Record linkage and analysis of linked data with application in French national health data system

Thanh Huan Vo

THESE DE DOCTORAT DE

L'INSTITUT NATIONAL DES SCIENCES APPLIQUEES RENNES

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Mathématiques et leurs Interactions*

Par

Thanh Huan VO

Record linkage and analysis of linked data with application in French national health data system

Thèse présentée et soutenue à l'INSA Rennes, le 01 Décembre 2022
Unité de recherche: IRMAR-UMR CNRS 6625
Thèse N° : 22ISAR 32 / D22 - 32

Rapporteurs avant soutenance:

Michael LARSEN	Professor, Saint Michael's College
Nicola SALVATI	Associate professor, University of Pisa

Composition du Jury:

Président du jury:	Jean-François DUPUY	Professeur, INSA Rennes
Examinateurs:	Michael LARSEN	Professor, Saint Michael's College
	Nicola SALVATI	Associate professor, University of Pisa
	Tiziana TUOTO	Researcher, Italian National Institute of Statistics
	Nicolas COURTY	Professeur, Université de Bretagne Sud
	Cécile CHEVRIER	Directrice de recherche, Inserm
Dir. de thèse:	Guillaume CHAUVET	Enseignant-chercheur, ENSAI
Co-enc. de thèse:	Valérie GARES	Maître de conférences, INSA Rennes

Acknowledgements

Three years ago, I started my PhD at INSA Rennes, France. Time flies and I am coming to an end of my PhD adventure. I would like to take this opportunity to thank you all who have been with me on this awesome journey.

First and foremost, I would like to express my sincere gratitude to my patient and supportive supervisors, Guillaume CHAUVET and Valérie GARES. Their guidance has helped me step by step to achieve the results of this research. Without their supports, I would not be able to complete the thesis. I am also thankful to Laurent LAUNAY and Stéphane PAQUELET who gave a lot of effort to help us on initiating this project and the administration procedure with b<>com. Thank you Stéphane for giving me valuable comments on developing and presenting this work. Besides, I would like to thank André HAPPE and Emmanuel OGER for their helpful discussion on the application of the projects. Special thanks to Li-Chun ZHANG for joining us on the second part of the project.

I greatly appreciate Michael LARSEN and Nicola SALVATI, the two reviewers of this thesis, who have been so kind to spend time to read, evaluate my works and give me valuable comments to improve it. I am also grateful to Jean-François DUPUY, Tiziana TUOTO, Nicolas COURTY and Cécile CHEVRIER for agreeing to be in my thesis committee.

Thanks should also go to all of my very nice colleagues in both research teams at IRMAR-INSA Rennes and at IRT b<>com. I would like to thank Mounir, Olivier and Camar-Eddine for always encouraging me during my PhD. I want to give a special thank to Justine, Patricia and Daviane for the enthusiastic administrative help. I also thank my friends Mériadec, El-Hassene and Majd for the motivational discussion. Many thanks to Bilel for supporting me throughout this journey.

Besides, I have also been happy to meet and make friends with many kind Vietnamese people in Rennes. Special thanks must go to my brother Van Trinh and my sister Minh Phuong for taking care of me from my very first days in Rennes. Thank you all my friends, brothers and sisters in "mathematics group" and "foot-

ball team" who have made this adventure unique. Especially, I am extremely grateful to brothers and sisters in "Mirabeau group" for their sharing and inspiration. The moments may have ended but the memories last forever.

Last but not least, I would like to express my deepest gratitude to my parents and my family in Vietnam who have supported me unconditionally and constantly from distance. Finally, I give a special thanks to my wife for her encouragement and endless love. I am grateful to have her on this journey with me.

Rennes, 01 December 2022

Thanh Huan VO

Acronyms/Abbreviations

AVC	Accident Vasculaire Cérébral (Stroke)
EM	Expectation Maximization
ECM	Expectation Conditional Maximization
GETBO	Groupe d'Étude de la Thrombose de Bretagne Occidental
	(Western Brittany Thrombosis Study Group)
i.i.d.	independent and identically distributed
PPV	Positive Predictive Value
SNDS	Système National des Données de Sante
	(National Health Data System of France)
TPR	True Positive Rate
VTE	Venous Thromboembolism

List of Figures

List of Tables

Contents

1 Introduction

De nos jours, de plus en plus de données sont collectées dans tous les domaines tels que l'administration publique, la finance, le commerce et la médecine. La capacité d'analyser ces ensembles de données peut améliorer les produits d'assurance, fournir un avantage concurrentiel pour un service commercial et des connaissances nouvelles pour les chercheurs. Cependant, dans de nombreuses situations, toutes les informations nécessaires à l'analyse peuvent ne pas être contenues dans une seule base de données, tandis que la collecte de variables supplémentaires est fastidieuse, longue et coûteuse. Le couplage d'enregistrements, aussi appelé appariement de données, est un processus qui consiste à combiner des données provenant de différentes bases et se rapportant à la même entité. C'est une méthode très utilisée pour enrichir, mettre à jour ou améliorer les informations stockées dans des bases de données, pour étudier la relation entre des variables provenant de différentes sources, ou encore pour éliminer les doublons dans une base de données.

Avec la disponibilité croissante de grandes bases de données sur les soins de santé provenant par exemple de sources administratives, la demande pour que ces bases soient fusionnées est également en augmentation, avec pour objectif d'améliorer les systèmes de santé. Par exemple, GETBO (Groupe d'Etude de la Thrombose de Bretagne Occidentale) est un registre des cas de thromboembolie veineuse (VTE) survenues entre 2013 et 2015 à Brest (France). Dans ce registre ne sont enregistrés que les informations démographiques de chaque patient (date de naissance, sexe, code de résidence) ainsi que quelques dates et types d'actes médicaux. Ces informations sont insuffisantes pour construire un modèle d'analyse tel qu'un modèle de prédiction permettant d'identifier précocement les VTE symptomatiques (Noboa et al., 2006). Le Système National des Données de Santé (SNDS) français collecte tous les dossiers de santé longitudinaux et les informations d'assurance pour la plus grande partie de la population française (Bezin et al., 2017). Les données du SNDS peuvent être utilisées pour enrichir le registre, ce qui devrait permettre aux chercheurs d'acquérir des connaissances nouvelles et précieuses. Cela nous motive à relier les données GETBO et les données SNDS afin de pouvoir obtenir plus d'informations médicales sur les patients pour l'analyse souhaitée.

Le processus d'appariement est simple si nous pouvons accéder aux identifiants uniques des patients. Cependant, l'utilisation de cet identifiant peut ne pas être autorisée pour des raisons éthiques, ou un tel identifiant peut simplement ne pas être disponible. Par conséquent, nous pouvons uniquement utiliser d'autres identifiants partiels qui sont communs à ces bases de données, tels que le sexe, le code postal ou les dates des traitements médicaux, pour identifier les paires appariées des deux bases de données. Ces variables sont souvent appelées variables d'appariement. Deux approches courantes pour l'appariement des données sont les méthodes déterministe et probabiliste de couplage d'enregistrements. Les méthodes déterministes exigent que toutes les variables d'appariement d'une paire appariée soient les mêmes. Bien que cette approche soit rapide et facile à mettre en œuvre, un grand nombre de paires appariées peuvent être manquées si des identifiants uniques ne sont pas disponibles et s'il y a des erreurs dans les données d'appariement. Dans ce contexte, le processus de couplage probabiliste des enregistrements s'avère être une approche intéressante (Zhu et al., 2015). Elle vise à fournir à chaque paire d'enregistrements un score ou une probabilité de correspondance. Fellegi and Sunter (1969) ont proposé un cadre mathématique qui est à la base de la plupart des méthodes de couplage probabiliste d'enregistrements, aujourd'hui encore. Dans ce cadre, nous calculons les scores de concordance pour toutes les paires d'enregistrements candidats sur la base de leurs vecteurs de comparaison. Le vecteur de comparaison d'une paire d'enregistrements est un vecteur binaire qui ne prend en compte que l'accord ou le désaccord des variables d'appariement.

1.1 Extension du modèle d'appariement de Fellegi-Sunter à des données mixtes - Application au SNDS

Bien que cette méthode soit largement utilisée et développée depuis plusieurs décennies, la comparaison binaire présente certaines limites. Tout d'abord, la simple comparaison binaire ne suffit pas à rendre compte des caractéristiques des variables d'appariement à faible prévalence, comme les codes de diagnostic de cancer (Hejblum et al., 2019). Par exemple, deux patients atteints tous deux d'un cancer du poumon auront probablement plus de chances d'être appariés que deux patients sans cancer. Cependant, la comparaison binaire conduit à la même probabilité de correspondance pour les deux cas. Deuxièmement, cette comparaison binaire n'est pas en mesure de tenir compte des tolérances dans les variables d'appariement con-

tinues, telles que le retard dans la date des traitements déclarés dans différents secteurs en raison de la procédure administrative. Ce sont les types de variables d'appariement les plus courants dans les données de santé ainsi que dans les bases de données GETBO ou SNDS.

Par conséquent, nous avons étendu le modèle de Fellegi-Sunter de manière à ce qu'il soit plus performant pour l'utilisation de différents types de variables d'appariement pour identifier les paires d'enregistrements. Deux approches de comparaison sont proposées. Pour les variables d'enregistrements catégorielles à faible prévalence, l'approche de comparaison proposée permet de distinguer la concordance des valeurs de faible prévalence avec les autres. Pour les variables de correspondance continues, la fonction de comparaison proposée peut tenir compte de décalages. Parallèlement à la comparaison proposée, la distribution hurdle gamma est pour la première fois utilisée pour modéliser des valeurs de comparaison continues. Les simulations montrent que le modèle étendu est plus performant que le modèle standard dans la plupart des scénarios. Ce modèle permet d'améliorer l'appariement entre les bases de données GETBO et SNDS. Il s'agit de la première contribution de la thèse.

Contributions du Chapitre:

- Extension du modèle de couplage d'enregistrements de Fellegi-Sunter pour des variables d'appariement de différents types.
- Deux nouvelles approches de comparaison pour des variables d'appariement catégorielles à faible prévalence (par exemple, code de diagnostic), et pour des variables d'appariement continues (par exemple, date des actes médicaux).
- Utilisation d'une distribution "hurdle gamma" pour modéliser des valeurs de comparaison continues.
- Application aux bases de données GETBO et SNDS.

1.2 Régression de Cox avec des données appariées

Dans la plupart des applications, le but du couplage d'enregistrements est d'obtenir un ensemble de données pour réaliser une analyse statistique. Dans certains cas, le couplage et l'analyse sont faites par la même équipe de personnes, et toute

l'information sur l'appariement est disponible et peut être utilisée dans l'analyse statistique. On parle alors d'analyse primaire ("primary analysis") du fichier de données apparié. Dans d'autres cas, l'appariement est confié à un tiers de confiance, et les équipes réalisant l'analyse statistique n'ont qu'une connaissance limitée du processus d'appariement, et en particulier les variables d'appariement ne sont pas connues. On parle alors d'analyse secondaire ("secondary analysis") du fichier de données apparié. Dans le cas d'une analyse secondaire, si des identifiants uniques ne sont pas accessibles et si les variables d'appariement contiennent des erreurs, alors les faux liens sont presque inévitables, quelles que soient les méthodes d'appariement. Par conséquent, les analystes secondaires qui travaillent avec des données appariées doivent être conscients de ces erreurs et choisir une stratégie d'analyse appropriée.

Neter et al. (1965) ont insisté sur les risques possibles de biais liés aux erreurs d'appariement, qui sont de deux types : les faux liens et les liens manqués. Les faux liens sont des enregistrements considérés comme liés, alors qu'ils ne se réfèrent pas réellement à la même entité. Les liens manqués sont des enregistrements qui se réfèrent à la même entité, mais qui ne sont identifiés comme étant liés lors de l'appariement. De nombreux auteurs (e.g. Lahiri and Larsen, 2005; Chambers, 2009; Zhang and Tuoto, 2021; Chambers et al., 2022) se sont penchés sur ce problème et ont proposé différentes méthodes pour tenir compte de ces erreurs de liaison. Cependant, elles sont principalement conçues pour les modèles de régression linéaire et logistique. Dans la recherche médicale, le modèle à risques proportionnels de Cox (1972) est l'un des plus utilisés pour étudier la relation entre des variables et une durée de vie. Cependant, dans la littérature, l'étude de l'effet du taux de faux liens dans l'estimation de la régression de Cox n'a pas suscité un grand intérêt.

S'il n'y a pas d'erreurs de couplage, l'estimation des coefficients du modèle de régression de Cox est simple et approximativement sans biais (Andersen and Gill, 1982). Cepndant, des simulations montrent qu'un petit taux d'erreurs de liaison peut entraîner une estimation biaisée des paramètres du modèle de Cox. En adoptant le modèle d'erreur de liaison hit-miss (Copas and Hilton, 1990), nous avons proposé une équation estimante ajustée pour la régression de Cox avec des données appariées. Ce modèle permet de corriger le biais de l'approche naïve qui ignore les erreurs dues aux faux liens. Nous proposons également un estimateur de variance pour le coefficient de régression estimé. Cet estimateur de variance permet de capturer l'ensemble de la variabilité, y compris celle due aux erreurs de liaion. Il s'agit de la deuxième contribution de cette thèse.

Contributions du Chapitre:

- Une équation estimante permettant d'obtenir une estimation sans biais des paramètres du modèle de Cox, en utilisant des données appariées.
- Un estimateur de variance pour ces paramètres estimés tenant compte des trois sources de variabilité :
  - la variabilité (usuelle) associée à la résolution d'une équation estimante basée sur un échantillon,
  - la variabilité associée au processus d'appariement,
  - la variabilité associée à l'estimation des vraies probabilités de couplage.

1.3 Plan de la thèse

Nous venons de présenter un bref aperçu des deux principales contributions de cette thèse ainsi que leurs motivations. Nous consacrons le chapitre 2 à la revue des méthodes d'appariement probabiliste d'enregistrements et à l'analyse de données appariées proposées dans la littérature. Tout d'abord, nous décrirons un processus général d'appariements d'enregistrements. Ensuite, nous présentons le cadre fondamental de l'appariement probabiliste d'enregistrements de Fellegi and Sunter (1969) et le développement de ce modèle dans la littérature. Ensuite, nous examinons deux modèles courants pour les erreurs d'appariement, et la manière dont les analystes les utilisent pour améliorer l'analyse de données appariées. Ce chapitre se termine par les préliminaires de l'analyse de survie et le modèle des risques proportionnels de Cox. Dans le chapitre 3, nous présentons la première contribution : "Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system". La deuxième contribution : "Cox regression with linked data" est présentée dans le chapitre 4.

Nos travaux de thèse ont donné lieu à deux articles et conférences dont:

Articles

✍ T.H. Vo, V. Garès, L-C. Zhang, A. Happe, E. Oger, S. Paquelet, G. Chauvet. Cox regression with linked data, *Under review at Statistics in Medicine*, 2022.

✍ T.H. Vo, G. Chauvet, A. Happe, E. Oger, S. Paquelet, V. Garès. Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system, *Computational Statistics & Data Analysis*, 2022.

Conférences

☞ T.H. Vo, V. Garès, A. Happe, E. Oger, S. Paquelet, G. Chauvet. Cox regression with linked data. *The 43rd Annual Conference of the International Society for Clinical Biostatistics (ISCB)*. Newcastle upon Tyne, UK, 21-25 August 2022.

☞ T.H. Vo, V. Garès, A. Happe, E. Oger, S. Paquelet, G. Chauvet. Cox regression with linked data. *Les 53èmes Journées de Statistique*. Lyon, France, 13-17 June 2022.

☞ T.H. Vo, G. Chauvet, A. Happe, E. Oger, S. Paquelet, V. Garès. An extension of Fellegi-Sunter record linkage model for mixed-type data with application to SNDS. *The 42rd Annual Conference of the International Society for Clinical Biostatistics (ISCB)*. Lyon, France, 18-22 July 2021.

☞ T.H. Vo, G. Chauvet, A. Happe, E. Oger, S. Paquelet, V. Garès. An extension of Fellegi-Sunter record linkage model for mixed-type data with application to SNDS. *Les 52èmes Journées de Statistique*. Nice, France, 7-11 June 2021.

Introduction (English version)

We are living in an era where data are collected everywhere and in every field, such as public administration, finance, commercial and medical fields. The ability of analyzing these datasets can improve insurance products, providing a competitive edge for a commercial service and novel knowledge for researchers. However, in many situations, a single dataset may not contain all necessary information for analysis, while collecting data on additional variables is burdensome, time consuming, and costly. Record linkage, a.k.a data matching, is a process of combining data from different databases that refer to the same entity. It is widely performed in order to enrich, update or improve the information stored in different sources, to study the relationship among variables reported in different sources, and to eliminate duplicates within a data frame, for example.

With increasing availability of large health care databases derived from administrative and other sources, scientists and healthcare workers increasingly demand for the successful linking of these databases, to provide rich sources of data for analysis and improve the healthcare systems. For example, the GETBO project (Groupe d'Étude de la Thrombose de Bretagne Occidentale) is a registry of venous thromboembolism (VTE) cases between 2013 and 2015 in Brest, France. This project only recorded the demographic information for each patient (date of birth, gender, residency code) along with some dates and types of medical acts. These informations are not sufficient to build an analysis model such as a prediction model which can early identify symptomatic VTE (Noboa et al., 2006). The French Système National des Données de Santé (SNDS) is the national health data system which collects all the longitudinal health records and insurance information of most of French population (Bezin et al., 2017). The valuable data in SNDS can be used to enrich the registry, which is expected to lead to valuable knowledge for researchers. This motivates us to link the GETBO and SNDS so that we can get more medical information for GETBO patients for the desired statistical analyses.

The record linkage process is straightforward if we can access to the unique identifiers of the patients. However, it is often not allowed due to ethical reasons. In such situation, we can only use other partial identifiers which are common to both

databases (e.g., gender, postal code or dates of medical treatments) to identify matched pairs from the two databases. These variables are often referred to as matching variables. The deterministic and the probabilistic method are two common approaches for record linkage. The deterministic methods requires that all matching variables of a matched pair are the same. Although this approach is fast and easy to carry out, a large amount of matched pairs can be missed if unique identifiers are unavailable and if there are some errors in the matching variables. In this context, the probabilistic record linkage approach is preferable (Zhu et al., 2015). This approach aims at providing each record pair with a matching score or a probability of matching. Fellegi and Sunter (1969) proposed a mathematical framework which laid the foundation for most of probabilistic record linkage methods even today. This framework enables to compute the matching scores for all candidate record pairs, based on their comparison vectors. The comparison vector of a record pair is a binary vector which only accounts for the agreement or disagreement of matching variables.

1.4 Extending the Fellegi-Sunter record linkage model with application to SNDS

Although the Fellegi-Sunter method has been widely used over the last decades, the binary comparison has some limits. Firstly, the simple binary comparison is not sufficient to account for the characteristics of low prevalence matching variables such as cancer diagnosis codes (Hejblum et al., 2019). For example, two patients which both have lung cancer are more likely to match than two patients without cancer. However, the binary comparison leads to the same matching probability for the two cases. Secondly, this binary comparison is not able to account for tolerances in the continuous matching variables, such as the delay in the date of treatments reported in different sectors due to administrative process. These are the common types of matching variables in health data as well as in GETBO and SNDS.

Therefore, we have extended the Fellegi-Sunter model to account for various types of matching variables for the identification of matched pairs. Two comparison approaches are proposed. For low prevalence categorical matching variables, we propose to distinguish the agreement of low prevalence values with others. For continuous matching variables, the proposed method enables to account for possible fluctuations. Along with the comparison proposed, the mixture of hurdle gamma distributions is, for the first time, used to model the continuous compari-

son values. Simulations show that the extended model outperforms the standard model in most scenarios. The improvement of this model in application of linking SNDS and GETBO has been recognized. This is the first contribution of the PhD thesis.

Contributions of the Chapter:

- Extending the Fellegi-Sunter record linkage model for mixed-type comparison values.
- Two novel comparison approaches for low prevalence categorical matching variables (e.g. diagnosis code) and continuous matching variables (e.g. date of medical acts).
- Using the mixture of hurdle gamma distributions to model the continuous comparison values.
- Application in SNDS and GETBO.

1.5 Cox regression with linked data

In most applications, the ultimate goal of record linkage is to obtain a linked dataset for a statistical analysis. In some cases, the persons performing the record linkage and the analysis may be the same, which is referred to as primary analysis. In other cases, the record linkage is performed by a trusted third party, and the person performing the statistical analysis has a limited knowledge about the record linkage. Since matching variables are likely to contain errors, false links are inevitable, regardless of the record linkage method used. This may enduce some bias in the analysis. Therefore, the secondary analysts should be aware of linkage errors, and choose a suitable strategy to avoid biased estimators.

Neter et al. (1965) first raise awareness of possible biases caused by linkage errors (false links or missed links). A false link corresponds to a couple of record identified as a match, while the two records actually do not refer to the same entity. A missed link corresponds to a couple of records which is not identified as a match, while the two records actually refer to the same entity. Many authors (e.g. Scheuren and Winkler, 1993, 1997; Lahiri and Larsen, 2005; Chambers, 2009; Zhang and Tuoto, 2021; Chambers et al., 2022) have considered this problem and have proposed different methods to deal with linkage errors. However, these

methods are mostly designed for linear and logistic regression models. In medical research, the Cox proportional hazard model (Cox, 1972) is one of the most used to study the relationship between covariates and time-to-event data. However, there has not been much interest in the literature in studying the effect of false links on Cox regression estimation.

If there is no linkage errors, the estimation for coefficients of the Cox regression model is straightforward and approximately unbiased (Andersen and Gill, 1982). However, some simulations results show that even a small amount of linkage errors can lead to biased estimators of the parameters of the Cox model. By adopting the hit-miss linkage error model (Copas and Hilton, 1990), we propose an adjusted estimating equation for Cox regression with linked data. This model corrects the bias obtained under the naive approach which ignores the linkage errors. An approximately unbiased variance estimator for the adjusted estimators of Cox regression coefficients is also proposed. This is the second contribution of this thesis.

Contributions of the Chapter:

- A bias-corrected estimating equation for Cox regression analysis with linked data.
- A variance estimator for the adjusted estimation of Cox regression coefficients accounts for three sources of variability:
  - the (usual) variability associated to solving a sample-based estimating equation.
  - the variability associated to the linkage process.
  - the variability associated to the estimation of the true linkage probabilities.

1.6 Plan of the thesis

We devote Chapter 2 to the review of probabilistic record linkage methods, and to the analysis of linked data in the literature. Firstly, a general record linkage process is described. Then we present the fundamental probabilistic record linkage framework of Fellegi and Sunter (1969), and the development of this model in literature. After that, we consider two common linkage errors models, and we explain how the analysts use them to improve the analysis of linked data. This

chapter ends with the preliminaries of survival analysis and Cox proportional hazard model. In Chapter 3, we present our first contribution, entitled "Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system". The second contribution, entitled "Cox regression with linked data", is presented in Chapter 4.

This thesis has contributed for 2 articles and 4 oral presentations at the conferences:

Articles

✍ T.H. Vo, V. Garès, L-C. Zhang, A. Happe, E. Oger, S. Paquelet, G. Chauvet. Cox regression with linked data, *Under review at Statistics in Medicine*, 2022.

✍ T.H. Vo, G. Chauvet, A. Happe, E. Oger, S. Paquelet, V. Garès. Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system, *Computational Statistics & Data Analysis*, 2022.

Conferences

☞ T.H. Vo, V. Garès, A. Happe, E. Oger, S. Paquelet, G. Chauvet. Cox regression with linked data. *The 43rd Annual Conference of the International Society for Clinical Biostatistics (ISCB)*. Newcastle upon Tyne, UK, 21-25 August 2022.

☞ T.H. Vo, V. Garès, A. Happe, E. Oger, S. Paquelet, G. Chauvet. Cox regression with linked data. *Les 53èmes Journées de Statistique*. Lyon, France, 13-17 June 2022.

☞ T.H. Vo, G. Chauvet, A. Happe, E. Oger, S. Paquelet, V. Garès. An extension of Fellegi-Sunter record linkage model for mixed-type data with application to SNDS. *The 42rd Annual Conference of the International Society for Clinical Biostatistics (ISCB)*. Lyon, France, 18-22 July 2021.

☞ T.H. Vo, G. Chauvet, A. Happe, E. Oger, S. Paquelet, V. Garès. An extension of Fellegi-Sunter record linkage model for mixed-type data with application to SNDS. *Les 52èmes Journées de Statistique*. Nice, France, 7-11 June 2021.

2 Literature review

Contents

2.1 Record linkage

Record linkage is the process of combining information about an individual, event or object in one or more databases. There exists several kinds of linkage algorithms, among which deterministic record linkage and probabilistic record linkage are two basic approaches. They both use available common information between databases such as name, age and postal code to identify matched pairs. The variables which are used for linking process are called matching variables. In deterministic record linkage, a pair of records is classified as a link if the two records agree exactly on some or all selected matching fields according to specified rule. Such methods are often applied when high discriminating matching variables are available with good quality. For example, deterministic linkage may be attempted using Social Security Number (SSN) for linking people. Since the SSNs may contain errors, one can use additional identifiers such as last name and date of birth to avoid incorrect links (see Grannis et al., 2002; Mao et al., 2019). However, due to confidentiality reasons, such identifiers are often unavailable for researchers.

Compared with the deterministic approach, the probabilistic record linkage only deems a record pair with certain probability of matching. This method can solve

problems caused by bad quality data, and can be helpful when no identifier information is available. It is more accurate with low quality data, while deterministic linkage is a comparable and faster method with high quality data (see Zhu et al., 2015). For example, Avoundjian et al. (2020) obtained better recalls (i.e., sensitivity, proportion of true matches identified by the method) with probabilistic linkage, and comparable results in terms of precision (i.e., positive predictive value, proportion of matches identified by the algorithm that were true matches) than with deterministic linkage. In some cases, depending on the aims of the whole linkage project, the deterministic and the probabilistic approaches can be combined in a two-step process (e.g. Larsen and Herning, 2023). Firstly, the deterministic method can be performed on the high quality variables. The probabilistic approach is then applied on the remaining individuals, which have not been linked during the first step (e.g. K Taylor et al., 2014).

In this thesis, we are interested in probabilistic record linkage. In Section 2.1.1, we describe a standard process of a record linkage problem. Then, a fundamental probabilistic record linkage model will be summarized in Section 2.1.2.

2.1.1 A general record linkage process

In general, there are five major steps in a record linkage process (Christen, 2012): data pre-processing, indexing (blocking), record pair comparison, record pair classification and evaluation of matching quality.

Data pre-processing

ID	Last name	Date of birth	Gender	Postal code	Lung cancer	Annual income ($)
a_1	Martin	04/07/1990	Male	35000	0	42000
a_2	Robert	06/08/1992	Female	35170	1	30000
a_3	Richard	05/03/1993	Male	35510	0	32000
a_4	Michel	08/03/1993	Male	35510	0	48000
a_5	Robert	07/04/1990	Female	35170	0	60000

(a) Database A

ID	Last name	Date of birth	Gender	Postal code	Lung cancer	Education
b_1	Simon	02 September 1993	F	35510	No	Doctoral
b_2	Martin	04 July 1990	M	35170	No	Master
b_3	Robern	07 August 1992	F	35170	Yes	Bachelor

(b) Database B

Table 2.1: An example of two raw databases

Databases which need to be linked often come from different sources. Therefore, they can have different designs or different type of errors, since matching variables may have different formats. To fix ideas, two small databases sharing information on last name, date of birth, gender and lung cancer with different formats are presented in Table 2.1. There is also information of income observed only in Database A and Education observed only in Database B. A researcher who is

interested in the relation between education and income is motivated to link these two databases.

Therefore, the first step in a record linkage process is pre-processing, which assures that data have a well-defined structure and the same format. This is a crucial step for an efficient record linkage. This step includes, for instance, converting dates into the same format, or removing unexpected punctuation and letters for string variables. Christen (2012) introduced various data cleaning and standardisation techniques for different types of matching variables. The databases A and B after a pre-procesing step are presented in Table 2.1.

ID	Last name	Date of birth	Gender	Postal code	Lung cancer	Annual income ($)
a_1	Martin	04/07/1990	1	35000	0	42000
a_2	Robert	06/08/1992	0	35170	1	30000
a_3	Richard	05/03/1993	1	35510	0	32000
a_4	Michel	08/03/1993	1	35510	0	48000
a_5	Robert	07/04/1990	0	35170	0	60000

(a) Database A

ID	Last name	Date of birth	Gender	Postal code	Lung cancer	Education
b_1	Simon	02/09/1993	0	35510	0	Doctoral
b_2	Martin	04/07/1990	1	35170	0	Master
b_3	Robern	07/08/1992	0	35170	1	Bachelor

(b) Database B

Table 2.2: Two databases after pre-processing to obtain the same formats for the common variables

Indexing or blocking

The second step is called indexing. It aims to reduce the number of record pairs that need to be compared afterwards, by removing pairs that are unlikely matches. At the same time, all record pairs that possibly correspond to true matches need to be kept for future evaluation. Potentially, all record pairs from two datasets are considered as matched candidates. However, this leads to a huge number of record pair comparison which is often impracticable. For example, matching two databases containing respectively 1 000 and 1 000 000 records results in $1\,000 \times 1\,000\,000 = 10^9$ possible record pair comparisons. There are various indexing techniques. One of the most common indexing techniques is called blocking. It consists in separating both files in blocks, according to the values of some so-called blocking variables. In this case, records are considered as matched candidates only if they belong to the same block (see Herzog et al., 2007; Christen, 2012). In the previous example from Table 2.1, we consider the postal code as a blocking variable. The candidate pairs are therefore:

Candidate pairs	Postal code
$(a_3, b_1), (a_4, b_1)$	35510
$(a_2, b_2), (a_2, b_3), (a_5, b_2), (a_5, b_3)$	35170

Table 2.3: Candidate record pairs for matching two databases in Table 2.2 if the postal code is used as a blocking key

If we choose a different blocking key such as the gender, we obtain a different set of candidate pairs as shown in Table 2.4. Overall, the numbers of candidate record pairs in Table 2.3 (6 pairs) or Table 2.4 (7 pairs) are all much smaller than the total number of possible pairs (15 pairs).

Candidate pairs	Gender
$(a_1, b_2), (a_3, b_2), (a_4, b_2)$	1
$(a_2, b_1), (a_2, b_3), (a_5, b_1), (a_5, b_3)$	0

Table 2.4: Candidate record pairs for matching two databases in Table 2.2 if the gender is the blocking key

However, blocking is a trade-off between the computation tile and the rate of false non-matches (i.e., the rate of true matched pairs which are classified as non-links). Indeed, records which disagree on the blocking keys are inside different blocks and are therefore automatically classified as non-links. Therefore, blocking fields should have high quality and contain a large number of possible values. It is also helpful that the blocking variables are distributed as uniformly as possible (Herzog et al., 2007).

Record pair comparison

In the third step, which is called record pair comparison, several attributes called matching variables are compared in detail for each candidate record pair remaining after the indexing step. The matching variables (which need to be chosen by the practitioner) contain the information that is in common between the two databases, such as the last name, the date of birth or the gender. A vector of comparison values is generated for each record pair. The set of all comparison vectors is our new working data (e.g. Table 2.5), where classification techniques will be applied to find matched and unmatched pairs.

The matching variables are of various types, depending on the application under consideration. They can be string fields such as last name, first name and ad-

dress; binary fields such as gender or disease presence; categorical fields such as postal code, code of hospital initial care; dates and numerical attributes, such as the date of treatment effects or salary. Each type of matching variables requires specific comparison methods. The most common method is a 0-1 comparison, considering only 1 for exact agreement and 0 for disagreement. However, even if matching variables have been cleaned and standardized, they may still contain errors, causing true matched pairs to have different attribute values. In the example presented in Table 2.1, the records a_2 and b_3 are likely to correspond to the same entity, but there is a slight difference on both the last name and the date of birth. Therefore, it is essential to introduce some similarity measures between comparison fields, instead of conducting matching in case of exact agreement only.

For string variables, there are various similarity measures such as the Levenshtein edit distance, or the Jaro and Winkler string comparison (Herzog et al., 2007). Some surveys have been conducted to evaluate the performances of these metrics (Yancey, 2005; Snae, 2007). Generally, these similarities return a value between 0 and 1, indicating how similar the strings are (1 standing for exact agreement, and 0 for total disagreement). For example, let $s = sim(s_1, s_2)$ be the Jaro string comparator metric defined as follows

$$s(s_1, s_2) = W_1 \cdot \frac{c}{L_1} + W_2 \cdot \frac{c}{L_2} + W_t \cdot \frac{c - \tau}{c} \tag{2.1}$$

where

- L_1, L_2 are the lengths of the first and second string, respectively,
- W_1, W_2 are some weights assigned to the first and second string, respectively,
- c is the number of characters that these two strings have in common. Two characters from s_1 and s_2 respectively, are considered common only if they are the same and not farther than $\frac{\max(L_1,L_2)}{2} - 1$,
- W_t is some weight assigned to the transposition,
- τ is the number of common characters which are transposed.

It is required that the weights sum to 1 ($W_1+W_2+W_t = 1$). We define $s(s_1, s_2) = 0$ if $c = 0$, i.e. if the two strings have no character in common.

For example, if we take $W_1 = W_2 = W_t = 1/3$ and $s_1 = "Robert"$, $s_2 = "Robetrn"$, we have

$$s(s_1, s_2) = \frac{1}{3}\frac{6}{6} + \frac{1}{3}\frac{6}{7} + \frac{1}{3}\frac{6-2}{6} \approx 0.84.$$

The similarity values are then often categorized into a pre-defined number of similarity levels (Winkler, 1990; Enamorado et al., 2019). For example, if we choose 0.85 as a cutoff, then a similarity value of two strings larger than 0.85 will be equal to 1, and to 0 otherwise. Other approaches are also possible for calculating similarities between numerical values such as salaries, expenses (Christen, 2012).

	Last name	Date of birth	Gender	Postal code	Lung cancer
γ_{11}	0	0	0	0	1
γ_{12}	1	1	1	0	1
γ_{13}	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
γ_{51}	0	0	1	0	1
γ_{52}	0	0	0	1	1
γ_{53}	1	0	1	1	0

Table 2.5: A comparison matrix for the two databases presented in Table 2.2 using Jaro metric with a 0.85 cutoff for the last name and exact comparison for the other matching variables

Record pair classification

Once the candidate pairs have been compared, we aim to classify them into two or three classes: matches, non-matches and possible matches, which depends on the linkage model (Fellegi and Sunter, 1969; Herzog et al., 2007). The first class contains record pairs that are predicted to refer to the same entity. In the second class, the two records in a pair are assumed not to refer to the same entity. If candidate record pairs have been classified as possible matches, they require an additional manual review to decide their final status (matches or non-matches). The classification step is mainly based on the comparison vectors that are generated in the comparison step. In general, the more similar two records are, the more reasonable they refer to the same real-world entity.

Over the past eight decades (Dunn, 1946), there have been various classification techniques developed for record linkage including supervised, unsupervised or semi-supervised approaches (Christen, 2012). When the candidate record pairs only need to be classified into matches and non-matches, we can consider this classification as a binary classification problem. If some training data are available under the form of record pairs with their true status (match or non-match), a supervised classification method can use them to train a classification model.

Then, this trained model is used to classify the remaining record pairs with unknown status. For example, one can mention two popular supervised classification techniques which have been applied in the field of record linkage: support vector machines (Bilenko and Mooney, 2003; Christen, 2008), and decision trees (Cochinwala et al., 2001).

Since training data are rarely available and are expensive to obtain, unsupervised methods which do not require the training data are the most preferable in practice. Fellegi and Sunter (1969) first proposed a probabilistic record linkage framework which can be employed as an unsupervised method. This model provides each candidate record pair with a matching score or a probability of matching which are also helpful for the analysis of linked data (Lahiri and Larsen, 2005). Since this model is widely used in applications and fundamental for most of probabilistic record linkage methods, we study it with more details in the next section. Besides, Larsen and Rubin (2001) suggested some clerical reviews and then re-estimating latent class model parameters.

Software and packages	API	GUI	Link	Dedup	Supervised learning	Unsupervised learning	Active learning
Atylmo	PySpark	No	Yes	Yes	?	?	?
Dedupe	Python	No	Yes	Yes	Yes	No	Yes
fastLink	R	No	Yes	?	No	Yes	No
FEBRL	Python	Yes	Yes	Yes	No	No	No
FRIL	Java	Yes	Yes	No	?	Yes	No
FuzzyMatcher	Python	No	Yes	No	No	Yes	No
JedAI	Java	Yes	Yes	?	Yes	?	?
PRIL	SQL	No	Yes	?	?	?	?
Python Record Linkage Toolkit	Python	No	Yes	Yes	Yes	Yes	No
RecordLinkage	R	No	Yes	Yes	Yes	Yes	No
RecLin2	R	?	Yes	Yes	Yes	No	No
RELAIS	No	Yes	Yes	?	?	Yes	No
ReMaDDer	No	Yes	Yes	Yes	No	Yes	No
RLTK	Python	No	Yes	Yes	Yes	No	No
Splink	PySpark	No	Yes	Yes	No	Yes	No

Table 2.6: List of freely available and open source data matching software and packages. The ? indicate for Unknown information (https://github.com/J535D165/data-matching-software/).

In practice, there are various software/packages which were developed for implementing record linkage. A comprehensive list of open source and freely software

and packages for data matching is available at: https://github.com/J535D165/data-matching-software/, and is attached in Table 2.6. The list provides a dense overview of data matching software properties: Application Programming Interface (API), Graphical User Interface (GUI), Linking, Deduplication, and the implemented record linkage approaches (Supervised Learning, Unsupervised Learning and Active Learning). In addition, some national statistical agencies have developed their own specialized record linkage and data editing systems such as Big Match[1] from U.S. Census Bureau , CAN LINK[2] from Statistics Canada

Evaluation of matching quality

Finally, we need to evaluate the linkage quality. We introduce four indicators which are commonly used to evaluate the performance of a binary classifier. Note that these indicators may be exactly computed only if the true matches are known.

