*Original Research Article*

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

# Privacy-preserving analysis of time-to-event data under nested case-control sampling

Statistical Methods in Medical Research
1–16

DOI: 10.1177/09622802231215804
journals.sagepub.com/home/smm


**Lamin Juwara[1,2], Yi Archer Yang[1,3], Ana M Velly[2,4] and Paramita Saha-Chaudhuri[5]**

**Abstract**

Analyses of distributed data networks of rare diseases are constrained by legitimate privacy and ethical concerns. Analytical centers (e.g. research institutions) are thus confronted with the challenging task of obtaining data from recruiting sites that are often unable or unwilling to share personal records of participants. For time-to-event data, recently popularized disclosure techniques with privacy guarantees (e.g. , etc.) are generally computationally expensive or inaccessible to applied researchers. To perform the widely used Cox proportional hazards regression, we propose an easy-to-implement privacy-preserving data analysis technique by pooling (i.e. aggregating) individual records of covariates at recruiting sites under the nested case-control sampling framework before sharing the pooled nested case-control subcohort. We show that the pooled hazard ratio estimators, under the pooled nested case-control subsamples from the contributing sites, are maximum likelihood estimators and provide consistent estimates of the individual level full cohort HRs. Furthermore, a sampling technique for generating pseudo-event times for individual subjects that constitute the pooled nested case-control subsamples is proposed. Our method is demonstrated using extensive simulations and analysis of the National Lung Screening Trial data. The utility of our proposed approach is compared to the gold standard (full cohort) and synthetic data generated using classification and regression trees. The proposed pooling technique performs to near-optimal levels comparable to full cohort analysis or synthetic data; the efficiency improves in rare event settings when more controls are matched on during nested case-control subcohort sampling.

**Keywords**

Survival analysis, data disclosure, privacy-preserving analysis, specimen pooling

## 1 Introduction

Large cohort biomarker studies of rare diseases or expensive-to-recruit studies often require merging datasets collected across multiple study centers or databases. Integrating individual data from contributing centers/nodes is a major challenge due to legitimate privacy and ethical concerns.[1] Analytical centers thus resort to disclosure control techniques such as intermediate statistics release or synthetic data generation to conduct etiologic studies or make predictions about relevant clinical endpoints.[2–5] Sampling[6] and noise perturbation[7] are also sometimes used.

[1]Quantitative Life Sciences, McGill University, Montreal, Canada
[2]Lady Davis Institute for Medical Research, Montreal, Quebec, Canada
[3]Department of Mathematics and Statistis, McGill University, Montreal, Quebec, Canada
[4]Department of Dentistry, McGill University, Montreal, Quebec, Canada
[5]Biogen Digital Health, Biogen Inc., Cambridge, MA, USA

**Corresponding author:**
Lamin Juwara, Quantitative Life Sciences, McGill University, Montreal, Canada; Lady Davis Institute for Medical Research, Montreal, Quebec, Canada.
Email: lamin.juwara@mail.mcgill.ca

In recent years, the strengths of proposed privacy-preserving techniques have improved considerably. Traditional attempts at inducing privacy, for example, k-anonymity, were primarily focused on masking quasi-identifiers using de-identification techniques such as generalization and suppression.[8,9] However, with recent advancements in computational power and our presence in multiple social network databases, the potential risk of re-identification is high.[10] More refined techniques like t-closeness promise better privacy guarantees but the strength of their effectiveness is largely dependent on the reliability of our assumptions about the intruder.[11] Techniques motivated from cryptography remain the strongest privacy guarantees; they often require different parties run known learning algorithms on a merged version of the local datasets without revealing individual data. Prediction from the learned model however requires the participation of each unit to implement a private scoring algorithm. Consequently, a major challenge of such techniques is the need for high communication and computational cost.[12]

Although not motivated by data privacy concerns, sample aggregation (referred to as pooling henceforth) was initially proposed as a method to circumvent the challenge of testing individual samples in a limited resource setting during World War II.[13] The method allowed practitioners to combine samples from multiple individuals for a single test in order to reduce the cost of resources needed to test for syphilis. The technique has seen some advancements in recent years, especially for identifying infectious agents in low prevalence settings[14] and more recently for testing Covid-19.[15] Alongside these developments, many variations of pooling have been proposed to infer associations between various clinical outcomes of interest and relevant covariates.[16–18] These techniques have become particularly appealing for the analysis of high-dimensional data where the strategy is utilized to compensate for limited samples or high biological variation.[19] Pooling is also sometimes used in the literature to mean combining records from various contributing data sites; however, we restrict our use of the term to mean aggregating individual-level covariate.

Of interest in the current article is the application of pooling to preserve patient privacy during the analysis of time-to-event (or survival time) data in a distributed data network. Survival analysis plays a fundamental role in biomedical research, where survival probabilities are routinely computed to inform clinical decisions.[20,21] Several approaches have been proposed for protecting patient privacy when dealing with survival time data.[22–25] For example, O'Keefe and Rubin[22] introduced a privacy technique based on data suppression (e.g. removal of censored events), smoothing, and data perturbation. Yu et al.[23] also proposed a method based on affine projections of the Cox model. However, these techniques have been challenging to adapt in practice for several reasons: (i) require sustained communication between the analytical center and data node to train the model, for example, distributed regression techniques, (ii) high computational cost to generate synthetic data at the study node, for example, classification and regression trees (CART) synthetic microdata generation,[4] PATE-GAN,[26] Differentially Private Generative Adversarial Network (DPGAN),[27] etc., and (iii) inaccessible to applied researchers without formal statistical training due to intricacies associated with sampling synthetic data from posterior predictive distributions. While pooling has been adapted to preserve patient privacy when analyzing binary outcome data in multi-center studies,[28] the framework has not been extended to survival-time data under privacy restrictions.

The objectives of the current study include (i) to propose an aggregation based, easy-to-implement, study design for time-to-event outcome under disclosure limitation, (ii) to estimate the hazard ratio of the postulated Cox proportional hazards (PH) model using the aggregate data from the proposed pooling design, and (iii) to estimate the full-cohort survival curves based on the subset of patients that were included in the pooling design. The rest of the manuscript is organized as follows. In Section 2, we first introduce the pooled study design for time-to-event outcome based on a nested case-control subset of the full cohort. We then introduce the full cohort (individual) and pooled subcohort Cox partial likelihood functions under nested case-control sampling. We show that the individual and pooled likelihoods are of the same form and inference could be carried out using existing analytic tools designed for individual likelihood functions. A sampling technique for generating the survival curve based on the subjects included in the pooled design is proposed, and compared to the full cohort and synthetic data. Section 3 discusses various performance assessment metrics relevant for our setting. We present simulations and a real-life example in Section 4 and close with a discussion in Section 5.

## 2 Methods

Consider a clinical study involving sensitive biomedical data, such as the time-to-death records of HIV/TB co-infected patients, distributed among multiple study centers. Assume that, due to confidentiality agreements and legitimate privacy concerns, the centers are unable to share individual data with analytical centers (see Figure 1 for a schematic representation of the distributed data setting). Our proposed design, based on nested case-control (NCC) subcohort sampling, provides an alternative approach to estimate the hazard ratio (HR) or overall survival curves of time-to-event data while limiting the disclosure of sensitive information from individual-level data. The approach is executed in three stages:

1. NCC design stage: NCC sampling of participants conditional on the event times,

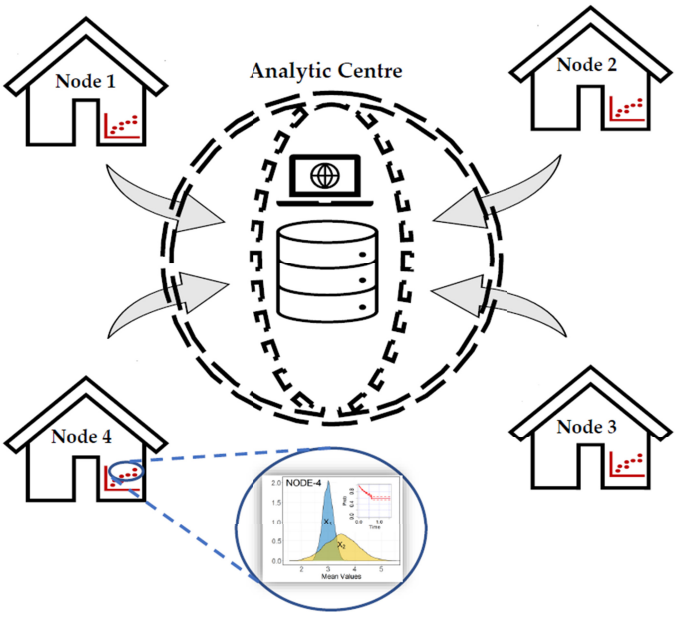


**Figure 1.** Problem setting—horizontally partitioned time-to-event data between study nodes 1–4. The nodes are only allowed to share aggregate/pooled data with the analytic center. Individual data cannot be shared between nodes or with the center.

2. Pooling design stage: aggregation/pooling of sampled subcohorts based on event status prior to data sharing, and
3. Estimation stage: estimation of model parameters from the pooled subcohorts and reconstructing approximate event times for pooled cases and controls that constitute the matched sets to estimate overall survival curves.

Steps (1) and (2) are sequentially conducted at each contributing data nodes before the final step (3) is performed at the central analyses node.

## 2.1 Cox PH model

*Notation and methods*

Let $N$ denote the number of participants in a cohort study at risk at baseline $t = 0$. We denote the time of failure (e.g. disease onset) and time of censoring of subject $i$ by $T_i$ and $W_i$, respectively, for $i = 1, 2, \ldots, N$. Define the censoring indicator $\delta_i = \mathbf{1}(T_i \leq W_i)$ and the observed follow-up time $Y_i = \min(T_i, W_i)$ for the $i$-th participant. Assume that $T_i$ and $W_i$ are conditionally independent given the collection of sensitive covariates $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})^T$. The Cox PH model for the covariate-outcome association is given by:

$$\lambda(t|Z_i) = \lambda_0(t)\exp(\beta_1 Z_{i1} + \cdots + \beta_p Z_{ip}) \equiv \lambda_0(t)\exp(\boldsymbol{\beta}^T Z_i) \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, the vector of unknown regression coefficients, are the log HRs characterizing the covariate-outcome association. $\lambda_0(t)$ denotes the unspecified baseline hazard.

