Cox regression with linked data

Thanh Huan Vo[1,3], Valérie Garès[1], Li-Chun Zhang[5], André Happe[4], Emmanuel Oger[4], Stéphane Paquelet[3], and Guillaume Chauvet *[2]

[1]Univ Rennes, INSA, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France
[2]Univ Rennes, ENSAI, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France
[3]IRT b<>com, Rennes, France
[4]EA 7449 REPERES, France
[5] University of Southampton, Southampton, UK and Statistics Norway

Abstract

Record linkage is increasingly used, especially in medical studies, to combine data from different databases that refer to the same entities. The linked data can bring analysts novel and valuable knowledge that is impossible to obtain from a single database. However, linkage errors are usually unavoidable, regardless of record linkage methods, and ignoring these errors may lead to biased estimates. While different methods have been developed to deal with the linkage errors in the generalized linear model, there is not much interest on Cox regression model, although this is one of the most important statistical models in clinical and epidemiological research. In this work, we propose an adjusted estimating equation for secondary Cox regression analysis, where linked data have been prepared by a third-party operator, and no information on matching variables is available to the analyst. Through a Monte Carlo simulation study, the proposed method is shown to lead to substantial bias reductions in the estimation of the parameters of the Cox model caused by false links. An asymptotically unbiased variance estimator for the adjusted estimators of Cox regression coefficients is also proposed. Finally, the proposed method is applied to a linked database from the Brest stroke registry in France.

*Corresponding author: Guillaume Chauvet, ENSAI, Rennes, France.
Email: guillaume.chauvet@ensai.fr

Keywords: adjusted estimating equation, cox regression, linkage error, secondary analysis, variance estimation

1 Introduction

Record linkage, also known as data matching, is a process of combining data from different sources that refer to the same individuals or entities. Nowadays, data are collected everywhere by different sectors, and the ability of combining information from several databases can lead to novel knowledge for analysts. For example, record linkage is widely used in epidemiology and medical studies to enrich data on clinical performance and other health-related information.[1,2] In national censuses, population data files obtained at different times can be linked to create longitudinal data sets.[3] Record linkage may also be applied early in a survey to link the sampling frame and administrative data.[4] The linked data allows for statistical analysis (e.g., Cox regression) which would not be possible with data collected solely by means of the survey.

The record linkage process is straightforward if unique identifiers (e.g. Social Security Number) are available and free of error in both databases. However, this information is often not available, or sometimes cannot be used due to ethical reasons. In such cases, record linkage methods may only use partial identifying information shared between databases, such as name, address, and gender. The variables used for comparison are called matching variables. Over the last decades, several methods have been developed to link data efficiently,[5,6] such as the frequentist approach[7–9] and the Bayesian approach.[10,11] However, because the matching variables are not unique and are likely to contain inaccuracies, linkage errors are unavoidable. The two kinds of record linkage errors are false links (false positives, i.e. a non-matched pair predicted as a link), and missed links (false negatives, i.e. a matched pair failed to be predicted as a link). Ignoring these errors may cause substantial bias in the analysis model,[12] causing misleading inference. It is therefore important to account for linkage errors in statistical analysis.

In published literature, two positions are usually considered to account for linkage errors in statistical analysis. Under the primary analysis framework, the data analyst is supposed to be granted access to the full linkage process, including knowledge of matching variables. From this perspective, Scheuren and Winkler[13] made use of the two highest matching weights of each record pair to reduce the bias of ordinary least square estimators under a linear regression model. However, the proposed estimators are not unbiased in full generality. Lahiri and Larsen[14] discussed this problem and proposed unbiased estimators in the same context, using the posterior matching probabilities obtained from the Fellegi-Sunter record linkage model. Hof and Zwinderman[15] extended the method by Lahiri and Larsen[14] for multiple links, and also proposed alternative estimators based on weighted least square methods, both for linear and logistic regression models. Recently, Han and Lahiri[16] adapted the approach by Lahiri and Larsen[14] to provide a system of estimating equations, which may lead to unbiased estimators under a generalized linear model.

In some applications, the analysis step is separated from the record linkage, e.g. when the matching variables contain confidential information. This is the secondary analysis framework, in which the data analyst is only provided access to the final linked data, whereas the (unknown) record linkage process has been performed by a third-party operator.[17] Starting from this perspective, Chambers[18] proposed the exchangeable linkage error (ELE) model, and bias-corrected estimating equations for both linear and logistic regression modeling. Under the ELE model, it is assumed that linked records may be split into distinct blocks inside which the probability of correct linkage and the probability of incorrect linkage are constant. Following this work, several authors[19–22] developed methods for secondary analysis of linked data. Recently, Zhang and Tuoto[23] proposed a pseudo ordinary least square method for secondary linkage-data linear regression analysis, which can accommodate heterogeneous linkage errors and incomplete match space problems.

Although the Cox proportional hazard model[24] is of routine use for survival analysis, comparatively very few papers have focused on accounting for record linkage errors in this context.[25] performed a simulation study emphasizing the impact of missing matches on the parameter estimation of the Cox model, but did not propose any solution to obtain unbiased estimators for the model parameters. Hof et al.[26] proposed a joint modeling for survival analysis and probabilistic record linkage. However, this analysis model is developed under a primary analysis viewpoint, while in many applications, a secondary analysis is more likely. In this work, we reason from the secondary analysis position. We propose a model to account for record linkage errors, and an estimation method to correct for the bias caused by false link errors in the Cox regression model.

The article is organised as follows. In Section 2, we propose a new estimating equation, which leads to an approximately unbiased estimation of the parameters for the Cox model with linked data. A variance estimator is also proposed. In Section 3, we evaluate the proposed estimator and the associated variance estimator through simulation studies. In Section 4, an application on a real dataset is presented. Finally, possible further research is discussed in Section 5.

2 Cox regression analysis with linked data

2.1 Cox regression model

The Cox proportional hazard model[24] is the most popular method to assess the effect of covariates $\mathbf{X}$ on a survival time. This is therefore one of the most important models in medical research. Suppose that a random sample of n units is available. For each unit $i = 1, \ldots, n$, we let $\widetilde{T}_i$ be a non-negative random variable, which denotes the duration

between a time origin and the time of occurrence of some event of interest. We suppose that $\widetilde{T}_i$ is right censored, which means that the event is observed only if it occurs before censoring time C_i. For units $i = 1, \ldots, n$, we therefore observe $T_i = \min(\widetilde{T}_i, C_i)$. We let $\delta_i = \mathbb{1}_{\{\widetilde{T}_i \leq C_i\}}$ denote the variable indicating whether the duration time is observed prior to censoring. The vector of covariates is denoted as $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)^T$. In this section, we first suppose that $\mathbf{X}_i$ is observed for any unit in the sample.

According to the Cox model, the hazard function of an event at time t is given by

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t) \exp\left(\mathbf{X}_i^T \boldsymbol{\beta}_0\right), \tag{1}$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ is a p-vector of unknown parameters and $\lambda_0(t)$ is a common baseline hazard function. Assuming that the survival times are observed on a finite interval, and that C is independent of $\widetilde{T}$ conditionally on $\mathbf{X}$, a consistent estimator $\hat{\beta}$ of β_0 may be obtained by solving the estimating equation:

$$H_0(\beta) \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_i \left\{ \mathbf{X}_i - \frac{\sum_{j=1}^n Y_j(T_i) \exp\left(\mathbf{X}_j^T \boldsymbol{\beta}\right) \mathbf{X}_j}{\sum_{j=1}^n Y_j(T_i) \exp\left(\mathbf{X}_j^T \boldsymbol{\beta}\right)} \right\} = 0, \tag{2}$$

where $Y_j(t) = \mathbb{1}_{(T_i \geq t)}$ is an at-risk indicator.[27] We call (2) the theoretical estimating equation. This is also the maximum partial likelihood (mpl) estimation. Under some mild assumptions, a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by[27]

$$\hat{\mathbb{V}}_{\text{mpl}}(\hat{\boldsymbol{\beta}}) = \left\{ -n \nabla H_0(\hat{\boldsymbol{\beta}}) \right\}^{-1}. \tag{3}$$

2.2 Linkage error model

Suppose that we have a dataset A of n_A time-to-event data. If the covariates $\mathbf{X}_i$ were known for any unit $i \in A$, the parameter of the Cox model would be estimated by solving the theoretical estimating equation (2). However, if the covariates are not known in database A, equation (2) may not be solved in practice.

In order to obtain the needed covariates, a linkage is performed with a dataset B of size $n_B \geq n_A$, containing in particular the auxiliary variables $\mathbf{X}_i$. For any unit i in A, we note $\mathbf{Z}_i$ for the vector of auxiliary values resulting from the linkage process. Reasoning from the secondary analysis perspective, we do not have access to the matching variables and do not know the actual linkage process.

We assume that the linkage error is non-informative of the regression model, i.e. may depend on the errors in the matching process, but not on the model covariates nor on the survival time.[28] This is the key assumption of most secondary analysis approaches in the literature, for which Zhang and Tuoto[23] have proposed a diagnostic test. Adopting the modelling approach of Copas and Hilton,[29] we suppose that both databases are partitioned into blocks A_v and B_v, $v = 1, \ldots, V$, and that the record linkage is performed independently in these blocks. Also, we suppose that for any entity $i \in A_v$, we have:

$$\mathbf{Z}_i = \begin{cases} \mathbf{X}_i & \text{with probability } \alpha_v, \\ \mathbf{X}_{(j)} & \text{with probability } 1 - \alpha_v, \end{cases} \tag{4}$$

where (j) stands for some unit randomly selected in database B_v. In other words, it is supposed that for any $i \in A_v$, the correct entity is linked to i with probability α_v, otherwise the unit j linked to i is randomly selected in B_v. It should be noted that we implicitly assume that A is a subset from B, and that all entities in A can therefore have some matching records in B. Also, we assume that there is at most one link for each record of both databases. In practice, there will often be some entities of A which remain unlinked after the linkage process. This may be due to errors in the matching variables, or to the fact they are not sufficiently discriminant for identifying links. Such incomplete record linkage can be problematic for further analysis if the missed links are not at random.[25] There are some discussions on this incomplete matching space problem.[19,23,30] This problem is out of the scope of our work. We therefore assume that the linkage is complete, or alternatively

that any missing links are independent on the time of event and model covariates.

