), baseline age (Age), $Gender$ and the 'expensive' covariates of interest ($Chol$,HDL,SBP,$Antihyp$ and $Smoke$). The 'gold-standard' cohort analysis uses this full information to estimate the hazard ratios and baseline hazard function. For methods that use only NCC data (i.e., L-B and weighted), only time of entry, censoring/onset time, baseline age and gender are available for all cohort members, while the 'expensive' covariates are only available for those subjects selected into the NCC study.

Within each of the 500 cohorts we generated NCC studies where 2 randomly chosen controls were matched with each case. Four different NCC studies with respect to the confounder matching procedure were performed. In the first study, controls were not matched to cases, the second study performed gender-matching and the third study performed gender and age-group matching where controls needed to be of the same gender and within the same baseline age group as the case, where baseline age was split into 4 categories of similar sizes using quartiles. In the fourth study, we performed fine matching in which controls were required to be of the same age (in years) and gender as the case.

The log hazard ratio estimates and absolute risk curves obtained under the L-B and weighted methods were compared to the cohort estimates. For log hazard ratio estimates, the empirical standard errors (SE) were calculated as the standard deviation of the estimates across 500 realisations, while the estimated SEs were calculated as the average of standard errors of estimates across 500 realisations. For absolute risk estimates, only empirical SEs were calculated. To examine the viability of NCC as an alternative to the cohort design, we also assessed the discriminant quality of the estimated 10-year absolute risk by calculating the percentage of cases whose risk was underestimated and the percentage of non-cases whose risk was overestimated by the L-B and weighted estimates, relative to the cohort estimates. The bounds for all 95% confidence intervals involving these percentages were calculated as the 2.5th and 97.5th percentiles from the 500 different realisations. All computations were conducted using R version 3.0.0.

4. SIMULATION RESULTS

4.1. No confounder-matching

The L-B and weighted approaches had very similar hazard ratio estimates and both sets of estimates were close to the cohort estimates and the true values, albeit with predictably larger standard errors than the cohort estimates (Table 1). The empirical and estimated standard errors agreed well for all parameters for all approaches. Figure 1 shows the absolute risk estimates for females, and although not shown here, the estimates for males had very similar patterns. Both approaches yielded unbiased absolute risk estimates at various timepoints up to 20 years (Figure 1a), with very similar precision in terms of empirical standard errors (Figure 2a).

Relative to the cohort estimates, the L-B approach overestimated the risk for 1.14% (95% CI: 1.13% , 1.15%) of subjects who did not develop CHD for at least 10 years, and also underestimated the risk for 3.76% (95% CI: 3.74% , 3.78%) of subjects who developed CHD within the first 10 years. The performance of the weighted approach was very similar: overestimation of the risk for 1.49% (95% CI: 1.47% , 1.49%) of subjects who did not develop CHD for at least 10 years and underestimation of the risk for 3.44% (95% CI: 3.42% , 3.46%) of subjects who developed CHD within the first 10 years.

[TABLE 1 AROUND HERE]

[FIGURE 1 AROUND HERE]

4.2. *Matching on gender*

Under this scenario, the log hazard ratio (HR) for gender is not estimable by the L-B approach. For the other parameters, the L-B and weighted approaches had very similar hazard ratio estimates and both were close to the cohort estimates and the true values (Table 2). The empirical and estimated standard errors agree well for all parameters for all approaches. Both approaches also yielded unbiased absolute risk estimates at various timepoints up to 20 years (Figure 1b), with the weighted estimates being slightly more precise (Figure 2b). For the L-B approach, we also examined the importance of modifying the weights to cater for matching, by estimating the absolute risk using naive (unstratified) weights. As can be seen from Figure 1b, the absolute risk estimates derived using the naive weights are clearly biased. This result highlights the need to use stratified weights when estimating absolute risk from matched NCC data.