*Likelihood for individual-level data*

Let the at-risk indicator for subject $i$ be defined as $\mathfrak{R}_i(t) = \mathbf{1}(Y_i \geq t)$, the riskset by $\mathcal{R}(t) = \{i : Y_i \geq t\}$ and the set of individuals surviving past time $t$ as $\mathcal{R}(t+) = \{i : Y_i > t\} \subset \mathcal{R}(t)$. If we denote the number of subjects that experience an event at $t$ by $n_t$ and those surviving past $t$ by $m_t$, then $n_t + m_t$ subjects make up the riskset $\mathcal{R}(t)$ whilst $m_t$ subjects make up $\mathcal{R}(t+)$ at time $t$. For the rest of the article, we assume the probability of ties for the continuous survival outcome $Y_i$ is negligible (i.e. $n_t = 1$) such that $\mathcal{R}(t)$ comprises $m_t + 1$ subjects. The likelihood of experiencing an event $\delta_i = 1$ at the observed time $Y_i$ is $f_i(Y_i) = S_i(Y_i)\lambda(Y_i)$ while the contribution from the censored individuals is $S_i(Y_i)$. Cox (1972) argued

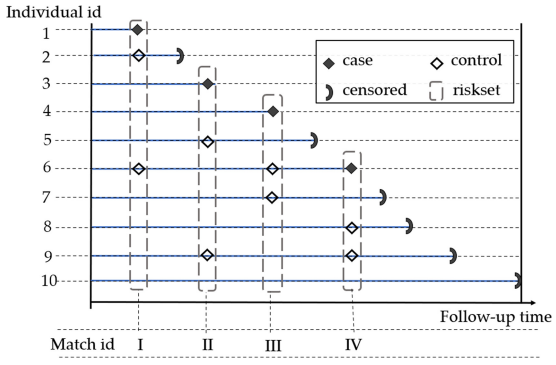


**Figure 2.** Nested case-control sampling example of two controls per case in a hypothetical cohort of 10 patients. At each event time, two controls are randomly selected in the riskset and matched to the case. Selected controls can become future cases or controls.

that the full likelihood $L_f(\boldsymbol{\beta}) = \prod_{i=1}^{N} \lambda(Y_i)^{\delta_i} S_i(Y_i)$ reduces to a partial likelihood under the PH assumption[29]:

$$L_f(\boldsymbol{\beta}) = \prod_{i=1}^{N} \left[ \frac{\exp\left(\boldsymbol{\beta}^T Z_i\right)}{\sum_{j \in \mathcal{R}(t_i)} \exp\left(\boldsymbol{\beta}^T Z_j\right)} \right]^{\delta_i} \tag{2}$$

where $\boldsymbol{\beta}$ is estimated by maximizing the partial likelihood. Parameter estimation and inference is readily available via standard statistical programs such as the `survival` package in R or the `PHREG` procedure in SAS.

### 2.1.1 NCC design stage

NCC subcohort sampling provides an attractive alternative to full cohort analysis, particularly when the analysis is constraint by the amount of available resources[30] or the amount of information that could be released.[31] This idea provides a key motivation for the proposed method where aggregate values of NCC subcohorts are utilized to preserve individual records during Cox regression modeling. NCC sampling is also sometimes referred to as riskset sampling or incidence density sampling. A hypothetical example is presented in Figure 2.

In NCC sampling, we embed a case-control study in a cohort such that for all individuals with the outcome of interest (cases), we randomly sample $\{m : m \geq 1\}$ controls for each case. In other words, $m$ controls are randomly sampled from the set $\mathcal{R}(t)$ at each event time for each case. Denote the set of controls selected from $\mathcal{R}(t+)$ by $C_i$ and the union of all controls by $C$, such that $C = \bigcup_{i:\delta_i=1} C_i$, then the index set of all the subjects included could be expressed as $T = \{i : \delta_i = 1\} \bigcup C$. The sampled riskset of subjects included in the NCC subcohort at each event time is $\tilde{\mathcal{R}}(t_i) = T \bigcap \mathcal{R}(t_i)$. Thus, the modified NCC subcohort partial likelihood is now expressed as

$$L_C(\boldsymbol{\beta}) = \prod_{i=1}^{N} \left[ \frac{\exp\left(\boldsymbol{\beta}^T Z_i\right)}{\sum_{j \in \tilde{\mathcal{R}}(t_i)} \exp\left(\boldsymbol{\beta}^T Z_j\right)} \right]^{\delta_i} \tag{3}$$

which preserves the properties of the full cohort likelihood. Samuelsen[32] was the first to give a proof for the consistency of the estimates obtained using the NCC partial likelihood in equation (3). As it would be shown later, equation (3) is of the same form as the conditional logistic likelihood for a $1 : m$ matched case-control subcohort.[33,31] Hence, model parameters and standard errors (SEs) could be estimated with standard statistical programs available for conditional logistic regression, for example, `clogit` function in the `survival` package.[34] The link between the two likelihoods is outlined in Section 2.1.3. Although the above likelihood is useful for estimating the log HRs under resource constraint settings and offers mild privacy protection through the random matching of select controls to cases, individual covariate records and event times still remain unprotected. The next step of the scheme thus involves matched-set pooling to mask individual records.

### 2.1.2 Pooling design stage: Pooling under NCC sampling

Aggregating individual-level data takes place at the contributing nodes following NCC sampling in stage 1. Upon reducing the Cox partial likelihood in (2) to the NCC subcohort likelihood given in equation (3), the nodes set out to follow the outcome dependent aggregation scheme described by Saha-Chaudhuri and Weinberg[18] and Saha-Chaudhuri and Juwara.[31]

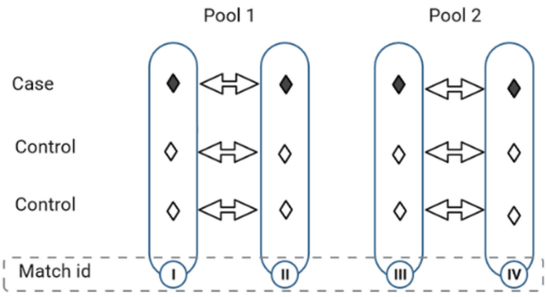


**Figure 3.** Pooling nested case-control matched sets of participants followed in Figure 2. Each pool is randomly formed from two matched sets by aggregating cases (with cases) and controls (with controls), independently.

**Table 1.** Pooled survival data created at each node under NCC subcohort sampling using the toy data of Figure 2. Information marked ⊘ is not retained during pooling. Only aggregate covariate values are shared with the analysis center. Riskset size ($|\mathcal{R}_i(t+)|$) is omitted from pooled data if the value is less than 5. The perturbation $\tau$ is sampled from $|N(0, (Y_{i+1} - Y_i)/2)|$. $Y_{\max}$ is the end of the study.

| Time $Y_i^{\text{pool}}$ | $\lvert\mathcal{R}_i(t+)\rvert$ | Status | $Z_{(e)}^{\text{pool}}$ | $Z_{(c)}^{\text{pool}}$ | Pool size $\kappa$ | Pool id |
|---|---|---|---|---|---|---|
| $\min(Y_1, Y_3) + \tau$ | 9 | 1 | $z_1^{\text{I}} + z_3^{\text{II}}$ | $z_1^{\text{I}} + z_3^{\text{II}}$ | 2 | 1 |
| ⊘ | ⊘ | 0 | $z_2^{\text{I}} + z_5^{\text{II}}$ | $z_2^{\text{I}} + z_5^{\text{II}}$ | 2 | 1 |
| ⊘ | ⊘ | 0 | $z_6^{\text{I}} + z_9^{\text{II}}$ | $z_6^{\text{I}} + z_9^{\text{II}}$ | 2 | 1 |
| $\min(Y_4, Y_6) + \tau$ | 6 | 1 | $z_4^{\text{III}} + z_6^{\text{IV}}$ | $z_4^{\text{III}} + z_6^{\text{IV}}$ | 2 | 2 |
| ⊘ | ⊘ | 0 | $z_6^{\text{III}} + z_8^{\text{IV}}$ | $z_6^{\text{III}} + z_8^{\text{IV}}$ | 2 | 2 |
| ⊘ | ⊘ | 0 | $z_7^{\text{III}} + z_9^{\text{IV}}$ | $z_7^{\text{III}} + z_9^{\text{IV}}$ | 2 | 2 |
| $Y_{\max}$ | ⊘ | ⊘ | ⊘ | ⊘ | ⊘ | ⊘ |

For simplicity, we assume individual records of the time-to-event data at each node are sensitive and comprise of individual records of event times, outcome status, exposure $Z_{(e)}$, and confounder $Z_{(c)}$. Figure 3 demonstrates pooled NCC subcohort sampling on the hypothetical cohort data provided in Figure 2. Covariate values of the matched sets (I vs. II and III vs. IV) are randomly aggregated via averaging while keeping the matching intact. For example, the cases and controls in matched sets I and II are independently pooled in order to maintain the outcome status of the pooled set (i.e. cases are pooled with cases and controls are pooled with controls).

A snippet of the node-level pooled subcohort that is anticipated for release at each node is presented in Table 1. In the example provided, the pooled exposure values of pool id 1 are obtained by adding the covariate values of cases in matched set I with cases in matched set II (i.e. $Z_{(e)}^{\text{pool}}$ for pooling individuals 1 and 3 = $z_1^{\text{I}} + z_3^{\text{II}}$) and controls with other controls in matched sets I and II, respectively (e.g. $Z_{(e)}^{\text{pool}}$ for individuals 2 and 5 = $z_2^{\text{I}} + z_5^{\text{II}}$). For the released event time $Y_i$ of the pooled cases, only the minimum perturbed event time of the cases pooled is released. The nodes do not release data for cells marked ⊘, including event times for cell where the riskset size is 5 or less.[35] A detailed description of disclosed aggregate data and the role they play in recovering some of the information lost to pooling is provided in Section 2.1.4.

The proposed pooling design can be extended to include pooled versions of additional confounders. Thus, if variables are available based on individual determinations, for example, by questionnaires, one could form set-based versions by summing the values across the individuals in the set. For categorical data, indicator variables of the categories could be used and pooled accordingly. Similarly, transformations of covariates (e.g. polynomial or log) and effect modifiers can be handled, but need more care.

### *2.1.3 Estimation stage*

Once the pooled NCC subcohorts have been created at the nodes, the next step is to share the aggregate values (as shown in Table 1) with the analytical center where they are combined into a single data file for analyses. We mainly focus on HR estimation under the postulated Cox model and estimation of overall survival curves.

*HR estimation with pooled NCC likelihood*

We now describe the likelihood contribution of individuals pooled at a single node. Consider a $1:m$ matched case-control sets at any arbitrary node following NCC subcohort sampling, along with measurements of a single exposure and confounder. Denote the outcome event indicators of the $i$-th matched set by $(\delta_i^1, \delta_i^2, \ldots, \delta_i^{m+1})$ and the random variable version by $(D_i^1, D_i^2, \ldots, D_i^{m+1})$, the observed survival times by $(Y_i^1, Y_i^2, \ldots, Y_i^{m+1})$ with $Y_i^j = Y_i$ for all $j$ representing the event time of the matched set, the exposure variable by $(Z_{i(e)}^1, Z_{i(e)}^2, \ldots, Z_{i(e)}^{m+1})$, and the confounder by $(Z_{i(c)}^1, Z_{i(c)}^2, \ldots, Z_{i(c)}^{m+1})$ for $i = 1, 2, \ldots, n$. The time matched samples are now randomly aggregated based on event status. For an arbitrary pool size $\kappa$ and $n$ matched sets, the pooled data generated is comprised of $n/\kappa$ matched samples of aggregate covariates and outcome events. The likelihood contribution of the pooled matched set $i' \in \{1, 2, \ldots, n/\kappa\}$ is expressed as