2.3 Adjusted estimating equation

By naively treating the linked covariates $\mathbf{Z}_i$ as if they were the true covariates $\mathbf{X}_i$ for the units $i \in A$, an estimator of $\boldsymbol{\beta}_0$ may be obtained by solving the following equation:

$$H_{naive}(\boldsymbol{\beta}) \equiv \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \delta_i \left\{ \mathbf{Z}_i - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) \exp(\mathbf{Z}_j^\top \boldsymbol{\beta}) \mathbf{Z}_j}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) \exp(\mathbf{Z}_j^\top \boldsymbol{\beta})} \right\} = 0. \tag{5}$$

We call (5) the naive estimating equation. Since some units are incorrectly linked, it may lead to biased estimates, see the simulation results in Section 3.

We propose a bias-corrected estimating equation, accounting for the fact that from the hit-miss model (4), the covariates may be incorrectly linked. We first introduce some notations. Let us define

$$g(\boldsymbol{\beta}, \mathbf{X}_i) = \exp(\mathbf{X}_i^\top \boldsymbol{\beta}) \quad \text{and} \quad h(\boldsymbol{\beta}, \mathbf{X}_i) = \exp(\mathbf{X}_i^\top \boldsymbol{\beta}) \mathbf{X}_i.$$

Also, let $\bar{\mathbf{X}}_{B_v}, \bar{g}_{B_v}(\boldsymbol{\beta})$ and $\bar{h}_{B_v}(\boldsymbol{\beta})$ denote the means of $\mathbf{X}_i, g(\boldsymbol{\beta}, \mathbf{X}_i)$ and $h(\boldsymbol{\beta}, \mathbf{X}_i)$ over B_v, respectively. The linkage-error *adjusted estimating equation* (AEE) is given by

$$\bar{H}(\boldsymbol{\beta}) \equiv \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \delta_i \left\{ \mathbf{X}_i^*(\alpha_v) - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^*(\alpha_v, \boldsymbol{\beta})}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^*(\alpha_v, \boldsymbol{\beta})} \right\} = 0 \tag{6}$$

where, for any $i \in A_v$,

$$\begin{aligned} \mathbf{X}_i^*(\alpha_v) &= \alpha_v^{-1} \mathbf{Z}_i - (\alpha_v^{-1} - 1) \bar{\mathbf{X}}_{B_v}, \\ g_j^*(\alpha_v, \boldsymbol{\beta}) &= \alpha_v^{-1} g(\mathbf{Z}_j, \boldsymbol{\beta}) - (\alpha_v^{-1} - 1) \bar{g}_{B_v}(\boldsymbol{\beta}), \\ h_j^*(\alpha_v, \boldsymbol{\beta}) &= \alpha_v^{-1} h(\mathbf{Z}_j, \boldsymbol{\beta}) - (\alpha_v^{-1} - 1) \bar{h}_{B_v}(\boldsymbol{\beta}). \end{aligned} \tag{7}$$

We prove in Appendix A that $\bar{H}(\boldsymbol{\beta})$ is an (approximately) conditionally unbiased estimator for the function $H_0(\boldsymbol{\beta})$ involved in the theoretical estimating equation. Solving the proposed AEE therefore leads to a consistent estimator of $\boldsymbol{\beta}$, see the simulation results in Section 3.

Since there is no closed-form solution for the estimating equations considered above, an iterative method like the Newton-Raphson algorithm is commonly used in practice. Also, the probabilities α_v may be (somewhat arbitrarily) specified by the record linkage practitioner, or estimated from a validation sample[18,23] if their true values are unknown.

2.4 Variance estimator

In this section, we discuss variance estimation for the estimator of the parameter β_0 obtained by solving the AEE given in (6). We first note that several sources of variance need to be accounted for: a) the (usual) variability associated to solving a sample-based estimating equation, b) the variability associated to the linkage process, and c) the variability associated to the estimation of the probabilities α_v, $v = 1, \ldots, V$. Using the variance estimator given in (3) fails to account for all these sources of variability, and therefore leads to an underestimation of the variance, see the simulation results in Section 3.

We propose a sandwich-like variance estimator, which reads as follows:

$$\hat{\mathbb{V}}_{\text{AEE}}(\hat{\boldsymbol{\beta}}) \equiv \{\nabla \bar{H}(\hat{\boldsymbol{\beta}})\}^{-1} \times \hat{\mathbb{V}}\{\bar{H}(\boldsymbol{\beta}_0)\} \times \{\nabla \bar{H}(\hat{\boldsymbol{\beta}})\}^{-1}, \tag{8}$$

$$\text{with } \hat{\mathbb{V}}\{\bar{H}(\boldsymbol{\beta}_0)\} = \hat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\} + \hat{\mathbb{V}}_2\{\bar{H}(\boldsymbol{\beta}_0)\}. \tag{9}$$

The first component $\hat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\}$ in (9) accounts for the variability in (c). Under the assumption that the validation samples S_v used for such estimation are selected in the datasets A_v through simple random sampling without replacement, this variance estimator

is

$$\hat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\} = \sum_{v=1}^{V} \bar{H}_{2,v}(\hat{\alpha}_v, \hat{\boldsymbol{\beta}})\{\bar{H}_{2,v}(\hat{\alpha}_v, \hat{\boldsymbol{\beta}})\}^\top \times \left(\frac{1}{n_{S_v}} - \frac{1}{n_{A_v}}\right) \frac{n_{S_v}}{n_{S_v}-1} \frac{1-\hat{\alpha}_v}{\hat{\alpha}_v^3},$$

where n_{S_v} is the sample size of the validation set S_v, and

$$\bar{H}_{2,v}(\alpha_v, \boldsymbol{\beta}) = \frac{1}{n_A} \sum_{i \in A_v} \delta_i \{(\mathbf{Z}_i - \bar{\mathbf{X}}_{B_v}) \\ - \frac{\sum_{j \in A_v} Y_j(T_i) \left\{\left\{h(\boldsymbol{\beta}, \mathbf{Z}_j) - \bar{h}_{B_v}(\boldsymbol{\beta})\right\} - R_i^*(\alpha_v, \boldsymbol{\beta}) \left\{g(\boldsymbol{\beta}, \mathbf{Z}_j) - \bar{g}_{B_v}(\boldsymbol{\beta})\right\}\right\}}{\sum_{j \in A_v} Y_j(T_i) g_j^*(\alpha_v, \boldsymbol{\beta})}\}.$$

with

$$R_i^*(\alpha_v, \boldsymbol{\beta}) = \frac{\sum_{j \in A_v} Y_j(T_i) h_j^*(\alpha_v, \boldsymbol{\beta})}{\sum_{j \in A_v} Y_j(T_i) g_j^*(\alpha_v, \boldsymbol{\beta})}.$$

The second component $\hat{\mathbb{V}}_2\{\bar{H}(\boldsymbol{\beta}_0)\}$ in (9) accounts for both the variability in (a) and (b). We have

$$\hat{\mathbb{V}}_2\{\bar{H}(\boldsymbol{\beta}_0)\} = \frac{s_H^2(\hat{\boldsymbol{\beta}})}{n_A}$$

where

$$s_H^2(\boldsymbol{\beta}) = \frac{1}{n_A - 1} \sum_{v=1}^{V} \sum_{i \in A_v} \left\{ H_i(\boldsymbol{\beta}) - \frac{1}{n_A} \sum_{v=1}^{V} \sum_{j \in A_v} H_j(\boldsymbol{\beta}) \right\}^2$$

and

$$H_i(\boldsymbol{\beta}) = \delta_i \left\{ \mathbf{X}_i^*(\hat{\alpha}_v) - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^*(\hat{\alpha}_v, \boldsymbol{\beta})}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^*(\hat{\alpha}_v, \boldsymbol{\beta})} \right\}.$$

The derivation of this variance estimator is explained in detail in Appendix B. It is evaluated empirically in the next section through a simulation study.

3 A simulation study

In this section, we evaluate the performance of the proposed estimator for the parameter of the Cox model, and the associated variance estimator. The data generation process is first presented in Section 3.1. The estimation methods that we evaluate are presented in Section 3.2, along with the performance indicators. The simulation results are given in Section 3.3. To facilitate interpretation and to study the influence of different simulation parameters, we first consider in Section 3.3.1 scenarios with a single block. Scenarios with multiple blocks and different levels of linkage quality are considered in Section 3.3.2.

3.1 Data generation

Assume that there are two datasets A with n_A individuals, and B with $n_B \geq n_A$ individuals. We first generate the n_B units in database B with $p = 2$ covariates, including a continuous variable $X_1 \sim \mathcal{N}(0, 1)$ and a binary variable $X_2 \sim$ Bernoulli(0.7). Given the p-vector of coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (0.5, -0.5)^\top$, the true survival time $\widetilde{T}^B$ is generated as

$$\widetilde{T}^B = -\frac{\log(U)}{\lambda \exp\left(\mathbf{X}^\top \boldsymbol{\beta}\right)} \tag{10}$$

where U follows a standard uniform distribution,[31] and λ is fixed as equal to 1 for simplicity. A constant censoring time is chosen (from 100 000 independent data generation runs) to yield a censoring rate of approximately 0.25 over all the simulation runs.