Relative to the cohort estimates, the L-B approach overestimated the risk for 2.76% (95% CI: 2.75% , 2.77%) of subjects who did not develop CHD for at least 10 years and underestimated the risk for 3.07% (95% CI: 3.05% , 3.09%) of subjects who developed CHD within the first 10 years. The weighted approach was slightly better, overestimating the risk for 1.08% (95% CI: 1.07% , 1.08%) of subjects who did not develop CHD for at least 10 years and underestimating the risk for 2.76% (95% CI: 2.74% , 2.78%) of subjects who developed CHD within the first 10 years.

[TABLE 2 AROUND HERE]

[FIGURE 2 AROUND HERE]

4.3. *Matching on gender and age-group*

Under this scenario, the log hazard ratios (HR) for gender is not estimable by the L-B approach. For the other parameters, the L-B and weighted approaches had very similar hazard ratio estimates and both sets of estimates were close to the cohort estimates and the true values (Table 3). The empirical and estimated standard errors agreed well for all parameters for all approaches. Both sets of estimates had very similar standard errors, except for age where the weighted estimate was much more precise. Both approaches yielded unbiased absolute risk estimates at various timepoints up to 20 years (Figure 1c), with the weighted approach being slightly more efficient (Figure 2c).

Relative to the cohort estimates, the L-B approach overestimated the risk for 4.78% (95% CI: 4.77% , 4.79%) of subjects who did not develop CHD for at least 10 years and also underestimated the risk for 8.46% (95% CI: 8.43% , 8.49%) of subjects who developed CHD within the first 10 years. The weighted approach was much better, where it only overestimated the risk for 0.92% (95% CI: 0.92% , 0.93%) of subjects who did not develop CHD in the first 10 years and only underestimated the risk for 1.29% (95% CI: 1.26% , 1.32%) of subjects who developed CHD within the first 10 years.

[TABLE 3 AROUND HERE]

4.4. *Fine matching*

Under this scenario, the log hazard ratios (HR) for gender and age are not estimable by the L-B approach. For the other parameters, the L-B and weighted approaches had very similar hazard ratio estimates and both sets of estimates were

close to the cohort estimates and to the true values (Table 4). The empirical and estimated standard errors agree well for all parameters for all approaches. Both approaches gave biased absolute risk estimates at various timepoints up to 20 years (Figure 1d), with the L-B approach exhibiting much greater bias (Figure 1d) and also larger standard errors, especially for long-term prediction beyond 15 years (see Figure 2d)

Relative to the cohort estimates, the L-B approach overestimated the risk for 5.99% (95% CI: 2.63% , 9.21%) of subjects who did not develop CHD for at least 10 years and also overestimated the risk for 15.11% (95% CI: 8.95% , 20.72%) of subjects who developed CHD within the first 10 years. The weighted approach was much better, where it only overestimated the risk for 1.87% (95% CI: -0.32% , 4.29%) of subjects who did not develop CHD in the first 10 years and only underestimated the risk for 1.52% (95% CI: -2.80% , 8.12%) of subjects who developed CHD within the first 10 years.

[TABLE 4 AROUND HERE]

5. Applications

We used a real cohort to evaluate the performance of the NCC design in predicting absolute risk of non-Hodgkins lymphoma (NHL) for siblings and children of NHL patients. The cohort that we used consists of 75856 siblings and children of all non-Hodgkin lymphoma (NHL) patients registered in the Swedish Cancer Register from 1958 to 2007. The time scale used is age and individuals were censored at emigration, death or the end of the study (31 December 2007), whichever occurred first. There were 293 cases during the follow-up period. We modeled the absolute risk to individuals (siblings and children of patients) as a function of four risk factors: year of birth of the individual, gender of the individual, gender of the patient and the type of relationship (sibling or child). We generated 100 nested case-control studies within the cohort, without matching and with gender matching, each with either two or five controls per case. The absolute risk was estimated using the L-B and weighted methods and compared to the cohort estimates. For the situation with no matching and with two controls per case, the absolute risk estimates (averaged over 100 realizations) at age 40, 50, 60 and 70 years, are shown in Figure 2. The weighted estimates are very close to the cohort estimates at all ages, while the L-B estimates are also quite close to the cohort estimates up to 60 years, but with somewhat noticeable bias beyond this point. In terms of precision, the weighted estimates are much more precise than the L-B estimates. The patterns of absolute risk estimates for other scenarios are qualitatively similar and are given in Supplementary Figures 1-3.