$$
\begin{aligned}
&\Pr\Big(D_{i'}^1 = 1 \Big| \sum_{j=1}^{m+1} \delta_{i'}^j = 1, \{Y_{i'}^j\}_{j=1}^{m+1}, \{Z_{i'(e)}^j\}_{j=1}^{m+1}, \{Z_{i'(c)}^j\}_{j=1}^{m+1}, \theta\Big) \\
&= \frac{\Pr\left(D_{i'}^1 = 1, D_{i'}^2 = 0, \ldots, D_{i'}^{m+1} = 0, \{Y_{i'}^j\}_{j=1}^{m+1} \middle| \{Z_{i'(e)}^j\}_{j=1}^{m+1}, \{Z_{i'(c)}^j\}_{j=1}^{m+1}, \theta\right)}{\Pr\left(D_{i'}^1 + D_{i'}^2 + \cdots + D_{i'}^{m+1} = 1, \{Y_{i'}^j\}_{j=1}^{m+1} \middle| \{Z_{i'(e)}^j\}_{j=1}^{m+1}, \{Z_{i'(c)}^j\}_{j=1}^{m+1}, \theta\right)} \\
&= \frac{\Pr\left(D_{i'}^1 = 1, D_{i'}^2 = 0, \ldots, D_{i'}^{m+1} = 0, \{Y_{i'}^j\}_{j=1}^{m+1} \middle| \{Z_{i'(e)}^j\}_{j=1}^{m+1}, \{Z_{i'(c)}^j\}_{j=1}^{m+1}, \theta\right)}{\sum_{\mathcal{D}} \Pr\left(D_{i'}^1 = \delta_{i'}^1, D_{i'}^2 = \delta_{i'}^2, \cdots + D_{i'}^{m+1} = \delta_{i'}^{m+1}, \{Y_{i'}^j\}_{j=1}^{m+1} \middle| \{Z_{i'(e)}^j\}_{j=1}^{m+1}, \{Z_{i'(c)}^j\}_{j=1}^{m+1}, \theta\right)} \\
&= \frac{\Pr\left(D_{i'}^1 = 1, Y_{i'}^1 \middle| Z_{i'(e)}^1, Z_{i'(c)}^1, \theta\right) \times \cdots \times \Pr\left(D_{i'}^{m+1} = 0, Y_{i'}^{m+1} \middle| Z_{i'(e)}^{m+1}, Z_{i'(c)}^{m+1}, \theta\right)}{\sum_{\mathcal{D}} \Pr\left(D_{i'}^1 = \delta_{i'}^1, Y_{i'}^1 \middle| Z_{i'(e)}^1, Z_{i'(c)}^1, \theta\right) \times \cdots \times \Pr\left(D_{i'}^{(m+)} = \delta_{i'}^{m+1}, Y_{i'}^{m+1} \middle| Z_{i'(e)}^{m+1}, Z_{i'(c)}^{m+1}, \theta\right)}
\end{aligned}
$$

where $Y_{i'}^j$ represents the minimum observed time of the cases or controls constituting that pool set, $Z_{i'(e)}^j = \kappa^{-1} \sum_{i=1}^{\kappa} Z_{i(e)}^j$, and $Z_{i'(c)}^j = \kappa^{-1} \sum_{i=1}^{\kappa} Z_{i(c)}^j$ for all $j$ ($j = 1, \ldots, m+1$).

$\mathcal{D}$ represents the set:

$$
(\delta_{i'}^1, \delta_{i'}^2, \ldots, \delta_{i'}^{m+1}) \in \{(1^{(1)}, 0^{(2)}, \ldots, 0^{(m+1)}), (0^{(1)}, 1^{(2)}, \ldots, 0^{(m+1)}), \ldots, (0^{(1)}, 0^{(2)}, \ldots, 1^{(m+1)})\}
$$

Assuming a PH model for all $j : \delta_{i'}^j = 0$, the probabilities give

$$
\Pr\left(D_{i'}^j = 0, Y_{i'}^j \middle| Z_{i'(e)}^j, Z_{i'j}^{(c)}, \theta\right) \overset{\theta \sim (\lambda_{0i'}, \beta_1, \beta_2)}{\propto} \underbrace{\exp\left(-\exp\left(\beta_1 Z_{i'(e)}^j + \beta_2 Z_{i'(c)}^j\right) \int_0^{Y_{i'}^j} \lambda_{0i'}(\tau) d\tau\right)}_{Q_{i'}^j(Y_{i'})}
$$

For all $j : \delta_{i'}^j = 1$,

$$
\Pr\left(D_{i'}^j = 1, Y_{i'}^j \middle| Z_{i'(e)}^j, Z_{i'(c)}^j, \theta\right) \overset{(\lambda_{0i'}, \beta_1, \beta_2)}{\propto} \lambda_{0i'}(Y_{i'}^j) \exp\left(\beta_1 Z_{i'(e)}^j + \beta_2 Z_{i'(c)}^j\right) \times Q_{i'}^j(Y_{i'})
$$

where $\lambda_{0i'}(Y_{i'}^j)$ denotes the baseline hazard specific to the $i'$-th observation of the matched subcohort for $j : \delta_{i'}^j = 1$ and $Q_{i'}^j(Y_{i'}) = \exp(-\exp(\beta_1 Z_{i'(e)}^j + \beta_2 Z_{i'(c)}^j) \int_0^{Y_{i'}^j} \lambda_{0i'}(\tau) d\tau)$ represent the survival terms common to the case and control matched

sets. We can thus rewrite the likelihood contribution to the pooled matched set as

$$\begin{aligned}
&\Pr\left(D^1_{i'}=1\Big|\sum_{j=1}^{m+1}\delta^j_{i'}=1,\{Y^j_{i'}\}_{j=1}^{m+1},\{Z^j_{i'(e)}\}_{j=1}^{m+1},\{Z^j_{i'(c)}\}_{j=1}^{m+1},\theta\right)\\
&=\frac{\lambda_{0i'}(Y^1_{i'})\exp\left(\beta_1 Z^1_{i'(e)}+\beta_2 Z^1_{i'(c)}\right)Q^1_{i'}\times Q^2_{i'}\times\cdots\times Q^{m+1}_{i'}}{\sum_D\left(\lambda_{0i'}(Y^1_{i'})\exp(\beta_1 Z^1_{i'(e)}+\beta_2 Z^1_{i'(c)})\right)^{\delta^1_{i'}}Q^1_{i'}\times\cdots\times\left(\lambda_{0i'}(Y^{m+1}_{i'})\exp\left(\beta_1 Z^{m+1}_{i'(e)}+\beta_2 Z^{m+1}_{i'(c)}\right)\right)^{\delta^{m+1}_{i'}}Q^{m+1}_{i'}}\\
&=\frac{\exp\left(\beta_1 Z^1_{i'(e)}+\beta_2 Z^1_{i'(c)}\right)}{\exp\left(\beta_1 Z^1_{i'(e)}+\beta_2 Z^1_{i'(c)}\right)+\exp\left(\beta_1 Z^2_{i'(e)}+\beta_2 Z^2_{i'(c)}\right)+\cdots+\exp\left(\beta_1 Z^{m+1}_{i'(e)}+\beta_2 Z^{m+1}_{i'(c)}\right)}
\end{aligned}$$

The product of the likelihood contribution $\Pr(D^1_{i'}=1|\cdot)$ over all the pooled matched sets $i'\in\{1,2,\ldots,n/\kappa\}$ gives the expression

$$L_{\text{Pooled}}(\boldsymbol{\beta})\;=\;\prod_{i'=1}^{n/\kappa}\left[\frac{\exp\left(\beta_1 Z^1_{i'(e)}+\beta_2 Z^1_{i'(c)}\right)}{\sum_{j'\in\mathcal{R}(t_{i'})}\exp\left(\beta_1 Z^{j'}_{i'(e)}+\beta_2 Z^{j'}_{i'(c)}\right)}\right]^{\delta_{i'}} \tag{4}$$

which results in the same likelihood form as the NCC subcohort likelihood in equation (3) with the same regression parameters. Thus, predictions and inference could be conducted using the pooled data instead of individual level data. The consistency of the pooled logistic likelihood in estimating the parameters of individual level likelihood has been shown by Saha-Chaudhuriet al.[36] Our derivation closely follows the conditional logistic likelihood derivation of Clayton and Hills[37] and Langholz and Goldstein[30] and Langholz and Clayton.[38]

Utilizing the well-established equivalence between likelihood of the NCC subcohort and the likelihood of conditional logistic regression,[39,36] inference on the MLEs could be carried out by using readily available packages for conditional logistic regression or stratified Cox regression.[34] The estimated parameters are interpreted as log HRs rather than the traditional log odds ratios derived from conditional logistic likelihood. Moreover, standard inference techniques applicable to conditional logistic likelihood can be employed to estimate the SEs of the pooled subcohort estimators. Of note, the units of analysis for the pooled NCC subcohort are the pools themselves, as opposed to individual measurements.

As mentioned earlier, the AC receives pooled NCC subcohorts (see Table 1) from each contributing node. This includes partial information on the observed event times of the pooled cases, the number of individuals making up each riskset at the node level if it is more than five individuals, the pool event status (1 if cases are pooled and 0 if controls are pooled), and their corresponding aggregate covariate values. While the log HRs associated with the covariates could be estimated using the pooled subcohorts without any need for the matched event times; to estimate the overall survival curves, the individual event times of subjects making up the pools would need to be recovered.

#### *2.1.4 Estimation of survival curves*

Suppose we want to reconstruct the overall survival curve $S(t)=\Pr(Y>t)$ ($t=0,1,\ldots$ are the event times) of the full cohort data using the pooled NCC subcohorts that are shared with the analytical center. When full cohort data is available, several estimators could be utilized for calculating the survival function. In biomedical research, the Kaplan–Meier (KM) estimator is usually the preferred method and is given by

$$\widehat{S}\Big(t\Big)=\prod_{i:Y_i\le t}\left(1-\frac{d_i}{n_i}\right) \tag{5}$$

where $Y_i$ represents the time at which one or more events happened, $d_i$ is the number of events (e.g. deaths) at $Y_i$, and $n_i$ is the number of individuals surviving past the event time. The KM estimator is easy to evaluate when individual-level data is accessible. However, with pooled NCC subcohorts under privacy restrictions, information about individual event time (and censoring time for the controls) is inaccessible to the analyst.

In order to accurately estimate the survival curve based on pooled NCC subcohorts, a procedure for regenerating the observed event/censoring times of the cases and controls making up the pooled matched sets is needed. To perform this reconstruction, the AC has access to the pooled survival times of cases, the size of the riskset at each disclosed event time, and the event status of the pools. Of note, only the minimum event time of the pooled cases from a given pooled matched

set is passed on to the AC. The survival time for the rest of the cases making up the pool and all of the controls would need to be re-generated to estimate the survival curves.

Let the pooled event times be ordered such that $Y_1^{\text{pool}} < Y_2^{\text{pool}} < \cdots < Y_\kappa^{\text{pool}} < Y_{\max}$, the observed survival time of the cases and controls that make up the pooled matched sets are derived using the following algorithm:

**Algorithm 1** : Reconstructing survival curves from pooled subcohort data.

**for** each of the pooled matched sets: **do**
  Sample an index case $Y_{i'} \sim Y_i^{\text{pool}}$ (i.e. sample a case at random among the pooled cases in that matched set and assign it the survival time of that pool).
**for** the remaining $\kappa - 1$ cases that made up the pooled matched set: **do**
  **if** $|\mathcal{R}_i(t+)| - |\mathcal{R}_{i+1}(t+)| > \kappa - 1$ **then**
    Sample $Y_{i'} \sim \text{Uniform}(Y_i^{\text{pool}} + \tau, Y_{i+1}^{\text{pool}})$
  **else**
    Sample the observed event time $Y_{i'} \sim \text{Uniform}(Y_i^{\text{pool}} + \tau, Y_{\max})$, where $\tau$ is a positive noise (e.g. absolute Gaussian) of the same unit and $Y_{\max}$ is the observed time at the end of study.
**for** the $\kappa \times m$ controls that make up the pooled matched set (see $\emptyset$ in Table 1): **do**
  Sample individual censoring times for the controls, $Y_{i'} \sim \text{Uniform}(Y_i^{\text{pool}} + \tau, Y_{\max})$.
Repeat for all pooled matched sets to generate the $\geq n$ survival outcomes.