Without loss of generality, we suppose that the units in dataset A are the n_A first ones in dataset B. In other words, a pair of individuals (a_i, b_j) for $i \in A$ and $j \in B$ is a match if $i = j = 1, \ldots, n_A$. The survival times T_i^A for $i \in A$ are therefore obtained as $T_i^A = T_i^B$ for $i = 1, \ldots, n$. Given the values of α, the linked values $\mathbf{Z}$ for covariates in database A are obtained according to the linkage error model (4).

If there are multiple blocks, data for each block were generated independently as follows. Firstly, for each block v, we generate n_{B_v} observations $(T, \delta, \mathbf{X})$ from the Cox model described in equation (10). Note that the value of the true parameters $\boldsymbol{\beta}$ and the distribution of $\mathbf{X}$ are the same over blocks v. Then, we choose randomly $n_{A_v} \leq n_{B_v}$ survival times (T, δ) for block A_v. All generated n_{B_v} values of $\mathbf{X}$ will be placed in block B_v. Secondly, given the value of α_v for block v, n_{A_v} linked values $\mathbf{Z}$ for block A_v are obtained by the linkage error model (4). Inside each block A_v, an audit sample of 10% of the units is selected by simple random sampling without replacement, and the proportion of correct links in the audit sample is used as the estimator $\hat{\alpha}_v$.

3.2 Methods and performance indicators

For each scenario, we consider the following estimation methods. The `Theoretical` is obtained by solving the theoretical estimating equation (2) with the true values of covariates $\mathbf{X}$. This is a benchmark estimation strategy, since it cannot be applied on linked data in practice. The `Naive` is obtained by solving the naive estimating equation (5) with linked data. The `Validation` is obtained by solving the theoretical estimating equation (2) with only correct linked pairs in the validation set. Note that, contrarily to `Theoretical`, this method may be used in practice if an audit sample is available. For each of these three methods, the variance of the estimator of the parameter in the Cox model is estimated by using the variance estimator $\hat{\mathbb{V}}_{\text{mpl}}(\hat{\boldsymbol{\beta}})$ in equation (3), implemented by means of R SURVIVAL package.

For each scenario, we also consider estimation methods making use of the proposed approach. The `TAEE` (theoretical adjusted estimating equation) is obtained by solving the proposed estimating equation (6) with the theoretical value of α_v. The `AEE` (adjusted estimating equation) is obtained by solving the proposed estimating equation (6), where

α_v is estimated by taking the proportion of correct links in the audit sample. For each method, the Newton-Raphson algorithm is applied with a maximum of 20 iterations and an initial parameter value $\boldsymbol{\beta} = (0, 0)^\top$. We also report the number of time (`Fails`) when the Newton-Raphson algorithm does not converge. For `AEE`, the variance is estimated by using $\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}})$ in equation (34). For `TAEE`, the variance is estimated by setting $\hat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\} = 0$ in $\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}})$. For both `TAEE` and `AEE`, we also compare to the variance estimator $\hat{\mathbb{V}}_{\text{mpl}}(\hat{\boldsymbol{\beta}})$ in equation (3).

The data generation and the estimation process are repeated $R = 1,000$ times. Over these simulations, we compare the estimation methods in terms of the Monte Carlo bias

$$\text{B}_{\text{MC}}(\hat{\boldsymbol{\beta}}) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\boldsymbol{\beta}}^{(r)} - \boldsymbol{\beta}\right),$$

with $\hat{\boldsymbol{\beta}}^{(r)}$ the estimator computed on the r-th sample. We also compute the Monte Carlo standard deviation:

$$\text{Sd}_{\text{MC}}(\hat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\boldsymbol{\beta}}^{(r)} - \bar{\hat{\boldsymbol{\beta}}}\right)^2}.$$

For the variance estimation methods, we compute the Monte Carlo estimates of standard deviation

$$\widehat{\text{Sd}} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\hat{\mathbb{V}}^{(r)}(\hat{\boldsymbol{\beta}}^{(r)})},$$

with $\hat{\mathbb{V}}^{(r)}$ a variance estimator computed on the r-th sample. The Monte Carlo estimate of standard deviation is compared to the true standard deviation $\text{Sd}(\hat{\boldsymbol{\beta}})$, approximated by $\text{Sd}_{\text{MC}}(\hat{\boldsymbol{\beta}})$.

3.3 Simulation results

3.3.1 One block situation

In this section, we consider the situation when the data sets are generated as presented in Section 3.1, with $V = 1$ block only. We consider two cases. In the first one, the sample sizes $n_A = 1,000$ and $n_B = 2,000$ are held fixed, and we let the probability of correct link α vary in $\{0.75, 0.85, 0.95\}$. In the second one, the probability of correct link is held fixed, equal to 0.85. We let n_A vary in $\{500, 1000, 2000\}$, with $n_B = 2n_A$.

The simulation results obtained in Case 1 are presented in Table 1. As expected, the `Theoretical` method leads to an unbiased estimation of the parameters. The `Naive` method leads to severely biased estimators, especially with the smaller value $\alpha = 0.75$. The bias ranges from 0.029 to 0.147, corresponding to an absolute relative bias between 5.8 % and 29.0 %. This bias decreases as the probability of correct link increases, as expected. The proposed methods `TAEE` and `AEE` lead to approximately unbiased estimation of the parameters, with a larger variability for `AEE` as expected. The bias under AEE ranges from 0.000 to 0.015, corresponding to a reduction of the relative bias (as compared to Naive) ranging between 5.0 % and 27.6 %. We note that the variability under both `TAEE` and `AEE` is but only moderately increased, as compared to `Theoretical`. The `Validation` method also leads to unbiased estimators of the Cox regression coefficients, but with a larger variability than both `TAEE` and `AEE`.

We now turn to the variance estimators. The variance estimator $\hat{\mathbb{V}}_{\text{mpl}}(\hat{\boldsymbol{\beta}})$ (3) performs well for `Theoretical`, `Naive` and `Validation`, but underestimates the variability of the estimators obtained under `TAEE` and `AEE`. This is due to the fact that this variance estimator only accounts for the variability of the sample-based estimating equation. On the other hand, the proposed variance estimator performs well, except for β_1 when $\alpha = 0.75$, in which

case the variance is underestimated. We have also computed coverage probabilities (`CP`) for normality-based confidence intervals with a nominal coverage of 95%. We note that the coverage probability is very poorly respected in case of `Naive`, even in situations when the bias is moderate.

			$\hat{\beta}_1$					$\hat{\beta}_2$				
α	Methods	Fails	B_{MC}	$\mathrm{Sd}_{\mathrm{MC}}$	$\widehat{\mathrm{Sd}}_{\mathrm{mpl}}$	$\widehat{\mathrm{Sd}}_{\mathrm{AEE}}$	CP	B_{MC}	$\mathrm{Sd}_{\mathrm{MC}}$	$\widehat{\mathrm{Sd}}_{\mathrm{mpl}}$	$\widehat{\mathrm{Sd}}_{\mathrm{AEE}}$	CP
*	Theoretical	0	0.000	0.039	0.040		0.961	0.003	0.080	0.080		0.950
0.75	Naive	0	0.147	0.041	0.039		0.050	0.143	0.081	0.081		0.577
	Validation	0	0.017	0.160	0.156		0.941	0.003	0.318	0.302		0.936
	TAEE	0	0.007	0.072	0.041	0.069	0.945	0.013	0.124	0.081	0.129	0.957
	AEE	4	0.009	0.082	0.041	0.085	0.962	0.015	0.131	0.081	0.138	0.962
0.85	Naive	0	0.092	0.040	0.039		0.347	0.088	0.081	0.080		0.799
	Validation	0	0.016	0.149	0.146		0.955	0.000	0.296	0.283		0.931
	TAEE	0	0.002	0.055	0.041	0.059	0.964	0.007	0.103	0.080	0.113	0.969
	AEE	0	0.005	0.063	0.041	0.066	0.969	0.010	0.110	0.080	0.118	0.972
0.95	Naive	0	0.033	0.041	0.040		0.862	0.029	0.083	0.080		0.928
	Validation	0	0.015	0.139	0.137		0.961	0.004	0.276	0.266		0.939
	TAEE	0	0.001	0.045	0.040	0.051	0.965	0.003	0.089	0.080	0.101	0.977
	AEE	0	0.000	0.048	0.040	0.054	0.973	0.004	0.090	0.080	0.103	0.981

Table 1: Simulation results in Case 1 with three different values for the probability of correct link $\alpha \in \{0.75, 0.85, 0.95\}$

The simulation results obtained in Case 2 are presented in Table 2. We observe no qualitative difference compared to Case 1. The `TAEE` and `AEE` lead to almost unbiased estimations for the regression coefficients, and the proposed variance estimator performs well for both methods. The bias obtained under the `Naive` method does not decrease as the sample size increases. As could be expected, the variability obtained under any estimation

method decreases as the sample size increases.