- True positives (TP): These are record pairs which are predicted as matches and which are indeed true matches.
- True negatives (TN): These are record pairs which are predicted as non-matches and which are indeed non-matches.
- False positives (FP): these are record pairs which are predicted as matches while they are actually non-matches.
- False negatives (FN): these are record pairs which are predicted as non-matches while they are actually matches.

A 2×2 table reporting the numbers for each of these four categories is called a confusion matrix, as shown in Table 2.7.

		Predicted status	
	Total (Matches + Non-matches)	**Matches**	**Non-matches**
Actual status	**Matches**	True matches (True positives: TP)	False non-matches (False negatives: FN)
	Non-matches	False matches (False positives: FP)	True non-matches (True negatives: TN)

Table 2.7: Confusion matrix for the outcome of the classification of record pairs

[1] https://www.census.gov/library/working-papers/2002/adrm/rrc2002-01.html
[2] https://www150.statcan.gc.ca/n1/en/catalogue/10H0036

Based on the confusion matrix, there are different evaluation criteria which can be calculated to assess the performance of the classification (Christen, 2012). Some common measures are:

- Accuracy (ACC): this is the proportion of pairs which are correctly predicted, computed as

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{Total}}.$$

- True negative rate (TNR): this is the proportion of non-matches which are correctly predicted as such. It is calculated as

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{True Predicted Non-matches}}{\text{True Non-matches}}.$$

  It is commonly referred to as *specificity* in medical fields.

- True positive rate (TPR): this is the proportion of matches which are correctly predicted as such. It is calculated as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{True Predicted Matches}}{\text{True Matches}}.$$

  This measure is also known as the *recall*, or *sensitivity* in epidemiological studies.

- Positive predictive value (PPV): this is the proportion of predicted matches which are indeed matches. It is commonly used in medical literature and is calculated as

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{True Predicted Matches}}{\text{Predicted Matches}}$$

- F-score: it is defined as the harmonic mean of the TPR and the PPV, and is calculated as

$$\text{F-score} = \frac{(\lambda^2 + 1) \times \text{PPV} \times \text{TPR}}{\lambda^2 \text{TPR} + \text{PPV}}.$$

  where λ is a value defined by the user, which is related to the importance of TPR over PPV in the harmonic mean (Dusetzina et al., 2014). If we wish to assign equal weight to TPR and PPV, then $\lambda = 1$.

The accuracy is not only widely used for binary classification problems, but also for multi-classes problems. However, it is usually effective when the classes are balanced, i.e. when the numbers of units per class are approximately the same.

In the record linkage problems, since the vast majority of record pairs are non-matches, the matches and non-matches are extremely imbalanced even if indexing or blocking is applied.

For example, let us consider a matching problem where two databases A and B have 100 and $1,000$ individuals respectively, which leads to $100,000$ possible record pairs. Assume that blocking is applied and leads to the decrease of the number of candidate record pairs to $10,000$, among which there are 100 true matches and $9,900$ true non-matches. Now, assume that the linkage method classified 120 record pairs as matches but that among them 70 pairs only are truly matches. Consequently:

$$TP = 70, \qquad FP = 50, \qquad TN = 9850 \quad \text{and} \quad FN = 30.$$

Then, the accuracy is calculated as $\text{ACC} = \frac{70+9850}{10\,000} = 99.2\%$ which is very good. Similarly to the accuracy, the true negative rate is also dominated by true negatives (true non-matches). For instance, we obtain a true negative rate of $\text{TNR} = \frac{9850}{9900} \approx 99.5\%$ for the above example. Thus, Christen and Goiser (2007) argued that linkage quality measures which include the true negatives (such as the accuracy and the true negative rate) are skewed and not suitable to evaluate a data matching method.

In record linkage problems, since we aim to find records that belong to the same entities, we are only interested in matches. Thus the TPR and PPV, which are not influenced by the domination of true negatives, are most commonly used to assess the matching quality (Grannis et al., 2003; Hejblum et al., 2019). While the TPR is calculated as the proportion of true matches which are correctly classified, the PPV is a measure of how often predicted matches are truly matches. In the above example, though accuracy and true negative rate are extremely high, the true positive rate is only $\text{TPR} = \frac{70}{100} = 70\%$, and the positive predictive value is only $\text{PPV} = \frac{70}{120} \approx 58.3\%$ only. This means that we are able to find only 70% of the true matched pairs, and that 58.3% of pairs classified by the algorithm as matches are indeed true matches.

Depending on the objectives of each problem, investigators may focus on different criteria. If they study a rare disease, they may want to emphasize true positive rate to maximize the sample size, while a researcher studying a more frequently occurring disease may seek to emphasize PPV to ensure that matches identified by the linkage methods are true matches. However, those who wish to maximise true

positive rate may increase their false positives, which leads to a smaller positive predictive value. The F-score can be considered as a trade-off between TPR and PPV. What are acceptable values for these measures depends on the context of the study. However, a good linkage algorithm is expected to have TPR, PPV, and F-score larger than 95 % (Dusetzina et al., 2014).

Knowing which pairs are indeed matches is rarely possible in practice. In some cases, a dataset where the matching status is known under reasonable conditions is available. It can therefore be used to evaluate the record linkage methods: this is referred to as a gold standard. For example, it can be obtained by a clerical review, but this is usually an expensive approach. Blakely and Salmond (2002) proposed a method to estimate the PPV when a gold standard is not feasible, assuming that there is only one match per record. Although there are some public data sources of data which are created to evaluate record linkage methods, there are some limitations such as their small size and the fact that these data are quite specific (Christen, 2012). Therefore, to evaluate and compare different methods in this work, synthetic datasets for which the true match status is known will be generated. These data will be generated so as to mimic as closely as possible real data where the methods are applied. In particular, the artificial data will match the real data in terms of types of variables, and frequencies for their modalities.

2.1.2 Fellegi-Sunter record linkage framework

Fellegi and Sunter (1969) proposed a framework which laid a mathematical foundation for many probabilistic record linkage methods, and has been widely used until today. Under this framework, all possible realizations in the comparison space are fitted by a mixture model of two classes, namely Matches and Non-matches. They are then ranked with respect to a defined matching score which is small (resp., large) for pattern more compatible with non-matches (resp., matches). Two thresholds are then defined to partition the records into three classes: those predicted as matches (link), those predicted as non-matches (non link), and those which remain undecided (possible link). The partition is defined so as to respect predetermined error levels, and so as to minimize the number of undecided record pairs.

Formally, let $a_i, i = 1, \ldots, n_A$ and $b_j, j = 1, \ldots, n_B$ be n_A and n_B elements of two databases A and B, respectively. We assume that some elements are common to

A and B. The product space

$$A \times B = \{(a, b); a \in A, b \in B)$$

is the union of two disjoint sets

$$M = \{(a, b); a = b, a \in A, b \in B\} \tag{2.2}$$

and

$$U = \{(a, b); a \neq b, a \in A, b \in B\} \tag{2.3}$$

which we call the true **M**atched and true **U**nmatched sets, respectively.

Let K be the number of matching variables and $\mathbf{X}_{A,i} = (X^1_{A,i}, \ldots, X^K_{A,i}), i = 1, \ldots, n_A$ be the records of n_A individuals in A, $\mathbf{X}_{B,j} = (X^1_{B,j}, \ldots, X^K_{B,j}), j = 1, \ldots, n_B$ be the records of n_B individuals in B. We define the comparison vector for each pair of individuals (a_i, b_j) as a vector function of $(\mathbf{X}_{A,i}, \mathbf{X}_{B,j})$:

$$\boldsymbol{\gamma}_{ij} = \left\{\gamma^1_{ij}, \ldots, \gamma^k_{ij}, \ldots, \gamma^K_{ij}\right\}, \tag{2.4}$$

where $\gamma^k_{ij} = h^k(X^k_{A,i}, X^k_{B,j})$ and h^k is a comparison function for the $k-$th matching variable. The set of all possible realizations of $\boldsymbol{\gamma}$ is called the comparison space and is denoted by Γ.

In some cases, the comparison vectors may not be computed for some records, due to missing data in matching variables. In such cases, one can choose to ignore or eliminate the corresponding records. Although this is the most simple approach, it fails to take advantage of other available data to match records. Another common approach assumes that if one of the two records being compared is missing, the comparison value for this variable could be an agreement or disagreement depending on each application. For example, if $X^k_{A,i}$ = "Male" and $X^k_{B,j}$ is missing, one can assign $\gamma^k_{ij} = 1$ for agreement or $\gamma^k_{ij} = 0$ for disagreement, which should be carefully decided depending on the application. If the amount of missing data is significant, more advanced techniques should be used to improve the linkage performance (Harron et al., 2015). For example, Ong et al. (2014) proposed three novel methods which are weight redistribution, distance imputation, and linkage expansion to improve record linkage performance in the presence of missing linkage data. In this work, unless explicitly specified, possible missing data problems are ignored.

Considering all candidate record pairs, Fellegi and Sunter (1969) proposed to order them with respect to a matching score, defined as

$$w_{ij} = \frac{\mathbb{P}\left[\boldsymbol{\gamma}_{ij} \in \Gamma | (a_i, b_j) \in M\right]}{\mathbb{P}\left[\boldsymbol{\gamma}_{ij} \in \Gamma | (a_i, b_j) \in U\right]}. \tag{2.5}$$

This ratio is large for values of $\boldsymbol{\gamma}$ that are found frequently among matches, but are rarely found among non-matches. This corresponds to record pairs that primarily consists of agreements in $\boldsymbol{\gamma}$. It follows that these record pairs are more likely designated as matches. On the other hand, record pairs which mainly consist of disagreements have small matching scores, and are therefore reasonably designated as non-matches. Instead of using the matching score, other authors (Larsen and Rubin, 2001) consider the posterior probability of matching. It is defined as

$$q_{ij} = \mathbb{P}\left[(a_i, b_j) \in M | \boldsymbol{\gamma}_{ij} \in \Gamma\right].$$

It follows that,

$$\begin{aligned} q_{ij} &= \frac{\mathbb{P}\left[\boldsymbol{\gamma}_{ij} \in \Gamma | (a_i, b_j) \in M\right] \mathbb{P}(M)}{\mathbb{P}\left[\boldsymbol{\gamma}_{ij} \in \Gamma | (a_i, b_j) \in M\right] \mathbb{P}(M) + \mathbb{P}\left[\boldsymbol{\gamma}_{ij} \in \Gamma | (a_i, b_j) \in U\right] \left[1 - \mathbb{P}(M)\right]} \\ &= \frac{w_{ij}\mathbb{P}(M)}{w_{ij}\mathbb{P}(M) + \left[1 - \mathbb{P}(M)\right]} = 1 - \frac{1 - \mathbb{P}(M)}{w_{ij}\mathbb{P}(M) + \left[1 - \mathbb{P}(M)\right]}. \end{aligned} \tag{2.6}$$

Therefore, q_{ij} is increasing with w_{ij} and it is equivalent to rank the record pairs by the matching scores or by the posterior probabilities.

In order to partition the records into three assignment classes, Fellegi and Sunter (1969) proposed an optimal decision rule which is as follows:

$$\begin{gathered} (a_i, b_j) \text{ is designated as a link if } T_\mu \leq w_{ij}, \\ (a_i, b_j) \text{ is designated as a possible link if } T_\lambda \leq w_{ij} < T_\mu, \\ (a_i, b_j) \text{ is designated as a non-link if } w_{ij} < T_\lambda, \end{gathered}$$

where T_μ and T_λ are two thresholds. They are determined by two given error bounds: the rate of false matches μ, and the rate of false non-matches λ. For more details on the construction of these thresholds, see Fellegi and Sunter (1969). This rule is optimal in the sense that it minimizes the number of units which are undecided (i.e. possible links) for given values of μ and λ. This seems a reasonable approach, since in applications handling the set of possible links would require expensive manual reviews. However, various simplifying assumptions are usually involved in the estimation of w_{ij} or q_{ij}, and may lead to estimation errors. The optimality may therefore not be attained in practice (Belin and Rubin, 1995; Binette and Steorts, 2022).

Because it is usually difficult and expensive to conduct some manual review to make a decision for possible links, other practitioners consider only two assignment classes (links or non-links). In such case, a single threshold needs to be determined (Grannis et al., 2003). In some cases, each individual in a database is expected to have at most one match in another database, which is known as one-to-one matching. For example, suppose that database A corresponds to a cohort sample, where each patient is unique and therefore with a single record. Each patient in the cohort can therefore match at most one patient from database B. In such cases, using a threshold can not guarantee for the one-one matching. If the optimal score is not demanded, a simple approach is to sort all candidate pairs according to their estimated posterior probabilities of matching, and to select matched pairs in a greedy approach (Christen, 2012). At each step, the greedy algorithm selects candidate record pairs with highest matching weights/matching probabilities, until no more un-assigned records can be matched. To optimize the sum of matching weights/matching probabilities, Jaro (1989) proposed a linear sum assignment problem.

We explain briefly the principles of the linear assignment problem. Formally, let $W = (w_{ij})_{n_A \times n_B}$ (respectively, $Q = (q_{ij})_{n_A \times n_B}$) be the matrix containing the matching scores (respectively, the posterior probabilities) of matching for all pairs. The problem may be formulated as

$$\max_{\boldsymbol{l}} \quad \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} w_{ij} l_{ij} \quad \text{or} \quad \max_{\boldsymbol{l}} \quad \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} q_{ij} l_{ij}$$

under the $n_A + n_B$ constrains:

$$\begin{aligned} \sum_{j=1}^{n_B} l_{ij} \leq 1 \quad & \text{for } i = 1, \ldots, n_A, \\ \sum_{i=1}^{n_A} l_{ij} \leq 1 \quad & \text{for } j = 1, \ldots, n_B, \end{aligned}$$

where l_{ij} is an indicator variable, equal to 1 if (a_i, b_j) is designated as a link and to 0 otherwise.

The most important step in a probabilistic record linkage procedure consists in estimating the probabilities $\mathbb{P}\left[\boldsymbol{\gamma}_{ij} \in \Gamma | (a_i, b_j) \in M\right], \mathbb{P}\left[\boldsymbol{\gamma}_{ij} \in \Gamma | (a_i, b_j) \in U\right]$ and $\mathbb{P}(M)$, see equation (2.5). The optimal property of Fellegi and Sunter's method heavily depends on the accuracy of the estimates of these matching parameters. In a

basic probabilistic record linkage approach when exact comparisons only are considered (binary values, 1 for exactly agreement and 0 for disagreement), there will be 2^K different patterns for the comparison vector $\boldsymbol{\gamma}$. For example, the possible comparison vectors for $K = 3$ matching variables are $(0, 0, 0)$ (all disagreements), $(0, 0, 1), (0, 1, 0), (1, 0, 0)$ (agree on 1 variable), $(0, 1, 1), (1, 0, 1), (1, 1, 0)$ (agree on 2 variables) and $(1, 1, 1)$ (all agreements). Thus, the number of parameters that need to be estimated (2^K) may be impracticable if the number of matching variables K is appreciable and if the databases are of moderate size. For ease of computation, it is usually assumed that the comparison patterns of matching variables are conditionally independent (Fellegi and Sunter, 1969; Herzog et al., 2007; Sayers et al., 2015), namely:

$$\mathbb{P}\left[\boldsymbol{\gamma}_{ij} \in \Gamma | (a_i, b_j) \in M\right] = \prod_{k=1}^{K} \mathbb{P}\left[\gamma_{ij}^k | (a_i, b_j) \in M\right], \tag{2.7}$$

$$\mathbb{P}\left[\boldsymbol{\gamma}_{ij} \in \Gamma | (a_i, b_j) \in U\right] = \prod_{k=1}^{K} \mathbb{P}\left[\gamma_{ij}^k | (a_i, b_j) \in U\right]. \tag{2.8}$$

In applications, it is possible to have some dependencies between matching variables. For example, records that have the same postal code are more likely to have the same street name. However, the method showed that a good matching quality can be achieved even though the conditional independence assumption is probably invalid in practice (Herzog et al., 2007; Sayers et al., 2015). Some authors proposed an extension of the model to allow for dependencies between matching variables (e.g. Larsen and Rubin, 2001; Schürle, 2005; Sadinle, 2017; Daggy et al., 2014), and Xu et al. (2019) investigated the implications of this assumption through a simulation study.

Under this assumption, $2K + 1$ parameters only need to be estimated for the binary comparison approach, namely:

$$\begin{aligned} m^k &\equiv \mathbb{P}\left[\gamma_{ij}^k = 1 | (a_i, b_j) \in M\right], k = 1, \ldots, K, \\ u^k &\equiv \mathbb{P}\left[\gamma_{ij}^k = 1 | (a_i, b_j) \in M\right], k = 1, \ldots, K, \\ \text{and} \quad p_M &\equiv \mathbb{P}(M). \end{aligned} \tag{2.9}$$

Fellegi and Sunter (1969) proposed two methods for the estimation of these parameters. The first approach considers detailed comparison vectors which involves both an agreement or disagreement of matching variable and a value of the matching variable in the case of an agreement. For example, "agreement on first name and the name is John". This was originally designed for simple realizations of

γ such as list of names. The method requires the frequency distribution of the matching variable and prior information about error rates to estimate m and u. For general matching variables, the authors proposed a second approach considering binary comparison. They provided formulas to estimate all parameters in a particular case of only three matching variables.

Winkler (1988) extended the above estimation methods by proposing an unsupervised approach using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977; Wu, 1983). It has now become widely used in most probabilistic record linkage methods. The EM algorithm is a widely used probabilistic algorithm for obtaining maximum likelihood estimates of unknown parameters under a latent class model. Given an observed dataset with a missing variable g, and a model for the incomplete data characterized by a parameter set θ, the fundamental goal of EM is to determine θ such that the probability $\mathbb{P}(\boldsymbol{\gamma}, g|\theta)$ is maximized. In case of record linkage, the observed dataset corresponds to the set of all comparison vectors $\boldsymbol{\gamma}$, the missing variable g corresponds to the vector with the true match or non-match statuses, the parameters to be estimated are given in equation 2.9. The EM algorithm makes use of an initial set of parameters θ to calculate an expected likelihood in the *expectation step*, providing estimates for the missing values. In the *maximization step*, the expectation is maximized with respect to θ. The E and M steps are repeated until convergence. More formally, at the $(t+1) - th$ iteration of the EM algorithm:

- *E-step*: the conditional expectation $\mathbb{E}_{g|\boldsymbol{\gamma},\theta^{(t)}}\left\{\ln \mathbb{P}(\boldsymbol{\gamma}, g|\theta)\right\}$ is determined,
- *M-step*: some value $\hat{\theta}^{(t+1)}$ maximizing the conditional expectation is determined.

Generally, the convergence is ensured because the likelihood is guaranteed to be non-decreasing at each iteration (McLachlan and Krishnan, 1996).

In the context of record linkage, because there are two classes (Matches and Unmatches) in the comparison space Γ, the distribution of the comparison vectors $\boldsymbol{\gamma}$ is assumed to follow a mixture model

$$\mathbb{P}(\boldsymbol{\gamma}) = \mathbb{P}(\boldsymbol{\gamma}|M)\mathbb{P}(\boldsymbol{\gamma} \in M) + \mathbb{P}(\boldsymbol{\gamma}|U)\left[1 - \mathbb{P}(\boldsymbol{\gamma} \in M)\right]. \tag{2.10}$$

The construction of the likelihood function requires the crucial assumption that all comparison vectors in Γ are independent. This assumption of independence usually does not hold in practice (Binette and Steorts, 2022). A counterexample is given by Tancredi and Liseo (2015): after comparing record pairs (a_1, b_1),

(a_1, b_2) and (a_2, b_1), the result for the comparison between a_2 and b_2 is often already known. Despite that, this assumption is widely used in most probabilistic record linkage models, and may provide acceptable results in practice. To avoid this assumption, Zhang and Tuoto (2020) have proposed a maximum entropy classification for record linkage.

Let

$$g_{ij} = \begin{cases} 1 & \text{if} \quad (a_i, b_j) \in M, \\ 0 & \text{if} \quad (a_i, b_j) \in U, \end{cases}$$

be a binary variable indicating whether the record is a true match or not. This is a latent variable, since this information is unknown. Under the assumption of independence between all comparison vectors, the full likelihood and the associated log likelihood may be written as follows:

$$\mathcal{L}(g, \boldsymbol{\gamma} | \boldsymbol{m}, \boldsymbol{u}, p_M) = \prod_{i=1}^{n_A} \prod_{j=1}^{n_B} \left[\mathbb{P}(\boldsymbol{\gamma}_{ij}|M)p_M\right]^{g_{ij}} \left[\mathbb{P}(\boldsymbol{\gamma}_{ij}|U)(1-p_M)\right]^{1-g_{ij}}, \tag{2.11}$$

where $\boldsymbol{m} = (m^1, ..., m^K)$ and $\boldsymbol{u} = (u^1, ..., u^K)$. Then,

$$\begin{aligned} \ln\left[\mathcal{L}(g, \boldsymbol{\gamma} | \boldsymbol{m}, \boldsymbol{u}, p_M)\right] &= \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} g_{ij} \ln\left[\mathbb{P}(\boldsymbol{\gamma}_{ij}|M)p_M\right] + (1-g_{ij}) \ln\left[\mathbb{P}(\boldsymbol{\gamma}_{ij}|U)(1-p_M)\right] \\ &= \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} g_{ij} \ln\left[\mathbb{P}(\boldsymbol{\gamma}_{ij}|M)\right] + \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (1-g_{ij}) \ln\left[\mathbb{P}(\boldsymbol{\gamma}_{ij}|U)\right] \\ &+ \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} g_{ij} \ln\left(p_M\right) + (1-g_{ij}) \ln(1-p_M). \end{aligned} \tag{2.12}$$

The first step in the implementation of the EM algorithm is to define initial estimates of the unknown parameters $\boldsymbol{m}, \boldsymbol{u}$ and p_M. The initial values for these parameters are often chosen by experience, depending on the linkage context and the data quality. For example, in the simulation study by Grannis et al. (2003), initial values for all m^k, u^k and p equal to 0.9, 0.1 and 0.5, respectively, were chosen. From here, the expectation (E) step, followed in turn by the maximization (M) step is implemented repeatedly, until the algorithm produces estimates that attain the desired tolerance. Because this is an iterative method, the algorithm may also stop when it attains a maximum number of iterations. For example, we can define a tolerance $\epsilon > 0$ (e.g., $\epsilon = 10^{-6}$) and stop the EM algorithm if the relative difference between estimated parameters from two successive steps is less

than ϵ, i.e. $\left|\theta_i^{(t+1)} - \theta_i^{(t)}\right| / \left|\theta_i^{(t)}\right| < \epsilon$ for all parameters $\theta_i \in \theta$.

In the expectation step, the EM algorithm wishes to find the estimates for the unknown values g_{ij}. The expectation of g_{ij} is

$$\mathbb{E}\left(g_{ij}|\boldsymbol{\gamma}_{ij}\right) = \mathbb{P}\left[(a_i, b_j) \in M|\boldsymbol{\gamma}_{ij}\right] = \frac{\mathbb{P}(\boldsymbol{\gamma}_{ij}|M)\mathbb{P}(M)}{\mathbb{P}(\boldsymbol{\gamma}_{ij}|M)\mathbb{P}(M) + \mathbb{P}(\boldsymbol{\gamma}_{ij}|U)\mathbb{P}(U)}. \tag{2.13}$$

From the conditional independence assumption between comparison fields (see equations 2.7 and 2.8), the EM algorithm replaces unknown parameters in (2.13) with their estimator at the current step, leading to

$$\widehat{g}_{ij} = \frac{\widehat{p}_M \prod_{k=1}^{K} (\widehat{m}^k)^{\gamma_{ij}^k} (1-\widehat{m}^k)^{1-\gamma_{ij}^k}}{\widehat{p}_M \prod_{k=1}^{K} (\widehat{m}^k)^{\gamma_{ij}^k} (1-\widehat{m}^k)^{1-\gamma_{ij}^k} + (1-\widehat{p}_M) \prod_{k=1}^{K} (\widehat{u}^k)^{\gamma_{ij}^k} (1-\widehat{u}^k)^{1-\gamma_{ij}^k}}.$$

For the maximization step, we find the maximum likelihood estimates for the log-likelihood function (2.12) where the value of latent variables is obtained from the expectation step. The estimates may be written as follows:

$$\widehat{m}^k = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \widehat{g}_{ij}\gamma_{ij}^k}{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \widehat{g}_{ij}}, \tag{2.14}$$

$$\widehat{u}^k = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} (1-\widehat{g}_{ij})\gamma_{ij}^k}{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} (1-\widehat{g}_{ij})}, \tag{2.15}$$

$$\widehat{p} = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \widehat{g}_{ij}}{n_A n_B}. \tag{2.16}$$

Once all parameters are estimated, the matching weights or the posterior probabilities of matching for all record pairs $\hat{q}_{ij}$ are computed from (2.6), and applied for linking decision. For analysis of linked data, the estimated matching probability matrix $\hat{Q} = (\hat{q}_{ij})$ can be helpful to reduce the bias in analysis models, such as linear or logistic regression (Lahiri and Larsen, 2005; Hof and Zwinderman, 2012; Chambers, 2009; Kim and Chambers, 2012c). An introduction to the way by which the matching probabilities may be used in analysis models is presented in Section 2.2.

2.2 Statistical analysis with linked data

In case of record linkage, because of the unavailability of unique identifiers and since the matching variables are likely to contain errors, linkage errors are unavoidable, regardless of record linkage methods. There are two types of linkage errors (Harron et al., 2015):

- Missed links: a true matched pair fails to be identified by the record linkage method. This may occur when there are errors in matching variables (e.g., typographical errors or missing data) which prevent records from agreeing.
- False links: a false matched pair is linked erroneously. This may occur when the matching variables are not sufficiently distinguishable (e.g., different individuals sharing the same gender, year of birth and postal code).

Neter et al. (1965) first raised awareness of substantial bias in response error analysis, which may be caused even by a small amount of incorrect links. It is therefore important to account for linkage errors in a statistical analysis.

Suppose that we are interested in analyzing the relationship between a response variable Y and a set of covariates $\mathbf{X}$, while Y and $\mathbf{X}$ are stored in two separate databases A and B, respectively. Since the true pairs $(Y_i, \mathbf{X}_i)$ are not observable, a record linkage process is needed to obtain data pairs for analysis. We note $(Z_i, \mathbf{X}_i)$ the linked pairs obtained from the record linkage step. While the objective of record linkage is to obtain $Z_i = Y_i$, this may not occur in some cases due to linkage errors. For the sake of simplicity, suppose that the two databases have the same size $n_A = n_B = n$, and have no duplicates. Also, suppose that the linkage is complete (i.e., all the units are linked) in a one-to-one matching.

There are two common positions for analysis of linked data: the primary analysis and the secondary analysis. In the primary analysis, the analysts can access to both linked data and information on the linking process. In the secondary analysis, they can access only to the linked data. In what follows, we review works on these two kinds of analysis, along with two corresponding linkage error models. We end with describing the Cox regression model in the context of linked data.

2.2.1 The primary analysis

In a primary analysis, people can not only access to the linked data, but also to some information about the record linkage process (matching variables, estimated matching probabilities, ...). Based on the knowledge of the record linkage process, Scheuren and Winkler (1993) proposed a linkage error model where

$$Z_i = \begin{cases} Y_i & \text{with probability} \quad q_{ii}, \\ Y_j & \text{with probability} \quad q_{ij} \text{ for } j \neq i = 1, \ldots, n, \end{cases} \tag{2.17}$$

with $\sum_{j=1}^{n} q_{ij} = 1$ for any $i = 1, \ldots, n$. This model allows the probability of being a true match to vary across record pairs. Depending on the available in-

formation about the record linkage process, specific assumptions can be made to facilitate the estimation of q_{ij}. For example, Scheuren and Winkler (1993) assumed that q_{ij} only depends on its matching weight w_{ij}. This matching weight can be transformed to estimate the probabilities q_{ij} by using a two-class Gaussian mixture model proposed by Belin and Rubin (1995). However, the transformation requires a clerical-review sample, which is not always available. Lahiri and Larsen (2005) simplify the estimation of q_{ij} by assuming that the matching probability of a record pair depends only on its comparison vector, which can be obtained from probabilistic record linkage methods.

Consider the following linear regression model:

$$Y_i = \mathbf{X}_i^\top \beta + \epsilon_i, \quad i =, \ldots, n \tag{2.18}$$

where $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)^\top$ is a column vector of p known covariates and $\boldsymbol{\beta}$ is a column vector of p unknown regression coefficients. In addition, we assume that $\mathbb{E}(\epsilon_i) = 0, \mathrm{Var}(\epsilon_i) = \sigma^2$ and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j = 1, \ldots, n$. Because the true pairs $(Y_i, \mathbf{X}_i)$ are unknown, we observe only $(Z_i, \mathbf{X}_i)$ where Z_i is a proxy value for Y_i, obtained from the record linkage process. The naive ordinary least squares estimator for $\boldsymbol{\beta}$, which ignores the linkage errors, is obtained as:

$$\hat{\boldsymbol{\beta}}_{naive} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z}. \tag{2.19}$$

This naive estimation may be seriously biased if the linkage is not perfect (Neter et al., 1965). Under a primary analysis with the linkage error model (2.17), Scheuren and Winkler (1993, 1997) proposed an unbiased estimator for β by adjusting the bias of the naive estimator (2.19). Lahiri and Larsen (2005) have adopted the approach of Scheuren and Winkler (1993), and proposed an exact unbiased estimator, assuming that linkage errors depends only on the comparison vectors. The estimator of the vector of regression coefficients is:

$$\hat{\boldsymbol{\beta}}_{LL} = \left(\mathbf{X}^\top \mathbf{Q}^\top \mathbf{Q} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Q}^\top \mathbf{Z}, \tag{2.20}$$

where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$ is an $n \times p$ matrix, $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ and $\mathbf{Q} = (q_{ij})_{n \times n}$ is a matrix of matching probabilities obtained from the Fellegi-Sunter record linkage model. A variance estimation for $\hat{\beta}_{LL}$ needs to capture both the variability of sample estimation and also the variability in estimating Q. Lahiri and Larsen (2005) proposed a bootstrap procedure for that. Hof and Zwinderman (2012) extended the method by Lahiri and Larsen (2005) for linked data obtained from more than two sources, and also proposed alternative estimators based on weighted

least square methods, both for linear and logistic regression models. Recently, Han and Lahiri (2019) adapted the approach by Lahiri and Larsen (2005) for a more practical scenario when $n_A \leq n_B$. They provide a system of estimating equations which may lead to unbiased estimators for a generalized linear model.

2.2.2 The secondary analysis

In a secondary analysis, the analysis step is separated from the record linkage because matching variables often contain confidential information. The linkage is carried out by an independent team, often referred to as a Trusted Third Party. Thus, the data analyst can access only to the linked data, and no information on the matching variables is available. In addition to the complete and one-to-one matching assumptions, Chambers (2009) assumed that records in two databases A and B can be partitioned error-free into V distinct blocks, and that the linkage errors only occur within blocks. For each block $v = 1, \ldots, V$, Chambers (2009) proposed the exchangeable linkage errors (ELE) model as

$$Z_i = \begin{cases} Y_i & \text{with probability} \quad \alpha_v \\ Y_j & \text{with probability} \quad \frac{1-\alpha_v}{n_v-1} \text{ for } j \neq i = 1, \ldots, n_v, \end{cases} \tag{2.21}$$

where n_v is the number of individuals in block v. In practice, we need access to a random audit sample to estimate α_v for each block. In this audit sample, we know whether the predicted links are correct or not. Under a secondary analysis, due to the unavailability of information on the record linkage, the non-informative linkage assumption is usually required. In linear and logistic regression models, this assumption states that the linkage are independent of Y, conditionally on $\mathbf{X}$. Formally, for any units belonging to block v, let $l_{ij}, i, j = 1, \ldots, n_v$ be an indicator variable, which is equal to 1 if units i and j are linked, and to 0 otherwise. Then,

$$\begin{aligned} \mathbb{E}(Z_i|\mathbf{X}) &= \mathbb{E}\left(\sum_{j=1}^{n_v} l_{ij} Y_j | \mathbf{X}\right) \\ &= \sum_{j=1}^{n_v} \mathbb{E}(l_{ij}|\mathbf{X})\mathbb{E}(Y_j|\mathbf{X}) \end{aligned} \tag{2.22}$$

Chambers (2009) proposed an approach of correcting estimating functions for linkage error. In this approach, we estimate the p-vector of unknown parameters $\boldsymbol{\beta}$ by solving a p-dimensional unbiased estimating equation:

$$H(\boldsymbol{\beta}) = 0, \tag{2.23}$$

where $H(\boldsymbol{\beta})$ is a function of observed data such that

$$\mathbb{E}(H(\boldsymbol{\beta}_0)|\mathbf{X}) \quad = \quad 0, \tag{2.24}$$

with $\boldsymbol{\beta}_0$ the true value of $\boldsymbol{\beta}$. Under appropriate smoothness conditions, the estimator $\hat{\boldsymbol{\beta}}$ obtained by solving equation (2.24) is approximately unbiased for $\boldsymbol{\beta}_0$. Indeed, from a one-term Taylor expansion around $\boldsymbol{\beta}_0$, we have

$$H(\hat{\boldsymbol{\beta}}) \approx H(\boldsymbol{\beta}_0) + \frac{\partial H(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = 0. \tag{2.25}$$

Assuming that $\frac{\partial H(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}$ is of full rank, it follows that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \approx -\left(\frac{\partial H(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right)^{-1} H(\boldsymbol{\beta}_0), \tag{2.26}$$

and therefore from equation (2.24)

$$\mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|\mathbf{X}) \approx -\left(\frac{\partial H(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right)^{-1} \mathbb{E}(H(\boldsymbol{\beta}_0)|\mathbf{X}) = 0. \tag{2.27}$$

In addition, we obtain from equation (2.26) that

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) \approx \left(\frac{\partial H(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right)^{-1} \mathrm{Var}(H(\beta_0)|\mathbf{X}) \left[\left(\frac{\partial H(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right)^{-1}\right]^{\top}.$$

We can therefore obtain a sandwich variance estimator for $\hat{\boldsymbol{\beta}}$, given as

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}) = \left(\frac{\partial H(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right)^{-1} \hat{\mathbb{V}}(H(\hat{\boldsymbol{\beta}})|\mathbf{X}) \left[\left(\frac{\partial H(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right)^{-1}\right]^{\top}, \tag{2.28}$$

where $\hat{\mathbb{V}}(H(\hat{\boldsymbol{\beta}})|\mathbf{X})$ is usually a plug-in estimate of $\mathrm{Var}(H(\beta_0)|\mathbf{X})$, i.e. $\mathrm{Var}(H(\beta_0)|\mathbf{X})$ evaluated at $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}$ (Chambers, 2009; Kim and Chambers, 2012c).

In his work, Chambers (2009) considered a typical form of $H(\boldsymbol{\beta})$ as:

$$\begin{aligned} H(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \boldsymbol{G}_i(\boldsymbol{\beta}) \left[Y_i - f_i(\beta)\right] \\ &= \sum_{v=1}^{V} \sum_{i=1}^{n_v} \boldsymbol{G}_i(\boldsymbol{\beta}) \left[Y_i - f_i(\beta)\right], \end{aligned} \tag{2.29}$$

where $f_i(\boldsymbol{\beta}_0) = \mathbb{E}(Y_i|\mathbf{X}_i)$ and $\boldsymbol{G}_i(\boldsymbol{\beta})$ is a p-vector function of $\mathbf{X}_i$ and $\boldsymbol{\beta}$, but not Y_i.

Clearly, this is a unbiased estimating equation for $\boldsymbol{\beta}_0$. Although f_i is arbitrary, the linear and logistic regression are of interest to the author.

In the context of secondary analysis for linked data, we can only observe the linked value Z_i instead of the true value Y_i. If we treat Z_i as if it was correctly linked, a naive estimating equation can be defined by simply replacing Y_i in (2.29) with Z_i as follows:

$$H_{naive}(\boldsymbol{\beta}) = \sum_{v=1}^{V}\sum_{i=1}^{n_v} \boldsymbol{G}_i(\boldsymbol{\beta})\left[Z_i - f_i(\beta)\right]. \tag{2.30}$$

Then,

$$\mathbb{E}\left[H_{naive}(\boldsymbol{\beta}_0)|\mathbf{X}\right] = \sum_{v=1}^{V}\sum_{i=1}^{n_v} \boldsymbol{G}_i(\boldsymbol{\beta}_0)\mathbb{E}\left[\sum_{j=1}^{n_v} l_{ij}Y_j - f_i(\boldsymbol{\beta}_0) \mid \mathbf{X}\right].$$

Under the non-informative linkage assumption (2.22) and the linkage error model (2.21), we have

$$\begin{aligned}
\mathbb{E}\left[H_{naive}(\boldsymbol{\beta}_0)|\mathbf{X}\right] &= \sum_{v=1}^{V}\sum_{i=1}^{n_v} \boldsymbol{G}_i(\boldsymbol{\beta}_0)\left[\sum_{j=1}^{n_v}\mathbb{E}(l_{ij}|\mathbf{X})\mathbb{E}(Y_j|\mathbf{X}) - f_i(\boldsymbol{\beta}_0)\right] \\
&= \sum_{v=1}^{V}\sum_{i=1}^{n_v} \boldsymbol{G}_i(\boldsymbol{\beta}_0)\left[\alpha_v\mathbb{E}(Y_i|\mathbf{X}) + \sum_{j\neq i, j=1}^{n_v}\frac{1-\alpha_v}{n_v-1}\mathbb{E}(Y_j|\mathbf{X}) - f_i(\boldsymbol{\beta}_0)\right] \\
&= \sum_{v=1}^{V}\sum_{i=1}^{n_v} \boldsymbol{G}_i(\boldsymbol{\beta}_0)\left[(\alpha_v-1)f_i(\boldsymbol{\beta}_0) + \sum_{j\neq i, j=1}^{n_v}\frac{1-\alpha_v}{n_v-1}f_j(\boldsymbol{\beta}_0)\right],
\end{aligned}$$

which differs from 0 if there are linkage errors, i.e. $\alpha_v < 1$. Therefore, the use of the naive estimating equation may lead to biased estimation. Given the value of α_v, Chambers (2009) proposed an adjusted estimating equation which can correct for this bias. The equation is defined as:

$$\begin{aligned}
H_{adj}(\boldsymbol{\beta}) &= H_{naive}(\boldsymbol{\beta}) - \sum_{v=1}^{V}\sum_{i=1}^{n_v} \boldsymbol{G}_i(\boldsymbol{\beta})\left[(\alpha_v-1)f_i(\boldsymbol{\beta}) + \sum_{j\neq i, j=1}^{n_v}\frac{1-\alpha_v}{n_v-1}f_j(\boldsymbol{\beta})\right] \\
&= \sum_{v=1}^{V}\sum_{i=1}^{n_v} \boldsymbol{G}_i(\boldsymbol{\beta})\left[Z_i - \left(\alpha_v f_i(\boldsymbol{\beta}) + \frac{1-\alpha_v}{n_v-1}\sum_{j\neq i, j=1}^{n_v} f_j(\boldsymbol{\beta})\right)\right].
\end{aligned} \tag{2.31}$$

In practice, suitable forms of $\boldsymbol{G}(\boldsymbol{\beta})$ and $f_i(\boldsymbol{\beta})$ would be defined depending on each application model. For example, for the linear regression model, we have $f_i(\boldsymbol{\beta}) = \mathbf{X}_i^T\boldsymbol{\beta}$ and by setting $\boldsymbol{G}_i(\boldsymbol{\beta})$ by an adjustment $\boldsymbol{G}_{i,adj}(\boldsymbol{\beta}) = \left(G^1_{i,adj}, \ldots, G^p_{i,adj}\right)^\top$ where $G^k_{i,adj} = \alpha_v X^k_i + \frac{1-\alpha_v}{n_v-1}\sum_{j\neq i, j=1}^{n_v} X^k_j$ for $k = 1, \ldots, p$, the adjusted estimating

equation would lead to the same unbiased estimator as (2.20) of Lahiri and Larsen (2005). This approach can also be applied on logistic regression models.

