Simultaneous Causal Inference and Probabilistic Record Linkage in Observational Studies with Covariates Spread Over Two Files

Sharmistha Guha and Jerome P. Reiter

November 23, 2021

Abstract

We consider observational studies with data spread over two files. One file includes the treatment, outcome, and some covariates measured on a set of individuals, and the other file includes additional covariates measured on a partially intersecting set of individuals. In absence of direct identifiers, researchers typically estimate causal effects in two stages: construct a linked database with probabilistic record linkage, then apply causal estimators on the linked data. This approach does not take advantage of relationships among the variables to improve the linkage quality. It also does not propagate uncertainty from imperfect linkages to the causal inferences. We address these shortcomings via a Bayesian joint modeling framework for simultaneous causal inference and probabilistic record linkage. The Markov chain Monte Carlo sampler generates multiple plausible linked data files as byproducts. We use these datasets for multiple imputation inferences with two causal estimators, one regression-adjusted and the other unadjusted, based on propensity score overlap weights. Using simulations and data from the Italian Survey on Household Income and Wealth, we show that the joint model with both estimators can improve the accuracy of estimated treatment effects compared to analogous two stage procedures. Supplementary material contains additional details about the causal estimators and additional simulation results.

Keywords: Matching; Treatment; Observational; Fusion; Propensity score.

1 Introduction

In many settings, researchers may be able to enhance the validity of causal inferences by using covariate information that is available across two databases. For example, in a causal study of a health intervention, a researcher with access to study subjects' health records may seek to account for additional causally-relevant covariates by linking subjects to their records in educational or financial databases. Similarly, in a causal study of a policy intervention, a researcher may seek to link study subjects from some survey to their records in administrative databases. These examples illustrate the scenario of interest in this article: one file contains the outcome variable, the treatment status and some causally-relevant covariates for a set of study subjects, and a different file contains additional causally-relevant covariates on some subset of the study subjects and other individuals.

When perfectly measured unique identifiers like social security numbers or patient IDs are available in both files, it is reasonably straightforward to link individuals across the files. However, often researchers do not have access to such direct identifiers. They may be missing from one or both files, or they may not be available due to privacy restrictions. In such situations, researchers have to link the files based on indirect identifiers, such as names, birth dates and address information. To do so, many researchers turn to probabilistic record linkage methods, often based on variants of the framework developed by Fellegi and Sunter (1969), which we review in Section 2.2.

Typically, researchers perform causal inference with linked files in a two-stage process, i.e., probabilistic record linkage is used to construct a single file comprising linked records, and then causal inference carried out on the linked file. This two-stage approach has two main drawbacks. First, the record linkage step does not take advantage of relationships among the variables in the two files. Several authors (e.g., Gutman *et al.*, 2013; Dalzell and Reiter, 2018; Steorts *et al.*, 2018; Tang *et al.*, 2020) have shown that leveraging these relationships in fact can improve the quality of the linkages. Second, the two-stage framework does not

propagate uncertainty arising from imperfect linkages to the causal inferences.

In this article, we address these shortcomings with a Bayesian joint modeling framework to perform simultaneous causal inference and probabilistic record linkage. To fix ideas, let File B contain the outcome variable, treatment status and some causally-relevant covariates on a set of individuals. Let File A contain an additional set of causally-relevant covariates measured on a different set of individuals, some of whom are in File B and some of whom are not. We specify models for (i) the conditional distribution of the outcome variable given the treatment status and all covariates, which we refer to as the *outcome model*, (ii) the conditional distribution of the treatment status given all covariates, which we refer to as the *propensity score model*, and (iii) the conditional distribution of the covariates in File B given the covariates in File A, which we refer to as the *covariate model*. We couple these with a probabilistic model for the unknown linkage statuses, i.e., which record pairs are links and which are not. We estimate the model using a Markov chain Monte Carlo (MCMC) sampler, which results in many draws of plausibly linked data files. In each plausibly linked dataset, we estimate the treatment effect using some causal estimator and combine the results using multiple imputation (Rubin, 1987). For the sake of illustrating this joint modeling approach, we estimate a weighted average treatment effect (WATE, Hirano *et al.*, 2003) using the propensity score overlap weights of Li *et al.* (2018). These have appealing features for causal inference, which we summarize in Section 2.1. We note that analysts could replace the overlap weights estimators with any other causal estimator.