			$\hat{\beta}_1$					$\hat{\beta}_2$				
n_A	Methods	Fails	B_{MC}	Sd_{MC}	$\widehat{Sd}_{mpl}$	$\widehat{Sd}_{AEE}$	CP	B_{MC}	Sd_{MC}	$\widehat{Sd}_{mpl}$	$\widehat{Sd}_{AEE}$	CP
500	Theoretical	0	0.002	0.056	0.057		0.954	0.005	0.113	0.114		0.955
	Naive	0	0.089	0.057	0.056		0.636	0.087	0.113	0.114		0.876
	Validation	0	0.033	0.222	0.215		0.951	0.024	0.435	0.419		0.949
	TAEE	0	0.009	0.078	0.058	0.085	0.963	0.010	0.145	0.114	0.161	0.972
	AEE	1	0.015	0.104	0.058	0.104	0.976	0.015	0.161	0.114	0.172	0.977
1000	Theoretical	0	0.000	0.039	0.040		0.961	0.003	0.080	0.080		0.950
	Naive	0	0.092	0.040	0.039		0.347	0.088	0.081	0.080		0.799
	Validation	0	0.016	0.149	0.146		0.955	0.000	0.296	0.283		0.931
	TAEE	0	0.002	0.055	0.041	0.059	0.964	0.007	0.103	0.080	0.113	0.969
	AEE	0	0.005	0.063	0.041	0.066	0.969	0.010	0.110	0.080	0.118	0.972
2000	Theoretical	0	0.000	0.028	0.028		0.945	0.000	0.056	0.057		0.960
	Naive	0	0.092	0.029	0.028		0.111	0.092	0.056	0.057		0.640
	Validation	0	0.006	0.103	0.100		0.932	0.003	0.197	0.197		0.948
	TAEE	0	0.001	0.039	0.029	0.041	0.953	0.000	0.071	0.057	0.080	0.971
	AEE	0	0.002	0.043	0.029	0.046	0.964	0.001	0.075	0.057	0.082	0.969

Table 2: Simulation results in Case 2 with three different values for the sample size n_A

3.3.2 Multiple blocks

In this section, we consider the situation when the data sets are generated as presented in Section 3.1, with $V = 3$ blocks only. We take $(n_{A_1}, n_{A_2}, n_{A_3}) = (250, 500, 250)$ and $(n_{B_1}, n_{B_2}, n_{B_3}) = (500, 1000, 500)$. Also, we consider a first scenario where $(\alpha_1, \alpha_2, \alpha_3) = (0.6, 0.7, 0.8)$; a second scenario where $(\alpha_1, \alpha_2, \alpha_3) = (0.7, 0.8, 0.9)$; a third scenario where $(\alpha_1, \alpha_2, \alpha_3) = (0.8, 0.9, 1.0)$.

Let $\bar{\alpha}$ be the weighted average of $\alpha_1, \ldots, \alpha_v$ defined as

$$\bar{\alpha} = \frac{\sum_{i=1}^{V} n_{A_v} \alpha_v}{\sum_{i=1}^{V} n_{A_v}}.$$

This leads to a percentage of correct links approximately equal to $\bar{\alpha} = 70\%$ in Scenario 1, $\bar{\alpha} = 80\%$ in Scenario 2 and $\bar{\alpha} = 90\%$ in Scenario 3. In this context, we also consider two additional versions of our proposed methods, when we are unable to access to the value α_v of each block, but we have only access to their weighted average: `TAEE`-$\bar{\alpha}$ where the `TAEE` is used with $V = 1$ and true value of $\bar{\alpha}$, and `AEE`-$\bar{\alpha}$ where the `AEE` is used with $V = 1$ and estimated value of $\hat{\bar{\alpha}}$.

The simulation results are presented in Table 3, and confirm the good results of the proposed methods observed in the situation of one block. Scenarios 1 and 2 are the cases when the behaviour of the `Naive` method is particularly poor, with a very large bias due to a larger number of false links, and very poor coverage for the confidence intervals. On the other hand, `AEE` performs well in reducing the estimation bias even in this situation. The proposed variance estimator also performs well in these cases. The standard errors of `TAEE` and `AEE` estimators decrease as $\bar{\alpha}$ increase, i.e. by going from Scenario 1 to Scenario 3 in Table 3. As explained in Section 2.4, there are three sources of variance in the estimation process: a) the variability associated to solving a sample-based estimating equation, b) the variability associated to the linkage process, and c) the variability associated to the estimation of the probabilities α_v. Since the sample size is kept constant, the term a) is likely not affected by the value of α_v. The term b) decreases as α_v increases, as the variance in the hit-miss model (4) does so. The term c) also decreases as α_v increases, as illustrated by the fact that $\hat{V}_1$ depends on $(1 - \alpha)/\alpha^3$, which is decreasing as $\alpha \to 1$. Concerning the coverage probability of normality-based confidence intervals, we note that they are well respected under the proposed methods, although the confidence intervals are

slightly conservative when the variance estimators are so.

			$\hat{\beta}_1$					$\hat{\beta}_2$				
Scenario	Methods	Fails	B_{MC}	Sd_{MC}	$\widehat{Sd}_{mpl}$	$\widehat{Sd}_{AEE}$	CP	B_{MC}	Sd_{MC}	$\widehat{Sd}_{mpl}$	$\widehat{Sd}_{AEE}$	CP
*	Theoretical	0	0.002	0.040	0.039		0.953	0.002	0.078	0.078		0.944
1	Naive	0	0.171	0.041	0.039		0.010	0.171	0.082	0.081		0.440
	Validation	0	0.021	0.161	0.161		0.961	0.002	0.322	0.315		0.945
	TAEE	1	0.018	0.097	0.042	0.136	0.944	0.013	0.143	0.081	0.144	0.947
	AEE	17	0.030	0.136	0.042	0.128	0.961	0.022	0.177	0.081	0.183	0.960
	TAEE-$\bar{\alpha}$	0	0.017	0.086	0.043	0.139	0.943	0.013	0.139	0.081	0.141	0.951
	AEE-$\bar{\alpha}$	9	0.021	0.108	0.042	0.144	0.964	0.016	0.153	0.081	0.168	0.963
2	Naive	0	0.118	0.041	0.039		0.167	0.120	0.084	0.080		0.660
	Validation	0	0.018	0.151	0.150		0.948	0.002	0.294	0.293		0.946
	TAEE	0	0.007	0.066	0.041	0.064	0.952	0.003	0.118	0.080	0.122	0.953
	AEE	1	0.015	0.086	0.041	0.081	0.969	0.010	0.129	0.080	0.135	0.961
	TAEE-$\bar{\alpha}$	0	0.007	0.064	0.041	0.063	0.955	0.003	0.116	0.080	0.120	0.959
	AEE-$\bar{\alpha}$	0	0.009	0.073	0.041	0.075	0.966	0.006	0.123	0.080	0.127	0.963
3	Naive	0	0.060	0.041	0.040		0.662	0.061	0.082	0.080		0.882
	Validation	0	0.016	0.143	0.140		0.945	0.006	0.272	0.275		0.950
	TAEE	0	0.005	0.052	0.041	0.056	0.965	0.004	0.097	0.080	0.108	0.967
	AEE	0	0.007	0.058	0.041	0.062	0.973	0.006	0.102	0.080	0.112	0.971
	TAEE-$\bar{\alpha}$	0	0.004	0.051	0.041	0.055	0.965	0.004	0.096	0.080	0.107	0.972
	AEE-$\bar{\alpha}$	0	0.005	0.055	0.041	0.060	0.970	0.005	0.099	0.080	0.109	0.973

Table 3: Simulation results with 3 blocks with different linkage quality

When the block-specific true link rate is not correlated with the block-specific distribution of T and $\mathbf{X}$, e.g. this multiple blocks simulation set up, a single-$\bar{\alpha}$ adjustment (`TAEE`-$\bar{\alpha}$ and `AEE`-$\bar{\alpha}$) can still perform well. The main result in Table 3 concerning AEE and AEE-$\bar{\alpha}$ is that they both lead to virtually unbiased estimators. The bias is indeed always smaller with

AEE-$\bar{\alpha}$, but the difference is no greater than 0.009, which is very small as compared to the value of the parameters ($\beta_1 = 0.5$ and $\beta_2 = -0.5$). Closeness between TAEE and TAEE-$\bar{\alpha}$ confirms the somewhat favourable simulation setup of non-informative linkage error. Reduced bias of AEE-$\bar{\alpha}$ compared to AEE-α may be due to the non-linearity of adjustment, such that the additional variance of AEE-α adjustment is manifested in terms of the bias of adjustment. Moreover, a single-$\bar{\alpha}$ adjustment can provide a smaller variance. In practice, this is very helpful when the analyst cannot conduct auditing, and when the linker can only provide a single overall estimate of α. However, the linkage error may be informative, such as when β and α vary across the blocks in a correlated manner. Block-specific adjustment would then be clearly more helpful at reducing the bias than adjustment by a single $\bar{\alpha}$.

Some additional simulation results are presented in the supplementary material. In particular, we have studied the situation when the non-informative assumption is not true. The simulation results in Tables S.1 and S.2 indicate that the Cox parameters estimated under `TAEE` and `AEE` wil be more biased when α is dependent on variables from the Cox model. This is especially true when α is small and dependent on $\tilde{T}$ (see Table S.2). This emphasizes the importance of the non-informative linkage error assumption. We note, however, that the proposed methods still perform better in this case than the `Naive` method. We have also performed a sensitivity analysis, evaluating the performance of TAEE with incorrect values for the parameter α. The results are presented in Table S.3. As could be expected, the bias in the estimated parameters increases with the error in α, but the estimator remain less biased than with the Naive method if the error is moderate.

4 Application

4.1 Data description

The proposed model is fitted to a linked dataset between a registry of strokes, denoted by AVC ("Accident Vasculaire Cérébral"), and an extraction of the national health information system of France, denoted by SNDS ("Système national des données de santé"). The AVC recorded all stroke cases of patients aged 15 years and older, who have lived in the Brest area from 2008 to the end of 2018. SNDS is an extraction from the French health information system, and contains patients for whom at least one medical service or hospitalization were recorded since 2008 while they were living in the Brest area. Due to the limited information in the registry, there is a demand of linking AVC and SNDS to enrich the registry for further analyses.