Following this work, Kim and Chambers (2012c,b) developed methods for multiple databases linkage and incomplete matching space. Chambers and Kim (2015) reviewed recent developments for the inference on linear or logistic regression parameters, using linked data. Chambers et al. (2019) studied the domain estimation under informative linkage context. Chambers and Diniz da Silva (2020) provided an empirical study of the effect of linkage errors in secondary analysis of linked data and new estimation methods which allow for linkage errors under the ELE framework. Recently, Zhang and Tuoto (2020) proposed a pseudo ordinary least square method for secondary linkage-data linear regression analysis, which can accommodate heterogeneous linkage errors and incomplete match space problem. Chambers et al. (2022) provided the robust inferences for linked data.

2.2.3 Cox regression analysis with linked data

The Cox proportional hazard model

Survival data, a.k.a. time-to-event data, is a common type of outcome in epidemiological and clinical studies. The Cox proportional hazard model (Cox, 1972) is the most popular method to assess the effect of covariates (e.g., age, gender or blood pressure) on the survival time. For example, this may be the duration of treatment until death. Survival data also occur in a variety of fields, such as the duration of unemployment for active people, or the duration until failure of an electronic device. An example of survival data in medical studies is presented in Table 2.8. The dataset registered survival time (expressed in days) of patients with lung cancer from different institution (`inst`) of the North Central Cancer Treatment Group. There are also informations about `age` (in years), `Gender` (Male = 1, Female = 2), calories consumed at meals (`meal.cal`), weight loss in the last six months (`wt.loss`) and different performance scores such that ECOG (`ph.ecog`, equal to 0 if the patient is asymptomatic, 1 if the patient is symptomatic but completely ambulatory, 2 if the patient is in bed $< 50\%$ of the day, 3 if the patient is in bed $> 50\%$ of the day but not bed bound, 4 if the patient is bed bound), Karnofsky performance score rated by a physician (`ph.karno`, from 0 (bad) to 100 (good)) and rated by the patient (`pat.karno`). Researchers may be interested in the relationship between the survival time and these covariates, in order to improve patients' healthcare.

A typical characteristic of survival data is censoring. Due to time or financial

id	inst	time	status	age	gender	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
1	3	306	2	74	1	1	90	100	1175	NA
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90	NA	15
4	5	210	2	57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90	NA	0
6	12	1022	1	74	1	1	50	80	513	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.8: Survival duration time (in days) of 228 patients with lung cancer from the North Central Cancer Treatment Group. Source: Loprinzi et al. (1994)

constraints, data are usually collected over a finite time period. This means that we can only observe the exact survival time for the subjects for which the event of interest occurs during the follow-up period. For the other subjects, we obtain so-called censored observations. For example, in Table 2.8, the `status` variable is equal to 1 if the patient is censored, and to 2 if the patient is dead. There are three major types of censoring: right, left and interval censoring. The most common type of censoring is right censoring, i.e. when we only know that the event happens posterior to the observed time. This may occur when a patient drops out or is lost to follow-up before the end of the study (e.g. patient 4 in Figure 2.1), or when a patient is event free during the observation period (e.g. patient 3 in Figure 2.1). In contrast to right censoring, an observation is left censoring if the true survival time is less than or equal to the observed time. A typical example for this situation is the virus testing, e.g. the testing of virus SARS-CoV-2 which caused COVID 19 pandemic. If an individual gets positive results, we can only know that they are exposed to the disease before the recorded time, which is the time of testing. However, suppose that the individual had the first test with negative results at time t_1, and was recorded positive at the second test of time t_2. In this case, we may know that the individual was exposed to the virus between t_1 and t_2, but we do not know the exact time of the disease. This case is referred to as interval censoring. Censored observations still provide partial information on the event time, which differs from missing observations. Ignoring this information implies bias in the inference.

Let $\tilde{T}$ denote a non-negative random variable, which stands for the duration between a time origin and the time of occurrence of some event of interest. It is assumed to be right censored: we observe the event only if it occurs before a certain time C. Suppose that we have a random sample of n observations, with $\tilde{T}_i$ and C_i the latent survival time and the censoring time for unit i. For units

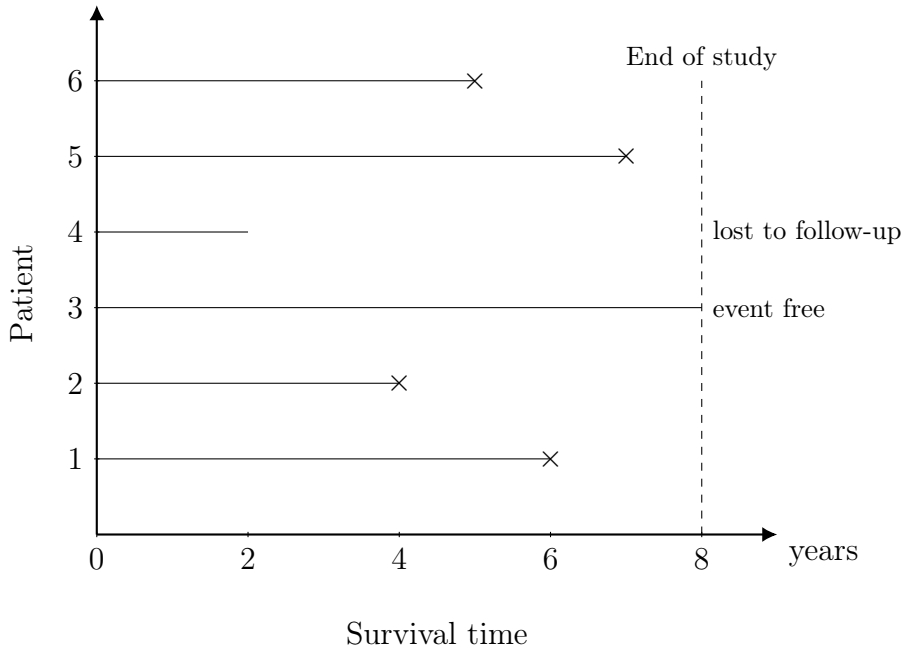


Figure 2.1: Example of right censoring observations (Patient 3 and 4)

$i = 1, \ldots, n$, we observe only the censored time $T_i = \min(\tilde{T}_i, C_i)$ and the non-censoring indicator $\delta_i = \mathbb{1}_{\{\tilde{T}_i \leq C_i\}}$, where $\mathbb{1}$ is the indicator function. The vector of covariates is denoted as $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)^T$.

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

$$\begin{aligned} \lambda(t|\mathbf{X}_i) &= \lim_{dt \to 0} \frac{\mathbb{P}\{t \leq T < t + dt \mid T \geq t\}}{dt} \\ &= \lambda_0(t) \exp\left(\mathbf{X}_i^T \boldsymbol{\beta}\right), \end{aligned} \tag{2.32}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of $p \times 1$ unknown parameters, and $\lambda_0(t)$ is a common baseline hazard function which can be interpreted as the hazard function for subjects with $\mathbf{X} = 0$. The baseline hazard function can take any shape as a function of t, but needs to be non-negative.

For any values $\mathbf{X}_0$ and $\mathbf{X}_1$ for the vector of covariates, the hazard ratio

$$\frac{\lambda\left(t \mid \mathbf{X}_1\right)}{\lambda\left(t \mid \mathbf{X}_0\right)} = \frac{\lambda_0(t) \mathrm{e}^{\mathbf{X}_1^T \boldsymbol{\beta}}}{\lambda_0(t) \mathrm{e}_0^{\mathbf{X}_0^T \boldsymbol{\beta}}} = \mathrm{e}^{(\mathbf{X}_1 - \mathbf{X}_0)^T \boldsymbol{\beta}}, \text{ for all } t \geq 0$$

is constant over time, hence the proportional hazard model denomination. In clinical trials, hazard ratios are often used to compare survival times of two different groups. For example, consider two groups differ only in treatment condition,

a hazard ratio of 2 means that a group has two times the probability that the event occurs, as compared to the comparison group. For continuous explanatory variable, the hazard ratio indicates the change in risk of event when this variable increases by 1 unit assuming all other covariates are fixed. For example, the risk of death if the patient is one year older given all other conditions are the same. Assume that the data are observed on a finite interval, and that C is independent of $\tilde{T}$ conditionally on $\mathbf{X}$. An estimator of $\boldsymbol{\beta}$ may be obtained by solving the estimating equation (Hu and Lin, 2002):

$$H_0(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{\mathbf{X}_i - \frac{\sum_{j=1}^{n} Y_j(T_i)\exp\left(\mathbf{X}_j^T\boldsymbol{\beta}\right)\mathbf{X}_j}{\sum_{j=1}^{n} Y_j(T_i)\exp\left(\mathbf{X}_j^T\boldsymbol{\beta}\right)}\right\} = 0, \tag{2.33}$$

where $Y_j(t) = \mathbb{1}_{(T_i \geq t)}$ is an at-risk indicator. We call (2.33) the theoretical estimating equation. This is also the partial likelihood score equation (Andersen and Gill, 1982). Since there is no closed-form solution, the Newton-Raphson algorithm is often used to solve this equation (Dennis and Schnabel, 1996).

Under some mild assumptions, the estimator $\hat{\boldsymbol{\beta}}$ obtained by solving equation (2.33) is consistent and asymptotically normal (see Andersen and Gill, 1982). This is also the maximum partial likelihood (mpl) estimation. A consistent estimator of the covariance matrix of $\hat{\beta}$ is given by

$$\hat{\mathbb{V}}_{\text{mpl}}(\hat{\boldsymbol{\beta}}) = \left\{-n\nabla H_0(\hat{\boldsymbol{\beta}})\right\}^{-1}, \tag{2.34}$$

see Andersen and Gill (1982).

Cox regression analysis with linked data

Although the Cox model has become the most used model for modeling the relationship of covariates to a survival outcome, there have been very few research works on using this model with linked data. Baldi et al. (2010) performed a simulation study emphasizing that incomplete record linkage is potentially leading to inefficient and biased estimation for parameters of Cox regression model, particularly in presence of medium or small sample sizes. However, there is no solution suggested.

Under the primary analysis situation, Hof et al. (2017) proposed a joint modeling for survival analysis and probabilistic record linkage framework of Fellegi and Sunter (1969). In their article, they considered a scenario under which the time-to-event variable has been registered in one database for a set of individuals, and

the covariates needed for modelling are registered in a separate database. These two databases also have some common partially identifying variables, which are used for matching purpose. In a two-stage approach, the set of matching indicators is firstly estimated by a record linkage method in which the probability of a match is assumed to depend only on the comparison vectors. After that, we fit the survival model to the linked data. In this approach, any errors in the first stage may lead to biased estimates in the regression model. Unlike this two-stage model, Hof et al. (2017) proposed a joint likelihood for both record linkage and survival model, which allows the matching distribution to depend on both the covariates and the time-to-event data. Simulation results show that the joint model gives unbiased regression parameter estimates for a Poisson process, with a good coverage of the confidence interval. However, their proposed variance estimation should be used as an acceptable approximation. A formal proof for the asymptotic normality remains challenging.

In other situations, where access to the partially identifying variables is restricted due to confidential issues, it is essential that the two stages are separated and performed by independent teams. In this case, the joint model of Hof et al. (2017) can not be applied. In Chapter 4, we propose a method to account for linkage errors in Cox regression model from this secondary analysis position.

3 Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system

Probabilistic record linkage is a process of combining data from different sources, when such data refer to common entities and identifying information is not available. Fellegi and Sunter have proposed a probabilistic record linkage framework that takes into account multiple non-identifying information, but is limited to simple binary comparison between matching variables. An extension of this method is proposed for mixed-type comparison vectors. A mixture model for handling comparison values of low prevalence categorical matching variables, and a mixture of hurdle gamma distribution for handling comparison values of continuous matching variables have been developed. The parameters are estimated by means of the Expectation Conditional Maximization (ECM) algorithm. Through a Monte Carlo simulation study, both the posterior probability estimation for a record pair to be a match and the prediction of matched record pairs are evaluated. The simulation results indicate that the proposed methods outperform existing ones in most considered cases. The proposed methods are applied on a real dataset, to perform linkage between a registry of patients suffering from venous thromboembolism in the Brest district area (GETBO) and the French national health information system (SNDS).

Contents

3.1 Introduction

Electronic health records have become more and more prevalent in medical fields, and the ability to exchange this information can help in providing better care for patients as well as richer sources for researchers. Record linkage is a process of combining data from different sources that refer to the same entity. The process is straightforward if each record contains a unique identifier such as Social Security Number (Zhu et al., 2015). However, some large health databases may not contain such identifying information. In other cases, this information is available but may contain errors, or may not be used for record linkage due to ethical reasons. Fellegi and Sunter (1969) proposed a probabilistic framework that takes into account multiple quasi-identifiers such as name, address and postal code. It has become widely used in applications when unique identifiers are unavailable or when data contain errors (e.g. Grannis et al., 2003; Sayers et al., 2015).

The French SNDS (Système National des Données de Santé) is the national health data system including the national health insurance information (SNIIRAM: Système National d'Information InterRégimes de l'Assurance Maladie) of around 99%

of the French population (Bezin et al., 2017). This data system also includes information on all health care expenses, as well as private and public hospital data collected in the medical information system (see Tuppin et al., 2017a). There is therefore an increasing demand of getting this information from SNDS, to enrich research datasets in epidemiology or public health. However, due to ethical reasons, the SNDS database is anonymous. This means that personal identifying information such as Social Security Number, Name or Address is not available. We are therefore interested in proposing a probabilistic record linkage model using other variables in common represented by the so-called matching variables. They can be of various types (categorical, binary, continuous) depending on the research study. For example, the matching variables may include postal code (categorical), date of treatment (continuous) and medical diagnosis (binary).

The Fellegi-Sunter probabilistic record linkage model laid the foundation for most record linkage models until now (Christen and Winkler, 2017). Although this model is useful for many applications in sample surveys and epidemiology, it has a limitation when some matching variables are binary and with a low prevalence (e.g., medical diagnoses). In that case, the simple binary comparison method proposed by Fellegi and Sunter (1969) can not distinguish the agreement of low prevalence values, which is much more informative than the agreement of high prevalence values. Such cases are considered in Hejblum et al. (2019), who propose a Bayesian linkage framework outperforming the Fellegi-Sunter model. However, their model is restricted to binary matching variables only.

Another limitation is that most probabilistic record linkage models only make use of simple binary or categorical comparison values (see Christen, 2012) even if the matching variables are continuous. Some authors introduced continuous similarity measures for comparing string data, but then comparison values are transferred to categorical values representing different levels of agreement (e.g., Herzog et al., 2007; Sadinle, 2017; Enamorado et al., 2019), which may result in a loss of information.

In this work, we propose a new linkage model adapted from the framework of Fellegi and Sunter, which handles such situations. We aim at better taking into account the nature of matching variables (e.g., low-prevalence binary, or continuous), so as to improve the performances of record linkage. The chapter is organized as follows. In Section 3.2, we review the Fellegi-Sunter probabilistic record linkage model and some relevant problems. We then propose two comparison strategies for low prevalence binary or continuous matching variables in

Section 3.3. An extended mixture model taking into account both categorical and continuous comparison values is also introduced in Section 3.3. In Section 3.4, we evaluate the proposed methods through simulation studies. In Section 3.5, a real data application is proposed, where we perform record linkage between SNDS and the GETBO (Groupe d'Etude de la Thrombose de Bretagne Occidentale) registry. Finally, possible further research is discussed in Section 3.6.

3.2 Probabilistic record linkage

Consider two databases A and B containing n_A and n_B records respectively, and with elements in common. Following the terminology in Fellegi and Sunter (1969), each possible pair of individuals (a_i, b_j) with $a_i \in A, i = 1, \ldots, n_A$ and $b_j \in B, j = 1, \ldots, n_B$ either belongs to the set of true matched pairs

$$M = \{(a, b); a = b, a \in A, b \in B\},$$

or to the set of true unmatched pairs

$$U = \{(a, b); a \neq b, a \in A, b \in B\}.$$

Because an identifying variable is not available, other less discriminant data are used in the probabilistic record linkage procedure, such as the name, date of birth, postal code, or some diagnosis codes. This information needs to be registered in both data sets and is referred to as matching variables. The matching variables in two databases are required to have the same format (Christen, 2012).

It is supposed that there is no prior knowledge on how likely the matches are, which is often the case in practice. The strategy therefore begins by comparing K matching variables for all records $\mathbf{X}_{A,i} = (X^1_{A,i}, \ldots, X^K_{A,i}), i = 1, \ldots, n_A$ of n_A individuals in A, with all records $\mathbf{X}_{B,j} = (X^1_{B,j}, \ldots, X^K_{B,j}), j = 1, \ldots, n_B$ of n_B individuals in B. This leads to $n_A \times n_B$ comparison vectors $\boldsymbol{\gamma}_{ij}$ such that

$$\boldsymbol{\gamma}_{ij} = \left\{\gamma^1_{ij}, \ldots, \gamma^k_{ij}, \ldots, \gamma^K_{ij}\right\}, \tag{3.1}$$

where $\gamma^k_{ij} = h^k(X^k_{A,i}, X^k_{B,j})$ and h^k is a comparison function for the $k-$th matching variable.

Because the number of all record pairs is quadratic in the number of individuals in each database, making the comparison for all possible record pairs is often impracticable in applications. One of the most popular methods to reduce the number

of record pairs that need to be compared is blocking, in which only records from the two databases that are in a same block (i.e., sharing the same values for the blocking variables) are compared with each other. Record pairs disagreeing on the blocking variable are automatically classified as non-matches. Therefore, blocking is a trade-off between computational cost and the proportion of missed matches (matched pairs are missed because of errors in the blocking variable), see Herzog et al. (2007).

The set of all possible realizations of $\boldsymbol{\gamma}$ is called the comparison space and denoted by Γ. The comparison function $\boldsymbol{\gamma}^k$ for the k-th matching variable can be defined in different ways depending on the type of matching variable (Christen, 2012). The most common way consists in a binary comparison, i.e.

$$\gamma_{ij}^k = h^k(X_{A,i}^k, X_{B,j}^k) = \begin{cases} 1 & \text{if} \quad X_{A,i}^k = X_{B,j}^k, \\ 0 & \text{if} \quad X_{A,i}^k \neq X_{B,j}^k. \end{cases} \tag{3.2}$$

If there is no error in the matching data, all components of a comparison vector of a matched pair are equal to 1. However, application data usually contain errors (e.g., typographical), and some similarity measures that can take them into account have been developed in the literature for string variables (Herzog et al., 2007).

Once all candidate pairs are compared, various approaches are possible to classify the set of comparison vectors into matches and non-matches (Christen, 2012). If training data where we observe the true matched status of record pairs is available, supervised classification methods (Christen, 2008) can be used to find a classification rule. If there is no training data but some clerical review is possible, some semi-supervised approaches (e.g. Enamorado, 2018) may be applied. However, the exact knowledge of matches is rarely possible in real world situations, and the clerical review is costly. Unsupervised methods (e.g. Winkler, 1988; Mamun et al., 2016) are therefore the more common approaches. From a Bayesian perspective, Tancredi and Liseo (2011) introduced a paradigm for probabilistic record linkage, and Steorts et al. (2016) proposed a Bayesian approach to graphical record linkage.

In the frequentist view, Fellegi and Sunter (1969) assumed that each record pair belongs to one of the two latent classes. The distribution of comparison vector $\boldsymbol{\gamma}$ for each pair is assumed to follow a mixture model

$$\mathbb{P}(\boldsymbol{\gamma}) = \mathbb{P}(\boldsymbol{\gamma}|M)\mathbb{P}(\boldsymbol{\gamma} \in M) + \mathbb{P}(\boldsymbol{\gamma}|U)\left[1 - \mathbb{P}(\boldsymbol{\gamma} \in M)\right]. \tag{3.3}$$

If we do not make additional assumptions on the joint agreement pattern, the comparison vector $\boldsymbol{\gamma}$ may take 2^K different values, each of which corresponds to a parameter that we need to estimate. To reduce this number, some authors (e.g. Fellegi and Sunter, 1969; Winkler, 1988) have proposed to make the so-called conditional independence assumption between fields of the comparison vector. Under this assumption, we obtain:

$$\mathbb{P}\left[\boldsymbol{\gamma} = (\gamma^1, \ldots, \gamma^K)|M\right] = \prod_{k=1}^{K} \mathbb{P}(\gamma^k|M), \tag{3.4}$$

$$\mathbb{P}\left[\boldsymbol{\gamma} = (\gamma^1, \ldots, \gamma^K)|U\right] = \prod_{k=1}^{K} \mathbb{P}(\gamma^k|U). \tag{3.5}$$

The conditional independence assumption is common in most probabilistic record linkage models (Winkler, 1988), although it may not hold in some practical cases. For example, if some records agree on a chronic disease, they are more likely to agree on the drug used. Although the assumption is invalid in some cases, the linkage result is still quite robust, in the sense that we may have a good linkage performance even if the conditional independence assumption does not hold (Winkler, 1988; Grannis et al., 2003; Sayers et al., 2015). Some authors (e.g. Xu et al., 2019) relaxed this assumption and showed better record linkage results in some specific scenarios.

Under the conditional independence assumption, we only need to estimate $2K+1$ parameters which are the marginal probabilities of agreement for matched and unmatched pairs $m^k \equiv \mathbb{P}(\gamma^k = 1|M)$ and $u^k \equiv \mathbb{P}(\gamma^k = 1|U)$, and the overall matching probability $p_M \equiv \mathbb{P}(\boldsymbol{\gamma} \in M)$. Winkler (1988) proposed to apply the expectation maximization (EM) algorithm (Dempster et al., 1977; Wu, 1983), to find the maximum likelihood estimates for the vector of parameters $\boldsymbol{\theta} \equiv \left\{p, m^k, u^k, k = 1, \ldots, K\right\}$. It has become widely used in probabilistic record linkage (Grannis et al., 2003; Christen, 2012). Once all the parameters are estimated, the record pairs may be ordered by either matching weights

$$\hat{w}_{ij} = \frac{\mathbb{P}(\boldsymbol{\gamma}_{ij}|M, \hat{\theta})}{\mathbb{P}(\boldsymbol{\gamma}_{ij}|U, \hat{\theta})},$$

see Fellegi and Sunter (1969); Belin and Rubin (1995), or by posterior probabilities of matching $\hat{q}_{ij} \equiv \mathbb{P}(M|\boldsymbol{\gamma}_{ij}, \hat{\theta})$ (Larsen and Rubin, 2001). Then, the pairs are classified into matches, non-matches or possible matches based on two defined thresholds (Fellegi and Sunter, 1969). Because the possible matches require manual review which is sometimes not available, Grannis et al. (2003) propose to

establish only a single threshold to avoid human review. Although the matching scores and the posterior probabilities produce the same ordering for record pairs (Larsen and Rubin, 2001), the posterior probabilities are preferable in our case because they may be useful for further analyses (Lahiri and Larsen, 2005; Kim and Chambers, 2012a; Hof and Zwinderman, 2012; Zhang and Tuoto, 2020).

In some applications, a one-to-one matching restriction may be needed; namely, that each record in B can be matched to one and only one record in A, and conversely. One possible approach to respect a one-to-one matching is to solve a linear sum assignment problem proposed by Jaro (1989). If the optimal score is not demanded, a simple approach is to sort all candidate pairs according to their estimated posterior probabilities of matching, and to select matched pairs in a greedy approach (Christen, 2012).

3.3 An extension of the Fellegi-Sunter model

In this section, we extend the Fellegi-Sunter model by making better use of low prevalence categorical matching variables and of continuous variables. Two new comparison approaches and a mixture model for mixed type of comparison values are introduced.

3.3.1 Comparison approaches

For a categorical matching variable, it is likely that the proportions for each category are different, and accounting for these differences in a record linkage model may help to improve the linkage results. This idea was proposed by Fellegi and Sunter (1969); Winkler (1989), and is applied on a real clinical data in Zhu et al. (2009). These authors use the same model for simple agreement/disagreement comparison, but the matching weights are rescaled a posteriori, using a frequency-based correction. We introduce a new comparison approach for categorical matching variables, which differs from simple binary comparison and may naturally handle different proportions for categories.

Let X^k be a categorical matching variable taking L different values, which means that the comparison function for this variable may take up to L^2 values. For example, the comparison for a binary matching variable may lead to four possible realizations $\{(0,0),(0,1),(1,0),(1,1)\}$ and a comparison function can be defined

as follows

$$h^k(0,0) = c_1, \quad h^k(0,1) = c_2, \quad h^k(1,0) = c_3 \quad \text{and} \quad h^k(1,1) = c_4, \tag{3.6}$$

where c_1, c_2, c_3 and c_4 stand for four different categories. It should be noted that the values taken by the comparison function have no ordinal meaning. If this is a low prevalence binary matching variable (e.g. a rare disease) such that only 5% (say) of the values in the dataset are equal to 1, the agreement on the value "1" is much more informative than the agreement on the value "0". Our comparison approach aims at using this information while the simple agreement comparison method does not, leading to poor performance. Hejblum et al. (2019) propose a Bayesian record linkage framework making use of a similar idea, and which is efficient in case of a large number of low-prevalence binary matching variables. However, their model is designed for binary variables only.

If the number of matching variables and/or the number of categories is large, the number of parameters to be estimated is $L^2 - 1$, which may be too large in practice. This number may be reduced by assigning the same comparison value for the agreement/disagreement of categories which have a close meaning. For instance, we may reduce the comparison values given in (3.6) as

$$h^k(0,0) = c_1, \quad h^k(0,1) = h^k(1,0) = c_2 \quad \text{and} \quad h^k(1,1) = c_3. \tag{3.7}$$

In general, the number of comparison values depends on which realizations we would like to distinguish. Suppose that we are interested in a categorical matching variable X^k with categories $1, 2, \ldots, L$. If the first category seems particularly meaningful, we may distinguish whether we have an agreement on the first category, an agreement on another category, or a disagreement. In such case, the comparison function would be defined as

$$h^k(i,j) \quad = \quad \begin{cases} c_1 & \text{if } i = j = 1, \\ c_2 & \text{if } i = j \neq 1, \\ c_3 & \text{if } i \neq j = 1, \ldots, L. \end{cases}$$

The objective of this comparison approach is to distinguish the agreement of low prevalence values from other agreements, which differs from multiple levels of agreement introduced in (Sadinle, 2017) and (Enamorado et al., 2019).

Now, let us consider the case of a continuous variable X^k. For example, date

variables (e.g., admission to the hospital, or medical act) are common in medical datasets. By converting each date into a duration from a specified origin, they may be treated as continuous counting variables. Even if an individual is present in both datasets, a lag between dates is likely to appear. The simple binary comparison is therefore not appropriate. In this work, if the k^{th} matching variable is continuous, we propose to consider

$$\gamma_{ij}^k = h^k(X_{A,i}^k, X_{B,j}^k) = d(X_{A,i}^k, X_{B,j}^k), \tag{3.8}$$

where d is a distance which can be used to measure the difference between two dates of events, in which case it can be interpreted as a time lag. By using the distance, the continuous comparison values γ^k of matching pairs $(X_{A,i}^k, X_{B,j}^k)$ can be described as

$$\gamma_{ij}^k | (X_{A,i}, X_{B,j}) \in M = \begin{cases} 0 & \text{with probability} \quad 1 - e^k, \\ \epsilon_{ij}^k > 0 & \text{with probability} \quad e^k, \end{cases}$$

where e^k is the proportion of error, and ϵ_{ij}^k is the error term of the k^{th} matching variable among matched pairs. For example, two patients who refer to the same individual should have the same day for a medical act, up to some errors in the registration process, and the distance should therefore be equal to 0 or to a small error term ϵ_{ij}^k. Therefore, $\gamma_{ij}^k | M$ follows a hurdle distribution in which the positive part depends only on the distribution of errors. On the other hand, the distribution of $\gamma_{ij}^k | U$ depends mostly on the distribution of the k^{th} matching variable, since ϵ_{ij}^k is often small compared to the distance between records for two unmatched units.

3.3.2 Estimation of parameters

Let

$$\boldsymbol{\gamma}_{ij} \quad = \quad \left(\gamma_{ij}^1, \ldots, \gamma_{ij}^{K_1}, \gamma_{ij}^{K_1+1}, \ldots, \gamma_{ij}^{K_1+K_2}\right) \tag{3.9}$$

be a mixed type comparison vector which includes K_1 categorical comparison values $\gamma_{ij}^1, \ldots, \gamma_{ij}^{K_1}$ and K_2 continuous distances $\gamma_{ij}^{K_1+1}, \ldots, \gamma_{ij}^{K_1+K_2}$. Following the Fellegi-Sunter framework, these comparison vectors are assumed to follow the mixture model (3.3).

Under the conditional independence assumption between the different fields in the

comparison vector for both the matched and the unmatched sets, we have

$$\mathbb{P}(\boldsymbol{\gamma}_{ij}|M) = \underbrace{\prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|M)}_{P_{ij}^{1M}} \underbrace{\prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|M)}_{P_{ij}^{2M}}, \tag{3.10}$$

$$\mathbb{P}(\boldsymbol{\gamma}_{ij}|U) = \underbrace{\prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|U)}_{P_{ij}^{1U}} \underbrace{\prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|U)}_{P_{ij}^{2U}}, \tag{3.11}$$

for $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$. For both equations (3.10) and (3.11), the first term in the right hand side involves K_1 categorical comparison values of the comparison vector γ_{ij}. We define

$$m_s^k = \mathbb{P}(\gamma_{ij}^k = s|M) \text{ and } u_s^k = \mathbb{P}(\gamma_{ij}^k = s|U) \text{ for } s \in S^k, \tag{3.12}$$

with S^k the set of all possible categorical comparison values for the k^{th} variable. Then

$$P_{ij}^{1M} = \prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|M) = \prod_{k=1}^{K_1} \prod_{s \in S^k} (m_s^k)^{\mathbb{1}_{\gamma_{ij}^k = s}},$$
$$P_{ij}^{1U} = \prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|U) = \prod_{k=1}^{K_1} \prod_{s \in S^k} (u_s^k)^{\mathbb{1}_{\gamma_{ij}^k = s}},$$

for $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$, and with $\sum_{s \in S^k} m_s^k = \sum_{s \in S^k} u_s^k = 1$.

The second part in the right hand side of equations (3.10) and (3.11) involves K_2 continuous values of the comparison vector $\boldsymbol{\gamma}$. We define

$$\begin{aligned} P_{ij}^{2M} &= \prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|M) \quad \text{with} \quad \mathbb{P}(\gamma_{ij}^k|M) \sim f_M^k(\phi_M^k), \\ P_{ij}^{2U} &= \prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|U) \quad \text{with} \quad \mathbb{P}(\gamma_{ij}^k|U) \sim f_U^k(\phi_U^k), \end{aligned} \tag{3.13}$$

for $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$. The distributions f_M^k and f_U^k need to be postulated, depending on the characteristics of the matching variables and on the chosen distance.

To find the maximum likelihood estimates for parameters, we apply the Expectation Maximization (EM) algorithm (Dempster et al., 1977) or the Expectation

Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993), depending on the distribution f^k. In Appendix A.1, we present the details of the ECM algorithm, when both f_M^k and f_U^k correspond to a hurdle gamma distribution (Cragg, 1971), which is used in the next part.

Once all parameters are estimated by means of the EM/ECM algorithm, the posterior probabilities $q_{ij} = \mathbb{P}(M|\boldsymbol{\gamma}_{ij})$ are estimated for all record pairs by the Bayes formula

$$\widehat{q}_{ij} = \frac{\widehat{p}_M \widehat{P}_{ij}^{1M} \widehat{P}_{ij}^{2M}}{\widehat{p}_M \widehat{P}_{ij}^{1M} \widehat{P}_{ij}^{2M} + (1-\widehat{p}_M)\widehat{P}_{ij}^{1U} \widehat{P}_{ij}^{2U}}. \tag{3.14}$$

These estimated posterior probabilities are then used to find proper matched pairs.

3.4 Simulation studies

In this section, our proposed approaches are evaluated and compared to other existing approaches. To facilitate interpretation, two simulation studies are performed to evaluate the properties of the proposed methods for binary and continuous variables separately. A simulation study for a combination of both categorical and continuous matching variables is presented in Appendix A.4. All the simulations are implemented in a R program, which is available on Github repository: https://github.com/thanhhuanVO/Extending-FellegiSunter-Record-linkage.git.

3.4.1 Simulation designs

In the following simulations, we consider two databases A and B containing $n_A = 500$ and $n_B = 200$ individuals and K matching variables. We assume that there is no duplicate in both databases, and that all individuals in B have corresponding individuals in A. The number of individuals in both databases remains fixed in our simulations. However, different sizes are considered in additional simulations available as a supplement in Appendix A.

We first generate the observations in A, and a random subset of n_B units is used to obtain the database B. For $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$ let us denote by

$$\mathbf{X}_{A,i} = \left(X_{A,i}^1, \ldots, X_{A,i}^K\right) \quad \text{and} \quad \mathbf{X}_{B,j} = \left(X_{B,j}^1, \ldots, X_{B,j}^K\right) \tag{3.15}$$

the i^{th} and j^{th} individual in A and B, respectively. Without loss of generality, we assume that the first unit in B is the first unit in A, ..., the n_B^{th} unit in B is the

n_B^{th} unit in A. The full comparison matrix $\gamma = \{\gamma_{ij}^k\}$ contains $n^A \times n^B = 100\,000$ lines and K columns.

Once the posterior matching probabilities are estimated for all possible record pairs, a pair is classified as a match if $\hat{q}_{ij}$ (see equation 3.14) is larger than a predefined threshold τ, and is classified as a non-match otherwise. The choice of the threshold depends on the objectives of the study, a higher threshold leading to a lower number of false matches.

Scenario 1: binary matching variables

Data generating process In this scenario, each variable $X_{A,i}^k$ is first generated according to a Bernoulli distribution with parameter p^k, for $k = 1, \ldots, K$. To account for possible errors in the matching variables, the variables $X_{B,j}^k$ in database B are then obtained as

$$X_{B,j}^k = \begin{cases} X_{A,j}^k & \text{with probability} \quad 1 - e^k, \\ 1 - X_{A,j}^k & \text{with probability} \quad e^k. \end{cases} \tag{3.16}$$

Simulation parameters Since the binary matching variables are less discriminant, all the methods tested require a large number K of matching variables, in order to have sufficient information for achieving acceptable linkage results. We therefore used $K \in \{30, 40, 50\}$. The probability of error is chosen as $e^k \in \{0.02, 0.04, 0.06\}$. For simplicity, the probability p^k for each Bernoulli variable is fixed to 0.2.

Methods Once the variables in the databases were generated, we considered four possible record linkage methods: `FS`, the Fellegi-Sunter model with simple binary comparison as described in (3.2); `FS3`, the Fellegi-Sunter model using a comparison with 3 categories, as described in (3.7); `FS4`, the Fellegi-Sunter model using a comparison with 4 categories, as described in (3.6); `Bayesian`, the bayesian method described in Hejblum et al. (2019). With the methods `FS`, `FS3` and `FS4`, the parameters p_M, m_s^k and u_s^k (see equation 3.12) are estimated by means of the EM algorithm, and some initial values are required. We initialize with $1/n_A$ for p_M. The formulas to compute the initial values for m_s^k and u_s^k and the stopping criteria are given in Appendix A.2. The `Bayesian` method is performed by means of the package ludic of Hejblum et al. (2019), where we used 0.01 as the discrepancy rates needed for the method.

Scenario 2: continuous matching variables

Data generating process In this scenario, each variable X_A^k is generated according to an exponential distribution with parameter λ^k, for $k = 1, \ldots, K$. To account for possible errors in the matching variables, the variables $X_{B,j}^k$ in database B are then obtained as

$$X_{B,j}^k = \begin{cases} X_{A,j}^k & \text{with probability} \quad 1 - e^k, \\ X_{A,j}^k + \epsilon_j^k & \text{with probability} \quad e^k, \end{cases} \tag{3.17}$$

where the ϵ_j^k's are iid, generated according to an exponential distribution of parameter λ_e^k.