These choices of sampling ensure the ranking of event status (of individuals in the full cohort) are appropriately reconstructed. In other words, the number of individuals that make up each riskset is correctly retrieved. Even though the sampling may not generate precise estimates of observed event times per se, it preserves their ranking and consequently the count of individuals in each riskset. Hence, the estimator in equation (5) could be employed to accurately estimate the survival risk of individuals given a reasonable time window.

Of note, the proposed reconstruction generates excess controls resulting from repeated sampling of controls during NCC sampling. Therefore, to ensure a fair comparison of survival curves generated from different pooled subcohorts, we recommend randomly sampling out the excess controls. For example, the pooled toy example of Figure 3 constitutes 12 data points whereas the original full cohort only included 10 individuals. Hence the reconstruction (of individual survival outcomes) will generate event times for two more controls than what we started with which would need to be randomly sampled out to ensure a fair comparison to full cohort data. We demonstrate the performance of the sampling technique for estimating absolute survival risk after time $t$ using simulations in Figure 4.

## 2.2 Alternative approaches for sharing microdata

In recent years, various synthetic data generation methods (e.g. CART synthesis, DP-GAN, etc.) have gained popularity as advancements over traditional anonymization techniques like suppression, aggregation, and noise perturbation.[3,4,26,27] These methods have emerged as effective approaches to facilitate the sharing of sensitive microdata while preserving data integrity. CART synthetic data generation, in particular, has attracted a lot of interest and is employed by various national agencies for data disclosure, for example, the Scottish Longitudinal Study Development and Support Unit (https://sls.lscs.ac.uk/guides-resources/synthetic-data/). We thus use it as a standard technique for generating synthetic data and compare its performance to the proposed method.

The CART algorithm partitions the predictor space into subsets, using unit partitions, in order to obtain subsets with consistent outcomes.[40,41] Homogeneity at each split is tracked in CART through the utilization of an impurity function. The best split is determined by thoroughly exploring all variables and split values, ultimately selecting the split that minimizes impurity the most. Examples of such functions include the entropy criterion and the gini index. The optimal size of the CART model is determined by a complexity measure that carefully balances the accuracy of the model and the size of the resulting tree. Initially, the tree is grown to its maximum size and subsequently pruned to refine its structure. We present a tree structure representation of the binary splitting of CART in the Supplemental Appendix.

CART is frequently used for imputing missing categorical variables; however, in the current setting, we are primarily interested in its adoption for synthesizing confidential data. The algorithm is employed to replace individuals at high risk of re-identification with records generated by sampling from the posterior predictive distribution of the outcome fitted with the CART model. In practice, we use the `synthpop` R package[42] to generate synthetic copies of the simulated datasets.

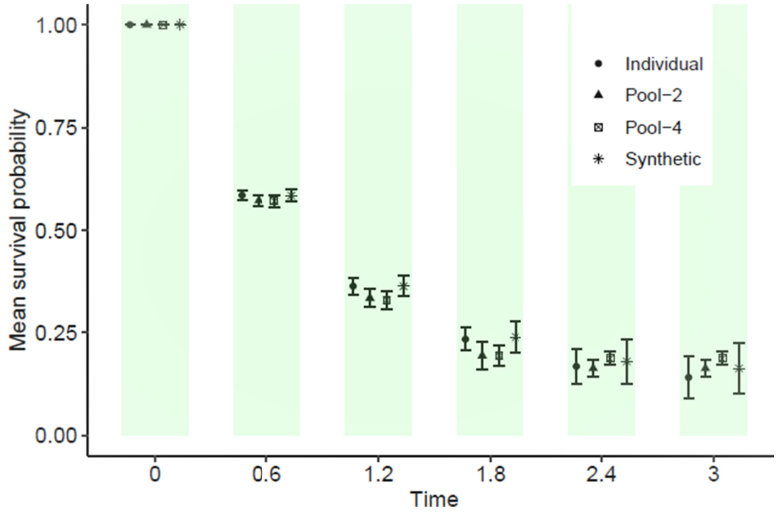


**Figure 4.** Average estimates of survival probabilities comparing individual, reconstructed pooled subcohorts, and classification and regression trees (CART)-generated synthetic data from 1000 simulated datasets each of size 5000. Mean survival probabilities and error bars are provided.

We assume all the variables are quasi-identifiers that could potentially aid information leakage. A detailed description of the method is given by Reiter.[4] For the rest of the article, we use the term "synthetic CART" to refer to this mechanism.

## 3 Performance metrics

The performance of the pooled subcohorts, under the postulated Cox PH regression framework, is assessed based on several metrics. Four main performance metrics are considered in assessing the usefulness of the released data. For the estimation of the log HR, we considered: (1) mean absolute error (bias) of the log HRs $\hat{\boldsymbol{\beta}}$, (2) SE estimates SE($\hat{\boldsymbol{\beta}}$), and (3) relative efficiency (Reff) of the log HRs. To compare the survival curves generated from pooled subcohorts to full cohort data, we considered (4) the mean survival probability and SEs. The first three are particularly relevant for conducting etiological studies while the fourth metric is useful for assessing the suitability of the released data for meaningful endpoint prediction.

1. The mean absolute bias estimate for each log HR $\hat{\beta}_j$ is expressed as $\mu_{bias}(\widehat{\beta}_j) = \sum_{r=1}^{k} |\beta_j - \widehat{\beta}_{jr}|/k$, where $k$ is the number of repetitions and $\hat{\beta}_{jr}$ is the maximum likelihood estimate (MLE) of the $j$-th log HR derived from the $r$-th repetition. This metric is particularly useful since covariate aggregation has been shown to lead to important underestimation of the exposure effect and systemic bias towards the null in Cox PH regression.[43] We compare the bias associated with the log HRs estimated from full cohort data, pooled NCC subcohorts, and synthetic data.
2. Similar to standard likelihood theory, the variance $\mathbb{V}(\hat{\boldsymbol{\beta}})$ can be estimated by inverting the second derivative of the partial likelihood function which is used to derive the SE. In practice, the SE is available from the `survival` packages.[30] If the SE($\widehat{\beta}_j$) of the $j$-th covariate is estimated over $k$ repetitions, we use the summary $\sum_{r=1}^{k}$ SE($\widehat{\beta}_{jr}$)$/k$.
3. The relative efficiency (Reff) of the log HR estimates compares the median model-based variance of the log HR to the empirical variance computed from log HR estimates of the re-sampled datasets.[44] This is expressed as Reff($\widehat{\boldsymbol{\beta}}_{emp}, \widehat{\boldsymbol{\beta}}_{mod}$) = $\mathbb{V}(\widehat{\boldsymbol{\beta}}_{mod})/\mathbb{V}(\widehat{\boldsymbol{\beta}}_{emp})$, where $\mathbb{V}(\widehat{\boldsymbol{\beta}}_{emp})$ is the empirical variance and $\mathbb{V}(\widehat{\boldsymbol{\beta}}_{mod})$ is the model-based variance, respectively.
4. The survival curves generated from the pooled subcohorts and synthetic data are each compared to the survival curve from the full cohort data. This comparison is based on their empirical mean survival probabilities and SEs obtained from multiple simulated datasets.

## 4 Applications

We used simulations and real data examples to evaluate the performance of the proposed method. The utility of pooled NCC data, under the proposed framework, was assessed using the performance metrics presented in previous sections. Our analysis considered varying combinations of pooled sizes, case-control matched sets, and censoring proportions. To do this,

we compared PH regression models for the full cohort (i.e. individual data), pooled NCC subcohorts, and CART-generated synthetic data.

## 4.1 Simulation study

The simulations were based on 1000 repetitions, each with a sample size $N = 5000$. We assumed the PH model $\lambda(t|Z) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$, with the covariates $\mathbf{Z} = (Z_1, Z_2)^T$ sampled from a bivariate normal distribution. We considered various parameter combinations and present the results for the following setting:

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \mathcal{N}\left( \begin{bmatrix} 1.5 \\ 2.8 \end{bmatrix}, \begin{bmatrix} 0.04 & -0.024 \\ -0.024 & 0.36 \end{bmatrix} \right)$$

where the correlation $\rho_{Z_1,Z_2} = -0.2$. The regression parameters $(\beta_1, \beta_2)$ were fixed to $(-1.5, 0.5)$. The true failure times $T$ and censoring times $C$ were generated from Weibull distributions scale parameters $\lambda_k \exp(-\beta_1 Z_1 - \beta_2 Z_2)$ and $\lambda_C$, respectively, whereas the shape parameter was fixed to 1 for both times. The observed survival time is taken to be the minimum of $T$ and $C$. The values of the fixed parameters $\lambda_k$ and $\lambda_C$ were varied to obtain different observed event or censoring prevalence. For each simulated full cohort, we generated a synthetic copy using CART as previously described. We also created NCC subcohorts for each full cohort by selecting all the cases and choosing $m = (2, 5, 10)$ controls per case. These subcohorts were then used for the pooling and estimation scheme outlined in Section 2.

For each constructed data type, we estimated the log HRs associated with the covariates, the model-based SE, mean absolute bias (bias), and relative efficiency (Reff), and compared them to the gold standard estimates of the full cohort. Specifically, the metrics were computed for full cohort, pooled NCC subcohorts of pool sizes 2 and 4, and CART-generated synthetic datasets. The empirical coverage (Cov) was also assessed for each data type and compared to the nominal coverage of 0.95. Additionally, the robustness of the subcohorts constructed to varying parameter values were assessed by alternatively fixing one of $\beta_1$ or $\beta_2$ to zero while varying the other parameter over the range of log HR values $(-2, 2)$.

Table 2 shows the estimates obtained for three different event prevalence rates (10%, 30%, and 50%) with each pooled NCC subcohort generated under five controls per case matching. The log HR estimates obtained from the pooled NCC subcohorts are accurate and comparable to the gold standard. However, the SE and bias estimates were generally worse for pooled subcohorts than individual data. We see an inflation of the SE estimates as more samples are pooled from individual to pool sizes 2 and 4, respectively. The confidence interval coverage of the pooled data are also consistent with the nominal 95% coverage of the full cohort. Upon fixing one log HR to zero (HR = 1) and varying the other parameter over the range $(-2, 2)$, the SE estimates obtained for pool-4 were consistently greater than those of pool-2 (shown in Section 1.2 of the Supplemental Appendix). Moreover, the power for a likelihood ratio test (with a type I error rate of 0.05) for the exposure effect (HR, $\hat{\beta}_1$ in Table 3) shows that the pooled subcohorts are consistently comparable to full cohort data and even outperform all data types when more controls are matched on in rare event settings. Compared to CART data synthesis, the pooled subcohorts consistently produces comparable log HR estimates, less bias, and higher 95% confidence interval coverage. The estimates for matching on two controls are shown in the Supplemental Appendix.