The linkage was performed by a separate team, and due to confidentiality restrictions, we were not allowed to access to the matching data and have limited knowledge about the linkage. A deterministic record linkage method was used. This is the simpler linkage approach, which ideally requires agreement on all matching variables, or otherwise on a (large) subset of these variables. In the linkage process, there are 9 matching variables, and the linkage is implemented sequentially. In the first step, it is required that the 9 matching variables agree for a pair to be viewed as a link. The corresponding pairs are then suppressed, and among the remaining ones it is asked that 8 matching variables agree for a pair to be viewed as a link. The procedure continues on similarly. The process is summarized in Table 4.

After performing the linkage process, a dataset of $3,535$ patients has been obtained. It contains the survival time, the censoring indicator and three covariates (age, gender, type of stroke). We suppose that these covariates were obtained from SNDS by the linkage

process, and may therefore be affected by linkage errors. A description of the dataset is presented in Table 5. In this application, we are interested in comparing the risk of death after the first stroke between males and females, taking into account the age and the type of stroke.

Steps	**Number of agreements among 9 matching variables**	**Number of record pairs**
1	9	1,792
2	8	170
3	7	11
4	6	1,500
5	5	58
6	4	4
	Total	3535

Table 4: Description of the linkage process

4.2 Cox regression analysis

In this application, we use the Cox regression model (1) to model the relationship between the survival time and three explanatory variables (age, gender, type of stroke). We consider AVC as database A and SNDS as database B in our proposed model. In the naive approach, we use the linked data as if it was directly observed. However, the simulation results in Section 3.3 show that linkage errors lead to biased estimators of the regression coefficients. Therefore, we also use the adjusted estimating equation (6).

Variable	**Description**	**Source**
Time	Time (in days) between the first stroke and death or end of follow-up (31/12/2018)	AVC
Censoring	If the patient died before 01/01/2019: 1 = Yes, 0 = No	AVC
Age	Age (in years) at the first stroke	SNDS
Gender	Sex: 0 = Male, 1 = Female	SNDS
Type AVC	Type of stroke (0 = Ischemic, 1 = Hemorrhagic)	SNDS

Table 5: Description of the linked database

For the record pairs obtained at each step, the percentage of matching variables which are in agreement are seen as a proxy of the probability that the matching is correct. For example, for the $1,500$ pairs obtained at step 4, the probability that the matching is correct is estimated as $6/9 = 0.667$. We suppose that the linked dataset is comprised of two blocks, so as to avoid the possibility of dependency between the linkage process performed into the different blocks. The estimates of α_v for each block v are obtained as follows:

- Block 1: 1,792 record pairs are obtained from Step 1, with $\hat{\alpha}_1 = 9/9 = 1$.
- Block 2: 1,743 remaining record pairs, with

$$\hat{\alpha}_2 = \frac{170 \times 8/9 + 11 \times 7/9 + 1500 \times 6/9 + 58 \times 5/9 + 4 \times 4/9}{1743} \simeq 0.694.$$

Besides, because the covariates are not available for any units in the SNDS, the adjustment terms in (7) cannot be computed since the proposed approach requires full access

to the set of covariates in database B. We therefore use the proxy solution suggested in equation (35), which requires that the covariates are known on the linked dataset only. Simulations in Appendix C show that if the database A may be seen as a random sample from the database B, or when the sampling leading to A is independent of the covariates, this method leads to comparable results as the method proposed in Section 2.3.

	Naive method			`AEE`		
	coef	sd	hr	coef	sd	hr
Age	0.059	0.002	1.061	0.070	0.001	1.073
Sex	-0.120	0.047	0.887	-0.145	0.067	0.865
Type AVC	0.773	0.058	2.165	0.846	0.082	2.330

Table 6: Estimated coefficients (coef), estimated standard deviation of the estimated coefficients (sd), and the hazard ratio (hr = exp(coef)) of the naive method and the `AEE` method from linked data.

In Table 6, we present the estimations arising from both the `Naive` and the `AEE` methods. The two methods decidedly lead to different estimations. If the `Naive` method is used, the hazard ratio of sex is 0.887, which means that given the same age and the same type of stroke, the female's risk of death after the first stroke is 0.887 times smaller than male's. On one hand, this ratio from the adjusted estimating equation approach is just 0.865.

5 Discussion

In this work, our simulations proved that the naive use of linked data may lead to substantial bias in a Cox regression model. Therefore, under the secondary analysis position where the

analyst can access to linked data only, we have proposed an adjusted estimating equation for linked data, which can correct the bias from the naive estimating equation. A variance estimator, which can capture three sources of variability has also been proposed. However, proving the asymptotic normality of the resulting estimators remains challenging.

Through various simulation scenarios with one block and also multiple blocks, the proposed adjusted estimating equation is shown to lead to substantial bias reductions as compared to the naive estimating equation. Additional simulations study the non-information linkage assumption and the sensitivity analysis of $\hat{\alpha}$ are also presented in the Supplementary material.

We have also proposed different variants of the approach for scenarios where information is limited. For example, when the block-specific linkage rate α_v is not available for each block, our method still works well by using the average true link rate $\bar{\alpha}$. If the analysts are not able to fully access the covariates in database B, we proposed to use the adjustments in (35) in the Appendix, which still maintain the good performance of the AEE if A is a random sample from B. Detailed simulation results are presented in Table S.4 of the Supplementary material. In addition, a linear approximated estimating equation (LAEE), which can provide better estimation than AEE with small sample size, is given in Table S.5 of the Supplementary material.

Although the proposed method has improved on the naive estimation, there are perspectives that need to be developed. In this work, we assumed that observations on survival time are already available and all explanatory variables are obtained from another database. In practice, there are some cases when a part of the covariates is also available in A, and only a part of the covariates is acquired from B by linkage. In addition, the covariates can be obtained from several sources with different linkage processes. The proposed model should be developed to adapt to these cases.

We also supposed that the survival time and the censoring indicator are observed in database A, while the explanatory variables are obtained from database B by a linkage process. However, the opposite situation may occur in practice: the covariates may be available for the units in A, while the survival time needs to be obtained from another database B by a linkage process. The proposed adjustment in equation (6) only accounts for the error associated to Z_i. If T_i and δ_i are linked from dataset B, they are prone to linkage errors which need to be accounted for in modifying the estimating equation. This requires a different adjustment approach.

A Expectation of the adjusted estimating equation

The proposed *adjusted estimating equation* is given by

$$\bar{H}(\boldsymbol{\beta}) \equiv \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \delta_i \left\{ \mathbf{X}_i^*(\alpha_v) - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^*(\alpha_v, \boldsymbol{\beta})}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^*(\alpha_v, \boldsymbol{\beta})} \right\} = 0. \tag{11}$$

Let $\mathcal{F} = \{(T_i, \delta_i), i = 1, \ldots, n_A \text{ and } \mathbf{X}_j, j = 1, \ldots, n_B\}$ denote the information related to the duration times and censoring indicators for the units in A, and to the true values of covariates for all the units in B. We have

$$\begin{aligned}
\mathbb{E}\{\bar{H}(\boldsymbol{\beta}) \mid \mathcal{F}\} &= \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \mathbb{E} \left\{ \delta_i \left[\mathbf{X}_i^* - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^*(\alpha_v, \boldsymbol{\beta})}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^*(\alpha_v, \boldsymbol{\beta})} \right] \,\middle|\, \mathcal{F} \right\} \\
&= \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \delta_i \left\{ \underbrace{\mathbb{E}(\mathbf{X}_i^* \mid \mathcal{F})}_{E_1} - \underbrace{\mathbb{E} \left(\frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^*(\alpha_v, \boldsymbol{\beta})}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^*(\alpha_v, \boldsymbol{\beta})} \,\middle|\, \mathcal{F} \right)}_{E_2} \right\}
\end{aligned} \tag{12}$$

For each $i \in A_v$ and $j \in B_v$, let l_{ij} be an indicator equal to 1 if unit i and j are linked,

and to 0 otherwise. Then for each $i \in A_v$, we have $\mathbf{Z}_i = \sum_{j \in B_v} l_{ij} \mathbf{X}_j$, and

$$\mathbb{E}(\mathbf{Z}_i \mid \mathcal{F}) = \sum_{j \in B_v} \mathbf{X}_j \mathbb{E}(l_{ij} \mid \mathcal{F}).$$

Under the non-informative assumption for the linkage process, we obtain from the hit-miss model (4) that

$$\begin{aligned}
\mathbb{E}(l_{ii} \mid \mathcal{F}) &= \alpha_v + (1 - \alpha_v)(n_B)^{-1}, \\
\mathbb{E}(l_{ij} \mid \mathcal{F}) &= (1 - \alpha_v)(n_B)^{-1} \text{ for } j \in B \setminus \{i\},
\end{aligned}$$

which leads to

$$\mathbb{E}(\mathbf{Z}_i \mid \mathcal{F}) = \alpha_v \mathbf{X}_i + (1 - \alpha_v) \bar{\mathbf{X}}_{B_v}$$