Simulation parameters We used $K = 3$ matching variables and $\lambda^k = 0.02$ for $k = 1, \ldots, K$. Because small lags are likely to happen in the registration process, we considered as possible proportions of errors $e^k \in \{0.1, 0.2, 0.3\}$ and $\lambda_e^k \in \{1/2, 1/3, 1/4\}$. This leads to a mean value of approximately 50 days for X^k, and a mean value of approximately $2, 3$ or 4 days for the lag value ϵ_j^k.

Methods Once the databases were generated, we compared three possible record linkage methods: `FS`, the Fellegi-Sunter model with simple binary comparison as described in (3.2); `FS3`, the Fellegi-Sunter model using a comparison with 3 categories defined as follows:

$$\gamma_{ij}^k = \begin{cases} 0 & \text{if } |X_{B,j}^k - X_{A,i}^k| = 0, \\ 1 & \text{if } 0 < |X_{B,j}^k - X_{A,i}^k| \leq 3, \\ 2 & \text{if } 3 < |X_{B,j}^k - X_{A,i}^k|, \end{cases} \tag{3.18}$$

for $k = 1, \ldots, K$; `FS-HGa`, the Fellegi-Sunter model using the absolute distance for comparison defined as:

$$\gamma_{ij}^k = d(X_{A,i}^k, X_{B,j}^k) = |X_{B,j}^k - X_{A,i}^k|. \tag{3.19}$$

For the `FS-HGa` method, we used the hurdle Gamma distribution

$$f(\gamma^k; p_0^k, \alpha^k, \beta^k) = \begin{cases} p_0^k & \text{if } \gamma^k = 0, \\ (1 - p_0^k) \dfrac{(\gamma^k)^{(\alpha^k - 1)} e^{-\gamma^k/\beta^k}}{(\beta^k)^{(\alpha^k)} \Gamma(\alpha^k)} & \text{if } \gamma^k > 0, \end{cases} \tag{3.20}$$

for both f_M^k, f_U^k in equation (3.13) where $\alpha^k, \beta^k \in \mathbb{R}^+$ and $\Gamma(\alpha^k)$ is the gamma function for $k = 1, \ldots, K$. This is the true distribution for $\gamma^k|M$ under our simulation set-up, since

$$\gamma_{j,j}^k|M = |X_{B,j}^k - X_{A,j}^k| = \begin{cases} 0 & \text{with probability} \quad 1 - e^k, \\ \epsilon_j^k & \text{with probability} \quad e^k, \end{cases} \tag{3.21}$$

and since ϵ_j^k follows an exponential distribution, which is a particular case of the Gamma distribution with parameters $\alpha^k = 1$ and $\beta^k = 1/\lambda^k$. On the other hand, it is more complicated to describe the true distribution of $\gamma^k|U$. For $j \neq i$, we have

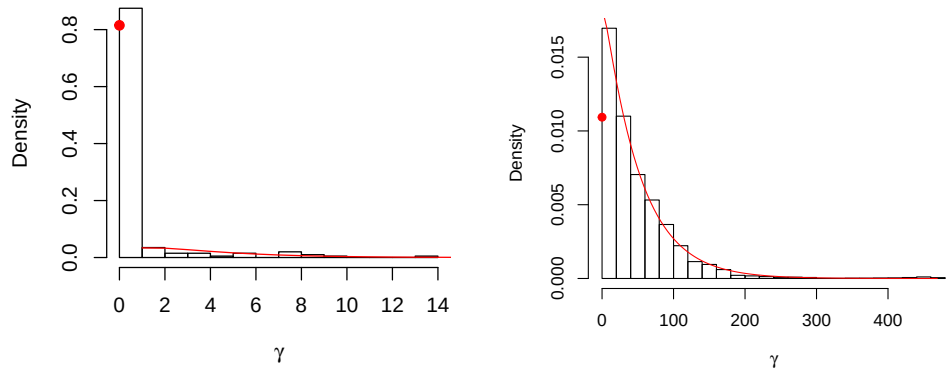
$$\gamma_{i,j}^k|U = |X_{B,j}^k - X_{A,i}^k| = \begin{cases} |X_{A,j}^k - X_{A,i}^k| & \text{with probability} \quad 1 - e^k, \\ |X_{A,j}^k - X_{A,i}^k + \epsilon_j^k| & \text{with probability} \quad e^k. \end{cases} \tag{3.22}$$

Since $X_{A,j}^k$ and $X_{A,i}^k$ are independent for $i \neq j$, $\gamma_{i,j}^k|U$ follows an exponential distribution with probability $1 - e^k$. With probability e^k, the distribution of $\gamma_{i,j}^k|U$ also involves that of the error ϵ^k. Since this error is typically small compared to the difference $X_{A,j}^k - X_{A,i}^k$, we may also consider that $\gamma^k|U$ approximately follows an exponential distribution.

With the `FS-HGa` method, we propose using the hurdle gamma distribution for f, since it adds more flexibility to the modeling. The histogram of $\gamma_{i,j}^k$ values on one sample is given in Figure 3.1a for the matched pairs, and in Figure 3.1b for the unmatched pairs. They decidedly indicate that the hurdle gamma distribution fits well to these values in this example. The robustness of our modeling with different families of distributions for X^k and ϵ^k is studied in Appendix A.3.6.

While the parameters for `FS` and `FS3` are estimated by the EM algorithm, the parameters for `FS-HGa` are estimated by the ECM algorithm, which is presented in Appendix A.1. The starting values and stopping criteria for all methods are presented in Appendix A.3.

Since the matching variables are interpreted as durations (in days) in the application presented in Section 3.5, the generated values X_A^k and ϵ^k values are rounded to the smallest larger integer in this simulation. For example, patients may get a medical act at different times (days, hours and minutes), but the durations are registered in days only.



(a) Histogram of $\gamma|M$ (b) Histogram of $\gamma|U$

Figure 3.1: Histogram of the positive values of $\gamma_{i,j}^k$ for the matched pairs (left side) and the unmatched pairs (right side) and fitted gamma density estimation (red curve) when $\lambda_e^k = 1/2$ and $e^k = 0.2$

3.4.2 Performance criteria

All the methods tested for record linkage are evaluated by means of the True Positive Rate

$$\text{TPR}_\tau \quad = \quad \mathbb{P}\left\{q_{ij} \geq \tau | (X_{A,i}, X_{B,j}) \in M\right\},$$

and the Positive Predictive Value

$$\text{PPV}_\tau \quad = \quad \mathbb{P}\left\{(X_{A,i}, X_{B,j}) \in M | q_{ij} \geq \tau\right\}.$$

The True Positive Rate (a.k.a sensitivity or recall) is the proportion of matched pairs which are correctly identified. The Positive Predictive Value (a.k.a. precision) is the proportion of predicted matched pairs which are correctly identified. These are the most common criteria in an imbalanced binary classification problem, which is the case when the overall set of record pairs is extremely dominated by non-matches. In this work, these criteria are estimated by means of $1,000$ independent Monte Carlo simulations. To save time, all the results are obtained by using the package simsalapar (Hofert and Mächler, 2016) for parallelizing the estimation of all combinations of simulation parameters. A server with 2 Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz with 12 cores in each has been used.

In the simulations, the EM/ECM algorithm is used for the estimation of parameters, with a convergence tolerance of 10^{-6} before reaching the maximum number

of iterations (set equal to 500). We observed convergence issues for record linkage methods, more particularly for the usual Fellegi-Sunter method in the Scenario 1: that is, there are some simulations for which the tolerance value is not reached after 500 iterations. The Monte Carlo approximation of the TPR and PPV for any method is therefore obtained from the subset of simulations for which the convergence is attained for all methods. The proportion of cases for which the convergence is attained is presented for all methods in Appendix A.2.2 and A.3.2.

The Monte Carlo approximation for the TPR and PPV are presented in Section 3.4.3 for the methods considered, with a threshold $\tau = 0.5$. This is the most natural threshold, since it is equivalent to classify a pair as a match if $\mathbb{P}(M|\gamma) \geq \mathbb{P}(U|\gamma)$. In practice, the choice of the threshold τ corresponds to a trade-off between TPR and PPV: a more stringent threshold may increase PPV, but decrease TPR. For a particular case of each scenario, we therefore plotted in Figures 3.3 and 3.5 the PPV-TPR curve (a.k.a. precision-recall curve) for different values of τ. Two types of curves are plotted. The "observed" curves correspond to the (theoretical) situation when the parameters are directly estimated by maximizing the full likelihood, assuming that the true status (matched/unmatched) is known for each pair. The "estimated" curves correspond to the (practical) situation when this status is not known, and the parameters are estimated by the EM/ECM algorithm as described in Section 3.4.1. The difference between an observed curve and its estimated counterpart is helpful to separate the effect in parameter estimation when using the EM/ECM algorithm.

3.4.3 Results

Scenario 1

The Monte Carlo estimates for the TPR and the PPV are presented in Figure 3.2. We first note that for all the methods considered, both criteria improve when the number of matching variables K increases and/or when the probability of error e decreases, as could be expected. In terms of TPR, `FS3` is preferable, followed by `FS4`; `Bayesian` and `FS` show comparable results for $K \leq 40$, but `FS` performs better for $K = 50$. In terms of PPV, `FS4` and `Bayesian` are preferable, with almost identical results; `FS3` performs slightly worse, while `FS` performs poorly, but both methods improve as K increases. Overall, `FS3` performs better than `FS` in both TPR and PPV. As explained in Hejblum et al. (2019), `FS` has many false matches, leading to the smallest PPV. In comparison to `FS4` and `Bayesian`, `FS3` improves the TPR substantially with a slight decrease of the PPV. `FS4` and `Bayesian` show a similar behavior when $e = 0.02$. However, when the error

increases, `FS4` has a better TPR with a minimal decrease in PPV as compared to `Bayesian`. In addition, we have also reported the proportion of convergence and the average execution time in Table A.1 in Appendix A. Generally, the `FS3` and `FS4` have a higher chance of convergence since their comparison vectors provide more information for the algorithm. However, they require a longer computation time than `FS`, since more parameters need to be estimated. In this specific scenario, the execution time of `Bayesian` is much faster than with other linkage methods, because it was performed by package ludic which is optimally designed for this specific scenario.

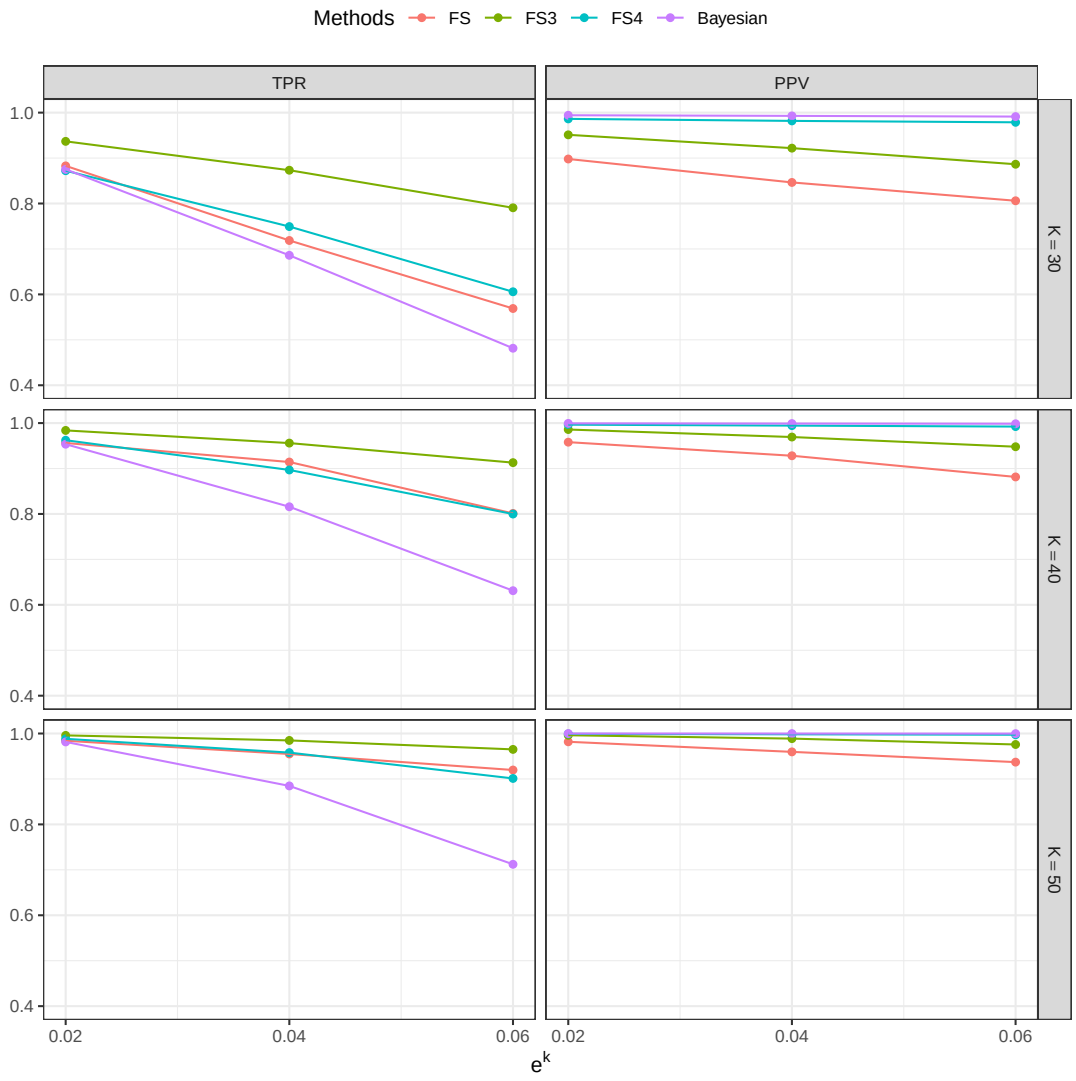


Figure 3.2: Monte-Carlo estimates of TPR and PPV with binary matching variables only and sample sizes $n_A = 500$ and $n_B = 200$, $p^k = 0.2$ for the parameter of the Bernoulli distribution, a number of matching variables $K \in \{30, 40, 50\}$, and a proportion of errors $e^k \in \{0.02, 0.04, 0.06\}$.

To evaluate the impact of the choice of the threshold τ in the performances of

the methods, we consider the particular scenario with the parameters $K = 40$, $p^k = 0.2$ and $e = 0.04$. We plot in Figure 3.3 the PPV in function of the TPR for different thresholds. The Figure 3.3 indicates that the observed `FS4` performs better among the observed methods, while the estimated `FS3` performs better among the estimated methods.

Additional simulations with a fixed number of matching variables $K = 40$ and different values for the probability p^k were performed in Appendix A.2.3. The results showed that all methods improve significantly when p^k rises from 0.1 to 0.3. Also, the results in Appendix A.2.4 indicate that all methods gradually improve as the ratio n_B/n_A increases.

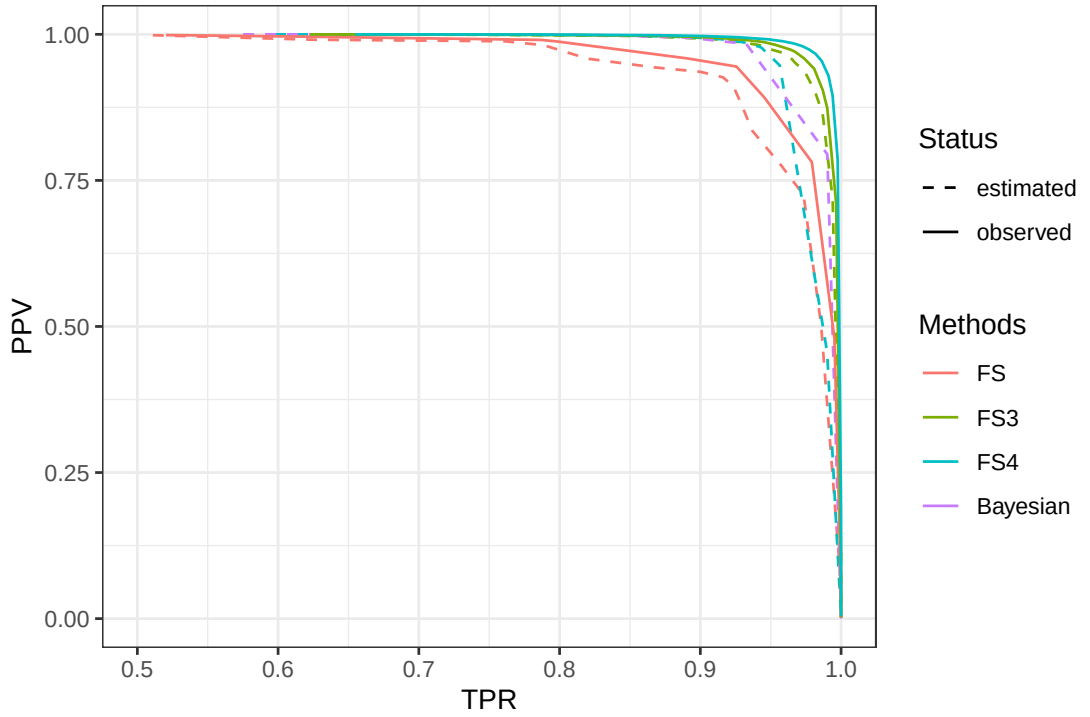


Figure 3.3: PPV-TPR curves for the observed/estimated version of the methods considered with binary matching variables only, with sample sizes $n_A = 500$ and $n_B = 200$, $K = 40$ matching variables, $p^k = 0.2$ for the parameter of the Bernoulli distribution, and a proportion of errors $e^k = 0.04$.

Scenario 2

The Monte Carlo estimates for the TPR and the PPV are presented in Figure 3.4. For each method, both the TPR and the PPV decrease as the proportion of errors e increases. We note that the slower decrease is observed for `FS-HGa`, which is also the method which gives both the best TPR and the best PPV in

all cases. We also observe that for `FS-HGa` and `FS3`, both the TPR and the PPV decrease with λ_e, but the decrease is very limited for `FS-HGa`. On the other hand, `FS` is not affected by λ_e: this is likely due to the fact that `FS` only considers exact agreement/disagreement in comparison step, while `FS3` accounts for an additional category when the time lag is no greater than 3 days. Therefore, `FS3` performs better than `FS` when the proportion of error is large ($e = 0.3$) and the mean value of the error is small ($\lambda_e = 1/2; 1/3$).

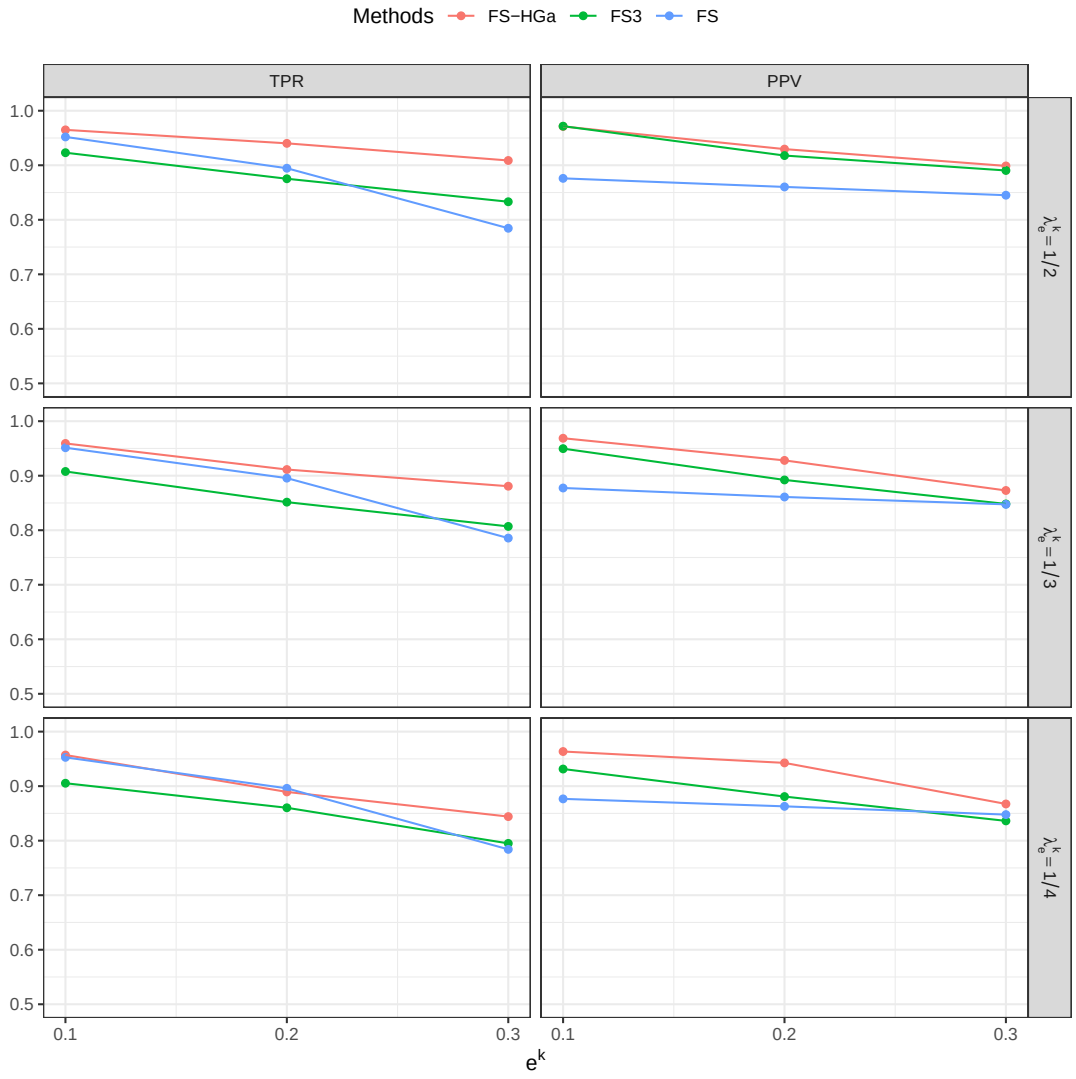


Figure 3.4: Monte-Carlo estimates of TPR and PPV over different simulation cases when there are only continuous matching variables with sample sizes $n_A = 500$ and $n_B = 200$, $K = 3$ matching variables, $\lambda^k = 0.02$ for the parameter of the Exponential distribution, a proportion of errors $e^k \in \{0.1, 0.2, 03\}$, and a parameter $\lambda_e^k \in \{1/2, 1/3, 1/4\}$ for the error lag.

We have also reported the proportion of convergence and the average execution time of each method in Appendix A.3.2. Since the implementation of the ECM

algorithm in `FS-HGa` has 2 maximization steps, it requires a longer computation time.

To evaluate the impact of the threshold τ, we consider the particular scenario with the parameters $K = 3$, $\lambda^k = 0.02$, $e = 0.2$ and $\lambda_e = 1/2$. We observe in Figure 3.5 that the PPV-TPR curves obtained for a given estimated method and for its observed counterpart are very similar. Also, `FS-HGa` performs significantly better than `FS3` and `FS`.

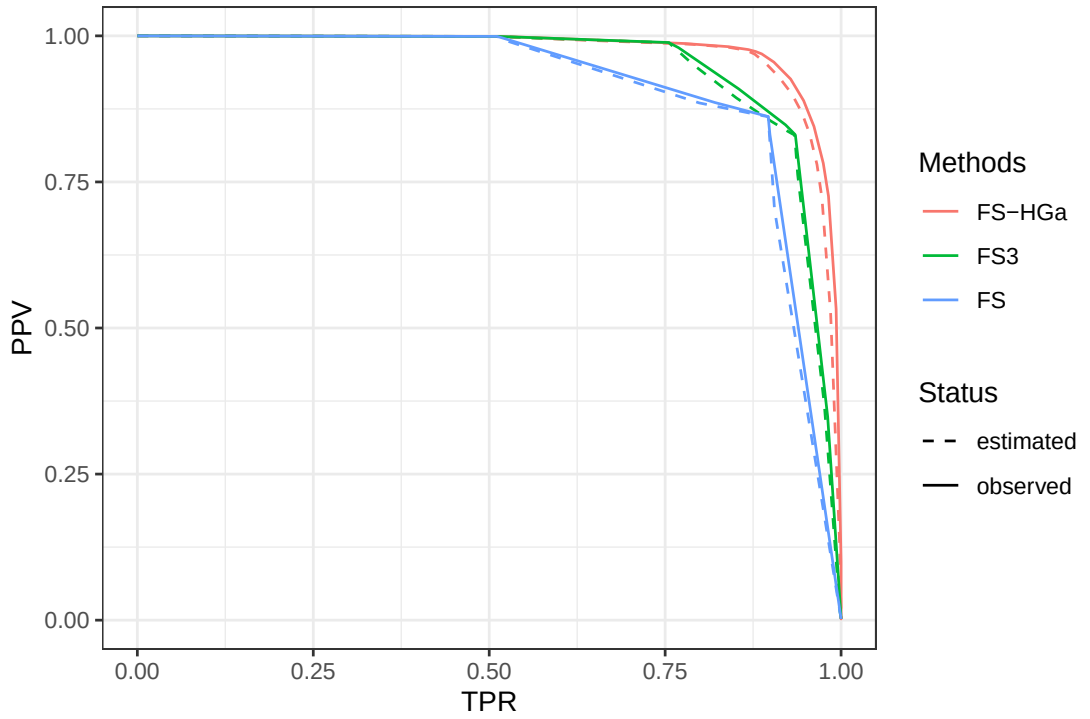


Figure 3.5: PPV-TPR curves for the observed/estimated version of the methods considered with continuous matching variables only, with sample sizes $n_A = 500$ and $n_B = 200$, $K = 3$ matching variables, $\lambda^k = 0.02$ for the parameter of the Exponential distribution, a proportion of errors $e = 0.2$, and a parameter $\lambda_e = 1/3$ for the error lag.

To evaluate the robustness of `FS-HGa`, we performed additional simulations presented in Appendix A.3.6. In these simulations, X^k is generated according to a uniform distribution and ϵ according to a normal distribution. The results indicate that even when the model is misspecified, `FS-HGa` is robust and performs better than the other methods. Also, we considered different values for K and λ^k in Appendix A.3.4 and A.3.3. In general, under the fixed sample sized n_A and n_B, all methods perform better with more matching variables (larger K) and/or when the matching variables are more informative (smaller λ^k). Finally, the results in Appendix A.3.5 indicate that all methods gradually improve as the ratio n_B/n_A increases.

3.5 Application

3.5.1 Description of SNDS and GETBO databases

The French national health information system SNDS was first created mainly based on the national register of health insurance information (SNIIRAM), which is currently one of the largest claims database in the world (Bezin et al., 2017). The SNDS includes information such as socio-demographic data, real-life use of drugs, chronic medical conditions (ICD10 codes), date and duration of hospital admissions. These databases are therefore of major interest, and their study has already led to several useful findings (e.g. Tuppin et al., 2017b,a). Because of this interest, there is an increasing demand for using this database to enrich existing cohorts or medical registers. However, most of the time, no common identifier is available in the database. Our objective is therefore to link the de-identified GETBO database to the SNDS, when no common individual identifier is available.

The GETBO database results from a data management process of the raw data of the GETBO registry. It is built as a list of documented cases of venous thromboembolism (VTE) recorded between 2013 and 2015 in Brest metropolitan area (Delluc et al., 2016). A given patient may have several events, and the database contains $1,404$ VTE events concerning $1,332$ distinct patients. For each documented case, the diagnostic or therapeutic medical acts were recorded with their type and the precise date, as well as the demographic information for each patient (date of birth, gender, residency code). Linked data consisting of VTE cases from GETBO and corresponding valuable health information from SNDS are used to build a prediction model, which can identify symptomatic VTE early for French people (Noboa et al., 2006; Delluc et al., 2016).

In this application, the so-called SNDS database results from a data extraction process of the raw data from SNDS, including the health insurance data from SNIIRAM and the national hospital discharge databases. The complete extraction was designed to select patients living in the Brest area, and having at least one care reimbursement between 2013 and 2015. It concerned $369,695$ distinct individuals. We selected patients having, during the studied period, at least one medical act either prescribed for diagnosis purposes of VTE (echodoppler, scintigraphy, tomoscintigraphy and angiography), or for therapeutic purposes (vena cava filter and thrombolysis) that were supposed to be recorded in the GETBO registry. This led to a list of $48,102$ timestamped medical acts concerning $32,382$ distinct patients with all the related demographic information (date of birth, gender, resi-

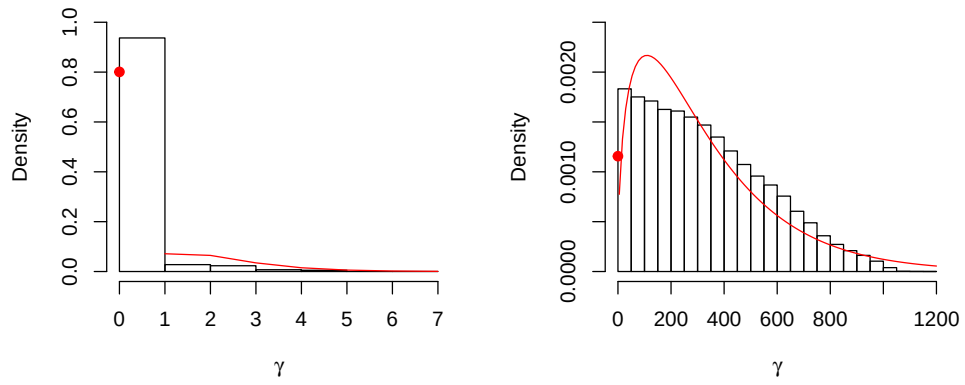
dency code). This database is expected to contain all medical acts in the GETBO database.

3.5.2 Probabilistic record linkage process

Since some VTE events in GETBO can relate to several medical acts, we first restructure this database such that each row contains only one medical act. This results in a new GETBO database with $1,919$ medical acts associated to $1,332$ patients. There are 6 available matching variables: year of birth, month of birth, residency code, gender, type and date of medical act. A full Cartesian product of the GETBO and SNDS databases requires computing $1,919 \times 48,102 = 92,307,738$ comparison vectors. Therefore, we need to choose a blocking variable to reduce computational time. A good blocking variable should have high quality, and multiple categories distributed as uniformly as possible (Herzog et al., 2007). The gender variable has only two categories and is therefore not very successful in reducing the dimension of the comparison space. Besides, the year of birth is not uniformly distributed and the residency code is likely to change due to moves, for example. Therefore, the month of birth seems the more reasonable choice. It should also be noted that only records with the same type of medical acts should be compared. By employing this scheme, there remains $4,308,847$ candidate pairs that need to be compared in terms of year of birth, residency code, gender and date of medical act.

We use the simple binary comparison function (3.2) for the year of birth and residency code variable. For the gender variable, since there is an imbalance between male and female in SNDS database (36.6% compared to 63.4%), we choose (3.7) as the comparison function. Finally, we choose the absolute distance (3.19) for the dates of medical acts variable. The comparison step results in a set of $4,308,847$ mixed-type comparison vectors. They are fitted by our proposed extension of Fellegi-Sunter model for mixed-type data, denoted by `FS-ext`. The ECM algorithm is applied to estimate all the model parameters. It stopped after 5 iterations when the relative difference of log-likelihood values of two successive steps was less than 10^{-7}. Once all parameters are estimated, we compute the estimated posterior probabilities of matching (3.14) for all record pairs of medical acts. Finally, we define a threshold $\tau = 0.5$, and a pair with a greater estimated posterior probability is predicted as a match.

In Figure 3.6, we present two histograms of comparison values of the dates of medical acts for our predicted matched/unmatched pairs. The red line is the



(a) Histogram of $\gamma|\hat{M}$ (b) Histogram of $\gamma|\hat{U}$

Figure 3.6: Histogram of the comparison values for dates of medical acts of predicted matched pairs (a) and unmatched pairs (b), and the fitted distribution (red line) of our model.

hurdle gamma distribution fitted by our model. Figure 3.6a indicates that there are more than 90% predicted matches with the same dates, and the others have 1 to 5 days in difference between dates.

3.5.3 Results

The observation unit is a medical act, and the matching variables are therefore observed at this level. On the other hand, the outcomes are needed at the patient level. We therefore report the application results in two steps. In the first one, we identify record pairs which refer to the same medical act, by applying the record linkage method on the observed data. In the second one, an ad-hoc procedure is performed to get corresponding pairs of patients from the pairs of medical acts.

Firstly, after performing the linkage method on the two databases of medical acts, we get $1,810$ pairs of medical acts that have estimated posterior matching probabilities no smaller than a threshold of 0.5. It is required that one patient in GETBO may be linked to one patient only in SNDS, and conversely. Therefore, if different pairs of medical acts lead to more than two candidates for one patient, we only keep the pairs of medical acts with the highest estimated probabilities, and suppress the others. Eventually, there remains $1,627$ pairs of medical act predicted as matches. The distribution of their (estimated) posterior matching probabilities is presented in Table 3.1. Among the predicted matched pairs, $1,410/1,627 = 87\%$ have estimated posterior probabilities larger than 0.9.
From the $1,627$ pairs of medical acts predicted as matches, we obtain $1,146$ cor-

$\hat{q}$	$(0.5, 0.6]$	$(0.6, 0.7]$	$(0.7, 0.8]$	$(0.8, 0.9]$	$(0.9, 0.95]$	$(0.95, 1]$
Number of pairs of medical acts	4	9	18	186	188	1,222

Table 3.1: Frequency distribution table of estimated posterior probability of matching for predicted matched pairs of medical acts

responding pairs of patients, since one patient may have several medical acts. Among them, 13 patients in GETBO have two different matched candidates in SNDS with the same probability. A random choice between two SNDS candidates is made for these patients. We also consider two different approaches for linking the two databases: the Fellegi-Sunter model `FS` with a binary comparison (3.2), and the deterministic method. Under the latter, a pair of medical acts is classified as a match if both records share the same type of medical act, month, year of birth, gender, residency code, while the date of medical act is compared with a tolerance of 3 days. Some manual review is required for pairs that link an individual in the database to more than two individuals in another database. We compare the three approaches in terms of predicted matched pairs of patients.

We summarize the linkage results of the different methods in Table 3.2. As could be expected, the set of predicted matched pairs of patients obtained under both `FS-ext` and `FS` include all the pairs identified by the deterministic record linkage. All 867 pairs predicted as matches by the deterministic method have a very high average posterior probability of matching for both `FS-ext` ($\bar{\hat{q}}_{\text{FS-ext}} = 0.993$) and `FS` ($\bar{\hat{q}}_{\text{FS-ext}} = 0.996$). Among the 247 remaining pairs which are classified as a match by `FS`, $245/247 \approx 99.2\%$ are also identified by `FS-ext`. Besides, 34 additional pairs of patients are identified as matches by `FS-ext`, with a high average probability ($\bar{\hat{q}}_{\text{FS-ext}} = 0.868$). From a look at the data, these pairs are not predicted by `FS` because they often correspond to a difference of 1 to 5 days in the date of medical acts. Consequently, the proposed method `FS-ext` predicts $1,146$ matched pairs for $1,332$ patients in GETBO, which represents 86% of the patients. On the other hand, the deterministic and `FS` only account for 65% and 83.6% respectively.

3.6 Discussion

In this contribution, we proposed two comparison approaches for low prevalence categorical and continuous matching variables. The proposed comparison functions aim to make a more extensive use of the matching variables in the compar-

Classified as a match by						
	`FS-ext`	`FS`	`Deterministic method`	Number of pairs of patients	$\overline{\hat{q}}_{\text{FS-ext}}$	$\overline{\hat{q}}_{\text{FS}}$
	X	X	X	867	0.993 (0.003)	0.996 (0)
	X	X		245	0.900 (0.045)	0.911 (0)
	X			34	0.868 (0.136)	
		X		2		0.911 (0)
Total	1146	1114	867			

Table 3.2: Comparison of three different record linkage methods with the number of pairs, the average estimated posterior probability of matching $\overline{\hat{q}}$ and the standard deviation (in parentheses)

ison vectors. We propose an extension of the Fellegi-Sunter probabilistic record linkage model, for comparison vectors containing both categorical and continuous comparison values. This model allows for using a variety of comparison functions, which can reflect matching data more accurately. We also suggest the use of a mixture of hurdle gamma distributions, for modeling the absolute difference between continuous variables such as dates. This distribution has never been formerly considered in the record linkage literature. In practice, the distribution for comparison values of continuous matching variables should be considered and validated a posteriori.

The simulation studies show that our proposed model outperforms the simple model with binary comparison in all the scenarios considered. For categorical matching variables, in Scenario 1, we have showed that the proposed model is more efficient than the standard model, especially when there are low prevalence values. However, if the frequencies of the different categories of a matching variable are similar, then there is not much difference between our approach and the standard one. In that case, the model with binary comparison should be considered due to its simplicity. For continuous matching variables, in Scenario 2, the proposed mixture of hurdle gamma distributions performs better than the standard model, and is robust to some misspecification of the distribution of the comparison function (see Appendix A.3.6). However, our evaluation remains specific to the fact that we are dealing with continuous time variables, which may be naturally modelled by Gamma distributions. A similar approach could be pursued

for other types of matching variables (e.g., string variables), but would require a different modelling for the similarity measure between strings.

We also conducted a simulation with mixed-type data in Appendix A.4. Consistently with the previous simulation results, the proposed model has a better performance than the standard model. In the application on real data, the performance is also better. We obtain a larger number of patients matched between the SNDS and the GETBO datasets, with high matching probabilities.

In practice, the matching variables that can be used for record linkage may include missing data. Also, dates of events may be censored. It would be of great practical interest to develop a joint modeling for record linkage and handling of missing values, to improve the performance of the record linkage process in this case. This is an important matter for further research. In a different approach, Copas and Hilton (1990) described a hit-miss model for record linkage which can accommodate the frequency distribution and missing values of the matching variables. However, this approach is not as commonly used in practice as the Fellegi-Sunter model due to its specific context (Goldstein et al., 2017). Besides, we did not consider matching variables varying over time. A study of Li et al. (2011) suggests that considering matching variables along with their time stamp (if applicable) may improve matching quality.