[FIGURE 3 AROUND HERE]

6. Discussion

Using simulated and real-life data, we have shown that nested case-control data can be used to obtain unbiased estimates of absolute risk, with the weighted method giving somewhat better precision when a matched design is used. When controls are not matched, the L-B and weighted approaches have very similar performance. When a combination of interval and exact matching on categorical variables is performed, the weighted approach outperforms the L-B approach in terms of discrimination quality and statistical efficiency of the variable involved in the interval matching (age in our illustration). We believe that the loss of efficiency in the L-B approach is due to the limited range of values for this variable as a consequence of interval matching. In contrast, the weighted approach can handle the interval matching well and the additional interval matching even improves the statistical efficiency slightly compared to exact matching. We also investigated the performance of the two approaches under fine matching. Both approaches estimate hazard ratios unbiasedly under fine matching. Stoer and Samuelsen reported bias in hazard ratio estimates for the weighted approach under fine matching [20]. We do not observe this bias because our fine matching (the same baseline age in years) was not

as strict as their fine-matching (blood collection date in the same week or month), and so our sampling strata were still relatively large compared to theirs. However, even with relatively large sampling strata we observed noticeable bias in the absolute risk estimates under the L-B approach, due to the unstable baseline hazard estimates in strata where there were only a few cases. On the other hand, the weighted approach does not suffer to the same extent as the baseline hazard is estimated using pool of all individuals selected into the NCC study. Although when the cohort size (and hence the NCC study size) is relatively small ($N = 1,000$), we have observed that weighted approach also yielded biased estimates of absolute risk (results not shown).

Our results have implications for designing studies that develop or validate risk prediction models. In addition to the traditional full cohort study and case-cohort study, researchers designing these studies now have the option of performing a NCC study with huge potential savings in cost and resources. Based on our findings here, we recommend that the weighted approach be used to obtain absolute risk estimates from NCC data. Interval and exact matching on categorical variable can be used to improve statistical efficiency, but we caution against the use of fine matching as it will tend to produce biased estimates of absolute risk, especially when the study size is small.

ACKNOWLEDGEMENT

This research was funded by National Health and Medical Research Council (NHMRC) Australia Grant No. 1108967 awarded to Agus Salim and National Medical Research Council (NMRC) Singapore Grant No. 1270/2010 awarded to Rob van Dam. Krystyn Villaflores received a vacation scholarship from the Australian Mathematical Sciences Institute (AMSI) that enabled her to work on this project and Bénédicte Delcoigne was partially funded by a grant from Cancerfonden (The Swedish Cancer Society), Contract No. 11 0343. The authors would like to thank Myeongjee Lee who prepared the data set for the real data application, and Nasheen Naidoo and Ye Sun for their help with data extractions.

REFERENCES

[1] National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation*. 2002; **106**:3143-3421.

[2] Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. British Medical Journal 2012; **24**, pp. e3318, DOI: 10.1136/bmj.e3318.

[3] Abbasi A, Peelen LM, Corpeleijn E, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. British Medical Journal 2012; **345**, pp. e5900, DOI: 10.1136/bmj.e5900.

[4] Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, Ingelsson E. Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. *American Journal of Epidemiology*. 2012; **175**:715-24.

[5] Sanderson J, Thompson SG, White IR, Aspelund T, Pennells L. Derivation and assessment of risk prediction models using case-cohort data. BMC Medical Research Methodolology 2013; **13**, pp. 113, DOI: 10.1186/1471-2288-13-113.