From equation (7) and under the non-informative linkage assumption, we have

$$\begin{aligned}
E_1 &= \mathbb{E}\left\{ \alpha_v^{-1} \mathbf{Z}_i - (\alpha_v^{-1} - 1) \bar{\mathbf{X}}_{B_v} \mid \mathcal{F} \right\} \\
&= \alpha_v^{-1} \mathbb{E}\left(\mathbf{Z}_i \mid \mathcal{F} \right) - (\alpha_v^{-1} - 1) \bar{\mathbf{X}}_{B_v} \\
&= \alpha_v^{-1} \left[\alpha_v \mathbf{X}_i + (1 - \alpha_v) \bar{\mathbf{X}}_{B_v} \right] - (\alpha_v^{-1} - 1) \bar{\mathbf{X}}_{B_v} \\
&= \mathbf{X}_i.
\end{aligned} \tag{13}$$

By using a first order Taylor approximation, we have up to negligible factors of order $O_p(n_A^{-1})$:

$$E_2 \approx \frac{\mathbb{E}\left\{ \sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^*(\alpha_v, \boldsymbol{\beta}) \middle| \mathcal{F} \right\}}{\mathbb{E}\left\{ \sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^*(\alpha_v, \boldsymbol{\beta}) \middle| \mathcal{F} \right\}} \tag{14}$$

where

$$\mathbb{E}\left\{\sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)h_j^*(\alpha_v,\boldsymbol{\beta})\middle|\mathcal{F}\right\} = \sum_{v=1}^{V}\sum_{j\in A_v}\mathbb{E}\left\{Y_j(T_i)h_j^*(\alpha_v,\boldsymbol{\beta})\mid\mathcal{F}\right\}$$
$$= \sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)\mathbb{E}\left\{h_j^*(\alpha_v,\boldsymbol{\beta})\mid\mathcal{F}\right\}$$
$$= \sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)h(\boldsymbol{\beta},\mathbf{X}_j).$$

Similarly:

$$\mathbb{E}\left(\sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)g_j^*(\alpha_v,\boldsymbol{\beta})\middle|\mathcal{F}\right) = \sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)g(\boldsymbol{\beta},\mathbf{X}_j).$$

Therefore,

$$E_2 \approx \frac{\sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)h(\beta,\mathbf{X}_j)}{\sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)g(\beta,\mathbf{X}_j)} \tag{15}$$

By plugging (13) and (15) into (12), we obtain

$$\mathbb{E}\left\{\bar{H}(\boldsymbol{\beta})\middle|\mathcal{F}\right\} \approx \frac{1}{n_A}\sum_{v=1}^{V}\sum_{i\in A_v}\delta_i\left\{\mathbf{X}_i - \frac{\sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)h(\boldsymbol{\beta},\mathbf{X}_j)}{\sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)g(\boldsymbol{\beta},\mathbf{X}_j)}\right\} \tag{16}$$
$$= H_0(\boldsymbol{\beta}).$$

B Variance estimation for the proposed adjusted estimator

In this appendix, the derivation of the variance estimator is explained. For simplicity, we focus on the case $V = 1$ when a single block is used. The extension to multiple blocks is straightforward.

We first recall the main notations. A database B of size n_B is first obtained, and the covariates $\mathbf{X}_i$ are observed for all the units in B. We use the notations

$$\begin{aligned}
\bar{\mathbf{X}}_B &= \frac{1}{n_B}\sum_{i=1}^{n_B}\mathbf{X}_i, \\
\bar{g}_B(\boldsymbol{\beta}) &= \frac{1}{n_B}\sum_{i=1}^{n_B} g(\boldsymbol{\beta},\mathbf{X}_i), \\
\bar{h}_B(\boldsymbol{\beta}) &= \frac{1}{n_B}\sum_{i=1}^{n_B} h(\boldsymbol{\beta},\mathbf{X}_i).
\end{aligned}$$

We also note $\mathbf{X}_B \equiv \{\mathbf{X}_i\}_{i\in B}$ for the set of auxiliary variables in B.

A subsample A of size n_A is then selected in B, and the variable T_i is obtained for any unit $i \in A$. We note $T_A \equiv \{T_i\}_{i\in A}$ for the set of outcome values in A. The auxiliary variables are obtained in A by using record linkage, leading to the pseudo auxiliary variables $\mathbf{Z}_i$ for any unit $i \in A$. We note $\mathbf{Z}_A \equiv \{\mathbf{Z}_i\}_{i\in A}$ for the set of pseudo values in A.

Finally, a validation sample V of size n_V is selected in A by simple random sampling, and the true auxiliary variables $\mathbf{X}_i$ are obtained for the units $i \in V$. By comparing the pseudo values $\mathbf{Z}_i$ and the true values $\mathbf{X}_i$ in V, we obtain an unbiased estimator $\hat{\alpha}$ for the parameter α.

B.1 Global estimating equation

Using the unbiased estimator $\hat{\alpha}$ for the parameter α (see equation 4), the global estimating equation for the parameter β is

$$\bar{H}(\boldsymbol{\beta}) \equiv \frac{1}{n_A}\sum_{i=1}^{n_A} \underbrace{\delta_i \left\{ \mathbf{X}_i^*(\hat{\alpha}) - \frac{\sum_{j=1}^{n_A} Y_j(T_i) h_j^*(\hat{\alpha},\boldsymbol{\beta})}{\sum_{j=1}^{n_A} Y_j(T_i) g_j^*(\hat{\alpha},\boldsymbol{\beta})} \right\}}_{H_i(\boldsymbol{\beta})} = 0, \tag{17}$$

where

$$\begin{aligned} \mathbf{X}_i^*(\hat{\alpha}) &= \frac{\mathbf{Z}_i}{\hat{\alpha}} - \frac{1-\hat{\alpha}}{\hat{\alpha}}\bar{\mathbf{X}}_B, \\ g_j^*(\hat{\alpha}, \boldsymbol{\beta}) &= \frac{g(\boldsymbol{\beta}, \mathbf{Z}_j)}{\hat{\alpha}} - \frac{1-\hat{\alpha}}{\hat{\alpha}}\bar{g}_B(\boldsymbol{\beta}), \\ h_j^*(\hat{\alpha}, \boldsymbol{\beta}) &= \frac{h(\boldsymbol{\beta}, \mathbf{Z}_i)}{\hat{\alpha}} - \frac{1-\hat{\alpha}}{\hat{\alpha}}\bar{h}_B(\boldsymbol{\beta}). \end{aligned} \tag{18}$$

Let us denote by $\boldsymbol{\beta}_0$ the true value of the parameter. Then we have

$$\bar{H}(\hat{\boldsymbol{\beta}}) - \bar{H}(\boldsymbol{\beta}_0) = -\bar{H}(\boldsymbol{\beta}_0) \simeq \{\mathbb{E}\nabla\bar{H}(\boldsymbol{\beta}_0)\}\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\},$$

with $\nabla\bar{H}(\boldsymbol{\beta})$ the differential of $\bar{H}(\boldsymbol{\beta})$. We obtain

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \simeq -\{\mathbb{E}\nabla\bar{H}(\boldsymbol{\beta}_0)\}^{-1} \times \bar{H}(\boldsymbol{\beta}_0).$$

It is thus sufficient to obtain a variance estimator for $\bar{H}(\beta_0)$, from which we can use the sandwich variance estimator

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}) = \{\nabla\bar{H}(\hat{\boldsymbol{\beta}})\}^{-1} \times \hat{\mathbb{V}}\{\bar{H}(\boldsymbol{\beta}_0)\} \times \{\nabla\bar{H}(\hat{\boldsymbol{\beta}})\}^{-1}. \tag{19}$$

The derivation of $\hat{\mathbb{V}}\{\bar{H}(\boldsymbol{\beta}_0)\}$ is explained in the next sections.

B.2 Accounting for the estimation of α

Since we have

$$\begin{aligned} \frac{1}{\hat{\alpha}} = \frac{1}{\alpha} \times \frac{1}{1 + \frac{\hat{\alpha}-\alpha}{\alpha}} &= \frac{1}{\alpha}\left[1 - \frac{\hat{\alpha}-\alpha}{\alpha} + o_p({n_V}^{-0.5})\right] \\ &= \frac{1}{\alpha} - \frac{\hat{\alpha}-\alpha}{\alpha^2} + o_p({n_V}^{-0.5}), \end{aligned}$$

we may rewrite the quantities in (18) as

$$\begin{aligned}
\mathbf{X}_i^*(\hat{\alpha}) &= \underbrace{\frac{1}{\alpha}(\mathbf{Z}_i - \bar{\mathbf{X}}_B) + \bar{\mathbf{X}}_B}_{\mathbf{x}_i^*(\alpha)} \\
&\quad - \frac{\hat{\alpha}-\alpha}{\alpha^2}(\mathbf{Z}_i - \bar{\mathbf{X}}_B) + o_p({n_V}^{-0.5}), \\
g_j^*(\hat{\alpha}, \boldsymbol{\beta}_0) &= \underbrace{\frac{1}{\alpha}\left\{g(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{g}_B(\boldsymbol{\beta}_0)\right\} + \bar{g}_B(\boldsymbol{\beta}_0)}_{g_j^*(\alpha, \boldsymbol{\beta}_0)} \\
&\quad - \frac{\hat{\alpha}-\alpha}{\alpha^2}\left\{g(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{g}_B(\boldsymbol{\beta}_0)\right\} + o_p({n_V}^{-0.5}),
\end{aligned} \tag{20}$$

$$\begin{aligned}
h_j^*(\hat{\alpha}, \boldsymbol{\beta}_0) &= \underbrace{\frac{1}{\alpha}\left\{h(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{h}_B(\boldsymbol{\beta}_0)\right\} + \bar{h}_B(\boldsymbol{\beta}_0)}_{h_j^*(\alpha, \boldsymbol{\beta}_0)} \\
&\quad - \frac{\hat{\alpha}-\alpha}{\alpha^2}\left\{h(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{h}_B(\boldsymbol{\beta}_0)\right\} + o_p({n_V}^{-0.5}).
\end{aligned} \tag{21}$$