A problem of most probabilistic record linkage models lies in the imbalance between matched and non-matched pairs in the set of all comparison vectors, which may cause bias in parameter estimation. Blocking methods have been introduced to reduce the number of non-matched pairs, along with the computational cost. However, some true matched pairs may be overlooked if the blocking variable contains errors. Recently, Fortini (2020) introduced a robust approach where the EM algorithm is modified to obtain unbiased estimates of parameters in this context. However, this approach is designed for binary comparison values only.

The construction of the complete likelihood function rests on the assumption that the comparison vectors are independent. Such assumption may not be valid in practice, especially when there are matching restrictions, such that each record in a database can be linked to only one record in another database. Lee et al. (2020) recently proposed a maximum entropy classification for record linkage which overcomes this assumption.

4 Cox regression with linked data

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 to obtain from a single database. However, linkage errors are usually unavoidable regardless of record linkage methods and ignoring these errors may lead to bias 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 someone else and no information on matching variables is available to the analyst. Through a Monte Carlo simulation study, the proposed method has significantly corrected the parameter estimate bias 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 will be applied to a linked database from the Brest stroke registry in France.

Contents

4.1 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 (e.g. Harron et al., 2016; Padmanabhan et al., 2018). In national censuses, population data files obtained at different times can be linked to create longitudinal data sets (Zhang and Campbell, 2012). Record linkage may also be applied early in a survey to link the sampling frame and administrative data (e.g. Winkler and Thibaudeau, 1987). 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 (Herzog et al., 2007; Christen, 2012), such as the frequentist approach (Fellegi and Sunter, 1969; Winkler, 1988; Vo et al., 2022) and the Bayesian approach (Tancredi and Liseo, 2010; Sadinle, 2017). 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 (Neter et al., 1965), 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 data. From this perspective, Scheuren and Winkler (1993) 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 (2005) 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 (2012) extended the method by Lahiri and Larsen (2005) 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 (2019) adapted the approach by Lahiri and Larsen (2005) 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, under 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 (see for example Zhang, 2019). Starting from this perspective, Chambers (2009) 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, Kim and Chambers (2012b,c); Chambers and Kim (2015); Chambers et al. (2019); Chambers and Diniz da Silva (2020) developed methods for secondary analysis of linked data. Recently, Zhang and Tuoto (2020) 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. Chambers et al. (2022) proposed robust estimation for linear regression with linked data.

Although the Cox proportional hazard model (Cox, 1972) is of routine use for survival analysis, comparatively very few papers have focused on accounting for record linkage errors in this context. Baldi et al. (2010) performed a simulation study emphasizing the impact of incomplete record linkage errors 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. (2017) 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.

This chapter is organised as follows. In Section 4.2, we propose a new estimating equation, which leads to an approximately unbiased parameters estimation of the Cox model with linked data. A variance estimator is also proposed. In Section 4.3, we evaluate the proposed estimator and the associated variance estimator through simulation studies. In Section 4.4, an application on a real dataset is presented. Finally, possible further research is discussed in Section 4.5.

4.2 Cox regression analysis with linked data

4.2.1 Cox regression model

The Cox proportional hazard model (Cox, 1972) 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

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t) \exp\left(\mathbf{X}_i^T \boldsymbol{\beta}_0\right), \tag{4.1}$$

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:

$$H_0(\beta) \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_i \left\{ \mathbf{X}_i - \frac{\sum_{j=1}^{n} Y_j(T_i) \exp\left(\mathbf{X}_j^T \boldsymbol{\beta}\right) \mathbf{X}_j}{\sum_{j=1}^{n} Y_j(T_i) \exp\left(\mathbf{X}_j^T \boldsymbol{\beta}\right)} \right\} = 0, \tag{4.2}$$

where $Y_j(t) = \mathbb{1}_{(T_i \geq t)}$ is an at-risk indicator (see for example Andersen and Gill, 1982). We call (4.2) 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

$$\hat{\mathbb{V}}_{\text{mpl}}(\hat{\boldsymbol{\beta}}) = \left\{ -n \nabla H_0(\hat{\boldsymbol{\beta}}) \right\}^{-1}, \tag{4.3}$$

see Andersen and Gill (1982).

4.2.2 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 (4.2). However, if the covariates are not known in database A, equation (4.2) 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. The notations are summarized in Table 4.1. Reasoning from the secondary analysis perspective, we do not have access to the matching variables and do not know the actual linkage process.

(a) File of interest A

	T	δ	$\mathbf{Z} \in R^p$
$i = 1$			$\mathbf{Z}_1$
$\vdots$			
$i = n_A$			$\mathbf{Z}_n$

(b) Linking data file B.

	$\mathbf{X} \in R^p$
$j = 1$	
$\vdots$	
$\vdots$	
$j = n_B$	

Table 4.1: Record linkage context.

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 (e.g. Chambers et al., 2019). This is the key assumption of most secondary analysis approaches in the literature, for which Zhang and Tuoto (2020) have proposed a diagnostic test. Adopting the modelling approach in Copas and Hilton (1990), 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:

$$\mathbf{Z}_i = \begin{cases} \mathbf{X}_i & \text{with probability } \alpha_v, \\ \mathbf{X}_{(j)} & \text{with probability } 1 - \alpha_v, \end{cases} \tag{4.4}$$

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. We suppose that the linkage is performed independently for any unit $i \in A_v$, conditionally on the $\mathbf{X}_j$'s for $j \in B_v$.

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 (Baldi et al., 2010). For more discussion on this incomplete matching space problem, see Kim and Chambers (2012b); Goldstein et al. (2012); Zhang and Tuoto (2020). 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.

4.2.3 Adjusted estimating equation

By naively treating the linked covariates 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:

$$H_{naive}(\boldsymbol{\beta}) \equiv \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \delta_i \left\{ \mathbf{Z}_i - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) \exp(\mathbf{Z}_j^\top \boldsymbol{\beta}) \mathbf{Z}_j}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) \exp(\mathbf{Z}_j^\top \boldsymbol{\beta})} \right\} = 0. \quad (4.5)$$

We call (4.5) the naive estimating equation. Since some units are incorrectly linked, it may lead to biased estimates, see the simulation results in Section 4.3.

We propose a bias-corrected estimating equation, accounting for the fact that from the hit-miss model (4.4), 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

$$\bar{H}(\boldsymbol{\beta}) \equiv \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \delta_i \left\{ \mathbf{X}_i^*(\alpha_v) - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^*(\alpha_v, \boldsymbol{\beta})}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^*(\alpha_v, \boldsymbol{\beta})} \right\} = 0 \quad (4.6)$$

where, for any $i \in A_v$,

$$\begin{aligned} \mathbf{X}_i^*(\alpha_v) &= \alpha_v^{-1} \mathbf{Z}_i - (\alpha_v^{-1} - 1) \bar{\mathbf{X}}_{B_v}, \\ g_j^*(\alpha_v, \boldsymbol{\beta}) &= \alpha_v^{-1} g(\mathbf{Z}_j, \boldsymbol{\beta}) - (\alpha_v^{-1} - 1) \bar{g}_{B_v}(\boldsymbol{\beta}), \\ h_j^*(\alpha_v, \boldsymbol{\beta}) &= \alpha_v^{-1} h(\mathbf{Z}_j, \boldsymbol{\beta}) - (\alpha_v^{-1} - 1) \bar{h}_{B_v}(\boldsymbol{\beta}). \end{aligned} \quad (4.7)$$

We prove in Appendix B.1 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 4.3.

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 (Chambers, 2009; Zhang and Tuoto, 2020) if their true values are unknown.

4.2.4 Variance estimator

In this section, we discuss variance estimation for the estimator of the parameter β_0 obtained by solving the AEE given in (4.6). 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 $\alpha_v,\ v = 1, \ldots, V$. Using the variance estimator given in (4.3) fails to account for all these sources of variability, and therefore leads to an underestimation of the variance, see the simulation results in Section 4.3.

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

$$\hat{\mathbb{V}}_{\text{AEE}}(\hat{\boldsymbol{\beta}}) \equiv \{\nabla \bar{H}(\hat{\boldsymbol{\beta}})\}^{-1} \times \hat{\mathbb{V}}\{\bar{H}(\boldsymbol{\beta}_0)\} \times \{\nabla \bar{H}(\hat{\boldsymbol{\beta}})\}^{-1}, \tag{4.8}$$

$$\text{with } \hat{\mathbb{V}}\{\bar{H}(\boldsymbol{\beta}_0)\} = \hat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\} + \hat{\mathbb{V}}_2\{\bar{H}(\boldsymbol{\beta}_0)\}. \tag{4.9}$$

The first component $\hat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\}$ in (4.9) 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

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

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 (4.9) 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.2. It is evaluated empirically in the next section through a simulation study.

4.3 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 4.3.1. The estimation methods that we evaluate are presented in Section 4.3.2, along with the performance indicators. The simulation results are given in Section 4.3.3. To facilitate interpretation and to study the influence of different simulation parameters, we first consider in Section 4.3.3 scenarios with a single block. Scenarios with multiple blocks and different levels of linkage quality are considered in Section 4.3.3. All R programs for simulation are available in https://github.com/thanhhuanVO/Cox-regression-with-linked-data.

4.3.1 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

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

where U follows a standard uniform distribution (Bender et al., 2005), 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 number of blocks v and the probabilities α_v for $V = 1, \ldots, V$, the linked values $\mathbf{Z}$ for covariates in database A are obtained according to the linkage error model (4.4). Inside each block A_v, an audit sample of 10% of the units is selected by simple random sampling without replacement, and used for the estimation of $\hat{\alpha}_v$.

4.3.2 Methods and performance indicators

For each scenario, we consider the following estimation methods. The `Theoretical` is obtained by solving the theoretical estimating equation (4.2) 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 (4.5) with linked data. The `Validation` is obtained by solving the theoretical estimating equation (4.2) 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 (4.3), 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 (4.6) with the theoretical value of α_v. The `AEE` (adjusted estimating equation) is obtained by solving the proposed estimating equation (4.6), 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 (B.23). 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 (4.3).

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

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

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

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

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

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

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 $\mathrm{Sd}(\hat{\boldsymbol{\beta}})$, approximated by $\mathrm{Sd}_{\mathrm{MC}}(\hat{\boldsymbol{\beta}})$.

4.3.3 Simulation results

One block situation

In this section, we consider the situation when the data sets are generated as presented in Section 4.3.1, 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 4.2. 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 decreases as the probability of correct link increases, as expected. The proposed methods `TAEE` and `AEE` lead to approximately unbi-

ased estimation of the parameters, with a larger variability for `AEE` as expected. We note that the variability 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`.

			$\hat{\beta}_1$					$\hat{\beta}_2$				
α	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	0.000	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 4.2: Simulation results in Case 1 with three different values for the probability of correct link $\alpha \in \{0.75, 0.85, 0.95\}$

We now turn to the variance estimators. The variance estimator $\hat{\mathbb{V}}_{mpl}(\hat{\boldsymbol{\beta}})$ (4.3) 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 one hand, the proposed variance estimator performs well, except for β_1 when $\alpha = 0.75$, in which case the variance is underestimated. Besides, by assuming that the estimated coefficients $\hat{\boldsymbol{\beta}}$ are asymptotically normally distributed, we reported the coverage probability for confidence intervals with a nominal level of 95%. The coverage probability of our proposed methods is close to 95% and it is a bit larger when $\alpha = 0.95$ because of the positive bias of the variance estimator.

The simulation results obtained in Case 2 are presented in Table 4.3. We observe no qualitative difference compared to Case 1. The `TAEE` and `AEE` lead to almost unbiased estimations of 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 4.3: Simulation results in Case 2 with three different values for the sample size n_A

Multiple blocks

In this section, we consider the situation when the data sets are generated as presented in Section 4.3.1, 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 $(\alpha_1, \alpha_2, \alpha_3) = (0.8, 0.9, 1.0)$; a second scenario where $(\alpha_1, \alpha_2, \alpha_3) = (0.7, 0.8, 0.9)$; a third scenario where $(\alpha_1, \alpha_2, \alpha_3) = (0.6, 0.7, 0.8)$.

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 $\bar{\alpha} = 90\%$ in Scenario 1, $\bar{\alpha} = 80\%$ in Scenario 2 and $\bar{\alpha} = 70\%$ 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 `AEE` 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 4.4, and confirm the good results of the proposed methods observed in the situation of one block. Scenario 2 and 3 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. 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.

			$\hat{\beta}_1$					$\hat{\beta}_2$				
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	0.060	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
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	0.171	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

Table 4.4: 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. Moreover, they can have 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. Although this is a favourable condition for secondary analysis, block-specific adjustment is the default approach as long as one cannot be sure whether such non-informativeness is the case in a given situation. Thus, obtaining block-specific α_v is much more demanding in the real world.

4.4 Application

4.4.1 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.

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 4.5: Description of the linkage process

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 4.5.

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 4.6. 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.

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 4.6: Description of the linked database

4.4.2 Cox regression analysis

In this application, we use the Cox regression model (4.1) 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 4.3.3 show that linkage errors lead to biased estimators of the regression coefficients. Therefore, we also use the adjusted estimating equation (4.6).

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, and 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 (4.7) 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 (B.26), which requires that the covariates are known on database A only. Simulations in Appendix B.3.2 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 4.2.3.

In Table 4.7, 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.

	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 4.7: 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.

4.5 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 have significantly corrected the bias of the naive estimating equation. 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 (B.26), which still maintain the good performance of the AEE if A is a random sample from B. In addition, a linear approximated estimating equation (LAEE), which can provide better estimation than AEE with small sample, is given in Appendix B.4.

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. In this case, the proposed estimating equation may not be applied and different adjustments need to be developed.

5 Conclusions and perspectives

Contents

5.1 Conclusions

In this thesis, we have extended the Fellegi-Sunter record linkage model for mixed-type data. The Fellegi-Sunter model has shown to be an effective method and is widely used in probabilistic record linkage. However, the simple binary comparison approach has some limitations on dealing with low prevalence categorical variables and continuous matching variables. We therefore proposed a mixture model which can accommodate both categorical and continuous comparison values. We have also proposed two comparison approaches for low prevalence categorical matching variables and continuous matching variables, which are common in health databases: e.g., diagnosis code (low prevalence binary) and date of medical acts (continuous).

The mixture of hurdle gamma distribution has been used for the first time to model the comparison value of dates. Throughout various simulation scenarios, the extended model has been shown to improve on the standard Fellegi-Sunter model, especially when there are low prevalence values in categorical matching variables and noise in continuous matching variables. It also showed a better performance on linking to real databases GETBO and SNDS.

Secondly, we proposed an adjusted estimating equation to correct the bias for Cox regression analysis of linked data. Due to the lack of unique identifiers in the matching process, errors are usual in linked data. We have shown that the naive use of such linked data may result in bias for the estimation of the parameters in the Cox model. Although the Cox model is one of the most used models in medical study, this problem has not been studied in the literature. Adopting a

secondary analysis viewpoint, we have proposed an adjusted estimating equation which can correct the bias from the naive estimating equation. In addition, a variance estimator for estimators of Cox regression coefficients which can account for the variability of the whole linkage process is also proposed.

5.2 Limitations and future works

Although our works improved on the existed methods for record linkage and analyzing of linked data, there are still several perspectives that can be developed. In this section, we discuss some remaining problems of the proposed methods, and suggest possible directions for future works.

One of the most common type of matching variables which has not been investigated is that of string variables, such as a name or an address, for example. Since this type of variable is not available in SNDS and GETBO, it has not been considered in this work. In the literature, various similarity measures were designed to compare string variables (Herzog et al., 2007). Generally, these measures return numerical comparison values between 0 and 1, where 1 indicates exact agreement and 0 means total disagreement. However, the numerical comparison values are often discretized before they reach the classification step. This discretization may result in a loss of information. Finding a continuous distribution which can model directly the continuous similarity measure between string variables would be a perspective for future research.

In addition, we have not considered the missing data problem in record linkage. In practice, if the proportion of missing data is significant, suitable techniques should be made to account for them (Ong et al., 2014). For example, imputation methods can be used to impute the comparison values corresponding to missing data. Besides, the implementation of the proposed record linkage model with large and dynamic databases is also a limit of our work. Dynamic databases are cases where new records can be added in the database, and existing records can be modified. For example, since data from both private and public services are regularly updated online, there is an increasing demand for real-time linkage such as online identity verification. In that case, specific techniques for efficient and fast record linkage are required (Ramadan et al., 2015).

In most applications, the main objective of doing record linkage is to obtain a database large enough for a statistical analysis. From this work and also from the literature, researchers should be aware of linkage errors which causes bias in any

statistical analysis (Harron et al., 2015). Although linked data become more and more popular in different fields, most works in the literature are mainly concerned with the case of a generalized linear model. Therefore, there remains various perspectives to develop methods for the statistical analysis of linked data.

Concerning Cox regression with linked data, we proposed an adjusted estimating equation which can account for linkage errors. However, there are still some limitations. For example, although we have proposed a variance estimator which can capture all sources of variability, the asymptotic properties of the proposed estimator need to be investigated conscientiously. Besides, we have assumed that the survival time is already available and only some explanatory variables are obtained from the linkage process. In practice, this may not be the case. There may be cases when the covariates are available, while the survival time is obtained by a linkage process. In that case, the current proposed adjusted estimating equation may not be used, and a different adjustment should be investigated. Also, the model needs to be developed if linkage is performed from more than two databases. Recently, Slawski et al. (2021) proposed a pseudo-likelihood approach for robust linear regression with shuffled data, which could be an alternative to the proposed method for Cox regression with linked data.

A Appendix for Chapter 3

This Appendix is divided into five Sections. The first provides details on ECM algorithm that has been used for the simulation with continuous comparison values in Chapter 3. Section A.2, A.3 and A.4 provides complementary simulation results for scenarios with only binary, continuous and mixed-type matching variables respectively. Finally, in Section A.5, a naive example for implementation of the proposed method has been introduced.

A.1 ECM algorithm

In Chapter 3, with the proposed comparison method for dates variables, f^k is chosen as the density function of a hurdle Gamma distribution. It is characterized by three parameter p_0^k and $\alpha^k, \beta^k \in \mathbb{R}^+$ as follows:

$$f(\gamma^k; p_0^k, \alpha^k, \beta^k) = \left[p_0^k\right]^{\mathbb{1}_{\{\gamma^k=0\}}} \left[(1-p_0^k)v(\gamma^k; \alpha^k, \beta^k)\right]^{\mathbb{1}_{\{\gamma^k>0\}}}, \tag{A.1}$$

where $v(\gamma^k; \alpha, \beta)$ is the density function of a Gamma distribution.

Let θ be the vector of all parameters, which are p_M (one parameter); m_s^k, u_s^k for $k = 1, \ldots, K_1$ and $s \in S^k$ is the set of all possible comparison values ($\sum_{k=1}^{K_1} 2(|S^k|-1)$ parameters with $|S^k|$ is the number of elements in S^k); $p_{0M}^k, p_{0U}^k, \alpha_M^k, \alpha_U^k, \beta_M^k, \beta_U^k$ for $k = K_1+1, \ldots, K_1+K_2$ ($6K_2$ parameters). By assuming the independence between all comparison vectors in Γ, the likelihood function for all observed comparison vectors is

$$\mathcal{L}(\theta|g,\gamma) = \prod_{i=1}^{n_A}\prod_{j=1}^{n_B}\left[p_M P_{ij}^{1M} P_{ij}^{2M} + (1-p_M) P_{ij}^{1U} P_{ij}^{2U}\right], \tag{A.2}$$

and

$$\ell(\theta) = \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \ln\left[p_M P_{ij}^{1M} P_{ij}^{2M} + (1-p_M) P_{ij}^{1U} P_{ij}^{2U}\right]$$

is the corresponding log-likelihood function. Let also

$$g_{ij} = \begin{cases} 1 & \text{if} \quad (X_{A,i}, X_{B,j}) \in M, \\ 0 & \text{if} \quad (X_{A,i}, X_{B,j}) \in U, \end{cases} \tag{A.3}$$

be a latent variable indicating whether the record is a true match or not. The complete likelihood may be written as follows:

$$\mathcal{L}_c(\theta|g,\gamma) = \prod_{i=1}^{n_A}\prod_{j=1}^{n_B} \left[p_M P_{ij}^{1M} P_{ij}^{2M}\right]^{g_{ij}} \left[(1-p_M) P_{ij}^{1U} P_{ij}^{2U}\right]^{1-g_{ij}}, \tag{A.4}$$

and the complete log-likelihood function $\ell_c(\theta)$ is

$$\begin{aligned} \ell_c(\theta) = &\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} g_{ij} \ln\left[P_{ij}^{1M}\right] + \sum_{i=1}^{n_A}\sum_{j=1}^{n_B}(1-g_{ij})\left[\ln P_{ij}^{1U}\right] \\ &+ \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} g_{ij} \ln\left[P_{ij}^{2M}\right] + \sum_{i=1}^{n_A}\sum_{j=1}^{n_B}(1-g_{ij})\left[\ln P_{ij}^{2U}\right] \\ &+ \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} g_{ij} \ln\left(p_M\right) + (1-g_{ij})\ln(1-p_M). \end{aligned}$$

It can be rewritten as

$$\begin{aligned} \ell_c(\theta) = &\sum_{k=1}^{K_1}\sum_{s\in S^k}\left[\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} g_{ij}\ln(m_s^k)\mathbb{1}_{\gamma_{ij}^k=s}\right] + \sum_{k=1}^{K_1}\sum_{s\in S^k}\left[\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}(1-g_{ij})\ln(u_s^k)\mathbb{1}_{\gamma_{ij}^k=s}\right] \\ &+ \sum_{k=K_1+1}^{K_2}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^k>0\}} g_{ij}\ln v(\gamma_{ij}^k;\alpha_M^k,\beta_M^k) \\ &+ \sum_{k=K_1+1}^{K_2}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^k>0\}} (1-g_{ij})\ln v(\gamma_{ij}^k;\alpha_U^k,\beta_U^k) \\ &+ \sum_{k=K_1+1}^{K_2}\left[\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^k=0\}} g_{ij}\ln\left(p_{0M}^k\right) + \sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^k>0\}} g_{ij}\ln\left(1-p_{0M}^k\right)\right] \\ &+ \sum_{k=K_1+1}^{K_2}\left[\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^k=0\}} (1-g_{ij})\ln\left(p_{0U}^k\right) + \sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^k>0\}} (1-g_{ij})\ln\left(1-p_{0U}^k\right)\right] \\ &+ \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} g_{ij}\ln\left(p_M\right) + (1-g_{ij})\ln(1-p_M). \end{aligned}$$

We iteratively maximize the above complete log-likelihood function. The expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993) is a variant of the EM algorithm. Young et al. (2019) showed that the ECM algo-

rithm needs to be rather used, because we do not simultaneously update all of the parameters, but rather make a conditional maximization to update the values of the shape and scale parameters. For simplicity of presentation and without loss of generality, we assume that $K_1 = K_2 = 1$.

E-step For iteration $t = 0, 1, \ldots$, we compute the expectation of the latent variable which is also the posterior matching probabilities as follows:

$$\begin{aligned} g_{ij} &= \frac{\mathbb{P}(\gamma_{ij}|M)\mathbb{P}(M)}{\mathbb{P}(\gamma_{ij}|M)\mathbb{P}(M) + \mathbb{P}(\gamma_{ij}|U)\mathbb{P}(U)} \\ &= \frac{p_M P_{ij}^{1M} P_{ij}^{2M}}{p_M P_{ij}^{1M} P_{ij}^{2M} + (1-p_M) P_{ij}^{1U} P_{ij}^{2U}} \\ &= \frac{p_M \left(\prod_{s\in S^k} \{m_s^1\}^{\mathbb{1}_{\gamma_{ij}^1=s}}\right) f(\gamma_{ij}^2, \boldsymbol{\theta}_M^2)}{p_M \left(\prod_{s\in S^k} \{m_s^1\}^{\mathbb{1}_{\gamma_{ij}^k=s}}\right) f(\gamma_{ij}^2, \boldsymbol{\theta}_M^2) + (1-p_M) \left(\prod_{s\in S^k} \{u_s^1\}^{\mathbb{1}_{\gamma_{ij}^1=s}}\right) f(\gamma_{ij}^2, \boldsymbol{\theta}_U^2)} \end{aligned}$$

where $\boldsymbol{\theta}_M^2 = (p_{0M}^2, \alpha_M^2, \beta_M^2)$ and $\boldsymbol{\theta}_U^2 = (p_{0U}^2, \alpha_U^2, \beta_U^2)$.
Therefore, we can estimate the posterior probabilities at step t as

$$g_{ij}^{(t)} = \frac{c_M^{(t)}}{c_M^{(t)} + c_U^{(t)}}$$

where

$$\begin{aligned} c_M^{(t)} &= p_M^{(t)} \left(\prod_{s\in S^k} m_s^{1,(t)^{\mathbb{1}_{\gamma_{ij}^1=s}}}\right) f(\gamma_{ij}^2, \boldsymbol{\theta}_M^{2,(t)}), \\ c_U^{(t)} &= (1-p_M^{(t)}) \left(\prod_{s\in S^k} u_s^{1,(t)^{\mathbb{1}_{\gamma_{ij}^1=s}}}\right) f(\gamma_{ij}^2, \boldsymbol{\theta}_U^{2,(t)}), \end{aligned}$$

and the superscripts (t) refer to the value of parameters at iteration t. The values at iteration $t = 0$ for all parameters are starting values specified by practitioners.

First CM-step We substitute the estimated value of g_{ij} in the E step into the log-likelihood function. Then we fix the value of β_M^2, β_U^2 and set all partial derivatives equal to zero. We obtain the estimates for other parameters as follows

$$\begin{aligned} p_M^{(t+1)} &= \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} g_{ij}^{(t)}, \\ m_s^{1,(t+1)} &= \frac{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} g_{ij}^{(t)} \mathbb{1}_{\gamma_{ij}^1=s}}{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} g_{ij}^{(t)}}, \end{aligned}$$

$$
\begin{aligned}
u_s^{1,(t+1)} &= \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}(1-g_{ij}^{(t)})\mathbb{1}_{\gamma_{ij}^1=s}}{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}(1-g_{ij}^{(t)})}, \\
p_{0M}^2 &= \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^2=0\}}g_{ij}}{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}g_{ij}}, \\
p_{0U}^2 &= \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^2=0\}}(1-g_{ij})}{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}(1-g_{ij})}.
\end{aligned}
$$

The updated value for the parameters $\alpha_M^{2,(t+1)}, \alpha_U^{2,(t+1)}$ are given by the solutions of the following equations (Young et al., 2019):

$$
0 = \frac{\partial \ell_c(\theta)}{\partial \alpha_M^2} = \sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^2>0\}}g_{ij}^{(t)}\left[\ln(\gamma_{ij}^2) - \ln(\beta_M^{(t)}) - \psi(\alpha_M^2)\right], \tag{A.5}
$$

$$
0 = \frac{\partial \ell_c(\theta)}{\partial \alpha_U^2} = \sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^2>0\}}(1-g_{ij}^{(t)})\left[\ln(\gamma_{ij}^2) - \ln(\beta_U^{(t)}) - \psi(\alpha_U^2)\right], \tag{A.6}
$$

where $\psi(\cdot)$ is the digamma function. The solutions may be found using iterative methods such as Newton-Raphson.

Second CM-step Maximizing $\ell_c(\theta)$ with respect to β_M^2, β_U^2 while all other parameters are fixed at their current values. The scale parameters are updated as follows:

$$
\beta_M^{2,(t+1)} = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^2>0\}}\gamma_{ij}^2 g_{ij}^{(t)}}{\alpha_M^{(t+1)}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^2>0\}}g_{ij}^{(t)}}, \tag{A.7}
$$

$$
\beta_U^{2,(t+1)} = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^2>0\}}\gamma_{ij}^2(1-g_{ij}^{(t)})}{\alpha_U^{(t+1)}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\mathbb{1}_{\{\gamma_{ij}^2>0\}}(1-g_{ij}^{(t)})}. \tag{A.8}
$$

The above ECM algorithm stops if the relative difference between estimated parameters from two successive steps is less than some $\epsilon > 0$, i.e. $\left|\theta_i^{(t+1)} - \theta_i^{(t)}\right| / \left|\theta_i^{(t)}\right| < \epsilon$ for all parameters $\theta_i \in \theta$ or if the relative difference between log-likelihood calculated at successive estimates of parameters is less than some $\epsilon' > 0$, i.e. $\left|\ell(\theta^{(t+1)}) - \ell(\theta^t)\right| / \left|\ell(\theta^t)\right| < \epsilon'$.

When all parameters are obtained from the ECM algorithm, the posterior proba-

bilities $q_{ij} = \mathbb{P}(M|\gamma_{ij})$ are estimated for all record pairs as

$$\widehat{q}_{ij} = \frac{\widehat{p}_M \widehat{P}_{ij}^{1M} \widehat{P}_{ij}^{2M}}{\widehat{p}_M \widehat{P}_{ij}^{1M} \widehat{P}_{ij}^{2M} + (1-\widehat{p}_M)\widehat{P}_{ij}^{1U} \widehat{P}_{ij}^{2U}}.$$

These posterior probabilities are then used to find proper matched pairs.

A.2 Evaluation of the model for binary matching variables

In this section, under the same setting at Section 3.4, we consider some additional simulation scenarios when there are only binary matching variables.

A.2.1 Initial values and stopping criteria for EM algorithm

To run the EM algorithm, we have to define a set on initial values for all parameters. We describe here the approach which has been used in the simulation to initiate the EM algorithm. We first consider the comparison method with all possible configuration of `FS4` as

$$\gamma_{ij}^k = \begin{cases} c_1 & \text{if} \quad (X_{A,i}^k, X_{B,j}^k) = (0,0), \\ c_2 & \text{if} \quad (X_{A,i}^k, X_{B,j}^k) = (0,1), \\ c_3 & \text{if} \quad (X_{A,i}^k, X_{B,j}^k) = (1,0), \\ c_4 & \text{if} \quad (X_{A,i}^k, X_{B,j}^k) = (1,1), \end{cases} \tag{A.9}$$

for $k = 1, \ldots, K$, $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$. The comparison vectors are then assumed to follow a mixture distribution of $\boldsymbol{\gamma}^k|M$ and $\boldsymbol{\gamma}^k|U$.

The $\boldsymbol{\gamma}^k|M$ follows a discrete distribution with parameters m_1^k, m_2^k, m_3^k and $m_4^k = 1 - m_1^k - m_2^k - m_3^k$. We have

$$\begin{aligned} m_1^k &= \mathbb{P}(X_{A,i}^k = 0, X_{B,i}^k = 0) \\ &= \mathbb{P}(X_{B,i}^k = 0 | X_{A,i}^k = 0)\mathbb{P}(X_{A,i}^k = 0) = (1-e^k)(1-p^k). \end{aligned}$$

By similar argument, we have $m_2^k = e^k(1-p^k), m_3^k = e^k p^k$.

The $\boldsymbol{\gamma}^k|U$ follows a discrete distribution with parameters u_1^k, u_2^k, u_3^k and $u_3^k = 1 - u_1^k - u_2^k - u_3^k$. As in this case $X_{A,i}^k$ and $X_{B,j}^k$, with $i \neq j$, are independent (correspond to two different individuals), we have

$$
\begin{aligned}
u_1^k &= \mathbb{P}(X_{A,i}^k = 0, X_{B,j}^k = 0) \\
&= \mathbb{P}(X_{A,i}^k = 0)\mathbb{P}(X_{B,j}^k = 0) \\
&= (1-p^k)\left[(1-e^k)\mathbb{P}(X_{A,j}^k = 0) + e^k\mathbb{P}(X_{A,j}^k = 1)\right] \\
&= (1-p^k)\left[(1-e^k)(1-p^k) + e^k p^k\right].
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
u_2^k &= (1-p^k)[(1-e^k)p^k + e^k(1-p^k)], \\
u_3^k &= p^k\left[(1-e^k)(1-p^k) + e^k p^k\right].
\end{aligned}
$$

To obtain initial values for parameters m_s^k and u_s^k of `FS4`, we substitute $e^k = 0.01$ and p^k by its empirical estimate from A. In a similar way, we find initial values for parameters m_s^k and u_s^k of `FS3` and `FS`.

In each method, the algorithm only stops when the relative difference between estimated parameters from two successive steps is less than 10^{-6} or when the number of iterations reaches 500.

A.2.2 Complementary results for the article

Proportion of convergence before the maximum number of iteration in EM algorithm

In Table A.1, we can see the proportion of convergence of each method over different simulation scenarios considered in Section 3.4. In this simulation, all the results are obtained by using the package simsalapar (Hofert and Mächler, 2016) for parallelly running all combinations of simulation parameters. A server with 2 Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz with 12 cores each has been used.

Boxplots

Figure A.1 shows the boxplots for all simulation results.

F-score

Let's define

$$
\text{f-score} = \frac{\text{TPR} + \text{PPV}}{2}
$$

K	e	Methods	Proportion of convergence	Average execution time (s)
30	0.02	FS	0.720	37.5
		FS3	0.995	63.0
		FS4	0.931	188.6
		Bayesian	1.00	3.3
	0.04	FS	0.385	61.6
		FS3	1.00	51.7
		FS4	0.922	186.1
		Bayesian	1.00	3.3
	0.06	FS	0.148	116.2
		FS3	0.997	70.9
		FS4	0.935	173.4
		Bayesian	1.00	3.3
40	0.02	FS	0.977	18.6
		FS3	1.00	30.8
		FS4	0.984	122.7
		Bayesian	1.00	3.4
	0.04	FS	0.853	32.8
		FS3	1.00	34.7
		FS4	0.979	130.2
		Bayesian	1.00	3.3
	0.06	FS	0.538	43.3
		FS3	1.00	42.7
		FS4	0.970	170.6
		Bayesian	1.00	3.2
50	0.02	FS	0.998	11.7
		FS3	1.00	30.7
		FS4	0.994	102.2
		Bayesian	1.00	3.4
	0.04	FS	0.983	17.8
		FS3	1.00	34.5
		FS4	0.989	138.8
		Bayesian	1.00	3.3
	0.06	FS	0.864	28.2
		FS3	1.00	39.6
		FS4	0.980	200.7
		Bayesian	1.00	3.2

Table A.1: Proportion of convergence of EM algorithm and average execution time (seconds) of each method over 1000 repeated simulation with $n_A = 500, n_B = 200, p^k = 0.2$ and $K \in \{30, 40, 50\}$, $e \in \{0.02, 0.04, 0.06\}$.

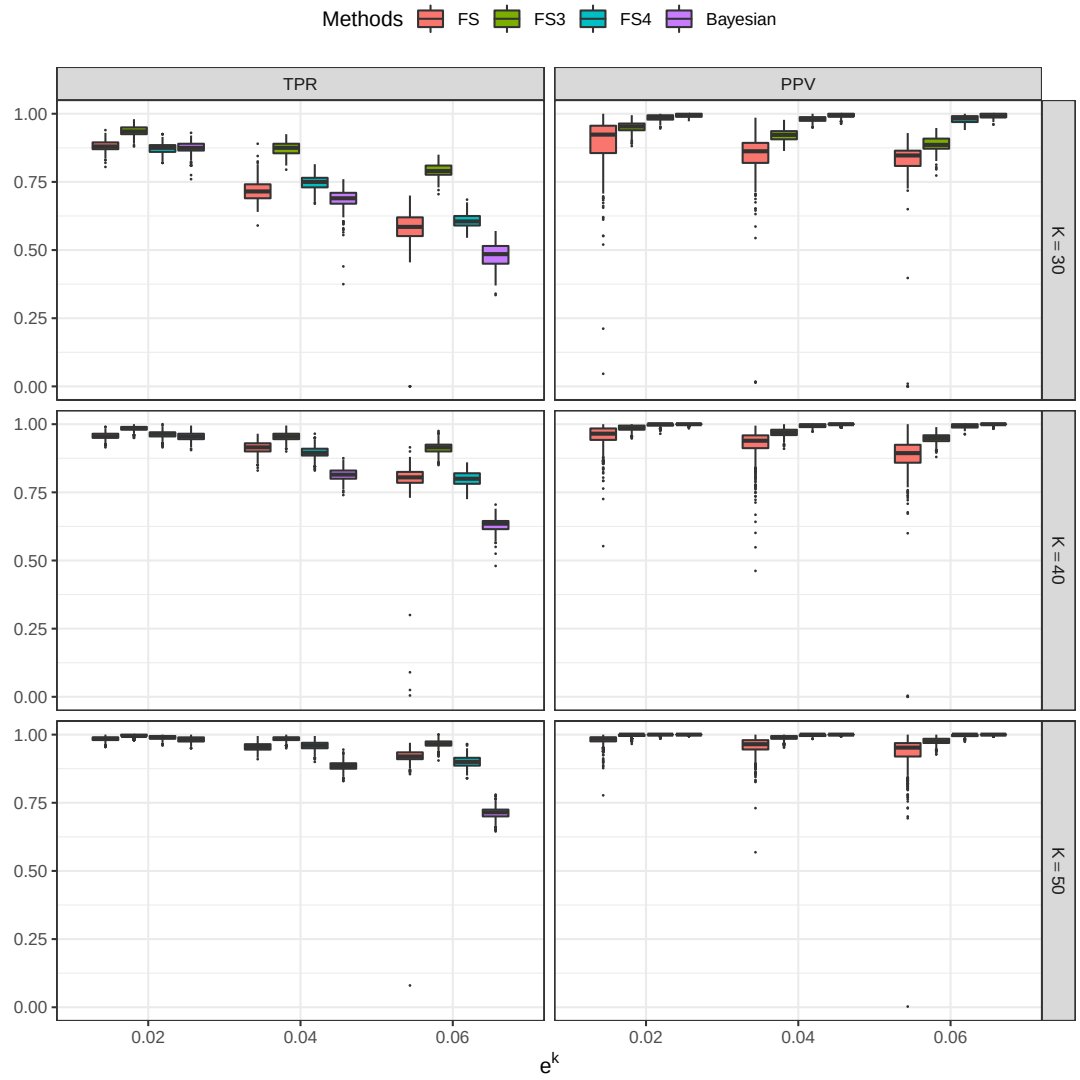


Figure A.1: Boxplots of TPR and PPV over different simulation cases when there are only binary matching variables with $n_A = 500, n_B = 200, p^k = 0.2$ and $K \in \{30, 40, 50\}$, $e \in \{0.02, 0.04, 0.06\}$.