[6] Cook NR, Paynter NP, Eaton CB, Manson JE, Martin LW, Robinson JG, Rossouw JE, Wassertheil-Smoller S, Ridker PM. Comparison of the Framingham and Reynolds Risk scores for global cardiovascular risk prediction in the multiethnic Women's Health Initiative. *Circulation*. 2012; **125**:1748-1756.

[7] Saarela O, Kulathinal S, Arjas E, Läärä E. Nested case-control data utilized for multiple outcomes: a likelihood approach and alternatives. *Statistics in Medicine*. 2008; **27**:5991-6008.

[8] Salim A, Hultman C, Sparen P, Reilly M. Combining data from 2 nested case-control studies of overlapping cohorts to improve efficiency. *Biostatistics*. 2009; **10**:70-79.

[9] Salim A., Yang Q, and Reilly M. The value of reusing prior nested case-control data in new studies with different outcome. *Statistics in Medicine*. 2012; **31**:1291-1302.

[10] Støer NC, Samuelsen SO. Inverse probability weighting in nested case-control studies with additional matching-a simulation study. *Statistics in Medicine*. 2013; **32**: 53285339.

[11] Støer NC, Samuelsen SO. Comparison of estimators in nested case-control studies with multiple outcomes. *Lifetime Data Analysis*. 2012; **18(3)**, pp. 261283, DOI: 10.1007/s10985-012-9214-8.

[12] Salim A, Ma X, Li J, Reilly M. A maximum likelihood method for secondary analysis of nested case-control data. *Statistics in Medicine*. 2014; **33**:18421852.

[13] Støer NC, Meyer HE, Samuelsen SO. Reuse of Controls in Nested Case-Control Studies. *Epidemiology*. 2014; **25**:315-317.

[14] Langholz B, Borgan O. Estimation of absolute risk from nested case-control data. *Biometrics*. 1997; **53**:767-774.

[15] Zhou QM, Zheng Y, Cai T. Assessment of biomarkers for risk prediction with nested case-control studies. *Clinical Trials*. 2013; **10**:677-679.

[16] Samuelsen SO. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*. 1997; **84**:379-394.

[17] Cai T, Zheng Y. Evaluating prognostic accuracy of biomarkers under nested case-control studies. *Biostatistics*. 2012; **13**: 89-100.

[18] Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams. *International Journal of Epidemiology*. 2013; **42(3)**, pp. 860-869, DOI: 10.1093/ije/dyt083.

[19] Hankin JH, Stram DO, Arakawa K, et al. Singapore Chinese Health Study: development, validation, and calibration of the quantitative food frequency questionnaire. *Nutrition and Cancer*. 2001; **39**:187-195.

[20] Støer NC, Samuelsen SO. Inverse probability weighting in nested case-control studies with additional matching–a simulation study. *Statistics in Medicine*. 2013; **32**:5328-5339.

Figure legends

Figure 1: The average of absolute risk estimates for females across 500 realisations and their 95% confidence intervals. Cohort estimates (solid lines), L-B estimates (dashed lines) and weighted estimates (dotted lines).

Figure 2: The empirical standard errors of absolute risk estimates for females across 500 different realisations. Cohort estimates (solid lines), L-B estimates (dashed lines) and weighted estimates (dotted lines).

Figure 3: The absolute risk of NHL at age 40, 50, 60 and 70 years obtained using the NCC design without matching and two controls per case for (a) female children of patients (b) female siblings of patients, (c) male children of patients and d) male siblings of patients. Estimates from the L-B method are dotted lines and estimates from the weighted method are dashed lines. The vertical lines indicate the 95% confidence intervals. The computed risk from the whole cohort (denoted as X) is included for comparison.

Table 1. Log hazard ratio estimates and their standard errors for NCC studies without matching. Estimates are averages across 500 simulated cohorts.