Let us denote $\epsilon = \frac{\hat{\alpha}-\alpha}{\alpha^2}$. By plugging (20) and (21) into equation (17), we have

$$\begin{aligned}
\frac{\sum_{j=1}^{n_A} Y_j(T_i) h_j^*(\hat{\alpha}, \boldsymbol{\beta})}{\sum_{j=1}^{n_A} Y_j(T_i) g_j^*(\hat{\alpha}, \boldsymbol{\beta})} &= \frac{\sum_{j=1}^{n_A} Y_j(T_i) h_j^*(\alpha, \boldsymbol{\beta}) - \epsilon \sum_{j=1}^{n_A} Y_j(T_i)\left[h(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{h}_B(\boldsymbol{\beta}_0)\right]}{\sum_{j=1}^{n_A} Y_j(T_i) g_j^*(\alpha, \boldsymbol{\beta}) - \epsilon \sum_{j=1}^{n_A} Y_j(T_i)\left[g(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{g}_B(\boldsymbol{\beta}_0)\right]} \\
&\quad + o_p({n_V}^{-0.5}).
\end{aligned}$$

After some algebra, this leads to:

$$\bar{H}(\boldsymbol{\beta}_0) \;=\; \bar{H}_1(\boldsymbol{\beta}_0) - \left(\frac{\hat{\alpha}-\alpha}{\alpha^2}\right)\bar{H}_2(\alpha, \boldsymbol{\beta}_0) + o_p({n_V}^{-0.5}), \tag{22}$$

where

$$\bar{H}_1(\boldsymbol{\beta}_0) \;=\; \frac{1}{n_A}\sum_{i=1}^{n_A} \underbrace{\delta_i\left\{\mathbf{X}_i^*(\alpha) - R_i^*(\alpha, \boldsymbol{\beta}_0)\right\}}_{H_{1i}(\boldsymbol{\beta}_0)} \tag{23}$$

with $R_i^*(\alpha, \boldsymbol{\beta}_0) = \frac{\sum_{j=1}^{n_A} Y_j(T_i) h_j^*(\alpha,\boldsymbol{\beta})}{\sum_{j=1}^{n_A} Y_j(T_i) g_j^*(\alpha,\boldsymbol{\beta})}$, and with

$$\bar{H}_2(\alpha, \boldsymbol{\beta}_0) = \frac{1}{n_A} \sum_{i=1}^{n_A} \delta_i \Bigg[(\mathbf{Z}_i - \bar{\mathbf{X}}_B) - \frac{\sum_{j=1}^{n_A} Y_j(T_i) \left\{ \left[h(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{h}_B(\boldsymbol{\beta}_0) \right] - R_i^*(\alpha, \boldsymbol{\beta}_0) \left[g(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{g}_B(\boldsymbol{\beta}_0) \right] \right\}}{\sum_{j=1}^{n_A} Y_j(T_i) g_j^*(\alpha, \boldsymbol{\beta}_0)} \Bigg]. \quad (24)$$

By neglecting the terms which are $o_p({n_V}^{-0.5})$, we obtain from (22) that

$$\begin{aligned} \mathbb{V}\left[\bar{H}(\boldsymbol{\beta}_0)\right] &= \mathbb{V}\left[\mathbb{E}\left\{ \bar{H}(\boldsymbol{\beta}_0) \middle| \mathbf{X}_B, T_A, \mathbf{Z}_A \right\}\right] + \mathbb{E}\left[\mathbb{V}\left\{ \bar{H}(\boldsymbol{\beta}_0) \middle| \mathbf{X}_B, T_A, \mathbf{Z}_A \right\}\right] \\ &\simeq \mathbb{V}\left[\bar{H}_1(\boldsymbol{\beta}_0)\right] + \mathbb{E}\left[\bar{H}_2(\boldsymbol{\beta}_0) \mathbb{V}\left\{ \frac{\hat{\alpha} - \alpha}{\alpha^2} \middle| \mathbf{X}_B, T_A, \mathbf{Z}_A \right\} \{\bar{H}_2(\boldsymbol{\beta}_0)\}^\top \right]. \end{aligned} \quad (25)$$

Under the assumption that the validation sample S_V is selected in A by simple random sampling without replacement, we have

$$\hat{\alpha} = \frac{1}{n_V} \sum_{i \in S_V} \mu_i \quad \text{where} \quad \mu_i = \begin{cases} 1 & \text{if linkage is correct,} \\ 0 & \text{otherwise.} \end{cases}$$

Since μ_i is a binary variable, it follows from standard results in survey sampling theory that an unbiased estimator for $\mathbb{V}\left\{ \hat{\alpha} \middle| \mathbf{X}_B, T_A, \mathbf{Z}_A \right\}$ is

$$\hat{\mathbb{V}}(\hat{\alpha}) = \left(\frac{1}{n_V} - \frac{1}{n_A} \right) \frac{n_V}{n_V - 1} \hat{\alpha}(1 - \hat{\alpha}).$$

Hence the second term in the right-hand side of (25) may be estimated by

$$\hat{\mathbb{V}}_1\left[\bar{H}(\boldsymbol{\beta}_0)\right] \;=\; \bar{H}_2(\hat{\alpha}, \hat{\boldsymbol{\beta}}) \{\bar{H}_2(\hat{\alpha}, \hat{\boldsymbol{\beta}})\}^\top \times \left(\frac{1}{n_V} - \frac{1}{n_A} \right) \frac{n_V}{n_V - 1} \frac{1 - \hat{\alpha}}{\hat{\alpha}^3}, \quad (26)$$

where $\bar{H}_2(\hat{\alpha},\hat{\boldsymbol{\beta}})$ is obtained from (24) by replacing $\boldsymbol{\beta}_0$ with $\hat{\boldsymbol{\beta}}$ and α with $\hat{\alpha}$. This is the component of the variance estimator which accounts for the estimation of α.

B.3 Accounting for the linkage and estimation error

In this section, we focus on the first term in the right-hand side of (25). We have

$$\mathbb{V}\left[\bar{H}_1(\boldsymbol{\beta}_0)\right] = \mathbb{V}\left[\mathbb{E}\left\{\bar{H}_1(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A\right\}\right] + \mathbb{E}\left[\mathbb{V}\left\{\bar{H}_1(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A\right\}\right]. \tag{27}$$

It follows from equation (16) in Appendix A that

$$\mathbb{E}\left\{\bar{H}_1(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A\right\} \simeq \frac{1}{n_A}\sum_{i=1}^{n_A}\underbrace{\delta_i\left\{\mathbf{X}_i - \frac{\sum_{j=1}^{n_A} Y_j(T_i)h(\boldsymbol{\beta}_0, \mathbf{X}_j)}{\sum_{j=1}^{n_A} Y_j(T_i)g(\boldsymbol{\beta}_0, \mathbf{X}_j)}\right\}}_{H_{ti}(\boldsymbol{\beta}_0)}, \tag{28}$$

which is the function associated to the theoretical estimating equation that we would solve if the covariates $\mathbf{X}_i$ were known without linkage error for the units $i \in A$. Secondly, note that conditionally on $\mathbf{X}_B$ and T_A, the terms $H_{1i}(\boldsymbol{\beta}_0)$ are approximately uncorrelated for $i = 1, \ldots, n_A$. More precisely, it can be proved after some algebra that for any $i \neq j = 1, \ldots, n_A$, we have

$$Cov\left(\delta_i\left\{\mathbf{X}_i^*(\alpha) - R_i^*(\alpha, \boldsymbol{\beta}_0)\right\}, \delta_j\left\{\mathbf{X}_j^*(\alpha) - R_j^*(\alpha, \boldsymbol{\beta}_0)\right\}\middle|\mathbf{X}_B, T_A\right) = O_p(n_A^{-1}).$$

Therefore, we obtain that

$$\mathbb{V}\left\{\bar{H}_1(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A\right\} \simeq \frac{1}{(n_A)^2}\sum_{i=1}^{n_A}\mathbb{V}\left\{H_{1i}(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A\right\}. \tag{29}$$

where $H_{1i}(\cdot)$ is defined in (23). From (27), (28) and (29), we obtain that

$$\mathbb{V}\left[\bar{H}_1(\boldsymbol{\beta}_0)\right] \simeq \mathbb{V}\left(\frac{1}{n_A}\sum_{i=1}^{n_A} H_{ti}(\boldsymbol{\beta}_0)\right) + \mathbb{E}\left[\frac{1}{(n_A)^2}\sum_{i=1}^{n_A}\mathbb{V}\left\{H_{1i}(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A\right\}\right]. \tag{30}$$

Now, we consider the sample dispersion term given by

$$\begin{aligned} s_H^2(\boldsymbol{\beta}_0) &= \frac{1}{n_A - 1}\sum_{i=1}^{n_A}\left\{H_i(\boldsymbol{\beta}_0) - \frac{1}{n_A}\sum_{j=1}^{n_A} H_j(\boldsymbol{\beta}_0)\right\}^2 \\ &= \frac{1}{2n_A(n_A-1)}\sum_{i\neq j=1}^{n_A}\left\{H_i(\boldsymbol{\beta}_0) - H_j(\boldsymbol{\beta}_0)\right\}^2. \end{aligned} \tag{31}$$

where $H_i(\cdot)$ is defined in (17).