To evaluate the impact of the choice of the threshold τ in the performances of the methods, we consider the particular scenario with the parameters $K = 40$, $p^k = 0.2$ and $e = 0.04$. We compute in Figure A.2 the Monte Carlo estimates of TPR (top left part **A**) and PPV (top right part **B**) for several values of τ. Both the TPR and PPV of observed `FS3` and `FS4` are close and far better than observed `FS`, as expected. Although the observed `FS4` has the best TPR and PPV, the estimated `FS4` has a slightly lower TPR and a slightly higher PPV. On the other hand, the estimated `FS3` and the observed `FS3` behave very similarly, in both TPR and PPV. We also plot in Figure A.2 the f-score (bottom part **C**). With the threshold $\tau = 0.5$, the observed `FS4` performs better among the observed methods, while the estimated `FS3` performs better among the estimated methods.

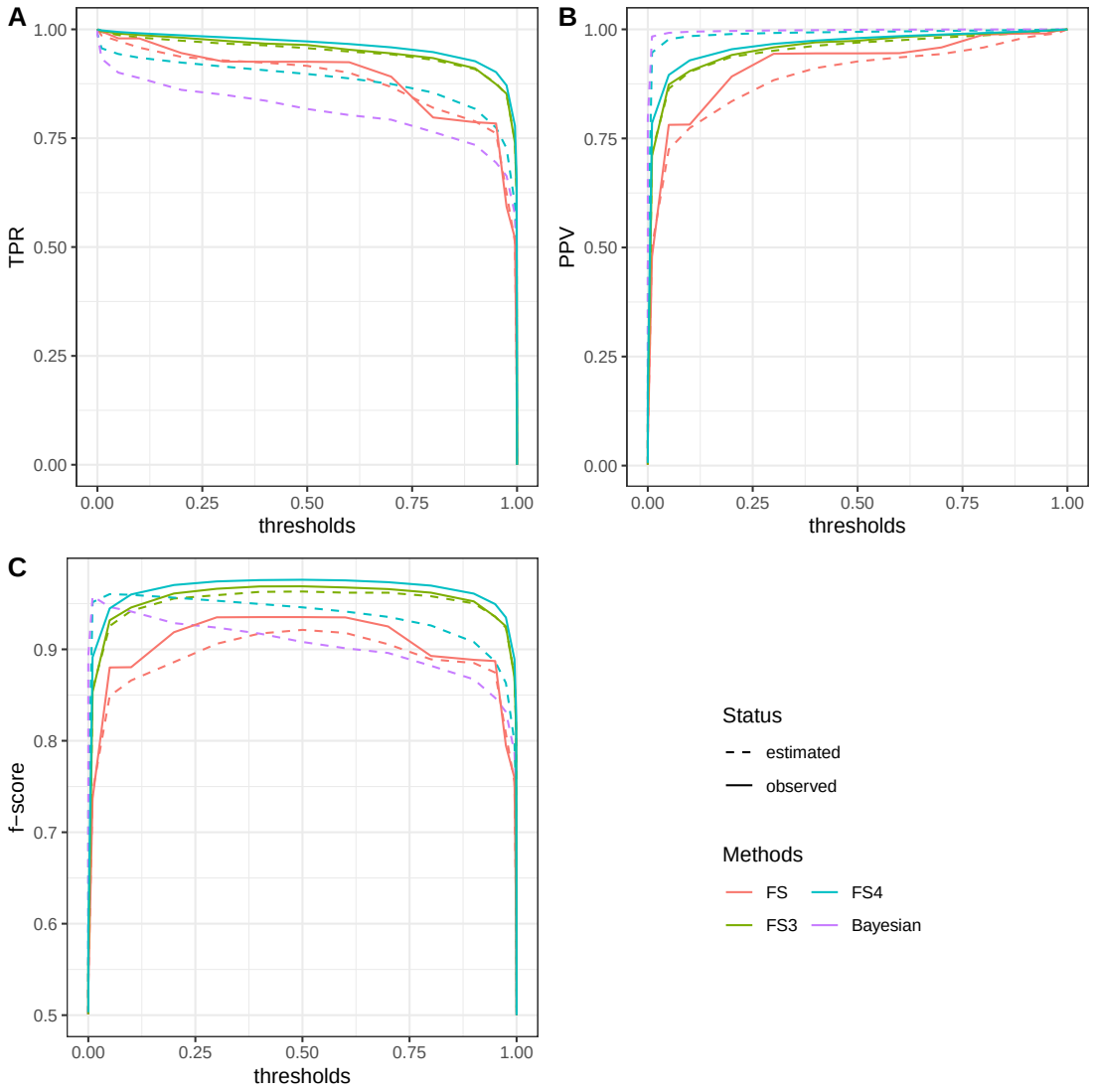


Figure A.2: TPR, PPV and f-score of estimated and observed methods over different thresholds when there are only binary matching variables with $K = 40, p^k = 0.2, e = 0.04$.

A.2.3 Affectation of prevalence

In this scenario, we vary prevalence keeping $n_A = 500, n_B = 200, K = 40, e = 0.04$ when the prevalence $p^k \in \{0.1, 0.2, 0.3\}$ for $k = 1, \ldots, K$. The boxplots of TPR and PPV obtained from 1000 repeated runs are presented in Figure A.3.

A.2.4 Affectation of ratio n_B/n_A

In this scenario, we vary $n_A \in \{400, 800, 1200\}$ when other simulation parameters are fixed as $n_B = 200, K = 40, e = 0.04, p^k = 0.2$ for $k = 1, \ldots, K$. Boxplots of TPR and PPV obtained from 1000 repeated runs are presented in Figure A.4.

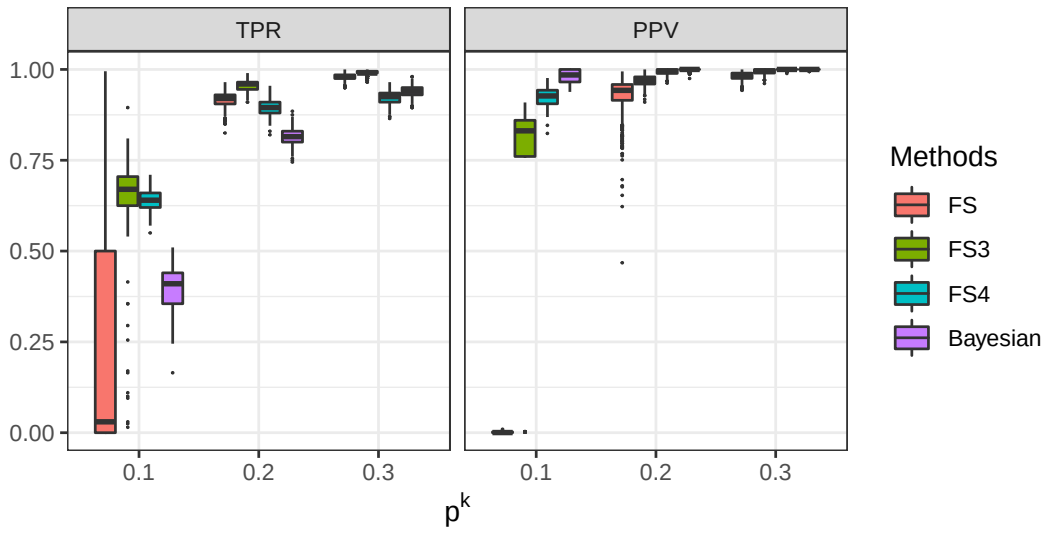


Figure A.3: Boxplots of TPR and PPV over different prevalence of matching variables $p^k \in \{0.1, 0.2, 0.3\}$ for $k = 1, \ldots, K$ keeping $n_A = 500, n_B = 200, K = 40, e = 0.04$

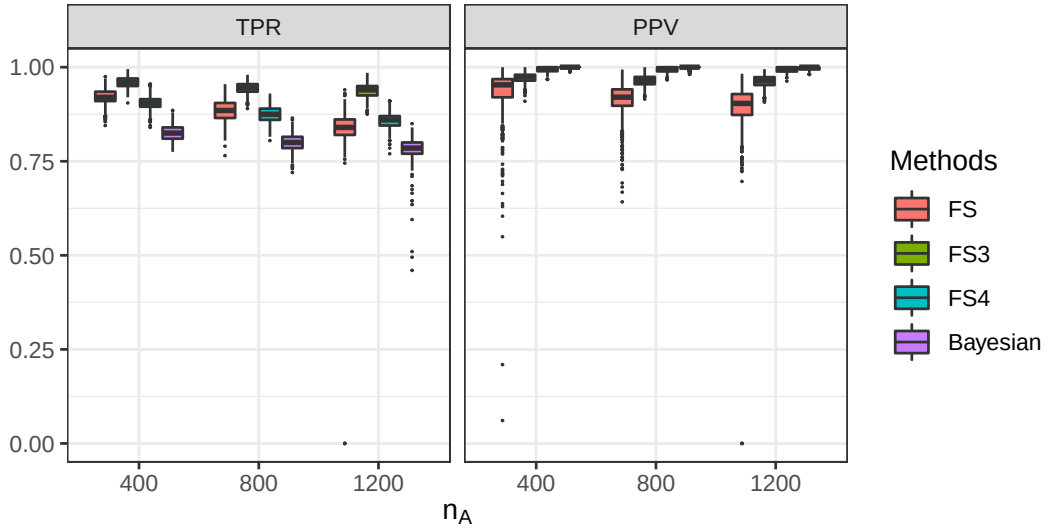


Figure A.4: Boxplots of TPR and PPV over three different ratio $n_B/n_A \in \{2/3, 2/5, 2/10\}$ keeping $n_B = 200, K = 40, e = 0.04, p^k = 0.2$ for $k = 1, \ldots, K$.

A.3 Evaluation of the model for continuous matching variables

In this section, under the same setting at Section 3.4, we consider some additional simulation scenarios when there are only continuous matching variables.

A.3.1 Initial values and stopping criteria for EM and ECM algorithm

In this scenario, the initial values for parameters of EM algorithm in `FS` and `FS3` are assigned as follows. In `FS`,

$$\begin{aligned} p_M &= \frac{n_B}{n_A n_B}, \\ m_1^k &= 0.8, m_0^k = 1 - m_1^k, \\ u_1^k &= 0.1, u_0^k = 1 - u_1^k. \end{aligned}$$

In `FS3`,

$$\begin{aligned} p_M &= \frac{n_B}{n_A n_B}, \\ m_0^k &= 0.8, m_1^k = m_2^k = 0.1, \\ u_0^k &= 0.1, u_1^k = 0.2, u_2^k = 1 - u_0^k - u_1^k. \end{aligned}$$

In `FS-HGa`, we first assign the initial values for $p_M = \frac{n_B}{n_A n_B}$ and $p_{0M}^k = 0.8$. Then, let's define

$$\begin{aligned} \Gamma_0^k &= \left\{\gamma^k : \gamma^k = 0\right\} \\ \Gamma_1^k &= \left\{\gamma^k : \gamma^k > 0\right\} \end{aligned}$$

and n_0^k, n_1^k are the number of elements in Γ_0^k, Γ_1^k respectively. Then the initial values for p_{0U}^k is computed as

$$p_{0U}^k = \frac{n_0^k - p_{0M}^k n_B}{n_A n_B - n_B}.$$

The set of positive distance Γ_1^k is divided into two sets: $\Gamma_{1,M}^k$ is the set of $n_B(1 - p_{0M})$ smallest positive values γ_{ij} and $\Gamma_{1,U}^k$ is the set of remained values. The initial values for (α_M^k, β_M^k) are the maximum likelihood estimates from $\Gamma_{1,M}^k$ and and (α_U^k, β_U^k) are the maximum likelihood estimates from $\Gamma_{1,U}^k$. If $\Gamma_{1,M}^k$ is constant, we assign $\alpha_M^k = 1, \beta_M^k = 1$.

In each method, the algorithm only stops when the relative difference between estimated parameters from two successive steps is less than 10^{-6} or when the number of iterations reaches 500.

A.3.2 Complementary results for the article

Proportion of convergence before the maximum number of iteration in EM and ECM algorithm

Table A.2 provides the proportion of convergence of each method in each simulation scenario and the average execution times. All the results are obtained by using the package simsalapar (Hofert and Mächler, 2016) for parallelly running all combinations of simulation parameters. A server with 2 Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz with 12 cores each has been used.

Boxplots

Figure A.5 shows the boxplots for all simulation results.

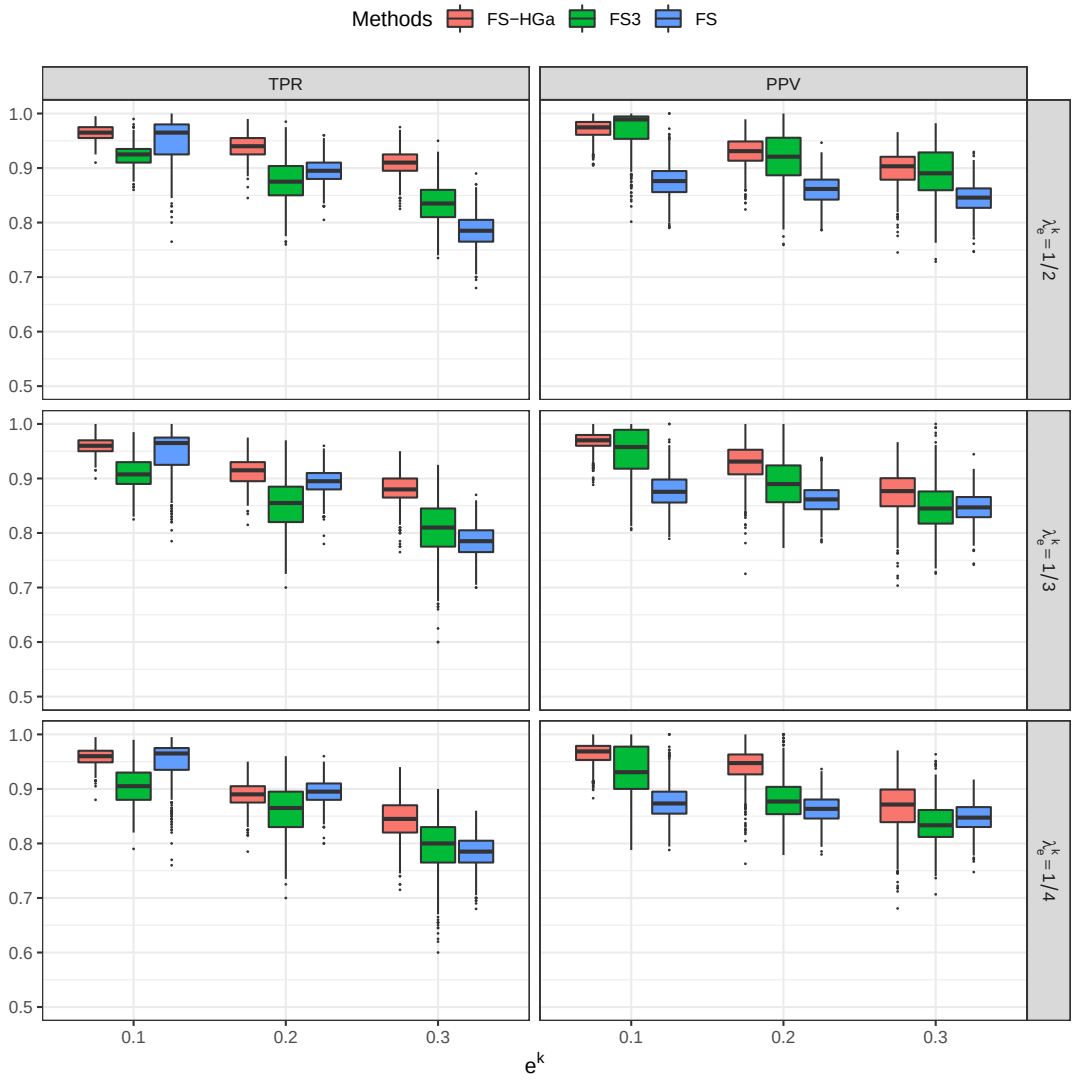


Figure A.5: Boxplots of TPR and PPV over different simulation cases when there are only continuous matching variables with $n_A = 500, n_B = 200, K = 3, \lambda^k = 0.02$ and $e \in \{0.1, 0.2, 03\}$, $\lambda_e \in \{1/2, 1/3, 1/4\}$.

e	λ_e	Methods	Proportion of convergences	Average execution time (s)
0.1	1/2	FS-HGa	0.906	183.9
		FS3	0.935	71.2
		FS	1.000	5.8
	1/3	FS-HGa	0.938	162.2
		FS3	0.973	46.9
		FS	1.000	6.2
	1/4	FS-HGa	0.955	173.3
		FS3	0.981	39.1
		FS	1.000	5.9
0.2	1/2	FS-HGa	0.950	189.5
		FS3	0.959	65.4
		FS	1.000	7.6
	1/3	FS-HGa	0.983	160.7
		FS3	0.993	36.9
		FS	1.000	7.7
	1/4	FS-HGa	0.979	150.9
		FS3	0.996	25.8
		FS	1.000	7.7
0.3	1/2	FS-HGa	0.953	175.1
		FS3	0.953	68.9
		FS	1.000	12.8
	1/3	FS-HGa	0.981	170.0
		FS3	0.993	39.1
		FS	1.000	12.6
	1/4	FS-HGa	0.970	183.4
		FS3	1.000	31.6
		FS	1.000	12.2

Table A.2: Proportion of convergence of EM/ECM algorithm and average execution time (seconds) of each method in different simulation with $n_A = 500, n_B = 200, K = 3, \lambda^k = 0.02$ and $e \in \{0.1, 0.2, 03\}$, $\lambda_e \in \{1/2, 1/3, 1/4\}$

F-score

To evaluate the impact of the choice of the threshold τ, we consider the particular scenario with the parameters $K = 3$, $\lambda^k = 0.02$, $e = 0.2$ and $\lambda_e = 1/2$. We compute in Figure A.6 the Monte Carlo estimates of TPR (top left part **A**) and PPV (top right part **B**) for several values of τ. We observe that for both the TPR and PPV, the results obtained for a given estimated method and for its observed counterpart are very similar. Also, we observe that `FS-HGa` performs significantly better than `FS3` and `FS`, both in terms of TPR and PPV. This is more clearly illustrated by the plot of f-score (bottom part **C**).

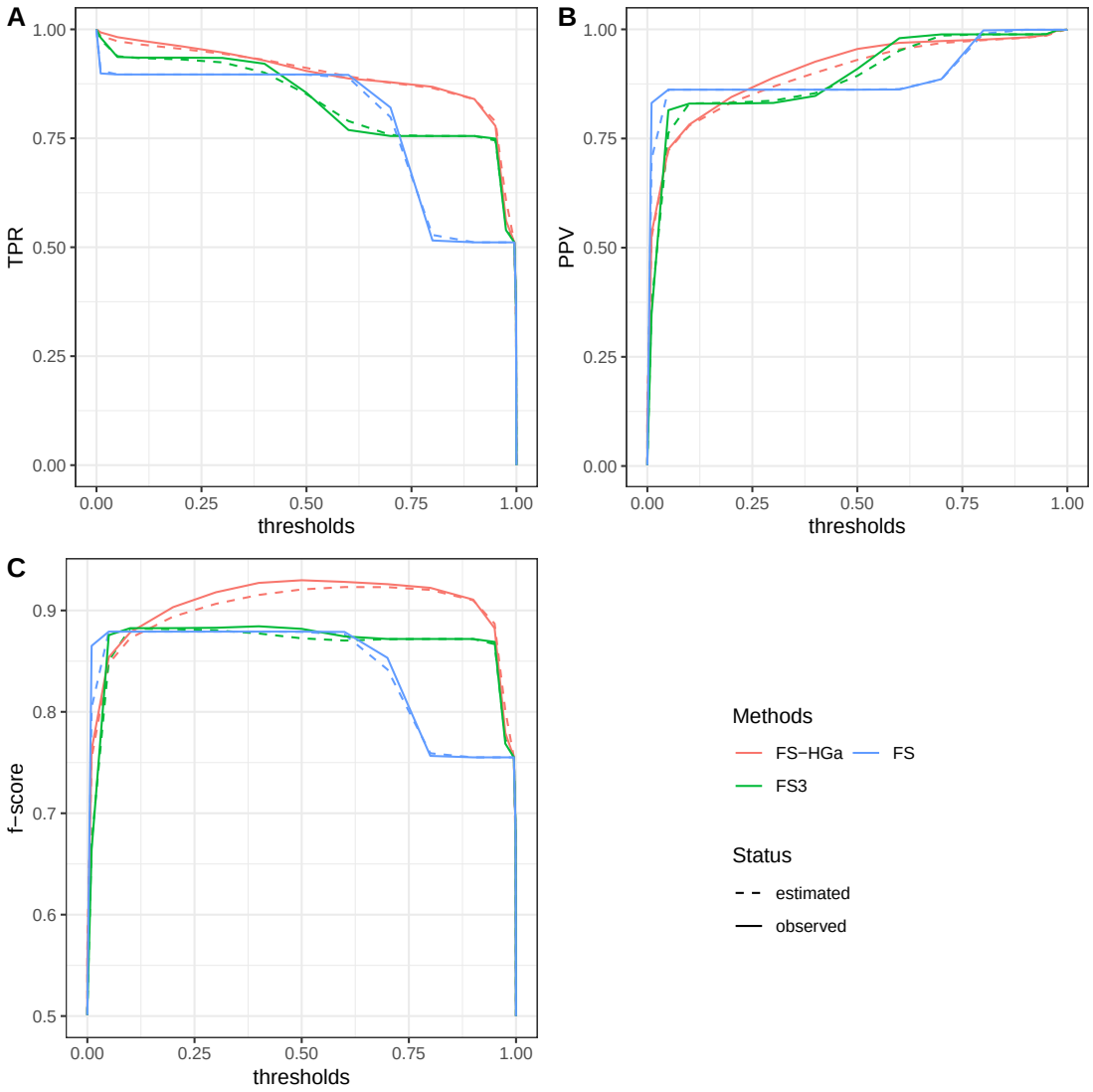


Figure A.6: TPR, PPV, f-score of estimated and observed methods over different thresholds when there are only continuous matching variables with $K = 3, \lambda^k = 0.02, e = 0.2$ and $\lambda_e = 1/3$.

A.3.3 Affectation of range of matching variables

In this scenario, we vary $\lambda^k \in \{0.01, 0.02, 0.03\}$ for $k = 1, \ldots, K$, which means the mean values for continuous variables are $100, 50, 25$ respectively. A matching variable having larger range is more distinguish and better for record linkage. Other simulation parameters are fixed as $n_A = 500, n_B = 200, K = 3, e = 0.2, \lambda^e = 1/3$. The boxplots of TPR and PPV over 1000 simulations is presented in Figure A.7 .

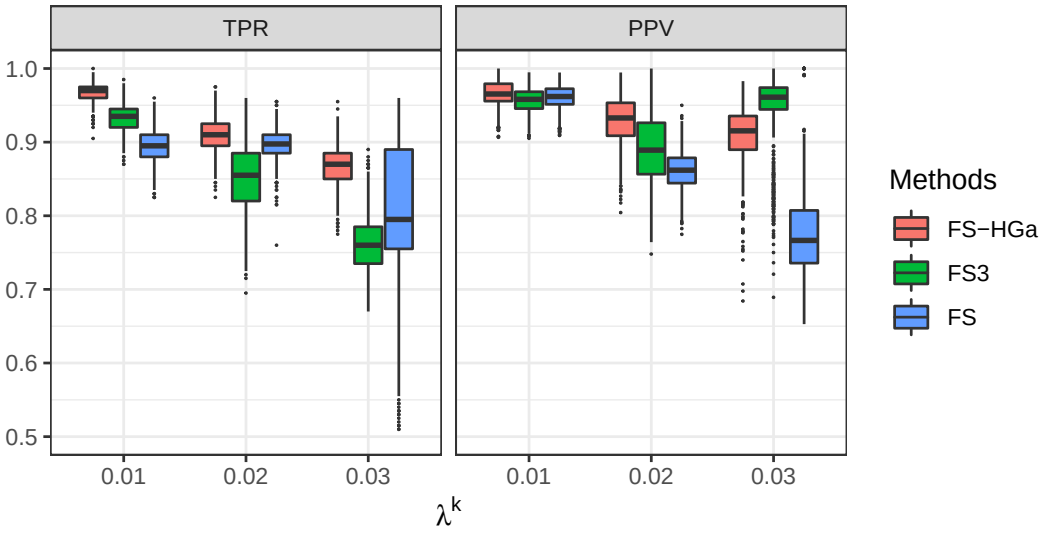


Figure A.7: Boxplots of TPR and PPV over three different range of matching variables with $\lambda^k \in \{0.01, 0.02, 0.03\}$ for $k = 1, \ldots, K$ keeping $n_A = 500, n_B = 200, K = 3, e = 0.2, \lambda^e = 1/3$.

A.3.4 Affectation of number of matching variables

In this simulation, we vary $K \in \{2, 3, 4\}$ when $n_A = 500, n_B = 200, \lambda^k = 0.005, e = 0.2, \lambda^e = 1/2$. Boxplots of TPR and PPV over 1000 simulations is presented in Figure A.8.

A.3.5 Affectation of ratio n_B/n_A

In this scenario, we vary $n_A \in \{400, 800, 1200\}$ keeping $n_B = 200, K = 3, \lambda^k = 0.02, e^k = 0.2, \lambda_e^k = 1/3$ for $k = 1, \ldots, K$. Boxplots of TPR and PPV over 1000 simulations is presented in Figure A.9.

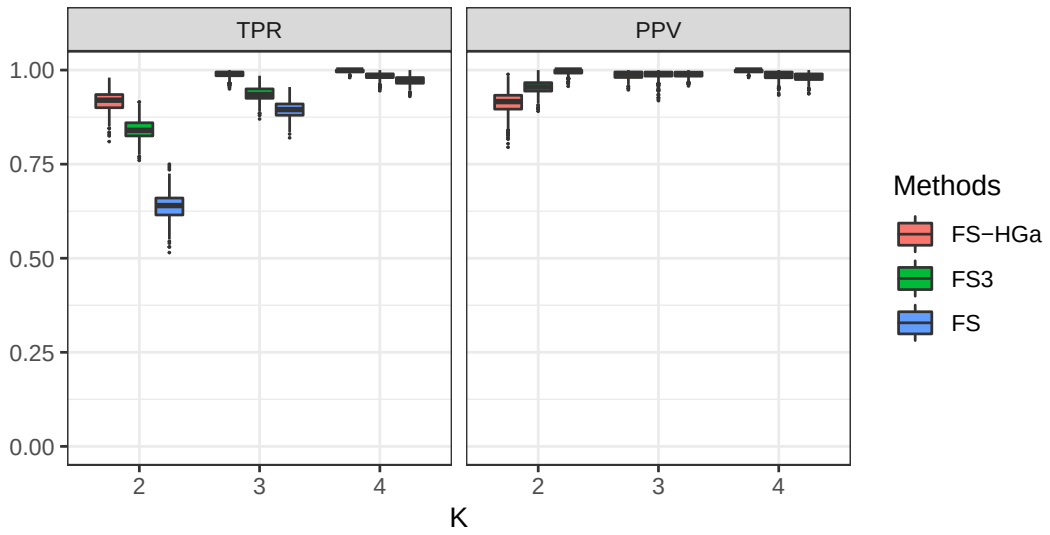


Figure A.8: Boxplots of TPR and PPV of different methods with different $K \in \{2, 3, 4\}$ when $n_A = 500, n_B = 200, \lambda^k = 0.005, e = 0.2, \lambda^e = 1/2$.

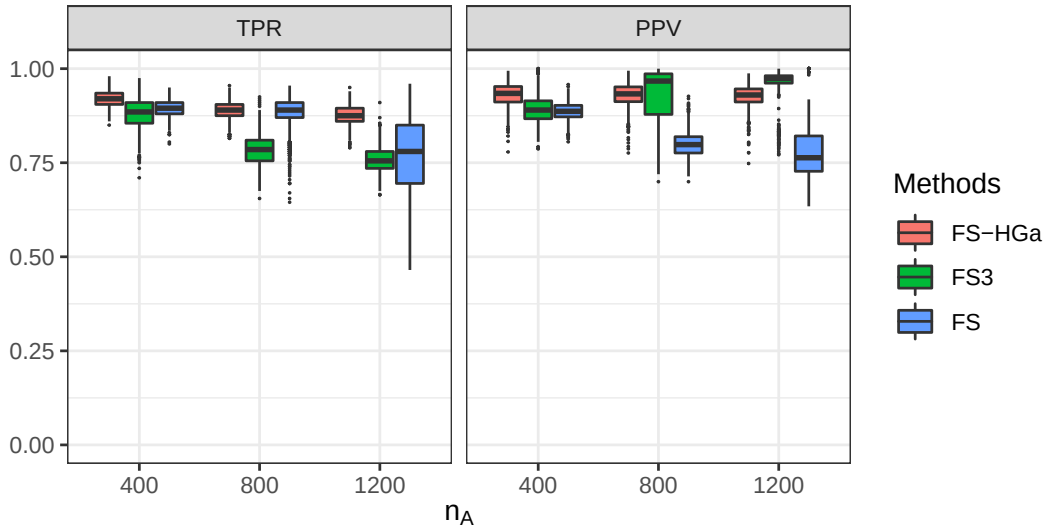


Figure A.9: Boxplots of TPR and PPV different ration of n_B/n_A with $n_A \in \{300, 500, 1000\}$ keeping $n_B = 200, K = 3, \lambda^k = 0.02, e^k = 0.2, \lambda_e^k = 1/3$ for $k = 1, \ldots, K$.

A.3.6 Robustness of the hurdle gamma mixture model

We evaluate the robustness of the hurdle gamma mixture model for different distribution of continuous matching variables.

Uniform distribution for X^k and exponential distribution for ϵ

In this simulation, we generate X_A^k follows an Uniform distribution $(0, z^k)$ and the error ϵ follows Exponential distribution with parameter λ_e. We consider different

case with $n_A = 500, n_B = 200, K = 3, z^k = 100, \lambda_e \in \{1/2, 1/3, 1/4\}$ and the proportion of error $e \in \{0.1, 0.2, 0.3\}$. Figure A.10 shows the Monte-Carlo estimates and Figure A.11 shows the boxplots of all methods over 1000 repeated runs.

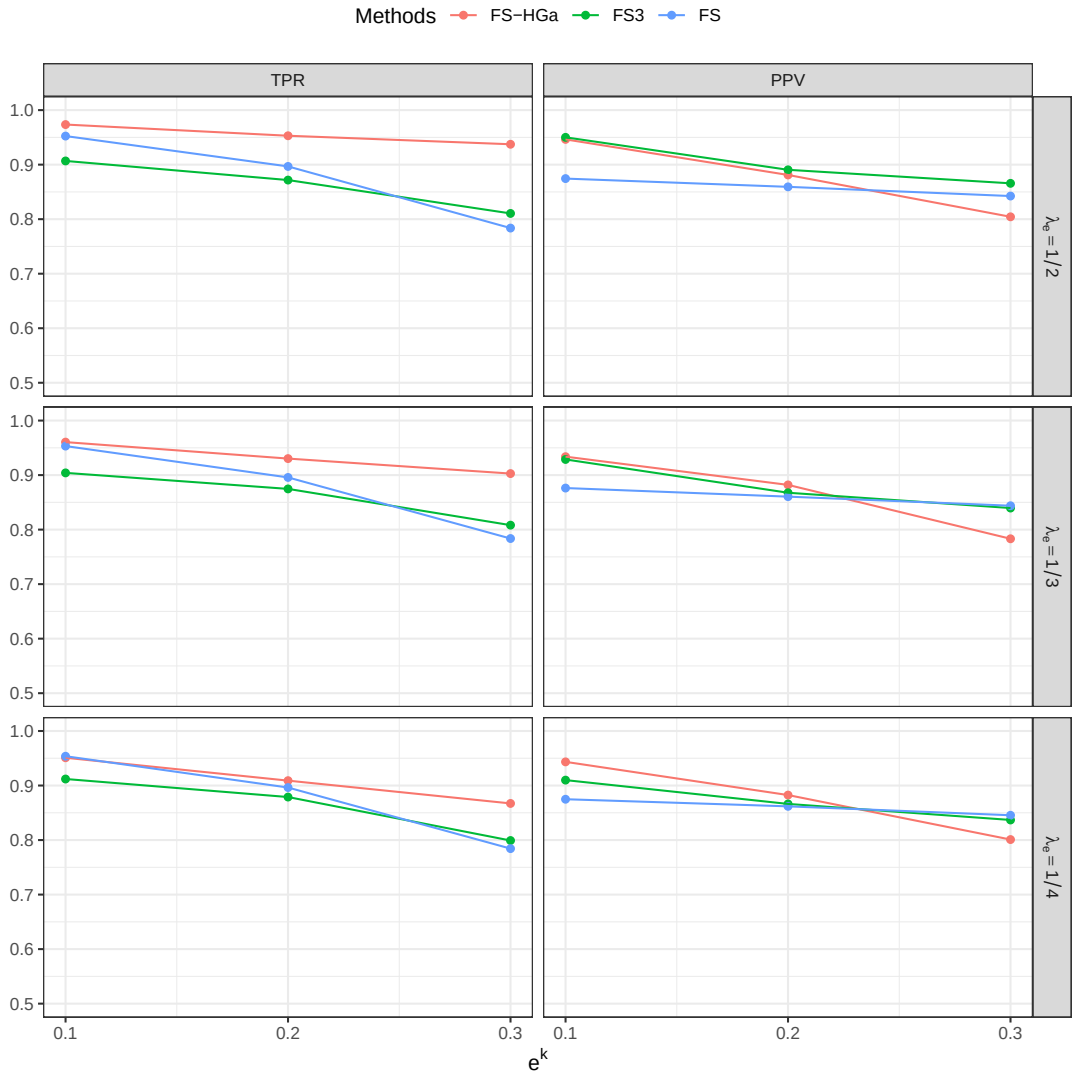


Figure A.10: Monte Carlo estimates of TPR and PPV when $n_A = 500, n_B = 200, K = 3, z^k = 100, \lambda_e \in \{1/2, 1/3, 1/4\}$ and the proportion of error $e \in \{0.1, 0.2, 0.3\}$

Exponential distribution for X^k and normal distribution for ϵ

In this simulation, we generate X_A^k follows an Exponential distribution (λ^k) and the error ϵ^k follows Normal distribution (μ_e^k, σ_e^k). We estimate TPR and PPV of different methods when $n_A = 500, n_B = 200, K = 3, \lambda^k = 0.02, \sigma_e^k = 1, \mu_e^k \in \{1, 2, 3\}$ and the proportion of error $e^k \in \{0.1, 0.2, 0.3\}$. Each case is repeated 1000 times.

Once the databases were generated, we compared four following record linkage

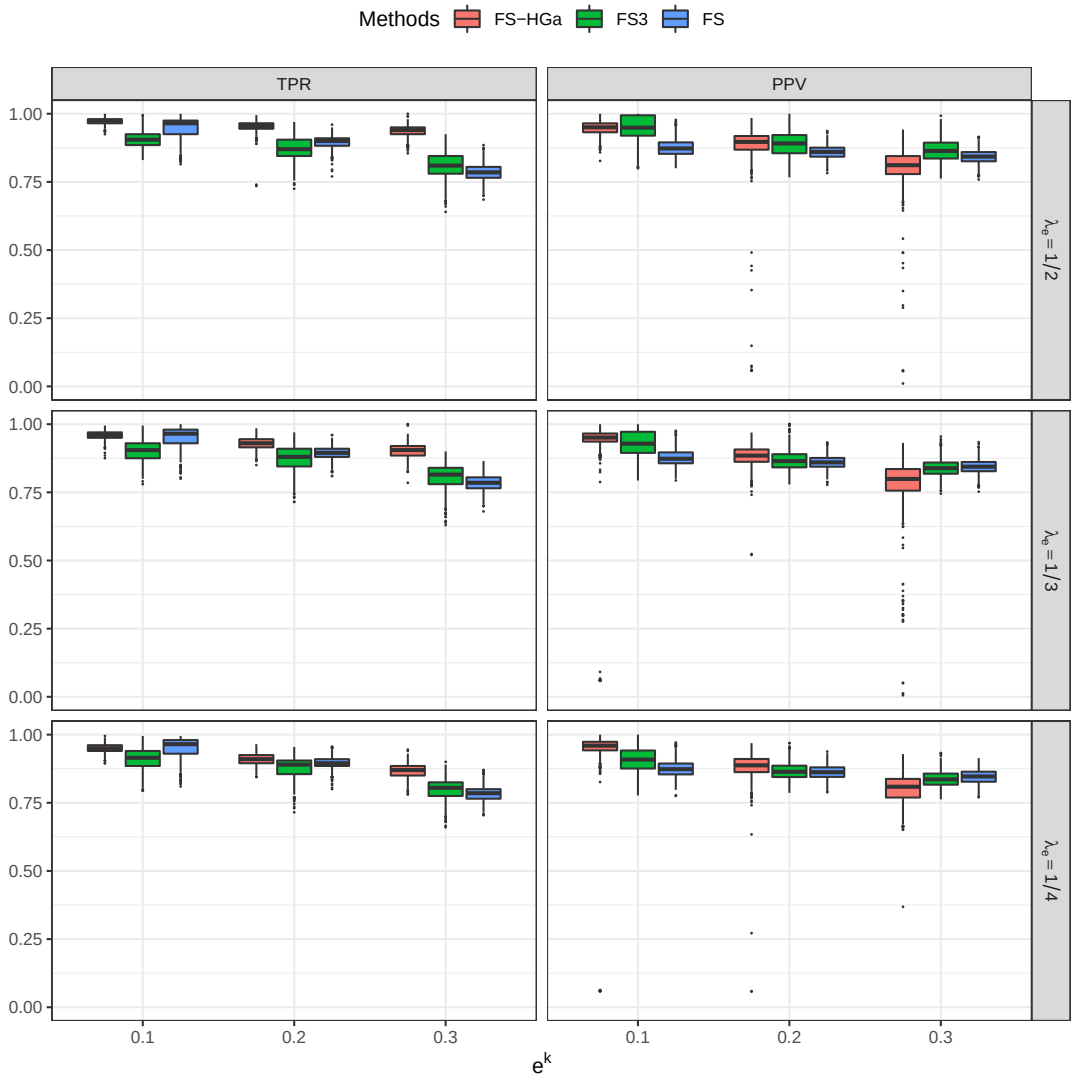


Figure A.11: Boxplots of TPR and PPV when $n_A = 500, n_B = 200, K = 3, z^k = 100, \lambda_e^k \in \{1/2, 1/3, 1/4\}$ and the proportion of error $e^k \in \{0.1, 0.2, 0.3\}$

methods: `FS`, the Fellegi-Sunter model with simple binary comparison; `FS3`, the Fellegi-Sunter model using a comparison with 3 categories described as follows:

$$\gamma_{ij}^k = \begin{cases} 0 & \text{if } |X_{B,j}^k - X_{A,j}^k| = 0, \\ 1 & \text{if } 0 < |X_{B,j}^k - X_{A,j}^k| \leq 2, \\ 2 & \text{if } 2 < |X_{B,j}^k - X_{A,j}^k| \end{cases} \tag{A.10}$$

for $k = 1, \ldots, K$; `FS-HGa`, the Fellegi-Sunter model with the 1-norm comparison method as

$$\gamma_{ij}^k = |X_{B,j}^k - X_{A,j}^k|,$$

and `FS-HGa2`, the Fellegi-Sunter model with the 2-norm comparison method as

$$\gamma_{ij}^k = |X_{B,j}^k - X_{A,j}^k|^2.$$

In `FS-HGa` and `FS-HGa2`, we keep using the mixture of hurdle gamma distribution model for the comparison data. Figure A.12 shows the Monte-Carlo estimates and Figure A.13 shows the boxplots of all methods over 1000 repeated runs.