Finally, pseudo-observed survival times were constructed for the full cohort data using the pooled NCC subcohorts created in Section 2. Mean survival probability plots were then generated for these reconstructions and compared to the exact curve derived from the original full cohort. Figure 4 shows a representative plot of mean survival probability by follow-up time. Overall, our results indicate that the pooled NCC subcohorts are comparable to the full cohort (considered the gold standard) and synthetic data when considering the mean survival probabilities across 1000 simulated samples. Summaries of the median survival time and the restricted mean survival time (a numeric expression of the area under the KM survival curve) are also presented in Table 4. In summary, the proposed reconstructions of pseudo-event times provide a good approximation for the number of individuals comprising the risksets during follow-up.

## 4.2 Lung cancer data example

The National Lung Screening Trial (NLST) lung cancer data[45] was used to assess the utility of the proposed method. NLST is a randomized controlled trial designed to determine whether screening for lung cancer with low-dose helical computed tomography reduces mortality in high-risk individuals relative to screening with chest radiography. The study enrolled ~54,000 participants, from 33 screening centers across the United States, between August 2002 and April 2004. The participants were followed up until 31 December 2009 when the follow-up period ended. The study contains information on survival of patients (e.g. whether lung cancer was the official cause of death and censored or dead status), days from randomization to death or censored (time in days), patient-reported pack-years, age of the patients, and other comorbidities. For simplicity, we focus on pack-years and age as the exposure and confounder of interest. For the 53,452 patients included

**Table 2.** Log HR ($\widehat{\beta}$) estimates of individual, pooled NCC subcohorts, and CART-generated synthetic data under the Cox PH model assumption. Estimates of SE, mean absolute bias (bias), relative efficiency (Reff), and coverage probability are shown. The pools were formed under 1:5 matched NCC subcohorts. Nominal coverage was 0.95.

| | | Estimate | SE | Bias | Reff | Coverage |
|---|---|---|---|---|---|---|
| Censoring | = 10% | | | | | |
| $\beta_1$: | Individual data | −1.50 | 0.15 | 0.12 | 1.13 | 0.95 |
| | Pool-2 subcohort | −1.51 | 0.17 | 0.13 | 1.06 | 0.97 |
| | Pool-4 subcohort | −1.50 | 0.19 | 0.15 | 0.94 | 0.94 |
| | Synthetic data | −1.47 | 0.15 | 0.18 | 0.65 | 0.85 |
| $\beta_2$: | Individual data | 0.50 | 0.05 | 0.04 | 0.79 | 0.96 |
| | Pool-2 subcohort | 0.50 | 0.05 | 0.05 | 0.89 | 0.95 |
| | Pool-4 subcohort | 0.49 | 0.06 | 0.05 | 0.98 | 0.97 |
| | Synthetic data | 0.50 | 0.05 | 0.06 | 0.68 | 0.86 |
| Censoring | = 30% | | | | | |
| $\beta_1$: | Individual data | −1.50 | 0.14 | 0.12 | 0.88 | 0.95 |
| | Pool-2 subcohort | −1.52 | 0.16 | 0.13 | 0.89 | 0.93 |
| | Pool-4 subcohort | −1.51 | 0.17 | 0.15 | 0.84 | 0.89 |
| | Synthetic | −1.50 | 0.14 | 0.16 | 0.71 | 0.81 |
| $\beta_2$: | Individual data | 0.49 | 0.05 | 0.04 | 0.86 | 0.95 |
| | Pool-2 subcohort | 0.51 | 0.05 | 0.05 | 0.95 | 0.96 |
| | Pool-4 subcohort | 0.50 | 0.06 | 0.05 | 0.97 | 0.97 |
| | Synthetic data | 0.48 | 0.05 | 0.06 | 0.60 | 0.83 |
| Censoring | = 50% | | | | | |
| $\beta_1$: | Individual data | −1.51 | 0.10 | 0.09 | 0.95 | 0.96 |
| | Pool-2 subcohort | −1.51 | 0.12 | 0.10 | 0.88 | 0.94 |
| | Pool-4 subcohort | −1.50 | 0.12 | 0.10 | 0.93 | 0.95 |
| | Synthetic data | −1.53 | 0.10 | 0.11 | 0.78 | 0.83 |
| $\beta_2$: | Individual data | 0.50 | 0.03 | 0.02 | 1.14 | 0.98 |
| | Pool-2 subcohort | 0.50 | 0.04 | 0.03 | 0.97 | 0.96 |
| | Pool-4 subcohort | 0.50 | 0.04 | 0.03 | 1.01 | 0.93 |
| | Synthetic data | 0.50 | 0.03 | 0.04 | 0.82 | 0.87 |

NCC: nested case-control; CART: classification and regression trees; HR: hazard ratio; PH: proportional hazards; SE: standard error.

in the cohort due to lung cancer, 98.08% were censored. The full cohort data was used to create pooled NCC subcohorts of pool sizes 2 and 4 based on the description given in Section 2. We considered two matched sets of 2 and 5 controls per case. We also generated a synthetic copy of the data using the CART technique introduced earlier. A summary of the baseline characteristics of the NLST data is presented in Table 5.

Cox PH regression was performed to assess the utility of the datasets in characterizing the association between the outcome and the exposure and confounder. The results presented in Table 6 compare the log HR estimates from six data sources: full cohort data of individual level observations (gold standard), Pool-2 and four subcohorts (each of 1:2 and 1:5 matched sets) that combines node-derived pooled data, and synthetic data generated based on the CART data synthesis.

Similar to the simulated data results, the HR estimates obtained for the full cohort NLST data are comparable to estimates obtained from the pooled NCC subcohorts and synthetic data. The SE estimates increase as the pool size is increased from 2 to 4 records per pool. As expected with the NCC sampling design, the precision of the parameter estimates also improve as more samples were matched on (e.g. 1:2 versus 1:5 matched case-control sets). For example, the 95% confidence interval of the HR associated with age tightens from $(1.69, 2.07)$ to $(1.74, 2.05)$ for matching on three additional controls in a subcohort of pool size 2 (i.e. moving from 1:2 to 1:5 matched sets of pool size 2); the confidence interval again decreased from $(1.57, 1.99)$ to $(1.59, 1.91)$ for the same additional controls when the pool size is 5. A comprehensive summary of all the estimates is presented in Table 6.

## 5 Discussion

In this article, we propose a privacy-preserving analysis technique for time-to-event data within the NCC sampling framework. The technique leverages the pooling scheme introduced by Weinberg and others[16,18,31] to randomly aggregate individual records across matched sets within the same NCC subcohorts, stratified by the outcome status. Only the aggregated covariate levels of the NCC subcohorts are shared with the analytical site. We employ several data utility metrics to

**Table 3.** Power calculations for likelihood ratio tests on $\hat{\beta}_1$ using 1000 resampled full cohorts with a sample size of 5000. Event prevalence ranged from 5% to 50%, with NCC subcohorts of pool sizes 2, 4, and 10.

| Fix HR($\beta_2$) = 1 | | | | Subcohort pool size | | |
|---|---|---|---|---|---|---|
| HR($\beta_1$) | Prevalence | Individual | Synthetic | $\kappa = 2$ | $\kappa = 4$ | $\kappa = 10$ |
| | 50% | 0.96 | 0.95 | 0.96 | 0.99 | 0.99 |
| | 30% | 0.94 | 0.94 | 0.95 | 0.98 | 0.99 |
| 1.50 | 12.5% | 0.94 | 0.92 | 0.94 | 0.94 | 0.97 |
| | 7.5% | 0.95 | 0.92 | 0.99 | 0.99 | 0.99 |
| | 5% | 0.97 | 0.90 | 0.97 | 0.98 | 0.99 |
| | 50% | 0.81 | 0.80 | 0.84 | 0.86 | 0.86 |
| | 30% | 0.80 | 0.75 | 0.81 | 0.85 | 0.84 |
| 1.30 | 12.5% | 0.73 | 0.96 | 0.79 | 0.83 | 0.88 |
| | 7.5% | 0.81 | 0.70 | 0.82 | 0.86 | 0.87 |
| | 5% | 0.85 | 0.70 | 0.85 | 0.86 | 0.90 |
| | 50% | 0.60 | 0.61 | 0.61 | 0.62 | 0.64 |
| | 30% | 0.50 | 0.56 | 0.57 | 0.55 | 0.60 |
| 1.10 | 12.5% | 0.51 | 0.55 | 0.48 | 0.52 | 0.56 |
| | 7.5% | 0.50 | 0.50 | 0.50 | 0.53 | 0.60 |
| | 5% | 0.54 | 0.48 | 0.56 | 0.58 | 0.68 |
| | 50% | 0.85 | 0.86 | 0.85 | 0.86 | 0.88 |
| | 30% | 0.82 | 0.81 | 0.84 | 0.85 | 0.84 |
| 0.75 | 12.5% | 0.80 | 0.76 | 0.82 | 0.83 | 0.83 |
| | 7.5% | 0.83 | 0.75 | 0.85 | 0.86 | 0.87 |
| | 5% | 0.83 | 0.72 | 0.75 | 0.86 | 0.90 |

HR represents the hazard ratio; NCC: nested case-control.

**Table 4.** KM curve summaries of median survival time ($T_{1/2}$), (2.5, 97.5)% credible intervals of $T_{1/2}$, and the mean values of the restricted mean survival time *rmean*(SD) for comparing the KM curves of Individual data, pooled subcohorts, and CART-generated synthetic data from 1000 repetitions.

| | Individual | Pool-2 | Pool-4 | Synthetic |
|---|---|---|---|---|
| Median $T_{1/2}$ | 0.79 | 0.75 | 0.75 | 0.79 |
| (2.5, 97.5)% $T_{1/2}$ | (0.73, 0.85) | (0.69, 0.82) | (0.68, 0.82) | (0.71, 0.89) |
| Mean *rmean* (SD) | 1.10 (0.05) | 1.06 (0.05) | 1.08 (0.05) | 1.11 (0.05) |

KM: Kaplan–Meier; CART: classification and regression trees; SD: standard deviation.

assess the usefulness of the shared aggregate records for Cox PH regression and survival curve estimation under data privacy restrictions. These metrics are applied to the following datasets: full cohort (individual) data, pooled NCC subcohorts with pool sizes of 2 and 4, and CART-generated synthetic data.

The results of our simulations and real data example demonstrate that the HR estimates and SEs obtained through pooled data analysis are comparable to those obtained from individual full cohort records. The estimators derived from the pooled NCC framework are MLEs of the HRs, providing consistent estimates of the individual-level HRs and asymptotically normal results. We also find that matching on more controls during NCC sampling, prior to pooling, leads to a substantial improvement in the efficiency of effect estimates (HRs), as previously reported by Langholz and Goldstein[30] and Kim.[46] Bias and relative efficiency estimates of pooled NCC subcohorts are similar to those of individual full cohort Cox regression. The empirical coverage, computed under pooled NCC subcohorts, closely approximates the 95% nominal coverage in most scenarios. Furthermore, survival times reconstructed based on the proposed sampling method are adequate for making meaningful clinical predictions of survival risk using the KM estimator.