	L-B Approach			Weighted Approach			Cohort			True
	Est	emp.SE	Est.SE	Est	emp.SE	Est.SE	Est	emp.SE	Est.SE	Value
Age	0.055	0.005	0.005	0.055	0.005	0.005	0.054	0.004	0.004	0.054
Gender	0.468	0.08	0.079	0.469	0.079	0.076	0.467	0.062	0.062	0.469
Chol	0.006	0.001	0.001	0.006	0.001	0.001	0.006	0.001	0.001	0.006
HDL	-0.023	0.003	0.003	-0.023	0.003	0.003	-0.023	0.002	0.003	-0.023
SBP	0.013	0.002	0.002	0.013	0.002	0.002	0.013	0.001	0.001	0.013
Treat	0.281	0.086	0.084	0.28	0.086	0.081	0.281	0.065	0.063	0.286
Smoke	0.541	0.09	0.091	0.543	0.089	0.091	0.54	0.064	0.062	0.537

Table 2. Log hazard ratio estimates and their standard errors for NCC studies with gender-matching. Estimates are averages across 500 simulated cohorts.

	L-B Approach			Weighted Approach			Cohort			True
	Est	emp.SE	Est.SE	Est	emp.SE	Est.SE	Est	emp.SE	Est.SE	Value
Age	0.055	0.005	0.006	0.055	0.005	0.006	0.054	0.004	0.004	0.054
Gender		N.A		0.465	0.081	0.071	0.469	0.062	0.063	0.469
Chol	0.006	0.001	0.001	0.006	0.001	0.001	0.006	0.001	0.001	0.006
HDL	-0.023	0.003	0.003	-0.023	0.003	0.003	-0.023	0.003	0.002	-0.023
SBP	0.013	0.002	0.002	0.013	0.002	0.002	0.013	0.001	0.001	0.013
Treat	0.291	0.085	0.083	0.291	0.085	0.081	0.286	0.065	0.065	0.286
Smoke	0.542	0.088	0.093	0.541	0.087	0.09	0.538	0.064	0.067	0.537

Table 3. Log hazard ratio estimates and their standard errors for NCC studies with age group and gender matching. Estimates are averages across 500 simulated cohorts.

	L-B Approach			Weighted Approach			Cohort			True
	Est	emp.SE	Est.SE	Est	emp.SE	Est.SE	Est	emp.SE	Est.SE	Value
Age	0.054	0.012	0.012	0.054	0.005	0.005	0.054	0.004	0.004	0.054
Gender		N.A		0.463	0.078	0.067	0.464	0.061	0.061	0.469
Chol	0.006	0.001	0.001	0.006	0.001	0.001	0.006	0.001	0.001	0.006
HDL	-0.023	0.003	0.003	-0.023	0.003	0.003	-0.023	0.002	0.002	-0.023
SBP	0.013	0.002	0.002	0.013	0.002	0.002	0.013	0.001	0.001	0.013
Treat	0.289	0.083	0.085	0.29	0.083	0.083	0.288	0.065	0.065	0.286
Smoke	0.541	0.085	0.086	0.543	0.084	0.086	0.538	0.064	0.066	0.537

Prepared using simauth.cls

Table 4. Log hazard ratio estimates and their standard errors for NCC studies with fine age and gender matching. Estimates are averages across 500 simulated cohorts.

	L-B Approach			Weighted Approach			Cohort			True
	Est	emp.SE	Est.SE	Est	emp.SE	Est.SE	Est	emp.SE	Est.SE	Value
Age		N.A		0.054	0.005	0.004	0.054	0.004	0.004	0.054
Gender		N.A		0.462	0.079	0.071	0.469	0.062	0.063	0.469
Chol	0.006	0.001	0.001	0.006	0.001	0.001	0.006	0.001	0.001	0.006
HDL	-0.023	0.003	0.003	-0.023	0.003	0.003	-0.023	0.002	0.002	-0.023
SBP	0.013	0.002	0.002	0.013	0.002	0.002	0.013	0.001	0.001	0.013
Treat	0.293	0.083	0.083	0.294	0.084	0.083	0.293	0.064	0.062	0.286
Smoke	0.539	0.085	0.083	0.541	0.085	0.08	0.536	0.064	0.059	0.537