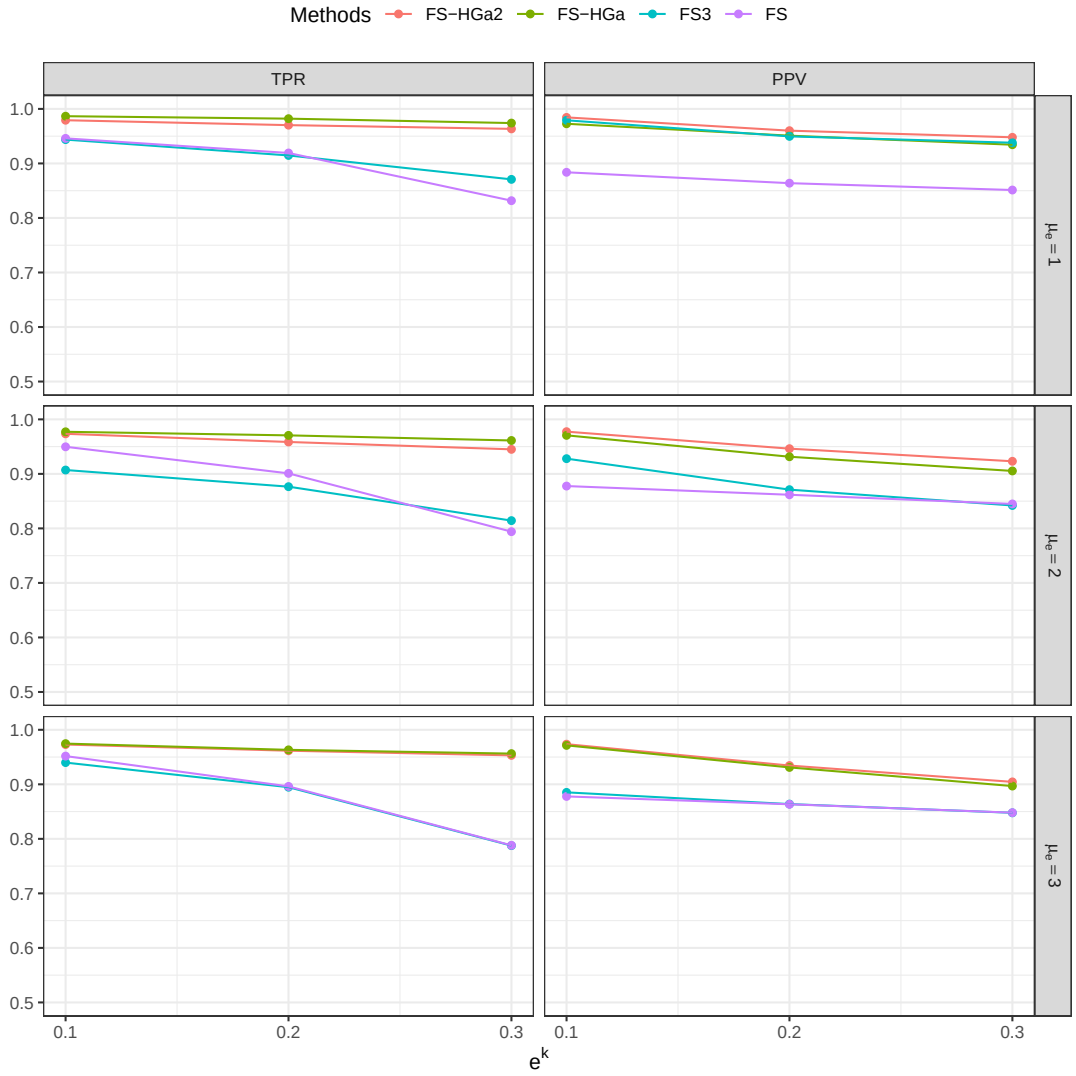


Figure A.12: Monte Carlo estimates of TPR and PPV of 4 methods when $n_A = 500, n_B = 200, K = 3, \lambda^k = 0.02, \sigma_e^k = 1, \mu_e^k \in \{1, 2, 3\}$ and the proportion of error $e^k \in \{0.1, 0.2, 0.3\}$

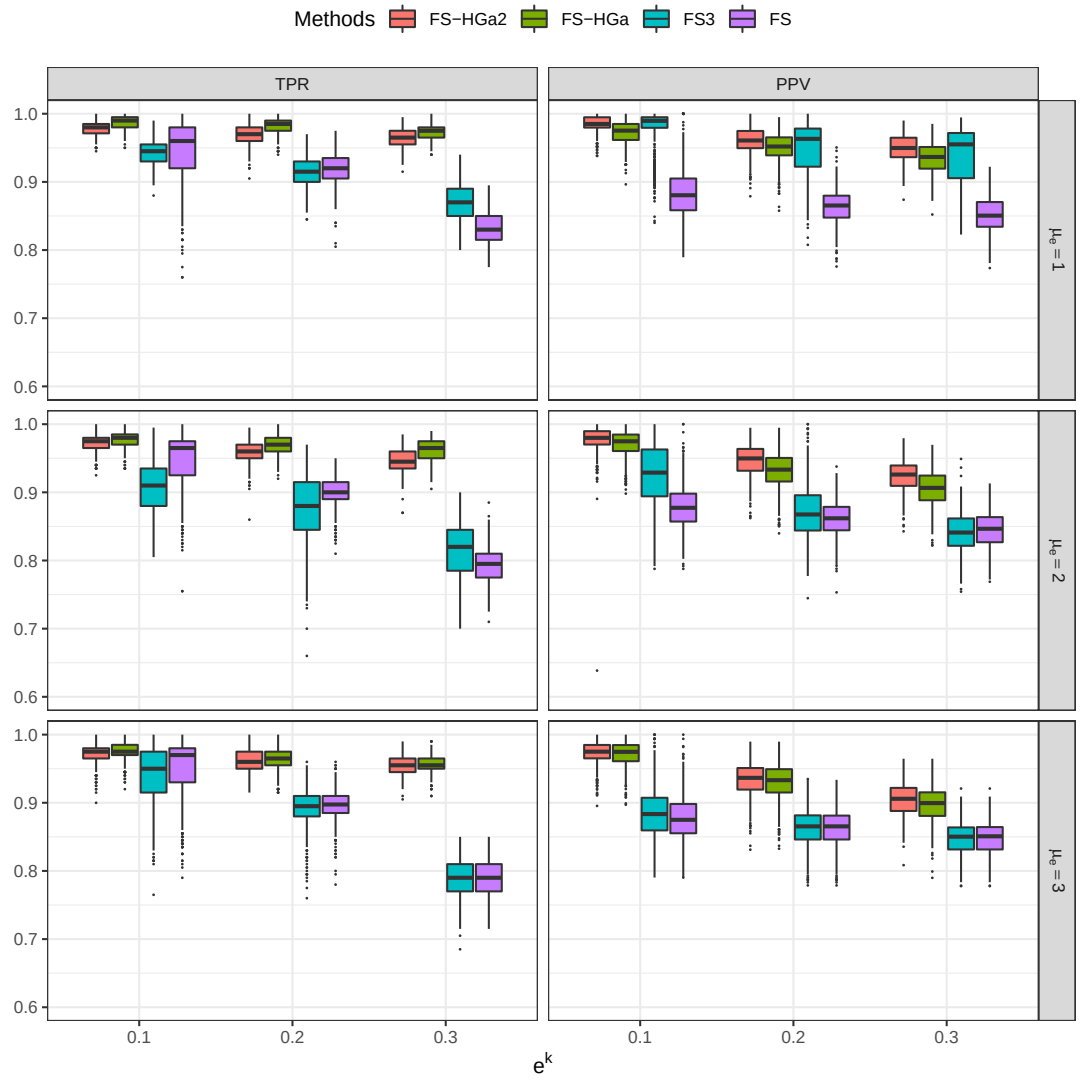


Figure A.13: Boxplots of TPR and PPV of 4 methods when $n_A = 500, n_B = 200, K = 3, \lambda^k = 0.02, \sigma_e^k = 1, \mu_e^k \in \{1, 2, 3\}$ and the proportion of error $e^k \in \{0.1, 0.2, 0.3\}$

.

A.4 Evaluation of the model for both categorical and continuous matching variables

In this section, we design a simulation study to compare the performance of the proposed model to the standard Fellegi-Sunter model, in case when both categorical and continuous matching variables are used.

We consider two databases A and B containing $n_A = 500$ and $n_B = 200$ individuals. There are $K = 6$ matching variables with $K_1 = 2$ categorical matching variables, $K_2 = 2$ low prevalence binary matching variables, and $K_3 = 2$ continuous matching variables. Following the same process as in the other simula-

tions presented in the article, the values in database A is first generated. The two categorical matching variables are generated according to a discrete uniform distribution, where the number of categories are randomly selected in $\{30, 50\}$. The two low prevalence binary matching variables are generated according to a Bernoulli distribution of parameter 0.01. The two continuous matching variables are generated from an exponential distribution with parameter 0.01.

A subset of 200 units is then selected in A to obtain the dataset B, and some random perturbations are then introduced in the generated values. In case of the binary matching variables, 2% of the records contain errors which are generated by randomly replacing the true modality by the other modality. In case of the categorical matching variables, 10% of the records contain errors, which are generated by randomly replacing the correct value with another category. In case of the continuous matching variables, the proportion of errors e varies in $\{10\%, 20\%, 30\%\}$, and these errors are generated according to an Exponential distribution with parameter 0.5.

The data generation process is repeated 1000 times. For each generated dataset, we perform two linkage methods: the standard Fellegi-Sunter model `FS`, and our proposed model `FS-ext`. In `FS`, the binary comparison only is applied. In `FS-ext`, the binary comparison is still applied for the two categorical matching variables, but the proposed comparison function (7) is applied for the binary matching variables, and the absolute distance (19) is applied for the continuous matching variables. The threshold $\tau = 0.5$ is used for the estimation of posterior probabilities of matching from `FS` and `FS-ext`.

In addition, we considered another implementation of the standard Fellegi-Sunter model by R RecordLinkage package, in order to compare our program to alternatives. This is denoted as `FS-RecLink`. Since this package orders the record pairs with respect to the matching weights, we have to compute a threshold for the matching weights equivalent to a threshold of 0.5 for the estimated posterior probabilities of matching.

The matching weight is

$$w = \frac{\mathbb{P}(\gamma|M)}{\mathbb{P}(\gamma|U)}, \tag{A.11}$$

for a value γ of the comparison vector, and the posterior probability of matching

is

$$q = \mathbb{P}(M|\gamma). \tag{A.12}$$

Using the Bayes's rule, we have

$$q = \frac{\mathbb{P}(\gamma|M)\mathbb{P}(M)}{\mathbb{P}(\gamma|M)\mathbb{P}(M) + \mathbb{P}(\gamma|U)(1-\mathbb{P}(M))} = \frac{w\mathbb{P}(M)}{w\mathbb{P}(M) + 1 - \mathbb{P}(M)}.$$

After some algebra, we obtain

$$w = \frac{1-\mathbb{P}(M)}{1-q}\frac{q}{\mathbb{P}(M)}. \tag{A.13}$$

In the RecordLinkage package, the matching weight is defined as $\log_2(w)$. Under our simulation setting, we have $\mathbb{P}(M) = \frac{n_B}{n_A n_B} = \frac{1}{500} = 0.002$. By replacing q in (A.13) with the threshold $\tau = 0.5$, the threshold value for the matching weight in `FS-RecLink` is $\log_2\left(\frac{1-0.002}{1-0.5}\frac{0.5}{0.002}\right) \approx 8.963$.

e	Methods	TPR	PPV	Execution time(s)
0.1	`FS`	0.948	0.994	3.73
	`FS-RecLink`	0.948	0.995	2.98
	`FS-ext`	0.982	0.987	66.20
0.2	`FS`	0.892	0.995	4.29
	`FS-RecLink`	0.892	0.995	3.38
	`FS-ext`	0.976	0.984	62.6
0.3	`FS`	0.825	0.995	5.01
	`FS-RecLink`	0.825	0.995	3.86
	`FS-ext`	0.967	0.981	61.40

Table A.3: Summary of 1,000 Monte-Carlo simulation in case of having both categorical and continuous matching variables.

The Monte-Carlo estimations of the TPR and the PPV are presented in Table A.3 for the three linkage methods, along with their standard error. It is clear that both `FS` and `FS-RecLink` produce approximately the same results. Concerning the proposed method `FS-ext`, the PPV is very similar to what is observed for both `FS` and `FS-RecLink`. The TPR is much improved, especially when the proportion of errors in the continuous matching variables increases.

In this simulation, we also compute the mean execution time under the three linkage models. The execution time is much larger for `FS-ext`, but remains very moderate (no more than 70 seconds). This is due to the fact that the proposed model requires more parameters to be estimated. Also, since the likelihood function is more complex, the ECM algorithm is needed. In addition, the ECM algorithm involves maximization steps, while the EM algorithm in `FS` involves only one.

A.5 Naive example for comparison step

In this section, we give an example of implementing the proposed comparison approaches. Let's consider in Tables A.4 and A.5 the case of patients for which two data sets are available. Both of them contain the postal code (categorical variable), an indicator for lung cancer (binary variable), and the date of admission in hospital.

	Postal code	Lung cancer	Date of admission
a_1	35170	1	05/01/2020
a_2	35510	0	22/01/2020
a_3	35170	0	12/01/2020

Table A.4: Naive example of database A

	Postal code	Lung cancer	Date of admission
b_1	35510	0	25/01/2020
b_2	35170	1	05/01/2020

Table A.5: Naive example of database B

Table A.6 is a comparison matrix associated to the example initiated in Tables A.4 and A.5. In Table A.6, each row is a comparison vector γ_{ij} of a record pair $(X_{A,i}, X_{B,j})$.

	γ^1	γ^2	γ^3
γ_{11}	0	0	0
γ_{12}	1	1	1
γ_{21}	1	1	0
γ_{22}	0	0	0
γ_{31}	0	1	0
γ_{32}	1	0	0

Table A.6: A simple comparison matrix for the data given in Tables A.4 and A.5

	γ^1	γ^2	γ^3
γ_{11}	0	1	20
γ_{12}	1	2	0
γ_{21}	1	0	3
γ_{22}	0	1	17
γ_{31}	0	0	13
γ_{32}	1	1	7

Table A.7: A variant for the comparison matrix given in Table A.6 with mixed type comparison values.

For example, Table A.7 is a new comparison matrix for the data given in Tables A.4 and A.5. The simple binary comparison function is applied for the first matching variable (postal code). The three categories comparison method is applied for the binary matching variable (lung cancer). The distance

$$d(X^k_{A,i}, X^k_{B,j}) = |X^k_{A,i} - X^k_{B,j}| \tag{A.14}$$

is applied for the continuous matching variable (duration of the hospital stay). The set of all observed comparison vectors is fitted by a mixture model for mixed type data.

B Appendix for Chapter 4

In this appendix, we present some complementary materials for Chapter 4. The appendix B.1 provides the expectation computation for the adjusted estimating equation. In appendix B.2, we show detail formulas for the variance estimator. Then, some additional simulation results are included in appendix B.3.

B.1 Expectation of the adjusted estimating equation

The proposed *adjusted estimating equation* is given by

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

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

$$\begin{aligned}
\mathbb{E}\{\bar{H}(\boldsymbol{\beta}) \mid \mathcal{F}\} &= \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \mathbb{E}\left\{ \delta_i \left[\mathbf{X}_i^* - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^*(\alpha_v, \boldsymbol{\beta})}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^*(\alpha_v, \boldsymbol{\beta})} \right] \middle| \mathcal{F} \right\} \\
&= \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \delta_i \left\{ \underbrace{\mathbb{E}(\mathbf{X}_i^* \mid \mathcal{F})}_{E_1} - \underbrace{\mathbb{E}\left(\frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^*(\alpha_v, \boldsymbol{\beta})}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^*(\alpha_v, \boldsymbol{\beta})} \middle| \mathcal{F} \right)}_{E_2} \right\}
\end{aligned} \tag{B.2}$$

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 (4.4) 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 (4.7) and under the non-informative linkage assumption, we have

$$\begin{aligned}E_1 &= \mathbb{E}\left\{\alpha_v^{-1}\mathbf{Z}_i - (\alpha_v^{-1}-1)\bar{\mathbf{X}}_{B_v} \mid \mathcal{F}\right\}\\ &= \alpha_v^{-1}\mathbb{E}\left(\mathbf{Z}_i \mid \mathcal{F}\right) - (\alpha_v^{-1}-1)\bar{\mathbf{X}}_{B_v}\\ &= \alpha_v^{-1}\left[\alpha_v \mathbf{X}_i + (1-\alpha_v)\bar{\mathbf{X}}_{B_v}\right] - (\alpha_v^{-1}-1)\bar{\mathbf{X}}_{B_v}\\ &= \mathbf{X}_i.\end{aligned} \tag{B.3}$$

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

$$E_2 \approx \frac{\mathbb{E}\left\{\sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)h_j^*(\alpha_v,\boldsymbol{\beta})\middle|\mathcal{F}\right\}}{\mathbb{E}\left\{\sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)g_j^*(\alpha_v,\boldsymbol{\beta})\middle|\mathcal{F}\right\}} \tag{B.4}$$

where

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

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,

$$E_2 \approx \frac{\sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)h(\beta,\mathbf{X}_j)}{\sum_{v=1}^{V}\sum_{j\in A_v} Y_j(T_i)g(\beta,\mathbf{X}_j)} \tag{B.5}$$

By plugging (B.3) and (B.5) into (B.2), we obtain

$$\begin{aligned}\mathbb{E}\left\{\bar{H}(\boldsymbol{\beta})\middle|\mathcal{F}\right\} &\approx \frac{1}{n_A}\sum_{v=1}^{V}\sum_{i\in A_v}\delta_i\left\{\mathbf{X}_i - \frac{\sum_{v=1}^{V}\sum_{j\in A_v}Y_j(T_i)h(\boldsymbol{\beta},\mathbf{X}_j)}{\sum_{v=1}^{V}\sum_{j\in A_v}Y_j(T_i)g(\boldsymbol{\beta},\mathbf{X}_j)}\right\} \\ &= H_0(\boldsymbol{\beta}).\end{aligned} \tag{B.6}$$

B.2 Computation of the variance estimator

In this section, 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.2.1 Global estimating equation

Using the unbiased estimator $\hat{\alpha}$ for the parameter α (see equation 4.4), the global estimating equation for the parameter β is

$$\bar{H}(\boldsymbol{\beta}) \equiv \frac{1}{n_A}\sum_{i=1}^{n_A} \underbrace{\delta_i \left\{ \mathbf{X}_i^*(\hat{\alpha}) - \frac{\sum_{j=1}^{n_A} Y_j(T_i) h_j^*(\hat{\alpha}, \boldsymbol{\beta})}{\sum_{j=1}^{n_A} Y_j(T_i) g_j^*(\hat{\alpha}, \boldsymbol{\beta})} \right\}}_{H_i(\boldsymbol{\beta})} = 0, \tag{B.7}$$

where

$$\begin{aligned} \mathbf{X}_i^*(\hat{\alpha}) &= \frac{\mathbf{Z}_i}{\hat{\alpha}} - \frac{1-\hat{\alpha}}{\hat{\alpha}}\bar{\mathbf{X}}_B, \\ g_j^*(\hat{\alpha}, \boldsymbol{\beta}) &= \frac{g(\boldsymbol{\beta}, \mathbf{Z}_j)}{\hat{\alpha}} - \frac{1-\hat{\alpha}}{\hat{\alpha}}\bar{g}_B(\boldsymbol{\beta}), \\ h_j^*(\hat{\alpha}, \boldsymbol{\beta}) &= \frac{h(\boldsymbol{\beta}, \mathbf{Z}_i)}{\hat{\alpha}} - \frac{1-\hat{\alpha}}{\hat{\alpha}}\bar{h}_B(\boldsymbol{\beta}). \end{aligned} \tag{B.8}$$

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

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}) = \{\nabla\bar{H}(\hat{\boldsymbol{\beta}})\}^{-1} \times \hat{\mathbb{V}}\{\bar{H}(\boldsymbol{\beta}_0)\} \times \{\nabla\bar{H}(\hat{\boldsymbol{\beta}})\}^{-1}. \tag{B.9}$$

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

B.2.2 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 (B.8) as

$$\begin{aligned}
\mathbf{X}_i^*(\hat{\alpha}) &= \underbrace{\frac{1}{\alpha}(\mathbf{Z}_i - \bar{\mathbf{X}}_B) + \bar{\mathbf{X}}_B}_{\mathbf{x}_i^*(\alpha)} \\
&\quad - \frac{\hat{\alpha}-\alpha}{\alpha^2}(\mathbf{Z}_i - \bar{\mathbf{X}}_B) + o_p({n_V}^{-0.5}), \\
g_j^*(\hat{\alpha}, \boldsymbol{\beta}_0) &= \underbrace{\frac{1}{\alpha}\left\{g(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{g}_B(\boldsymbol{\beta}_0)\right\} + \bar{g}_B(\boldsymbol{\beta}_0)}_{g_j^*(\alpha, \boldsymbol{\beta}_0)} \\
&\quad - \frac{\hat{\alpha}-\alpha}{\alpha^2}\left\{g(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{g}_B(\boldsymbol{\beta}_0)\right\} + o_p({n_V}^{-0.5}), \\
h_j^*(\hat{\alpha}, \boldsymbol{\beta}_0) &= \underbrace{\frac{1}{\alpha}\left\{h(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{h}_B(\boldsymbol{\beta}_0)\right\} + \bar{h}_B(\boldsymbol{\beta}_0)}_{h_j^*(\alpha, \boldsymbol{\beta}_0)} \\
&\quad - \frac{\hat{\alpha}-\alpha}{\alpha^2}\left\{h(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{h}_B(\boldsymbol{\beta}_0)\right\} + o_p({n_V}^{-0.5}).
\end{aligned} \tag{B.10}$$

Let us denote $\epsilon = \frac{\hat{\alpha}-\alpha}{\alpha^2}$. By plugging (B.10) into equation (B.7), 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:

$$\bar{H}(\boldsymbol{\beta}_0) \;=\; \bar{H}_1(\boldsymbol{\beta}_0) - \left(\frac{\hat{\alpha}-\alpha}{\alpha^2}\right)\bar{H}_2(\alpha, \boldsymbol{\beta}_0) + o_p({n_V}^{-0.5}), \tag{B.11}$$

where

$$\bar{H}_1(\boldsymbol{\beta}_0) \;=\; \frac{1}{n_A}\sum_{i=1}^{n_A} \underbrace{\delta_i\left\{\mathbf{X}_i^*(\alpha) - R_i^*(\alpha, \boldsymbol{\beta}_0)\right\}}_{H_{1i}(\boldsymbol{\beta}_0)} \tag{B.12}$$

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

$$\begin{aligned}
\bar{H}_2(\alpha, \boldsymbol{\beta}_0) = \frac{1}{n_A}\sum_{i=1}^{n_A} \delta_i \Bigg[&(\mathbf{Z}_i - \bar{\mathbf{X}}_B) \\
&- \frac{\sum_{j=1}^{n_A} Y_j(T_i)\left\{\left[h(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{h}_B(\boldsymbol{\beta}_0)\right] - R_i^*(\alpha, \boldsymbol{\beta}_0)\left[g(\boldsymbol{\beta}_0, \mathbf{Z}_j) - \bar{g}_B(\boldsymbol{\beta}_0)\right]\right\}}{\sum_{j=1}^{n_A} Y_j(T_i) g_j^*(\alpha, \boldsymbol{\beta}_0)} \Bigg].
\end{aligned} \tag{B.13}$$

By neglecting the terms which are $o_p({n_V}^{-0.5})$, we obtain from (B.11) that

$$\begin{aligned}\mathbb{V}\left[\bar{H}(\boldsymbol{\beta}_0)\right] &= \mathbb{V}\left[\mathbb{E}\left\{\bar{H}(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A, \mathbf{Z}_A\right\}\right] + \mathbb{E}\left[\mathbb{V}\left\{\bar{H}(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A, \mathbf{Z}_A\right\}\right] \\ &\simeq \mathbb{V}\left[\bar{H}_1(\boldsymbol{\beta}_0)\right] + \mathbb{E}\left[\bar{H}_2(\boldsymbol{\beta}_0)\mathbb{V}\left\{\frac{\hat{\alpha}-\alpha}{\alpha^2}\middle|\mathbf{X}_B, T_A, \mathbf{Z}_A\right\}\{\bar{H}_2(\boldsymbol{\beta}_0)\}^\top\right]. \qquad \text{(B.14)}\end{aligned}$$

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 $\mathbb{V}\{\hat{\alpha}|\mathbf{X}_B, T_A, \mathbf{Z}_A\}$ 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 (B.14) may be estimated by

$$\hat{\mathbb{V}}_1\left[\bar{H}(\boldsymbol{\beta}_0)\right] = \bar{H}_2(\hat{\alpha}, \hat{\boldsymbol{\beta}})\{\bar{H}_2(\hat{\alpha}, \hat{\boldsymbol{\beta}})\}^\top \times \left(\frac{1}{n_V} - \frac{1}{n_A}\right)\frac{n_V}{n_V-1}\frac{1-\hat{\alpha}}{\hat{\alpha}^3}, \qquad \text{(B.15)}$$

where $\bar{H}_2(\hat{\alpha},\hat{\boldsymbol{\beta}})$ is obtained from (B.13) 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.2.3 Accounting for the linkage and estimation error

In this section, we focus on the first term in the right-hand side of (B.14). We have

$$\mathbb{V}\left[\bar{H}_1(\boldsymbol{\beta}_0)\right] = \mathbb{V}\left[\mathbb{E}\left\{\bar{H}_1(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A\right\}\right] + \mathbb{E}\left[\mathbb{V}\left\{\bar{H}_1(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A\right\}\right]. \qquad \text{(B.16)}$$

It follows from equation (B.6) in Appendix B.1 that

$$\mathbb{E}\left\{\bar{H}_1(\boldsymbol{\beta}_0)\middle|\mathbf{X}_B, T_A\right\} \simeq \frac{1}{n_A}\sum_{i=1}^{n_A}\underbrace{\delta_i\left\{\mathbf{X}_i - \frac{\sum_{j=1}^{n_A}Y_j(T_i)h(\boldsymbol{\beta}_0, \mathbf{X}_j)}{\sum_{j=1}^{n_A}Y_j(T_i)g(\boldsymbol{\beta}_0, \mathbf{X}_j)}\right\}}_{H_{ti}(\boldsymbol{\beta}_0)}, \qquad \text{(B.17)}$$

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

$$\mathbb{V}\left\{\bar{H}_1(\boldsymbol{\beta}_0) \middle| \mathbf{X}_B, T_A\right\} \;\simeq\; \frac{1}{(n_A)^2} \sum_{i=1}^{n_A} \mathbb{V}\left\{H_{1i}(\boldsymbol{\beta}_0) \middle| \mathbf{X}_B, T_A\right\}. \tag{B.18}$$

where $H_{1i}(\cdot)$ is defined in (B.12). From (B.16), (B.17) and (B.18), we obtain that

$$\begin{aligned}
&\mathbb{V}\left[\bar{H}_1(\boldsymbol{\beta}_0)\right] \\
&\quad \simeq \mathbb{V}\left(\frac{1}{n_A} \sum_{i=1}^{n_A} H_{ti}(\boldsymbol{\beta}_0)\right) + \mathbb{E}\left[\frac{1}{(n_A)^2} \sum_{i=1}^{n_A} \mathbb{V}\left\{H_{1i}(\boldsymbol{\beta}_0) \middle| \mathbf{X}_B, T_A\right\}\right].
\end{aligned} \tag{B.19}$$

Now, we consider the sample dispersion term given by

$$\begin{aligned}
s_H^2(\boldsymbol{\beta}_0) &= \frac{1}{n_A - 1} \sum_{i=1}^{n_A} \left\{H_i(\boldsymbol{\beta}_0) - \frac{1}{n_A} \sum_{j=1}^{n_A} H_j(\boldsymbol{\beta}_0)\right\}^2 \\
&= \frac{1}{2n_A(n_A - 1)} \sum_{i \neq j = 1}^{n_A} \left\{H_i(\boldsymbol{\beta}_0) - H_j(\boldsymbol{\beta}_0)\right\}^2.
\end{aligned} \tag{B.20}$$

where $H_i(\cdot)$ is defined in (B.7). We have

$$\begin{aligned}
\mathbb{E}\left\{\frac{s_H^2(\boldsymbol{\beta}_0)}{n_A}\right\} &= \mathbb{E}\mathbb{E}\left\{\frac{s_H^2(\boldsymbol{\beta}_0)}{n_A} \middle| \mathbf{X}_B, T_A\right\} \\
&= \mathbb{E}\left[\frac{1}{2n_A^2(n_A - 1)} \sum_{i \neq j = 1}^{n_A} \mathbb{E}\{H_i(\boldsymbol{\beta}_0) - H_j(\boldsymbol{\beta}_0) \mid \mathbf{X}_B, T_A\}^2\right] \\
&\quad + \mathbb{E}\left[\frac{1}{2n_A^2(n_A - 1)} \sum_{i \neq j = 1}^{n_A} \mathbb{V}\{H_i(\boldsymbol{\beta}_0) - H_j(\boldsymbol{\beta}_0) \mid \mathbf{X}_B, T_A\}\right] \\
&\simeq \mathbb{E}\left[\frac{1}{2n_A^2(n_A - 1)} \sum_{i \neq j = 1}^{n_A} \{H_{ti}(\boldsymbol{\beta}_0) - H_{tj}(\boldsymbol{\beta}_0)\}^2\right] \\
&\quad \text{(where } H_{ti}(\cdot) \text{ is defined in (B.17))} \\
&\quad + \mathbb{E}\left[\frac{1}{2n_A^2(n_A - 1)} \sum_{i \neq j = 1}^{n_A} \mathbb{V}\{H_i(\boldsymbol{\beta}_0) \mid \mathbf{X}_B, T_A\} + \mathbb{V}\{H_j(\boldsymbol{\beta}_0) \mid \mathbf{X}_B, T_A\}\right]
\end{aligned} \tag{B.21}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{1}{n_A(n_A-1)}\sum_{i=1}^{n_A}\left\{H_{ti}(\boldsymbol{\beta}_0) - \frac{1}{n_A}\sum_{j=1}^{n_A}H_{tj}(\boldsymbol{\beta}_0)\right\}^2\right. \\
&\qquad \left. + \frac{1}{n_A^2}\sum_{i=1}^{n_A}\mathbb{V}\{H_i(\boldsymbol{\beta}_0)|\,\mathbf{X}_B, T_A\}\right] \\
&\simeq \mathbb{V}\left[\bar{H}_1(\boldsymbol{\beta}_0)\right],
\end{aligned}
$$

where the last line in (B.21) follows from a comparison with equation (B.19). Therefore, $\mathbb{V}\left[\bar{H}_1(\boldsymbol{\beta}_0)\right]$ may be approximately unbiasedly estimated by replacing in (B.20) the unknown parameter $\boldsymbol{\beta}_0$ with $\hat{\boldsymbol{\beta}}$, which leads to

$$
\hat{\mathbb{V}}_2\left[\bar{H}(\boldsymbol{\beta}_0)\right] = \frac{s_H^2(\hat{\boldsymbol{\beta}})}{n_A}. \tag{B.22}
$$

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

B.2.4 Global variance estimator

By plugging (B.15) and (B.22) into (B.14), 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 (B.9) as:

$$
\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}) = \{\nabla\bar{H}(\hat{\boldsymbol{\beta}})\}^{-1} \times \left\{\hat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\} + \hat{\mathbb{V}}_2\{\bar{H}(\boldsymbol{\beta}_0)\}\right\} \times \{\nabla\bar{H}(\hat{\boldsymbol{\beta}})\}^{-1} \tag{B.23}
$$

B.3 Additional simulation studies

We conduct several additional simulations to study the performance of the proposed method. In Section B.3.1, we consider scenarios where linkage errors are informative. In Section B.3.2, we study the affectation of non-random sample. A sensitivity analysis of $\hat{\alpha}$ is present in Section B.3.3.

For ease of interpretation, we assume there is only one block $V = 1$. Data are also generated by the scheme described in Section 4.3.1.

B.3.1 Informative linkage error

In this section, we study the proposed methods when linkage errors are informative of Cox model. We fix the sample size $n_A = 1000, n_B = 2000$ and let α vary in $\{0.75, 0.85, 0.95\}$ in the following simulations.

When α is dependent on X_1

In this case, we generate $\mathbf{Z}_i$ according to linkage model (4.4) as

$$\mathbf{Z}_i = \begin{cases} \mathbf{X}_i & \text{if } X_{i,1} \geq F^{-1}_{\mathcal{N}(0,1)}(1-\alpha), \\ \mathbf{X}_{(j)} & \text{otherwise,} \end{cases} \tag{B.24}$$

where $F^{-1}_{\mathcal{N}(0,1)}(\cdot)$ is the quantile function of the standard normal distribution. By this way, there is a dependency between the linkage and the covariate.

			$\hat{\beta}_1$			$\hat{\beta}_2$		
α	Methods	Fails	B_{MC}	Sd_{MC}	$\widehat{\text{Sd}}$	B_{MC}	Sd_{MC}	$\widehat{\text{Sd}}$
*	Theoretical	0	0.000	0.039	0.040	0.003	0.080	0.080
0.75	Naive	0	0.131	0.047	0.046	0.142	0.081	0.081
	Validation	0	0.021	0.195	0.190	0.033	0.308	0.291
	TAEE	0	0.095	0.088	0.085	0.001	0.119	0.128
	AEE	0	0.099	0.122	0.116	0.001	0.128	0.138
0.85	Naive	0	0.085	0.048	0.045	0.091	0.082	0.080
	Validation	0	0.023	0.167	0.166	0.016	0.281	0.276
	TAEE	0	0.033	0.066	0.067	0.004	0.102	0.113
	AEE	0	0.035	0.076	0.078	0.003	0.104	0.116
0.95	Naive	0	0.033	0.043	0.043	0.032	0.082	0.080
	Validation	0	0.016	0.145	0.144	0.012	0.267	0.264
	TAEE	0	0.002	0.048	0.054	0.001	0.088	0.101
	AEE	0	0.003	0.051	0.057	0.000	0.090	0.102

Table B.1: Summary of 1000 Monte-Carlo simulations when there is only 1 block with 2 covariates and α is dependent on X_1.

When α is dependent on $\widetilde{T}$

In this case, we generate $\mathbf{Z}_i$ according to linkage model (4.4) as

$$\mathbf{Z}_i = \begin{cases} \mathbf{X}_i & \text{if } \tilde{T}_i \geq F_{\widetilde{T}}^{-1}(1-\alpha), \\ \mathbf{X}_{(j)} & \text{otherwise,} \end{cases} \tag{B.25}$$

where $F_{\widetilde{T}}^{-1}(\cdot)$ is the quantile function of $\widetilde{T}_B$. By this way, there is a dependency between the linkage and the survival time.

			$\hat{\beta}_1$			$\hat{\beta}_2$		
α	Methods	Fails	B_{MC}	Sd_{MC}	$\widehat{\text{Sd}}$	B_{MC}	Sd_{MC}	$\widehat{\text{Sd}}$
*	Theoretical	0	0.000	0.039	0.040	0.003	0.080	0.080
0.75	Naive	0	0.146	0.038	0.040	0.140	0.078	0.082
	Validation	0	0.014	0.181	0.173	0.019	0.352	0.336
	TAEE	10	0.065	0.112	0.096	0.080	0.136	0.146
	AEE	56	0.079	0.177	0.239	0.089	0.197	0.285
0.85	Naive	0	0.103	0.039	0.040	0.097	0.078	0.081
	Validation	0	0.015	0.154	0.153	0.019	0.304	0.299
	TAEE	0	0.001	0.055	0.064	0.010	0.100	0.117
	AEE	0	0.006	0.065	0.077	0.015	0.108	0.125
0.95	Naive	0	0.040	0.039	0.040	0.036	0.080	0.081
	Validation	0	0.014	0.142	0.139	0.020	0.277	0.271
	TAEE	0	0.006	0.043	0.052	0.002	0.087	0.102
	AEE	0	0.006	0.045	0.055	0.002	0.088	0.103

Table B.2: Summary of 1000 Monte-Carlo simulations when there is only 1 block with 2 covariates and α is dependent on $\widetilde{T}$.