The proposed method offers several advantages over competing techniques. For example, matching cases to controls before aggregating the matched sets based on event status is computationally inexpensive and readily accessible via standard statistical software. The communication cost for transferring pooled data across the network is also fairly low, as pooling reduces the overall sample size. The cost efficiency improves as the pool size increases. Moreover, the technique is intuitively simple and accessible to researchers in other fields. In comparison, methods for synthetic data generation are often quite expensive, especially for a large quantity of attributes (e.g. three or more attributes), and might require special

**Table 5.** Baseline characteristics of the lung cancer data (full cohort), pooled NCC subcohorts, and synthetic data.

| | | Dataset | | | |
|---|---|---|---|---|---|
| Variable | | Full cohort | Pool-2 | Pool-4 | Synthetic-Cart |
| Prevalence | N \| died | 5,3452 \| 1021 | 3060 \| 510 | 765 \| 255 | 53,452 \| 1004 |
| Time (days) | Median (IQR) | 2428 (2270, 2553) | 2316 (2193, 2436) | 2199 (1643, 2313) | 2430 (2272, 2554) |
| Pack years | Mean (SD) | 56.0 (23.9) | 57.0 (17.5) | 59.00 (13.40) | 56.2 (24.1) |
| | Median (IQR) | 48.0 (39.0, 66.0) | 53.5 (44.0, 66.5) | 56.75 (49.0, 66.5) | 48.8 (39.0, 67.5) |
| Age (years) | Mean (SD) | 61.4 (5.02) | 62.00 (3.74) | 62.0 (2.86) | 61.4 (5.03) |
| | Median (IQR) | 60.0 (57.0, 65.0) | 59.0 (61.5, 64.5) | 61.7 (60.0, 64.0) | 60.0 (57.0, 65.0) |

IQR denotes the interquartile range; SD: standard deviation; NCC: nested case-control.

**Table 6.** HR and 95% CI estimates for the associations of pack years and age with lung cancer deaths across the full cohort (individual), pooled NCC subcohorts (1:2 and 1:5 matched sets), and CART synthetic datasets.

| Variable | Data source | HR | 95% CI |
|---|---|---|---|
| Pack years | Individual | 1.14 | (1.12, 1.16) |
| | Pool-2 (1:2 NCC) | 1.14 | (1.11, 1.18) |
| | Pool-2 (1:5 NCC) | 1.14 | (1.11, 1.17) |
| | Pool-4 (1:2 NCC) | 1.15 | (1.11, 1.19) |
| | Pool-4 (1:5 NCC) | 1.14 | (1.11, 1.17) |
| | Synthetic | 1.13 | (1.09, 1.17) |
| Age of patient | Individual | 1.84 | (1.72, 1.96) |
| | Pool-2 (1:2 NCC) | 1.88 | (1.69, 2.07) |
| | Pool-2 (1:5 NCC) | 1.90 | (1.74, 2.05) |
| | Pool-4(1:2 NCC) | 1.77 | (1.57, 1.99) |
| | Pool-4 (1:5 NCC) | 1.75 | (1.59, 1.91) |
| | Synthetic | 1.63 | (1.51, 1.76) |

HR: hazard ratio; 95% CI: 95% confidence interval; NCC: nested case-control; CART: classification and regression trees.

and difficult-to-program algorithms. Another noteworthy advantage is that NCC sampling minimizes both selection and recall biases, reducing the risk of re-identification before matched set aggregation. Inclusion of binary effect modifiers and additional confounders is also easily attainable without major modifications to the underlying likelihood.[18] For example, outcome pooling followed by stratifying by an effect modifier could improve efficiency of estimation. Sometimes, such further stratification is recommended to ensure a fair comparison of measurements. For instance, when the exposures of interest are biomarker measurements from frozen biological material, the cases and controls should be matched based on factors such as storage time or condition. However, the current study recommends random pooling of covariates based on only the outcome status, as the privacy gain in inferential or re-identification risk outweighs the efficiency gain when the patients are further stratified.

Inferential and re-identification disclosure risk are fundamental concerns in assessing privacy. The former is still an active area of research and as such assessing the associated disclosure risk of individuals in the released microdata remains a challenging task.[47–49] In principle, an intruder (e.g. the analyst or data steward) could infer or learn information about a participant even if exact records were not disclosed. Inferential disclosure risk has been reported to be a major concern in traditional techniques such as suppression of quasi-identifiers or noise perturbation, partly because disclosed records often retain local data structures potentially aiding inference. Pooled NCC subcohorts, in comparison, smooth out underlying local perturbations via summation rendering the risk of inferential disclosure extremely small.[50]

While a comparison of the risk of patient re-identification disclosure of individual records is not covered in the current manuscript, we believe the risk is low in NCC subcohort data for several reasons:[51] (1) under the generous assumption that all attributes in the database are quasi-identifiers, pooled NCC subcohort data is non-overlapping with individual records of external data sources because aggregates do not retain underlying structures of individual elements, (2) any unique combination of pooled covariates records cannot be an identifying class of the individual components due to random pooling, and (3) any pattern of population overlap between the sensitive individual record and an external source is destroyed via pooling.[50,52] Furthermore, the distribution of any single aggregate attribute is not an accurate indicator of re-identification disclosure risk, or in the case of a categorical variable the results may be falsely reassuring. In the instances the combination

of aggregate attributes leads to a unique combination, individuals remain protected as the aggregates are non-overlapping with external databases of individual records.

Our method has some limitations. First, data aggregation decreases statistical power, potentially reducing the utility of pooled NCC subcohorts compared to individual full cohort data. Consequently, meaningful data patterns could be distorted or completely destroyed via aggregation. Second, selecting a suitable pool size requires the analyst to strike a balance between privacy and utility. While a pool size of 4 has been shown to offer good privacy protection and data utility, even with small to moderate samples,[36,28] it may be helpful for individual data nodes to assess various pool sizes before sharing the pooled data. For instance, each node could perform a pool-level assessment and share data when the estimates from the largest pool size are comparable to the site's individual-level estimate. Third, another limitation of publishing pooled NCC subcohort data is the lack of a provable privacy guarantee for the randomized summation mechanism used to generate pools. In recent years, synthetic data generation methods with differential privacy guarantees[53] have gain popularity. Various classical and modern (deep generative) approaches have since been developed to generate differentially private synthetic datasets. In the future, the development of a pooling mechanism with a mathematically provable privacy guarantee warrants exploration. Lastly, the current method cannot be implemented in heterogeneous data settings where the data nodes harbor different distributions[54] or with vertically partitioned databases where the attributes are split among the data nodes. In the former case, our approach assumes that the datasets stored at the nodes follow the same distributions and that site-specific subcohorts can be combined without taking between-node differences into account.[55] A potential modification to account for these differences might be to introduce a weighted likelihood function for evaluating the pooled subcohorts. The latter case remains an active area of research.

Despite these limitations, the trade-off between disclosure risk and utility of pooled disclosed data makes the technique appealing. Pooling individual records at contributing data sites result in minimal utility loss whilst preserving the confidentiality of patients. The additional efficiency gain when more controls are matched-on makes the technique particularly desirable, especially in rare disease settings or in emerging infectious disease studies where the outcome prevalence is still low. In conclusion, the proposed pooling technique of survival time data under NCC subcohort sampling preserves patient privacy while ensuring consistent estimation of effects, suitable standard errors, and accurate survival curves comparable to individual full cohort data.

## Acknowledgements

The authors are grateful to the editor, associate editor, and two referees for their insightful comments and suggestions, which significantly improved the quality of the paper. We thank the funding agencies Mitacs Accelerate Fellowship, Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-06100), and Fonds de la recherche en santé du Québec (Salary Award for PS-C) for supporting our research.

## Declaration of conflicting interests

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

## Data availability

The NLST data is available for request from https://www.cancer.gov/types/lung/research/nlst. All of the codes used to generate the results presented in this manuscript are available on the gitpage of the corresponding author.

## Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

## ORCID iD

Lamin Juwara https://orcid.org/0000-0001-9934-6623

## Supplemental material

Supplemental materials for this article are available online.

## References

1. Mostert M, Bredenoord AL, Biesaart MC, et al. Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach. *Eur J Human Genet* 2016; **24**: 956–960.
2. Fienberg SE, Fulp WJ, Slavkovic AB, et al. "Secure" log-linear and logistic regression analysis of distributed databases. In: *International Conference on Privacy in Statistical Databases*. Springer; 2006. pp.277–290.
3. Raghunathan TE, Reiter JP and Rubin DB. Multiple imputation for statistical disclosure limitation. *J Offi Stat* 2003; **19**: 1.

4. Reiter JP. Using CART to generate partially synthetic public use microdata. *J Off Stat* 2005; **21**: 441.
5. Saha-Chaudhuri P and Weinberg C. Addressing data privacy in matched studies via virtual pooling. *BMC Med Res Methodol* 2017; **17**: 136.
6. Fienberg SE and Sanil AP. A Bayesian approach to data disclosure: optimal intruder behavior for continuous data. *J Off Stat* 1997; **13**: 75.
7. Fienberg SE and Steele RJ. Disclosure limitation using perturbation and related methods for categorical data. *J Off Stat* 1998; **14**: 485.
8. Narayanan A and Shmatikov V. Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy (sp 2008). IEEE; 2008. pp.111–125.
9. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *Int J Uncertain, Fuzz Knowl-Based Syst* 2002; **10**: 571–588.
10. Heffetz O and Ligett K. Privacy and data-based research. *J Econo Perspect* 2014; **28**: 75–98.
11. Soria-Comas J and Domingo-Ferrert J. Differential privacy via t-closeness in data publishing. In: 2013 Eleventh Annual Conference on Privacy, Security and Trust. IEEE; 2013. pp.27–35.
12. Nikolaenko V, Weinsberg U, Ioannidis S, et al. Privacy-preserving ridge regression on hundreds of millions of records. In: 2013 IEEE Symposium on Security and Privacy. IEEE; 2013. pp.334–348.
13. Dorfman R. The detection of defective members of large populations. *Ann Math Stat* 1943; **14**: 436–440.
14. Du D, Hwang FK and Hwang F. *Combinatorial group testing and its applications*. 2nd ed. London: World Scientific, 1999.
15. Mutesa L, Ndishimye P, Butera Y et al. A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature* 2021; **589**: 276–280.
16. Weinberg CR and Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics* 1999; **55**: 718–726.
17. Vansteelandt S, Goetghebeur E and Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* 2000; **56**: 1126–1133.
18. Saha-Chaudhuri P and Weinberg CR. Specimen pooling for efficient use of biospecimens in studies of time to a common event. *Am J Epidemiol* 2013; **178**: 126–135.
19. Sham P, Bader JS, Craig I, et al. DNA pooling: a tool for large-scale association studies. *Nature Rev Genet* 2002; **3**: 862.
20. Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat Med* 1992; **11**: 1871–1879.
21. Henderson R, Jones M and Stare J. Accuracy of point predictions in survival analysis. *Stat Med* 2001; **20**: 3083–3096.
22. O'Keefe CM and Rubin DB. Individual privacy versus public good: protecting confidentiality in health research. *Stat Med* 2015; **34**: 3081–3103.
23. Yu S, Fung G, Rosales R et al. Privacy-preserving Cox regression for survival analysis. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining; 2008. pp.1034–1042.
24. Mohammed N, Fung BC, Hung PC, et al. Anonymizing healthcare data: a case study on the blood transfusion service. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining; 2009. pp.1285–1294.
25. Duan R, Luo C, Schuemie MH et al. Learning from local to global—an efficient distributed algorithm for modeling time-to-event data. bioRxiv. 2020.
26. Jordon J, Yoon J and der Schaar M van. PATE-GAN: generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations, 2018.
27. Xie L, Lin K, Wang S, et al. Differentially private generative adversarial network. arXiv preprint arXiv:180206739. 2018.
28. Juwara L and Saha-Chaudhuri P. A hybrid covariate microaggregation approach for privacy-preserving logistic regression. *J Surv Stat Methodol* 2022; **10**: 568–595.
29. Cox DR. Partial likelihood. *Biometrika* 1975; **62**: 269–276.
30. Langholz B and Goldstein L. Risk set sampling in epidemiologic cohort studies. *Stat Sci* 1996; **11**: 35–53.
31. Saha-Chaudhuri P and Juwara L. Survival analysis under the Cox proportional hazards model with pooled covariates. *Stat Med* 2021; **40**: 998–1020.
32. Samuelsen SO. A psudolikelihood approach to analysis of nested case-control studies. *Biometrika* 1997; **84**: 379–394.
33. Langholz B and Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *Am J Epidemiol* 1990; **131**: 169–176.
34. Therneau T et al. A package for survival analysis in S. *R package version* 2015; **2**.
35. Cox LH. Suppression methodology and statistical disclosure control. *J Am Stat Assoc* 1980; **75**: 377–385.
36. Saha-Chaudhuri P, Umbach DM and Weinberg CR. Pooled exposure assessment for matched case-control studies. *Epidemiology (Cambridge, Mass)* 2011; **22**: 704.
37. Clayton D and Hills M. *Statistical models in epidemiology*. Oxford: Oxford University Press, 2013.
38. Langholz B and Clayton D. Sampling strategies in nested case-control studies. *Environ Health Perspect* 1994; **102**: 47–51.
39. Goldstein L and Langholz B. Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann Stat* 1992; **20**: 1903–1928.
40. Loh WY. Classification and regression trees. *Wiley Interdiscipl Rev: Data Mining Knowl Discov* 2011; **1**: 14–23.
41. James G, Witten D, Hastie T, et al. *An introduction to statistical learning*. **112**. New York: Springer, 2013.