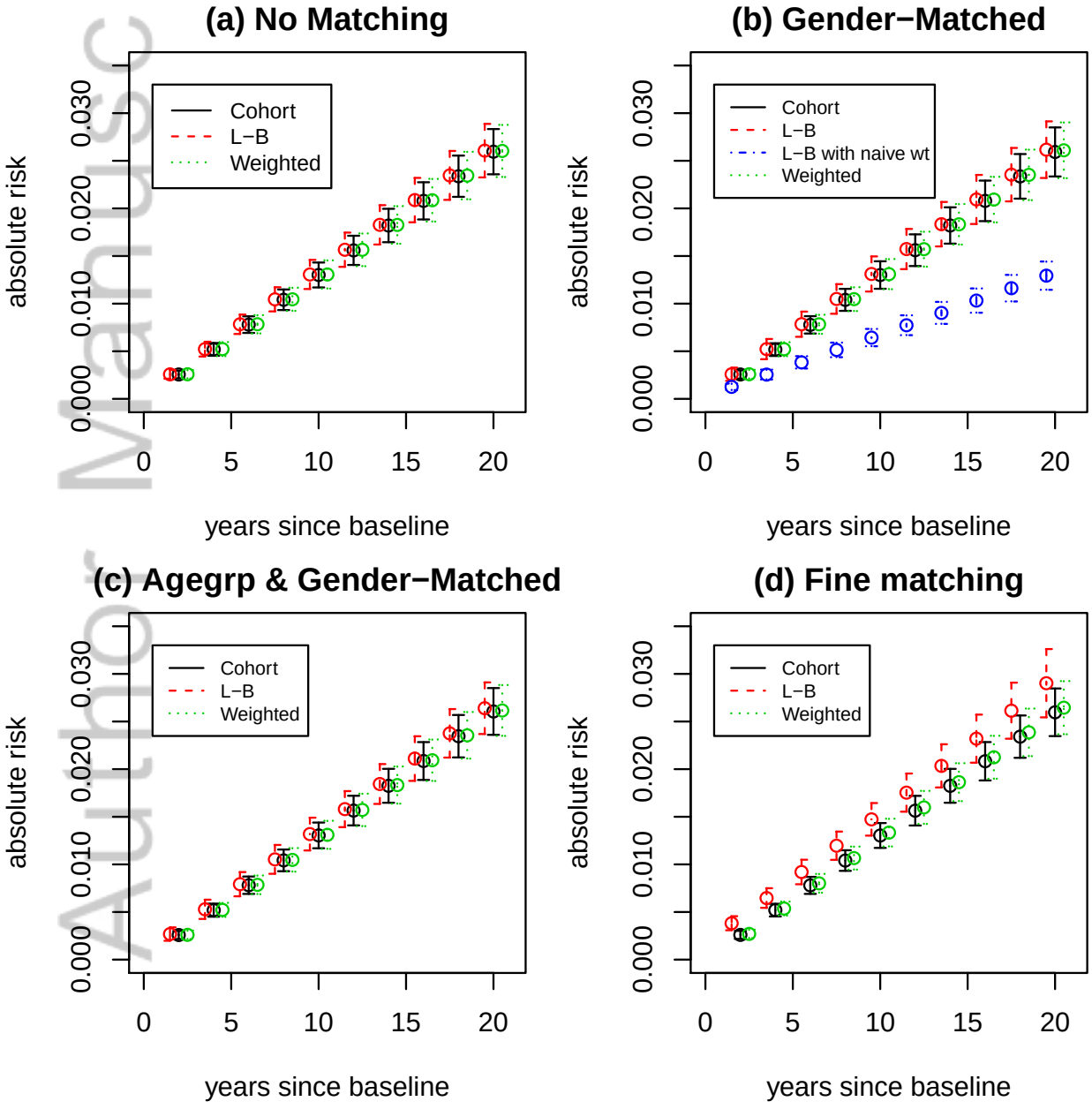

Figure 1