Simulation results in Table B.1 and Table B.2 indicate that there will be more bias in `TAEE` and `AEE` when α is dependent on Cox model, especially when α is small and dependent on $\tilde{T}$ (Table B.2). It reassures the necessity of the non-informative linkage error assumption. However, in case the assumption may not hold, the proposed methods are still better than the `Naive` method.

B.3.2 Sampling affectations

To compute $\bar{\mathbf{X}}_{B_v}$, $\bar{g}_{B_v}(\boldsymbol{\beta})$ and $\bar{h}_{B_v}(\boldsymbol{\beta})$ in (4.7), 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

$$\begin{aligned} \bar{\mathbf{X}}_{B_v} \quad &\text{with} \quad \bar{\mathbf{Z}}_{A_v} = \frac{1}{n_{A_v}} \sum_{i \in A_v} \mathbf{Z}_i, \\ \bar{g}_{B_v}(\boldsymbol{\beta}) \quad &\text{with} \quad \bar{g}_{A_v}(\boldsymbol{\beta}) = \frac{1}{n_{A_v}} \sum_{i \in A_v} \exp(\mathbf{Z}_i^\top \boldsymbol{\beta}), \\ \bar{h}_{B_v}(\boldsymbol{\beta}) \quad &\text{with} \quad \bar{h}_{A_v}(\boldsymbol{\beta}) = \frac{1}{n_{A_v}} \sum_{i \in A_v} \exp(\mathbf{Z}_i^\top \boldsymbol{\beta}) \mathbf{Z}_i. \end{aligned} \tag{B.26}$$

Under this situation, we perform two methods `TAEE-A` and `AEE-A` using the same equations as `TAEE` and `AEE` respectively except that (4.7) is replaced with (B.26). If A_v is a random sample of B_v, (B.26) can be a good approximation. In the following simulations, we consider different type of sampling of A_v. We first generate $n_B = 2000$ units in database B with $p = 2$ covariates, a continuous variable $X_1 \sim \mathcal{N}(0, 1)$ and a binary variable $X_2 \sim$ Bernoulli(0.7). The value of $\alpha = 0.85$ is fixed.

Suppose that each individual i in B_v can be selected with probability p_i following the logistic regression model as

$$p_i = \mathbb{P}(\mathbf{X}_i \text{ is chosen for } A_v) = \frac{\exp(\mathbf{X}_i^\top \boldsymbol{\omega})}{1 + \exp(\mathbf{X}_i^\top \boldsymbol{\omega})} \tag{B.27}$$

where $\boldsymbol{\omega}$ is a pre-defined vector of coefficients.

- Type 0: $p_i = 0.5$ for all $i = 1, \ldots, n_B$. A_v is random selected from B_v
- Type 1: $p_i = \dfrac{\exp(0.5X_1)}{1 + \exp(0.5X_1)}$
- Type 2: $p_i = \dfrac{\exp(0.4X_1 + 0.1X_2)}{1 + \exp(0.4X_1 + 0.1X_2)}$
- Type 3: $p_i = \dfrac{\exp(0.4Y_1 + 0.4Y_2)}{1 + \exp(0.4Y_1 + 0.4Y_2)}$ where $Y_1 \sim \mathcal{N}(-0.5, 1)$ and $Y_2 \sim$ Bernoulli(0.6) are independent of X_1 and X_2.

From the simulation results in Table B.3, we can see that when the sampling process is independent of model covariates (X_1, X_2) (Type 0 and 3), the `AEE-A` (or `TAEE-A`) gives the same performance as `AEE` (or `TAEE`).

		$\hat{\beta}_1$				$\hat{\beta}_2$			
Type	Methods	B_{MC}	Sd_{MC}	$\widehat{Sd}_{mpl}$	$\widehat{Sd}_{AEE}$	B_{MC}	Sd_{MC}	$\widehat{Sd}_{mpl}$	$\widehat{Sd}_{AEE}$
0	Theoretical	0.000	0.039	0.040		0.003	0.080	0.080	
	Naive	0.092	0.040	0.039		0.088	0.081	0.080	
	Validation	0.016	0.149	0.146		0.000	0.296	0.283	
	TAEE	0.002	0.055	0.041	0.059	0.007	0.103	0.080	0.113
	AEE	0.005	0.063	0.041	0.066	0.010	0.110	0.080	0.118
	TAEE-A	0.001	0.055	0.040	0.059	0.006	0.102	0.080	0.113
	AEE-A	0.004	0.062	0.041	0.066	0.010	0.108	0.080	0.118
1	Theoretical	0.001	0.040	0.040		0.003	0.081	0.078	
	Naive	0.092	0.040	0.040		0.087	0.080	0.079	
	Validation	0.013	0.151	0.145		0.013	0.286	0.277	
	TAEE	0.001	0.054	0.041	0.058	0.005	0.101	0.079	0.108
	AEE	0.002	0.060	0.041	0.065	0.006	0.104	0.079	0.112
	TAEE-A	0.006	0.056	0.041	0.058	0.012	0.102	0.079	0.108
	AEE-A	0.007	0.063	0.041	0.065	0.013	0.106	0.079	0.112
2	Theoretical	0.001	0.039	0.040		0.003	0.081	0.078	
	Naive	0.091	0.039	0.039		0.089	0.079	0.078	
	Validation	0.021	0.144	0.143		0.018	0.283	0.277	
	TAEE	0.002	0.053	0.040	0.057	0.004	0.100	0.078	0.108
	AEE	0.002	0.059	0.040	0.063	0.004	0.103	0.078	0.112
	TAEE-A	0.007	0.054	0.040	0.057	0.009	0.101	0.078	0.108
	AEE-A	0.007	0.061	0.040	0.063	0.009	0.105	0.078	0.112
3	Theoretical	0.002	0.037	0.040		0.003	0.081	0.079	
	Naive	0.090	0.038	0.039		0.091	0.079	0.080	
	Validation	0.019	0.156	0.144		0.019	0.288	0.281	
	TAEE	0.004	0.052	0.040	0.058	0.003	0.101	0.080	0.112
	AEE	0.008	0.060	0.040	0.066	0.007	0.106	0.080	0.117
	TAEE-A	0.004	0.051	0.040	0.058	0.003	0.101	0.080	0.112
	AEE-A	0.008	0.059	0.040	0.066	0.006	0.105	0.080	0.117

Table B.3: Simulation results with different sampling type

However, when sampling process depends on the model covariates (Type 1 and

2), the `AEE-A` (or `TAEE-A`) has larger bias and variability estimation than `AEE` (or `TAEE`). However, the different is very small.

B.3.3 Sensitivity analysis

In this section, we study the affectation of $\hat{\alpha}$ on the performance of the proposed methods. We consider a same above data generation with a fixed value of $\alpha = 0.85$ and different sample size $n_A \in \{500, 1000, 2000\}$. Then we examine following methods:

- `TAEE`: Using the AEE with true $\alpha = 0.85$
- `TAEE-1`: Using the AEE with $\alpha = \alpha_1 = 0.75$
- `TAEE-2`: Using the AEE with $\alpha = \alpha_2 = 0.8$
- `TAEE-3`: Using the AEE with $\alpha = \alpha_3 = 0.9$
- `TAEE-4`: Using the AEE with $\alpha = \alpha_4 = 0.95$

From the Table B.4, we can see that there will be more biased when α is poor specified, especially in `TAEE-1` and `TAEE-4` . With a moderate bias of α, `TAEE-2` and `TAEE-3` still have smaller bias than the Naive method. The simulation results also indicate that with a same level of bias in $\hat{\alpha}$, the larger estimate value of α is better for the proposed method. For example, `TAEE-4` is better than `TAEE-1`, `TAEE-3` is better than `TAEE-2`.

B.4 Linearly approximated estimating equation

We propose another version of the adjusted estimating equation. For each block v and $i \in A_v$, the expectation of the 2nd term inside the big parentheses of the naive estimating equation (4.5) can be approximately given as $\Delta_{i,v}$ below:

$$\begin{aligned}\Delta_{i,v} &:= \frac{\sum_v \sum_{j\in A_v} Y_j(T_i)\mathbb{E}(h(\boldsymbol{\beta}, \mathbf{Z}_j))}{\sum_v \sum_{j\in A_v} Y_j(T_i)\mathbb{E}(g(\boldsymbol{\beta}, \mathbf{Z}_j))} \\ &= \frac{\sum_v \sum_{j\in A_v} Y_j(T_i)\left[\alpha_v h(\boldsymbol{\beta}, \mathbf{X}_i) + (1-\alpha_v)\bar{h}_{B_v}(\boldsymbol{\beta})\right]}{\sum_v \sum_{j\in A_v} Y_j(T_i)\left[\alpha_v g(\boldsymbol{\beta}, \mathbf{X}_i) + (1-\alpha_v)\bar{g}_{B_v}(\boldsymbol{\beta})\right]}\end{aligned}$$

By linear expansion of $\Delta_{i,v}$ around $\alpha_v = 1$ for any v, we have

$$\Delta_{i,v} \approx \frac{\sum_v \sum_{j\in A_v} Y_j(T_i)h(\beta, \mathbf{X}_j)}{\sum_v \sum_{j\in A_v} Y_j(T_i)g(\beta, \mathbf{X}_j)} - \sum_{v=1}^{V} \nabla_{i,v}(1-\alpha_v), \tag{B.28}$$

			$\hat{\beta}_1$			$\hat{\beta}_2$		
n	Methods	Fails	B_{MC}	$\mathrm{Sd}_{\mathrm{MC}}$	$\widehat{\mathrm{Sd}}$	B_{MC}	$\mathrm{Sd}_{\mathrm{MC}}$	$\widehat{\mathrm{Sd}}$
500	Naive	0	0.090	0.055	0.054	0.091	0.110	0.111
	TAEE	0	0.009	0.078	0.085	0.010	0.145	0.161
	TAEE-1	21	0.124	0.146	0.117	0.122	0.214	0.192
	TAEE-2	1	0.058	0.094	0.099	0.059	0.164	0.175
	TAEE-3	0	0.029	0.069	0.078	0.028	0.132	0.152
	TAEE-4	0	0.061	0.062	0.072	0.060	0.122	0.144
1000	Naive	0	0.094	0.040	0.038	0.092	0.080	0.078
	TAEE	0	0.002	0.055	0.059	0.007	0.103	0.113
	TAEE-1	2	0.106	0.083	0.077	0.112	0.136	0.132
	TAEE-2	0	0.047	0.065	0.065	0.053	0.116	0.121
	TAEE-3	0	0.035	0.049	0.054	0.030	0.094	0.107
	TAEE-4	0	0.066	0.044	0.050	0.061	0.087	0.101
2000	Naive	0	0.094	0.028	0.027	0.095	0.054	0.055
	TAEE	0	0.001	0.039	0.041	0.000	0.071	0.080
	TAEE-1	0	0.102	0.057	0.051	0.100	0.092	0.091
	TAEE-2	0	0.045	0.046	0.045	0.044	0.079	0.085
	TAEE-3	0	0.035	0.035	0.038	0.036	0.065	0.075
	TAEE-4	0	0.065	0.031	0.035	0.066	0.060	0.071

Table B.4: Sensitivity analysis of $\hat{\alpha}$

where $\nabla_{i,v}$ is the partial derivative $\partial\Delta_{i,v}/\partial\alpha_v$ evaluated at $\alpha_v = 1$ and

$$\frac{\partial\Delta_{i,v}}{\partial\alpha_v} = \frac{\sum_{j\in A_v} Y_j(T_i)\left[h(\boldsymbol{\beta},\mathbf{X}_j) - \bar{h}_{B_v}(\boldsymbol{\beta})\right]\sum_{j\in A_v} Y_j(T_i)\mathbb{E}(g(\boldsymbol{\beta},\mathbf{Z}_j))}{\left[\sum_{j\in A_v} Y_j(T_i)\mathbb{E}(g(\boldsymbol{\beta},\mathbf{Z}_j))\right]^2} - \frac{\sum_{j\in A_v} Y_j(T_i)\mathbb{E}(h(\boldsymbol{\beta},\mathbf{Z}_j))\sum_{j\in A_v} Y_j(T_i)\left[g(\boldsymbol{\beta},\mathbf{X}_j) - \bar{g}_{B_v}(\boldsymbol{\beta})\right]}{\left[\sum_{j\in A_v} Y_j(T_i)\mathbb{E}(g(\boldsymbol{\beta},\mathbf{Z}_j))\right]^2}$$

When evaluated at $\alpha_v = 1$, we have $\mathbb{E}(g(\boldsymbol{\beta}, \mathbf{Z}_j)) = g(\boldsymbol{\beta}, \mathbf{X}_j) = g(\boldsymbol{\beta}, \mathbf{Z}_j)$ and $\mathbb{E}(h(\boldsymbol{\beta}, \mathbf{Z}_j)) = h(\boldsymbol{\beta}, \mathbf{X}_j) = h(\boldsymbol{\beta}, \mathbf{Z}_j)$, as well as $\mathbf{X}_i = \mathbf{Z}_i$. Thus, we obtain

$$\begin{aligned}
\nabla_{i,v} = & \frac{\sum_{j\in A_v} Y_j(T_i)\left[h(\boldsymbol{\beta}, \mathbf{X}_j) - \bar{h}_{B_v}(\boldsymbol{\beta})\right] \sum_{j\in A_v} Y_j(T_i) g(\boldsymbol{\beta}, \mathbf{Z}_j)}{\left[\sum_{j\in A_v} Y_j(T_i) g(\boldsymbol{\beta}, \mathbf{Z}_j)\right]^2} \\
& - \frac{\sum_{j\in A_v} Y_j(T_i) h(\boldsymbol{\beta}, \mathbf{Z}_j) \sum_{j\in A_v} Y_j(T_i)\left[g(\boldsymbol{\beta}, \mathbf{X}_j) - \bar{g}_{B_v}(\boldsymbol{\beta})\right]}{\left[\sum_{j\in A_v} Y_j(T_i) g(\boldsymbol{\beta}, \mathbf{Z}_j)\right]^2} \\
= & \frac{\sum_{j\in A_v} Y_j(T_i)\left[h(\boldsymbol{\beta}, \mathbf{X}_j) - \bar{h}_{B_v}(\boldsymbol{\beta})\right]}{\sum_{j\in A_v} Y_j(T_i) g(\boldsymbol{\beta}, \mathbf{Z}_j)} \\
& - \frac{\sum_{j\in A_v} Y_j(T_i) h(\boldsymbol{\beta}, \mathbf{Z}_j)}{\sum_{j\in A_v} Y_j(T_i) g(\boldsymbol{\beta}, \mathbf{Z}_j)} \frac{\sum_{j\in A_v} Y_j(T_i)\left[g(\boldsymbol{\beta}, \mathbf{X}_j) - \bar{g}_{B_v}(\boldsymbol{\beta})\right]}{\sum_{j\in A_v} Y_j(T_i) g(\boldsymbol{\beta}, \mathbf{Z}_j)}.
\end{aligned}$$

Hence, we obtain

$$\frac{\sum_v \sum_{j\in A_v} Y_j(T_i) h(\boldsymbol{\beta}, \mathbf{X}_j)}{\sum_v \sum_{j\in A_v} Y_j(T_i) g(\boldsymbol{\beta}, \mathbf{X}_j)} \approx \frac{\sum_v \sum_{j\in A_v} Y_j(T_i) h(\boldsymbol{\beta}, \mathbf{Z}_j)}{\sum_v \sum_{j\in A_v} Y_j(T_i) g(\boldsymbol{\beta}, \mathbf{Z}_j)} + \sum_{v=1}^{V} \nabla_{i,v}(1 - \alpha_v).$$

Writing $\frac{\mathbf{Z}_i}{\alpha_v} - (\frac{1}{\alpha_v} - 1)\bar{\mathbf{X}}_{B_v} = \mathbf{Z}_i + (\frac{1}{\alpha_v} - 1)(\mathbf{Z}_i - \bar{\mathbf{X}}_q)$, we obtain the *linearly approximated estimating equation (LAEE)*

$$\sum_{v=1}^{V} \sum_{i\in A_v} \delta_i \left[\mathbf{Z}_i - \frac{\sum_{j\in A} Y_j(T_i) h(\boldsymbol{\beta}, \mathbf{Z}_j)}{\sum_{j\in A_v} Y_j g(\beta, \mathbf{Z}_j)} + W_i(\alpha_v, \boldsymbol{\beta})\right] = 0 \tag{B.29}$$

where $W_i(\alpha_v, \boldsymbol{\beta})$ an adjustment of the naive estimating equation for $i \in A_v$, given by

$$W_i(\alpha_v, \boldsymbol{\beta}) = \left(\frac{1}{\alpha_v} - 1\right)(\mathbf{Z}_i - \bar{\mathbf{X}}_{B_v}) - \sum_{v=1}^{V} \nabla_{i,v}(1 - \alpha_v).$$

B.4.1 Simulation results

In the simulation, we consider the following methods

- Theoretical: Using the theoretical estimating equation (4.2) with true values of covariates $\mathbf{X}$.
- Naive: Using the naive estimating equation (4.5) with linked data.
- AEE: Uisng the AEE with estimated value of α and when covariates $\mathbf{X}$ in database B are avaialble.
- LAEE: Using the LAEE with estimated value of α and when covariates $\mathbf{X}$ in database B are avaialble.

				$\hat{\beta}_1$		$\hat{\beta}_2$	
n	α	Methods	Fails	B_{MC}	Sd_{MC}	B_{MC}	Sd_{MC}
500	*	Theoretical	0	0.003	0.055	0.005	0.109
	0.75	Naive	0	0.144	0.052	0.147	0.116
		AEE	24	0.033	0.168	0.024	0.219
		LAEE	1	0.015	0.088	0.020	0.168
	0.85	Naive	0	0.090	0.055	0.091	0.110
		AEE	5	0.015	0.094	0.015	0.159
		LAEE	0	0.002	0.078	0.004	0.141
	0.95	Naive	0	0.029	0.055	0.028	0.110
		AEE	0	0.007	0.066	0.009	0.124
		LAEE	0	0.005	0.064	0.006	0.122
1000	*	Theoretical	0	0.000	0.039	0.002	0.078
	0.75	Naive	0	0.149	0.040	0.146	0.079
		AEE	6	0.011	0.085	0.017	0.133
		LAEE	0	0.026	0.065	0.023	0.113
	0.85	Naive	0	0.094	0.040	0.092	0.080
		AEE	0	0.006	0.065	0.011	0.111
		LAEE	0	0.009	0.057	0.006	0.103
	0.95	Naive	0	0.034	0.040	0.031	0.081
		AEE	0	0.001	0.048	0.004	0.089
		LAEE	0	0.001	0.047	0.002	0.088
2000	*	Theoretical	0	0.000	0.028	0.000	0.054
	0.75	Naive	0	0.149	0.028	0.147	0.055
		AEE	0	0.007	0.058	0.011	0.092
		LAEE	0	0.027	0.045	0.026	0.079
	0.85	Naive	0	0.094	0.028	0.095	0.054
		AEE	0	0.002	0.044	0.002	0.074
		LAEE	0	0.011	0.040	0.012	0.069
	0.95	Naive	0	0.033	0.028	0.033	0.055
		AEE	0	0.001	0.033	0.001	0.061
		LAEE	0	0.000	0.033	0.000	0.060

Table B.5: Simulation studies for comparing the linear approximated estimating equation to the adjusted estimating equation and the naive estimating equation.

From the simulation results in Table B.5, we can see that LAEE is "easier" to solve than AEE (only 1 fail when $n = 500$ and $\alpha = 0.75$ with LAEE). In general, the estimated parameters from LAEE have smaller variance than AEE. When α is close to 1 (e.g. 0.95), the LAEE and AEE have the same bias. When α is far from 1 (e.g. 0.75 and 0.85), LAEE has larger bias than AEE due to the approximation (B.28). However, with small sample size ($n = 500$), LAEE is better than AEE in both bias and standard deviation with any $\alpha \in \{0.75, 0.85, 0.95\}$.

Bibliography

P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982. ISSN 00905364.

Tigran Avoundjian, Julia C Dombrowski, Matthew R Golden, James P Hughes, Brandon L Guthrie, Janet Baseman, and Mauricio Sadinle. Comparing methods for record linkage for public health action: Matching algorithm validation study. *JMIR Public Health Surveill*, 6(2):e15917, Apr 2020. ISSN 2369-2960. doi: 10.2196/15917.

Ileana Baldi, Antonio Ponti, Roberto Zanetti, Giovannino Ciccone, Franco Merletti, and Dario Gregori. The impact of record inkage bias in the cox model. *Journal of evaluation in clinical practice*, 16:92–6, 02 2010.

T. R. Belin and D. B. Rubin. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90(430):694–707, 1995.

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11): 1713–1723, 2005.

Julien Bezin, Mai Duong, Régis Lassalle, Cécile Droz, Antoine Pariente, Patrick Blin, and Nicholas Moore. The national healthcare system claims databases in france, sniiram and egb: Powerful tools for pharmacoepidemiology. *Pharmacoepidemiology and Drug Safety*, 26(8):954–962, 2017.

Mikhail Bilenko and Raymond Mooney. Adaptive duplicate detection using learnable string similarity measures. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 07 2003. doi: 10.1145/956750.956759.

Olivier Binette and Rebecca C. Steorts. (almost) all of entity resolution. *Science advances*, 8 12:eabi8021, 2022.

Tony Blakely and Clare Salmond. Probabilistic record linkage and a method to calculate the positive predictive value. *International journal of epidemiology*, 31 6:1246–52, 2002.

Ray Chambers. Regression analysis of probability-linked data. *Statistics New Zealand*, 2009.

Ray Chambers and Andrea Diniz da Silva. Improved secondary analysis of linked data: a framework and an illustration. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183, 06 2020. doi: 10.1111/rssa.12477.

Ray Chambers, Nicola Salvati, Enrico Fabrizi, and Andrea Diniz da Silva. Domain estimation under informative linkage. *Statistical Theory and Related Fields*, 3 (2):90–102, 2019. doi: 10.1080/24754269.2019.1653158.

Raymond Chambers and Gunky Kim. *Secondary analysis of linked data*, chapter 5, pages 83–108. John Wiley & Sons, Ltd, 2015.

Raymond L. Chambers, Enrico Fabrizi, Maria Giovanna Ranalli, Nicola Salvati, and Suojin Wang. Robust regression using probabilistically linked data. *WIREs Computational Statistics*, n/a(n/a):e1596, 2022. doi: https://doi.org/10.1002/wics.1596.

P. Christen and Karl Goiser. Quality and complexity measures for data linkage and deduplication. In *Quality Measures in Data Mining*, 2007.

Peter Christen. Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 151–159, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934.

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

Peter Christen and William E. Winkler. *Record Linkage*, pages 1066–1075. Springer US, Boston, MA, 2017.

Munir Cochinwala, Verghese Kurien, Gail Lalk, and Dennis Shasha. Efficient data reconciliation. *Inf. Sci.*, 137:1–15, 2001.

J. B. Copas and F. J. Hilton. Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3):287–320, 1990. ISSN 09641998, 1467985X.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 00359246.

John G. Cragg. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844, 1971. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1909582.

Joanne Daggy, Huiping Xu, Siu Hui, and Shaun Grannis. Evaluating latent class models with conditional dependence in record linkage. *Statistics in Medicine*, 33(24):4250–4265, 2014.

A. Delluc, C. Tromeur, F. Ven, M. Gouillou, N. Paleiron, L. Bressollette, M. Nonent, P.Y. Salaun, K. Lacut, G. Leroyer, C.and Le Gal, F. Couturaud, D. Mottier, and EPIGETBO study group. Current incidence of venous thromboembolism and comparison with 1998: a community-based study in western france. *Thromb Haemost*, 116:967–974, 2016.

Arthur Dempster, Natalie Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 01 1977.

J. E. Dennis and Robert B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics, 1996. doi: 10.1137/1.9781611971200. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611971200.

H. L. Dunn. Record linkage. *American journal of public health and the nation's health*, 36(12):1412–1416, 1946.

Stacie B Dusetzina, Seth Tyree, Anne-Marie Meyer, Adrian Meyer, Laura Green, and William R Carpenter. Linking data for health services research: A framework and instructional guide, 2014.

Ted Enamorado. Active learning for probabilistic record linkage. *Social Science Research Network (SSRN)*, 2018.

Ted Enamorado, Benjamin Fifield, and Kosuke Imai. Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113:353–371, 2019.

Ivan Fellegi and Alan Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 12 1969.

Marco Fortini. An improved fellegi-sunter framework for probabilistic record linkage between large data sets. *Journal of Official Statistics*, 36(4):803–825, 2020.

H. Goldstein, K. Harron, and M. Cortina-Borja. A scaling approach to record linkage. *Statistics in Medicine*, 36:2514–2521, 2017.

Harvey Goldstein, Katie Harron, and Angie Wade. The analysis of record-linked data using multiple imputation with data value priors. *Statistics in Medicine*, 31(28):3481–3493, 2012.

Shaun J. Grannis, J. Marc Overhage, and Clement J. McDonald. Analysis of identifier performance using a deterministic linkage algorithm. *Proceedings. AMIA Symposium*, pages 305–9, 2002.

Shaun J. Grannis, J. Marc Overhage, Siu L. Hui, and Clement J. McDonald. Analysis of a probabilistic record linkage technique without human review. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 259–63, 2003.

Ying Han and Partha Lahiri. Statistical analysis with linked data. *International Statistical Review*, 87(S1):S139–S157, 2019.

Katie Harron, Harvey Goldstein, and Chris Dibben. *Introduction*, chapter 1, pages 1–7. John Wiley & Sons, Ltd, 2015. ISBN 9781119072454. doi: https://doi.org/10.1002/9781119072454.ch1.

Katie Harron, Ruth Gilbert, David Cromwell, and Jan van der Meulen. Linking data for mothers and babies in de-identified electronic health data. *PLOS ONE*, 11(10):1–18, 10 2016. doi: 10.1371/journal.pone.0164667.

B. Hejblum, G. Weber, K. Liao, N. Palmer, S. Churchill, N. Shadick, P. Szolovits, S. Murphy, I. Kohane, and T. Cai. Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. *Scientific Data*, 6, 2019.

Thomas Herzog, Fritz Scheuren, and William Winkler. *Data Quality and Record Linkage Techniques*. Springer-Verlag New York, 2007.

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

Michel H. Hof, Anita C. Ravelli, and Aeilko H. Zwinderman. A probabilistic record linkage model for survival data. *Journal of the American Statistical Association*, 112(520):1504–1515, 2017.

Marius Hofert and Martin Mächler. Parallel and other simulations in r made easy: An end-to-end study. *Journal of Statistical Software*, 69(4):1–44, 2016. doi: 10.18637/jss.v069.i04.

Chengcheng Hu and D. Y. Lin. Cox regression with covariate measurement error. *Scandinavian Journal of Statistics*, 29(4):637–655, 2002. ISSN 03036898, 14679469.

Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.

Lee K Taylor, Katie Irvine, Renee Iannotti, Taylor Harchak, and Kim Lim. Optimal strategy for linkage of datasets containing a statistical linkage key and datasets with full personal identifiers. *BMC medical informatics and decision making*, 14:85, 09 2014.

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

Gunky Kim and Ray Chambers. Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica*, 66, 02 2012b. doi: 10.1111/j.1467-9574.2011.00509.x.

Gunky Kim and Raymond Chambers. Regression analysis under incomplete linkage. *Computational Statistics & Data Analysis*, 56(9):2756 – 2770, 2012c.

P. Lahiri and Michael D. Larsen. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230, 2005.

Michael D. Larsen and Alan Herning. Record linkage for establishments: Background, challenges, and an example. In Ger Snijkers, Stefan Bender Mojca Bavdaž, Jacqui Jones, Steve MacFeely, Joseph W. Sakshaug, Katherine Jenny Thompson, and Arnout van Delden, editors, *Advances in Business Statistics, Methods and Data Collection*, chapter 33, pages 743–765. John Wiley & Sons, Inc., 2023.

Michael D Larsen and Donald B Rubin. Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453):32–41, 2001.

D. Lee, L.-C. Zhang, and J.K. Kim. Maximum entropy classification for record linkage, 2020.

P. Li, X. Dong, A. Maurino, and D. Srivastava. Linking temporal records. In *Proceedings of the VLDB Endowment*, volume 4, pages 956–967, 2011.

C L Loprinzi, J A Laurie, H S Wieand, J E Krook, P J Novotny, J W Kugler, J Bartel, M Law, M Bateman, and N E Klatt. Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group. *Journal of Clinical Oncology*, 12(3):601–607, 1994. doi: 10.1200/JCO.1994.12.3.601. PMID: 8120560.

Abdullah-Al Mamun, Robert Aseltine, and Sanguthevar Rajasekaran. Efficient record linkage algorithms using complete linkage clustering. *PloS one*, 11: e0154446, 04 2016.

Jialin Mao, Caryn D Etkin, David G Lewallen, and Art Sedrakyan. Creation and validation of linkage between orthopedic registry and administrative data using indirect identifiers. *The Journal of Arthroplasty*, 02 2019.

Geoffrey McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, New York, 1996.

X.L. Meng and D. Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80:267–278, 1993.

John Neter, E. Scott Maynes, and R. Ramanathan. The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60(312):1005–1027, 1965. doi: 10.1080/01621459.1965.10480846.

S. Noboa, D. Mottier, E. Oger, and THE EPI-GETBO STUDY GROUP. Estimation of a potentially preventable fraction of venous thromboembolism: a community-based prospective study. *Journal of Thrombosis and Haemostasis*, 4(12):2720–2722, 2006.

Toan C. Ong, Michael V. Mannino, Lisa M. Schilling, and Michael G. Kahn. Improving record linkage performance in the presence of missing linkage data. *Journal of Biomedical Informatics*, 52:43–54, 2014. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2014.01.016. Special Section: Methods in Clinical Research Informatics.

Shivani Padmanabhan, Lucy Carty, Ellen Cameron, Rebecca Elisabeth Ghosh, Rachael Williams, and Helen Strongman. 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*, 34:91 – 99, 2018.

Banda Ramadan, Peter Christen, Huizhi Liang, and Ross W. Gayler. Dynamic sorted neighborhood indexing for real-time entity resolution. *J. Data and Information Quality*, 6(4), oct 2015. ISSN 1936-1955. doi: 10.1145/2816821. URL https://doi.org/10.1145/2816821.

Mauricio Sadinle. Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518):600–612, 2017.

A. Sayers, Y. Ben-Shlomo, A. Blom, and F. Steele. Probabilistic record linkage. *International journal of epidemiology*, 45:954–964, 2015.

Fritz Scheuren and William Winkler. Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 1993. URL https://www150.statcan.gc.ca/n1/pub/12-001-x/1993001/article/14476-eng.pdf.

Fritz Scheuren and William Winkler. Regression analysis of data files that are computer matched - part ii. *Survey Methodology*, 23, 01 1997. URL https://www150.statcan.gc.ca/n1/pub/12-001-x/1997002/article/3613-eng.pdf.

Josef Schürle. A method for consideration of conditional dependencies in the fellegi and sunter model of record linkage. *Statistical Papers*, 46:433–449, 07 2005.

Martin Slawski, Guoqing Diao, and Emanuel Ben-David. A pseudo-likelihood approach to linear regression with partially shuffled data. *Journal of Computational and Graphical Statistics*, 30(4):991–1003, 2021. doi: 10.1080/10618600.2020.1870482.

Chakkrit Snae. A comparison and analysis of name matching algorithms. *Proceedings of World Academy of Science, Engineering and Technology*, 21, 01 2007.

Rebecca C. Steorts, Rob Hall, and Stephen E. Fienberg. A bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672, 2016.

Andrea Tancredi and Brunero Liseo. A hierarchical bayesian approach to matching and size population problems. *Annals of Applied Statistics - ANN APPL STAT*, 5, 11 2010.

Andrea Tancredi and Brunero Liseo. A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2B): 1553 – 1585, 2011.

Andrea Tancredi and Brunero Liseo. Regression analysis with linked data: problems and possible solutions. *Statistica*, 75(1):19–35, 2015.

P. Tuppin, J. Rudant, P. Constantinou, C. Gastaldi-Ménager, A. Rachas, L. Roquefeuil, G. Maura, H. Caillol, A. Tajahmady, J. Coste, C. Gissot, A. Weill, and A. Fagot-Campagna. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'assurance maladie (sniiram) to the système national des données de santé (snds) in france. *Revue d'Épidémiologie et de Santé Publique*, 65 Suppl 4:146–167, 2017a.

Philippe Tuppin, Laurence Pestel, Solène Samson, Anne Cuerq, Sébastien Rivière, Stéphane Tala, Pierre Denis, Jérôme Drouin, Claude Gissot, Christelle Gastaldi-Ménager, and Anne Fagot-Campagna. Poids humain et économique des cancers en france en 2014, les données du sniiram. *Bulletin du Cancer*, 104(6):524 – 537, 2017b. ISSN 0007-4551.

Thanh Huan Vo, Guillaume Chauvet, André Happe, Emmanuel Oger, Stephane Paquelet, and Valérie Garès. Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system. *Computational Statistics & Data Analysis*, 2022.

William E. Winkler. 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*, pages 667–671, 1988.

William E. Winkler. Frequency-based matching in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 778–783, 1989.

William E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, 01 1990.

William E. Winkler and Yves Thibaudeau. 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.

C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 03 1983.

Huiping Xu, Xiaochun Li, Changyu Shen, Siu Hui, and Shaun Grannis. Incorporating conditional dependence in latent class models for probabilistic record

linkage: Does it matter? *The Annals of Applied Statistics*, 13:1753–1790, 09 2019.

William Yancey. Evaluating string comparator performance for record linkage. *Technical Report Statistical Research Report Series RRS2005/05, Washington, DC: US Bureau of the Census*, 01 2005.

Derek Young, Xi Chen, Dilrukshi Hewage, and Ricardo Nilo Poyanco. Finite mixture-of-gamma distributions: estimation, inference, and model-based clustering. *Advances in Data Analysis and Classification*, 13, 05 2019.

Guangyu Zhang and Paul Campbell. Data survey: Developing the statistical longitudinal census dataset and identifying its potential uses. *Australian Economic Review*, 45(1):125 – 133, 3 2012. doi: 10.1111/j.1467-8462.2011.00673.x.

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

Li-Chun Zhang. On secondary analysis of datasets that cannot be linked without errors. In Li-Chun Zhang and Ray Chambers, editors, *Analysis of integrated data*. London: CRC/Chapman and Hall, 2019.

Li-Chun Zhang and Tiziana Tuoto. Linkage-data linear regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):522–547, 2021.

Vivienne J. Zhu, Marc J. Overhage, James Egg, Stephen M. Downs, and Shaun J. Grannis. An Empiric Modification to the Probabilistic Record Linkage Algorithm Using Frequency-Based Weight Scaling. *Journal of the American Medical Informatics Association*, 16(5):738–745, 09 2009.

Ying Zhu, Yutaka Matsuyama, Yasuo Ohashi, and Soko Setoguchi. When to conduct probabilistic linkage vs. deterministic linkage? a simulation study. *Journal of Biomedical Informatics*, 56:80 – 86, 2015.

Titre: Couplage d'enregistrements et analyse des données couplées avec application dans le système national des données de santé français.

Mots clés: couplage d'enregistrements, analyse secondaire, régression de Cox, données de santé

Résumé: Cette thèse a deux contributions principales. Nous considérons le modèle d'appariement probabiliste de Fellegi et Sunter, et nous l'étendons à des données de type mixte. L'appariement probabiliste consiste à combiner des données de différentes sources, quand elles correspondent à une même entité mais qu'une variable d'identification n'est pas disponible. Le modèle de Fellegi et Sunter utilise des variables partiellement identifiantes, mais se limite à une comparaison binaire pour ces variables. Dans la première contribution, nous proposons une extension du modèle pour les vecteurs de comparaison de type mixte. Nous développons un modèle de mélange pour comparer les valeurs des variables d'appariement catégorielles présentant des prévalences faibles, et un mélange de distributions "hurdle gamma" pour les valeurs des variables d'appariement continues. Nous appliquons ce modèle pour apparier les données du SNDS avec un registre de patients de l'aire urbaine de Brest, souffrant de thromboembolie veineuse. Dans le second travail, nous proposons un modèle pour une régression de Cox avec des données appariées. Des erreurs d'appariement sont presque inévitables quelle que soit la méthode utilisée, et ignorer ces erreurs peut conduire à des estimations biaisées. Nous proposons une équation estimante ajustée adaptée au modèle de Cox, quand l'appariement a été réalisé par un opérateur tiers et que l'analyste ne connaît pas les variables d'appariement. Nous proposons un estimateur de variance asymptotiquement sans biais pour l'estimateur des paramètres du modèle de Cox. Le modèle proposé est appliqué à une base de données appariées, correspondant à des AVC survenus à Brest.

Title: Record linkage and analysis of linked data with application in French national health data system.

Keywords: record linkage, secondary analysis, Cox regression, health data

Abstract: This thesis has two main contributions. Firstly, we extend the Fellegi-Sunter probabilistic record linkage model for mixed-type data. Probabilistic record linkage is a process of combining data from different sources, when such data refer to common entities and identifying information is not available. Fellegi and Sunter proposed a probabilistic record linkage framework that takes into account multiple non-identifying information, but is limited to simple binary comparison between matching variables. In the first contribution, we propose an extension of this model for mixed-type comparison vectors. We develop a mixture model for handling comparison values of low prevalence categorical matching variables, and a mixture of hurdle gamma distribution for handling comparison values of continuous matching variables. The proposed model is applied to perform linkage between a registry of patients suffering from venous thromboembolism in the Brest and the French national health data system. Secondly, we propose a model for Cox regression with linked data. The linked data can bring analysts novel and valuable knowledge which is unable to obtain from a single database. However, linkage errors are usually unavoidable regardless of record linkage methods and ignoring these errors may lead to bias estimates. In this work, we propose an adjusted estimating equation for secondary Cox regression analysis, where linked data have been prepared by someone else and no information on matching variables are available to the analyst. An asymptotically unbiased variance estimator is also proposed. The proposed model is applied to a linked database from the Brest stroke registry.