42. Nowok B, Raab GM, Dibben C et al. synthpop: bespoke creation of synthetic data in R. *J Stat Softw* 2016; **74**: 1–26.
43. Abrahamowicz M, Du Berger R, Krewski D et al. Bias due to aggregation of individual covariates in the Cox regression model. *Am J Epidemiol* 2004; **160**: 696–706.
44. Mittleman MA, Maclure M and Robins JM. Control sampling strategies for case-crossover studies: an assessment of relative efficiency. *Am J Epidemiol* 1995; **142**: 91–98.
45. Loprinzi CL, Laurie JA, Wieand HS et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. *North Central Cancer Treatment Group J Clin Oncol* 1994; **12**: 601–607.
46. Kim RS. A new comparison of nested case-control and case-cohort designs and methods. *Eur J Epidemiol* 2015; **30**: 197–207.
47. Dalenius T. Towards a methodology for statistical disclosure control. *Statistik Tidskrift* 1977; **15**: 2–1.
48. Duncan GT and Lambert D. Disclosure-limited data dissemination. *J Am Stat Assoc* 1986; **81**: 10–18.
49. Loong B, Zaslavsky AM, He Y, et al. Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS. *Stat Med* 2013; **32**: 4139–4161.
50. Boyens C, Krishnan R and Padman R. On privacy-preserving access to distributed heterogeneous healthcare information. In: 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the IEEE; 2004. p.10.
51. Paass G. Disclosure risk and disclosure avoidance for microdata. *J Bus Econ Stat* 1988; **6**: 487–500.
52. Simon GE, Shortreed SM, Coley RY et al. Assessing and minimizing re-identification risk in research data derived from health care records. *eGEMs* 2019; **7**: 6.
53. Dwork C. Differential privacy: A survey of results. In: International conference on theory and applications of models of computation. Springer, 2008. pp.1–19.
54. Luo C, Duan R, Naj AC, et al. ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data. *Sci Report* 2022; **12**: 6627.
55. Xue Y, Schifano ED and Hu G. Geographically weighted Cox regression for prostate cancer survival data in Louisiana. *Geogr Anal* 2020; **52**: 570–587.

# Appendix for "Privacy-preserving analysis of time-to-event data under nested case-control sampling"

September 24, 2023

## 1 Simulated data

Consider a survival dataset with a total size of $n = 5000$. Assume the proportional hazards (PH) model $\lambda(t|X) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_2)$, where the covariates $\mathbf{X} = (X_1, X_2)^T$ are generated from a bivariate normal distribution given by:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \mathcal{N}\left( \begin{bmatrix} 1.5 \\ 2.8 \end{bmatrix}, \begin{bmatrix} 0.04 & -0.024 \\ -0.024 & 0.36 \end{bmatrix} \right) \equiv \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

Where the correlation $\rho_{X_1,X_2} = -0.2$. The regression parameters $(\beta_1, \beta_2)$ are set as $(-1.5, 0.5)$. The true failure times $T$ and censoring times $C$ are generated from Weibull distributions with scale parameters $\lambda_k \exp(-\beta_1 X_1 - \beta_2 X_2)$ and $\lambda_C$, respectively, while the shape parameter is fixed at 1 for both times. The observed survival time is determined as the minimum of $T$ and $C$. The values of the fixed parameters $\lambda_k$ and $\lambda_C$ are adjusted to achieve varying observed event or censoring prevalence.

### 1.1 Simulation results: 1:2 nested case-control matched sets

We generate pooled NCC subcohorts using a 1:2 NCC matching for three different event prevalence rates (10%, 30%, and 50%). We assess the same Cox PH model on the full cohort data (Individual), the pooled NCC subcohorts, and synthetic data generated by CART. We present results for the Log HR estimates ($\widehat{\boldsymbol{\beta}}$), the standard error (SE), mean absolute bias (Bias), relative efficiency (Reff), and coverage probability (with a nominal coverage probability set at 0.95).

| | | Estimate | SE | Bias | Reff | Coverage |
|---|---|---|---|---|---|---|
| Censoring | = 10% | | | | | |
| $\beta_1$: | Individual data | -1.50 | 0.13 | 0.11 | 0.87 | 0.94 |
| | Pool-2 subcohort | -1.51 | 0.14 | 0.12 | 0.73 | 0.86 |
| | Pool-4 subcohort | -1.50 | 0.12 | 0.10 | 0.73 | 0.89 |
| | Synthetic data | -1.49 | 0.10 | 0.13 | 0.65 | 0.82 |
| $\beta_2$: | Individual data | 0.50 | 0.04 | 0.03 | 1.05 | 0.95 |
| | Pool-2 subcohort | 0.50 | 0.04 | 0.04 | 0.92 | 0.94 |
| | Pool-4 subcohort | 0.50 | 0.04 | 0.03 | 0.90 | 0.93 |
| | Synthetic data | 0.50 | 0.03 | 0.04 | 0.80 | 0.89 |
| Censoring | = 30% | | | | | |
| $\beta_1$: | Individual data | -1.50 | 0.14 | 0.12 | 0.90 | 0.95 |
| | Pool-2 subcohort | -1.48 | 0.19 | 0.16 | 0.85 | 0.93 |
| | Pool-4 subcohort | -1.48 | 0.17 | 0.14 | 0.94 | 0.93 |
| | Synthetic data | -1.48 | 0.18 | 0.17 | 0.66 | 0.82 |
| $\beta_2$: | Individual data | 0.50 | 0.05 | 0.04 | 0.93 | 0.96 |
| | Pool-2 subcohort | 0.51 | 0.06 | 0.06 | 0.88 | 0.95 |
| | Pool-4 subcohort | 0.50 | 0.06 | 0.04 | 1.14 | 0.97 |
| | Synthetic data | 0.51 | 0.06 | 0.06 | 0.72 | 0.81 |
| Censoring | = 50% | | | | | |
| $\beta_1$: | Individual data | -1.51 | 0.13 | 0.09 | 1.10 | 0.97 |
| | Pool-2 subcohort | -1.52 | 0.14 | 0.12 | 0.89 | 0.96 |
| | Pool-4 subcohort | -1.52 | 0.12 | 0.10 | 0.95 | 0.94 |
| | Synthetic data | -1.51 | 0.14 | 0.17 | 0.66 | 0.80 |
| $\beta_2$: | Individual data | 0.51 | 0.04 | 0.03 | 0.95 | 0.95 |
| | Pool-2 subcohort | 0.50 | 0.05 | 0.04 | 1.04 | 0.97 |
| | Pool-4 subcohort | 0.51 | 0.04 | 0.03 | 0.97 | 0.94 |
| | Synthetic data | 0.48 | 0.05 | 0.06 | 0.60 | 0.78 |

Table 1: Log HR ($\widehat{\boldsymbol{\beta}}$) estimates of individual, pooled NCC subcohorts, and CART-generated synthetic data under the Cox PH model assumption. Estimates of standard error (SE), mean absolute bias (Bias), relative efficiency (Reff), and coverage probability are shown. The pools were formed under 1:2 matched NCC subcohorts. Nominal coverage was 0.95.

## 1.2 Simulation results: 1:5 nested case-control matched sets

### 1.2.1 Distributions of the Pooled subcohorts *vs* Synthetic data

We present the distributions of the simulated covariates $X_i,\ i = 1, 2$. The NCC subcohorts were created by selecting all the cases and 5 controls per case.

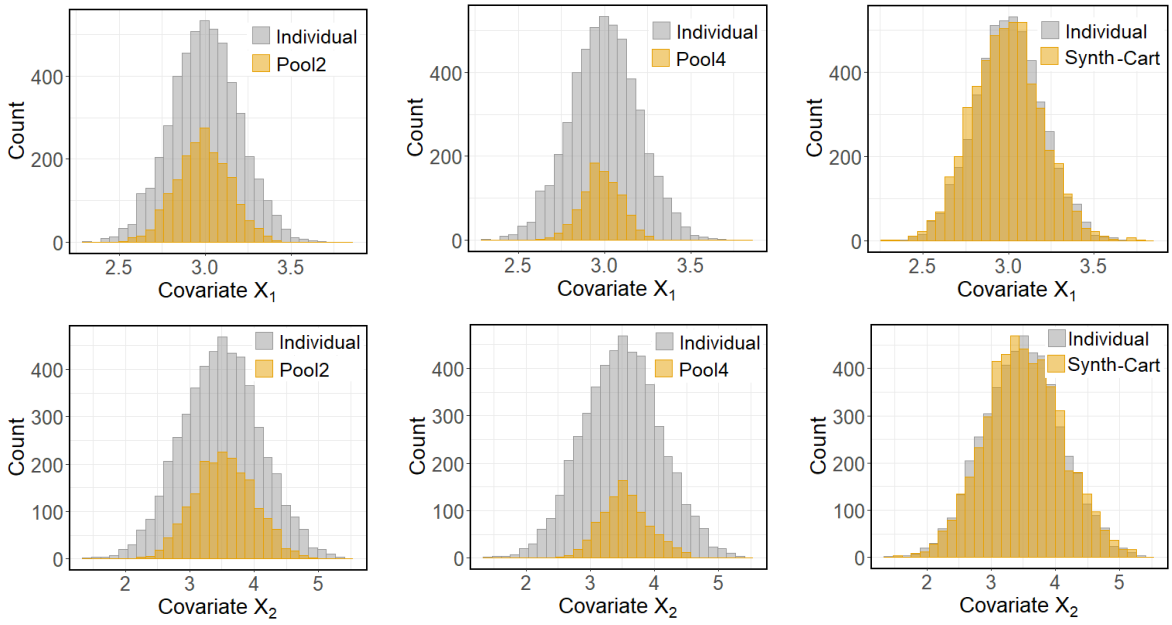


Figure 1: Histogram of pooled NCC subcohorts (of 5 controls per case matched sets) and CART synthetic data overlaid on top of the original dataset. Specifically, Pools of size 2, 4 and synthetic data were plotted on top the full cohort.