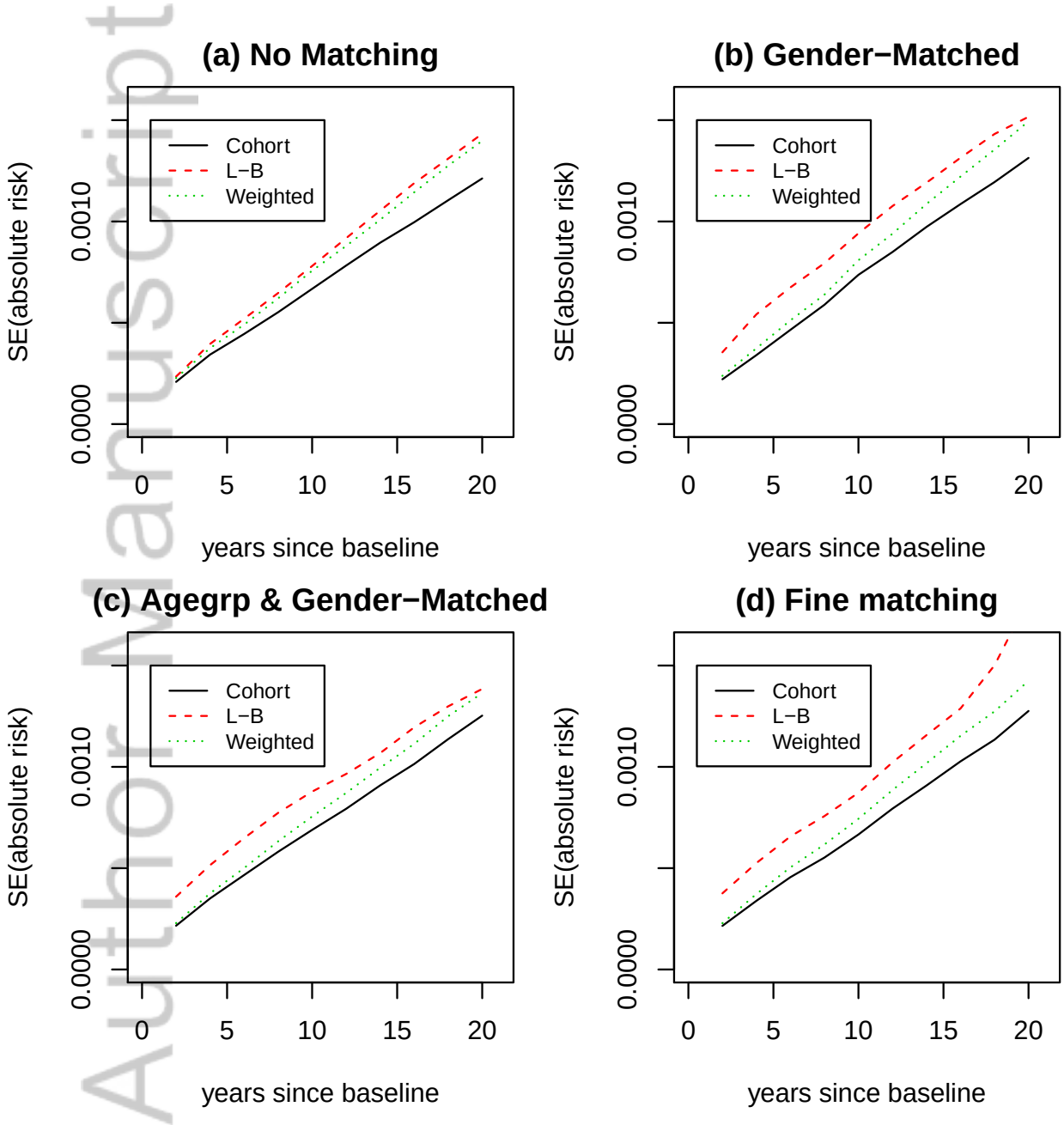

Figure 2