We have

$$\begin{aligned}
\mathbb{E}\left\{\frac{s_H^2(\boldsymbol{\beta}_0)}{n_A}\right\} &= \mathbb{E}\mathbb{E}\left\{\frac{s_H^2(\boldsymbol{\beta}_0)}{n_A}\middle|\, \mathbf{X}_B, T_A\right\} \qquad (32)\\
&= \mathbb{E}\left[\frac{1}{2n_A^2(n_A-1)}\sum_{i\neq j=1}^{n_A}\mathbb{E}\{\,H_i(\boldsymbol{\beta}_0)-H_j(\boldsymbol{\beta}_0)|\,\mathbf{X}_B, T_A\}^2\right]\\
&+ \mathbb{E}\left[\frac{1}{2n_A^2(n_A-1)}\sum_{i\neq j=1}^{n_A}\mathbb{V}\{\,H_i(\boldsymbol{\beta}_0)-H_j(\boldsymbol{\beta}_0)|\,\mathbf{X}_B, T_A\}\right]\\
&\simeq \mathbb{E}\left[\frac{1}{2n_A^2(n_A-1)}\sum_{i\neq j=1}^{n_A}\{H_{ti}(\boldsymbol{\beta}_0)-H_{tj}(\boldsymbol{\beta}_0)\}^2\right]\\
&\quad (\text{where } H_{ti}(\cdot) \text{ is defined in (28)})\\
&+ \mathbb{E}\left[\frac{1}{2n_A^2(n_A-1)}\sum_{i\neq j=1}^{n_A}\mathbb{V}\{\,H_i(\boldsymbol{\beta}_0)|\,\mathbf{X}_B, T_A\}+\mathbb{V}\{\,H_j(\boldsymbol{\beta}_0)|\,\mathbf{X}_B, T_A\}\right]\\
&= \mathbb{E}\left[\frac{1}{n_A(n_A-1)}\sum_{i=1}^{n_A}\left\{H_{ti}(\boldsymbol{\beta}_0)-\frac{1}{n_A}\sum_{j=1}^{n_A}H_{tj}(\boldsymbol{\beta}_0)\right\}^2\right.\\
&\quad \left.+\frac{1}{n_A^2}\sum_{i=1}^{n_A}\mathbb{V}\{\,H_i(\boldsymbol{\beta}_0)|\,\mathbf{X}_B, T_A\}\right]\\
&\simeq \mathbb{V}\left[\bar{H}_1(\boldsymbol{\beta}_0)\right],
\end{aligned}$$

where the last line in (32) follows from a comparison with equation (30). Therefore, $\mathbb{V}\left[\bar{H}_1(\boldsymbol{\beta}_0)\right]$ may be approximately unbiasedly estimated by replacing in (31) the unknown parameter $\boldsymbol{\beta}_0$ with $\hat{\boldsymbol{\beta}}$, which leads to

$$\hat{\mathbb{V}}_2\left[\bar{H}(\boldsymbol{\beta}_0)\right] \;=\; \frac{s_H^2(\hat{\boldsymbol{\beta}})}{n_A}. \qquad (33)$$

This is the component of the variance estimator, which accounts for both the linkage and estimation errors.

B.4 Global variance estimator

By plugging (26) and (33) into (25), we obtain:

$$\hat{\mathbb{V}}\{\bar{H}(\boldsymbol{\beta}_0)\} = \hat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\} + \hat{\mathbb{V}}_2\{\bar{H}(\boldsymbol{\beta}_0)\}.$$

The global variance estimator is therefore obtained from (19) as:

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}) \;=\; \{\nabla\bar{H}(\hat{\boldsymbol{\beta}})\}^{-1} \times \left\{\hat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\} + \hat{\mathbb{V}}_2\{\bar{H}(\boldsymbol{\beta}_0)\}\right\} \times \{\nabla\bar{H}(\hat{\boldsymbol{\beta}})\}^{-1} \tag{34}$$

C Sampling affectations

To compute $\bar{\mathbf{X}}_{B_v}$, $\bar{g}_{B_v}(\boldsymbol{\beta})$ and $\bar{h}_{B_v}(\boldsymbol{\beta})$ in (7), the `AEE` requires access to all the $\mathbf{X}$-vectors in B. In some cases, this may not be possible due to confidentiality reasons. In that case, we have access to only the linked dataset A. In this situation, we propose to approximate

$$\begin{aligned}
\bar{\mathbf{X}}_{B_v} \quad &\text{with} \quad \bar{\mathbf{Z}}_{A_v} = \frac{1}{n_{A_v}} \sum_{i \in A_v} \mathbf{Z}_i, \\
\bar{g}_{B_v}(\boldsymbol{\beta}) \quad &\text{with} \quad \bar{g}_{A_v}(\boldsymbol{\beta}) = \frac{1}{n_{A_v}} \sum_{i \in A_v} \exp(\mathbf{Z}_i^\top \boldsymbol{\beta}), \\
\bar{h}_{B_v}(\boldsymbol{\beta}) \quad &\text{with} \quad \bar{h}_{A_v}(\boldsymbol{\beta}) = \frac{1}{n_{A_v}} \sum_{i \in A_v} \exp(\mathbf{Z}_i^\top \boldsymbol{\beta}) \mathbf{Z}_i.
\end{aligned} \tag{35}$$

If A_v is a random sample of B_v, (35) can be a good approximation. A simulation study is presented in Table S.4 of the Supplementary material.

Data availability statement

Our R programs for simulation results are available at Github, https://github.com/thanhhuanVO/Cox-regression-with-linked-data. The data used in the application of

the methods may be obtained from a third party and are not publicly available. For all interested researchers, data are available via SNDS, https://www.snds.gouv.fr/SNDS/Accueil, subject to the authorization of CNIL (National Commission on Informatics and Liberty of France).

References

1. Harron K, Gilbert R, Cromwell D, Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. *PLOS ONE.* 2016;11(10):1-18.

2. Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. *European Journal of Epidemiology.* 2018;34:91 - 99.

3. Zhang G, Campbell P. Data Survey: Developing the Statistical Longitudinal Census Dataset and Identifying Its Potential Uses. *Aust Econ Rev.* 2012;45(1):125 - 133.

4. Winkler W, Thibaudeau Y. An Application Of The Fellegi-Sunter Model Of Record Linkage To The 1990 U.S. Decennial Census. in *Technical report, US Bureau of the Census* 1987.

5. Herzog T, Scheuren F, Winkler W. *Data Quality and Record Linkage Techniques.* Springer-Verlag New York 2007.

6. Christen P. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Springer Publishing Company, Incorporated 2012.

7. Fellegi I, Sunter A. A Theory for Record Linkage. *Journal of the American Statistical Association.* 1969;64:1183-1210.

8. Winkler WE. Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. in *Proceedings of the Section on Survey Research Methods, American Statistical Association*:667–671 1988.

9. Vo TH, Chauvet G, Happe A, Oger E, Paquelet S, Garès V. Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system. *Computational Statistics & Data Analysis.* 2023;179:107656.

10. Tancredi A, Liseo B. A hierarchical Bayesian approach to matching and size population problems. *Annals of Applied Statistics.* 2010;5.

11. Sadinle M. Bayesian Estimation of Bipartite Matchings for Record Linkage. *Journal of the American Statistical Association.* 2017;112(518):600-612.

12. Neter J, Maynes ES, Ramanathan R. The Effect of Mismatching on the Measurement of Response Errors. *J Am Stat Assoc.* 1965;60(312):1005-1027.

13. Scheuren F, Winkler W. Regression Analysis of Data Files that are Computer Matched. *Survey Methodology.* 1993;19.

14. Lahiri P, Larsen MD. Regression Analysis with Linked Data. *Journal of the American Statistical Association.* 2005;100(469):222–230.

15. Hof M, Zwinderman A. Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Statistics in medicine.* 2012;31:4231–4242.

16. Han Y, Lahiri P. Statistical Analysis with Linked Data. *International Statistical Review.* 2019;87(S1):S139-S157.

17. Zhang LC. On secondary analysis of datasets that cannot be linked without errors. in *Analysis of integrated data*London: CRC/Chapman and Hall 2019.

18. Chambers R. Regression Analysis of Probability-Linked Data. *Statistics New Zealand.* 2009.

19. Kim G, Chambers R. Regression Analysis under Probabilistic Multi-Linkage. *Statistica Neerlandica.* 2012;66.

20. Kim G, Chambers R. Regression analysis under incomplete linkage. *Computational Statistics & Data Analysis.* 2012;56(9):2756 - 2770.

21. Chambers R, Kim G. *Secondary analysis of linked data*ch. 5, :83-108. John Wiley & Sons, Ltd 2015.

22. Chambers R, Silva A. Improved secondary analysis of linked data: a framework and an illustration. *J R Stat Soc Ser A Stat Soc.* 2020;183.

23. Zhang LC, Tuoto T. Linkage-data linear regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* 2020;100:222–230.

24. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological).* 1972;34(2):187–220.

25. Baldi I, Ponti A, Zanetti R, Ciccone G, Merletti F, Gregori D. The impact of record inkage bias in the Cox model. *J Eval Clin Pract.* 2010;16:92-6.

26. Hof M, Ravelli A, Zwinderman A. A Probabilistic Record Linkage Model for Survival Data. *Journal of the American Statistical Association.* 2017;112(520):1504-1515.

27. Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics.* 1982;10(4):1100–1120.

28. Chambers R, Salvati N, E.Fabrizi , Silva AD. Domain estimation under informative linkage. *Statistical Theory and Related Fields.* 2019;3(2):90-102.

29. Copas JB, Hilton FJ. Record Linkage: Statistical Models for Matching Computer Records. *J R Stat Soc Ser A Stat Soc.* 1990;153(3):287–320.

30. Goldstein H, Harron K, Wade A. The analysis of record-linked data using multiple imputation with data value priors. *Statistics in Medicine.* 2012;31(28):3481-3493.

31. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005;24(11):1713-1723.