### 1.2.2 Mean Absolute Bias and Standard Error Estimation for 1:5 case-control matched sets

For a more plausible real-life hazard ratio range (e.g., HR estimates between 0.5 and 2), the pooled NCC subcohorts exhibit practically identical estimated standard errors (SEs) to both the full cohort and synthetic data. The SE values remain generally comparable across the entire range of simulated log hazard ratios. Moreover, we observe a sharp increase in the estimated SE when $\beta_1 \to -2, \beta_2 = 0$ due to the asymmetry introduced by taking the exponential of $\beta_1$. Conversely, a similar increase occurs when $\beta_2 \to -2, \beta_1 = 0$.

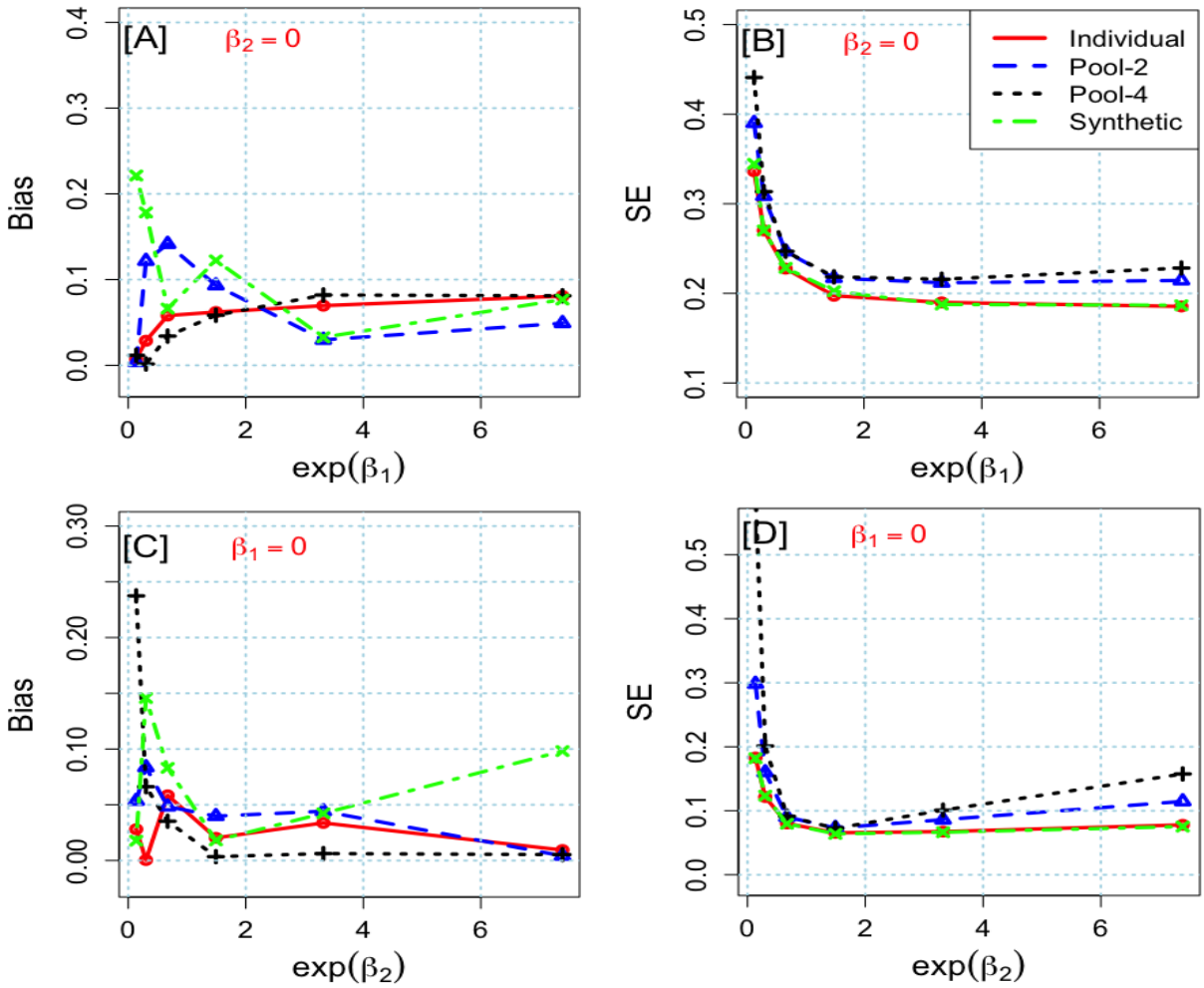


Figure 2: Mean Absolute Bias and Standard Error Estimation: Cox PH Model Applied to Individual Cohorts, Pooled Subcohorts (Size 2 and 4), and CART Synthetic Datasets. The Estimates are Computed over 1000 Simulated Datasets, Each of Size n=5000, for Six Equally Spaced Log Hazard Ratio (log HR) Values in the Interval (-2, 2). Plots are in Terms of Hazard Ratios. The pools were formed under 1:5 matched NCC subcohorts.

### 1.2.3 Kaplan-Meier survival curves

We reconstructed the survival curve using a randomly selected simulated dataset and present plots for the full cohort (Individual), Pool-2 subcohort, Pool-4 subcohort, and synthetic data generated using CART.

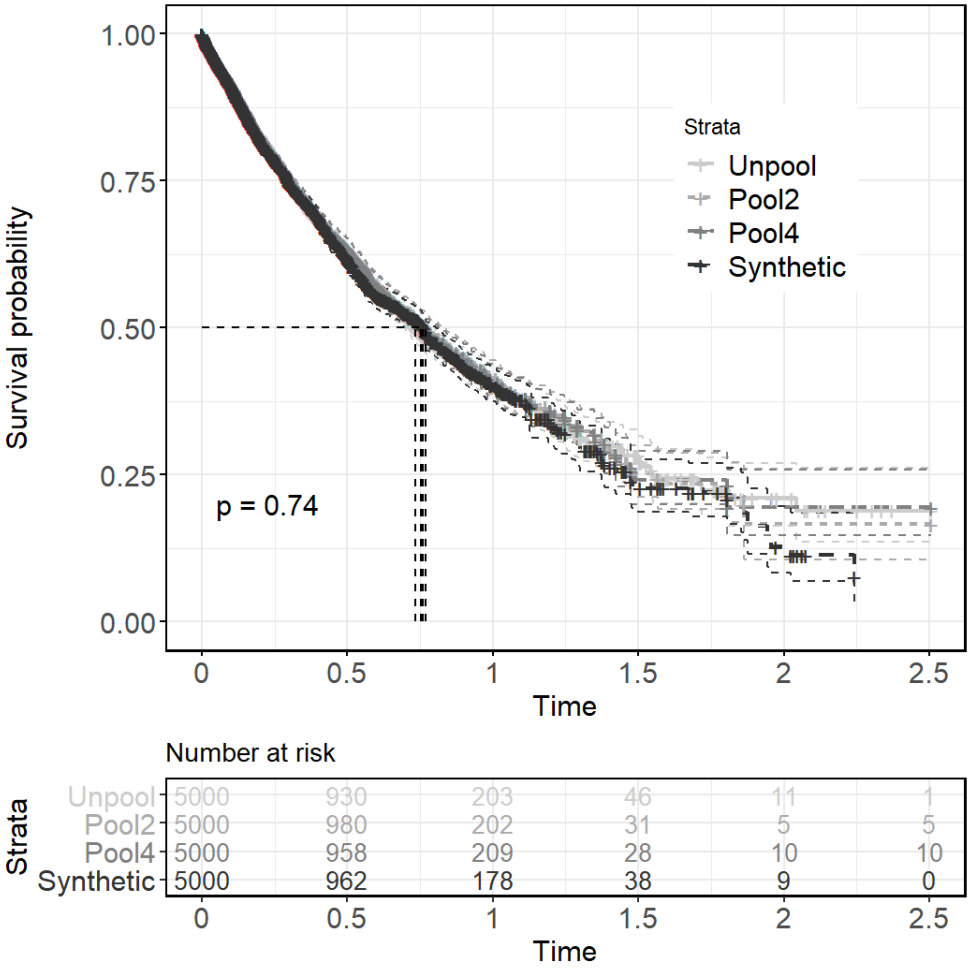


Figure 3: Comparison of Kaplan-Meier Survival Curves: Full Cohort (Unpool), Reconstructed Pooled Subcohorts, and Synthetic Data Generated by CART from a Randomly Sampled Simulated Dataset. The pools were formed under 1:5 matched NCC subcohorts.

# 2 Classification And Regression Trees for Synthetic Data Generation

Classification And Regression Trees (CART) is a modeling technique that recursively splits the dataset into subsets with more homogeneous outcomes (see Breiman et al., 1984). The splits in the explanatory variable space are typically represented by a tree structure. In Figure 4, we illustrate a tree structure for a univariate outcome $Y$ and two predictors, $X_1$ and $X_2$, which was grown using the algorithms proposed by Clark and Pregibon in 1992. The values within each of the final groups (leaves $L_1$-$L_3$) approximate a conditional distribution of the predicted variable when the criteria governing that group are met by the predictors. Synthetic data copies are subsequently sampled from these groups. In this manuscript, synthetic data were generated using the *synthpop* package in R.

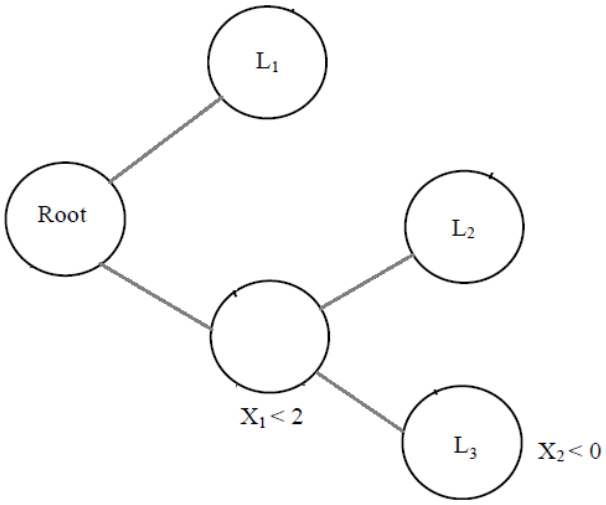


Figure 4: Hypothetical tree structure involving a single outcome and two continuous predictors.