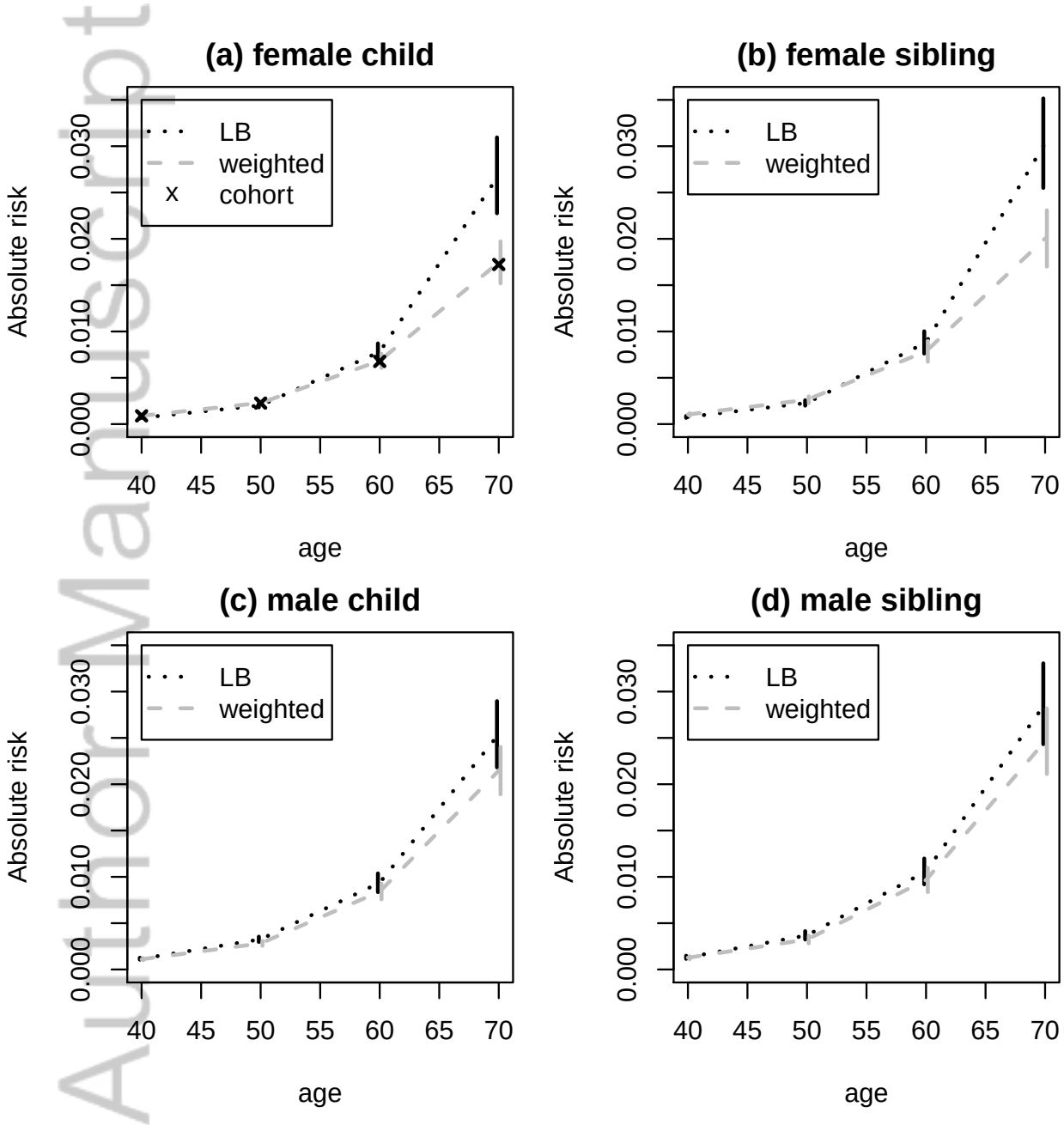

Figure 3