Our work contributes to existing methods for simultaneous record linkage and statistical inference (e.g., Scheuren and Winkler, 1993; Lahiri and Larsen, 2005; Chipperfield *et al.*, 2011; Tancredi and Liseo, 2011; Kim and Chambers, 2012; Gutman *et al.*, 2013; Ventura and Nugent, 2014; Dalzell and Reiter, 2018; Sadinle, 2018; Solomon, 2019; Tancredi *et al.*, 2020; Tang *et al.*, 2020), though none of these works consider causal inference as the analysis goal. Heck Wortman and Reiter (2018) present a version of simultaneous causal inference and record linkage. They use point estimates of average causal effects from propensity score stratification to determine the thresholds at which record pairs are declared links in a Fellegi

and Sunter (1969) algorithm. They do not use relationships among the variables to determine the record pairs to consider as possible links in the first place, which our approach does. Guha *et al.* (2020) propose a model for Bayesian causal inference and record linkage when the treatment and all covariates reside in one file and the outcome in another. This different setting demands different model specification tasks; for example, one need not include the propensity score model nor the covariate model as components of a joint model. Additionally, their framework relies on a fully Bayesian approach to causal inference, estimating an average treatment effect by imputing counterfactual outcomes from the outcome model. Thus, both the causal inference and record linkage quality are highly dependent on the quality of the fit of the outcome model. In contrast, we apply causal estimators based on balancing scores like the overlap weights, which reduces sensitivity to the fit of the outcome model.

The remainder of this article is organized as follows. In Section 2, we review the causal inference and probabilistic record linkage procedures that form the basis of the methodology. In Section 3, we present the joint model for simultaneous causal inference and probabilistic record linkage. Here, we also describe the regression-adjusted estimator exploiting overlap weights, which we believe itself has not appeared previously in the literature. In Section 4, we present results of simulation studies comparing the joint model to two-stage approaches. In Section 5, we illustrate the methodology using partially simulated data based on an Italian household survey to assess the effect of debit card possession on household spending. Both sets of simulation results demonstrate the potential of the joint model to improve on the two-stage approaches in terms of both record linkage quality and causal inference accuracy. Finally, in Section 6, we conclude with a discussion.

2 Background on Causal Inference and Record Linkage

We first define a few key concepts and assumptions related to the causal inference procedures in Section 2.1. We describe the Bayesian probabilistic record linkage model that we utilize in Section 2.2.

2.1 Causal Inference: Overview of the Weighted Average Treatment Effect and Propensity Score Overlap Weights

For any unit in the study population, let $\boldsymbol{x}$ represent its $p \times 1$ vector of covariates. Let $z \in \{0, 1\}$ represent a binary treatment, where $z = 1$ and $z = 0$ indicate assignment to the treatment and control conditions, respectively. Each unit has two potential outcomes (Rubin, 1974), one under each value of the treatment. Let $y(1)$ and $y(0)$ be the potential outcomes for the individual when $z = 1$ or $z = 0$, respectively. For any unit, we observe only one of $y(1)$ and $y(0)$. Thus, the observed outcome for any unit can be written as $y = zy(1) + (1 - z)y(0)$.

We make the following assumptions:

1. *Stable unit treatment value assumption* (SUTVA): The SUTVA contains two sub-assumptions, no interference between units (i.e., the treatment applied to one unit does not affect the outcome for another unit) and no different versions of a treatment (Rubin, 1974).

2. *Strong ignorability*: Strong ignorability stipulates that $(y(0), y(1)) \perp z|\boldsymbol{x}$ for all units, i.e., there is no confounded effect in treatment assignment, and that $0 < P(z = 1|\boldsymbol{x}) < 1$, i.e., the probability of assigning treatment is positive for every unit.

We also utilize propensity scores, defined as $e(\boldsymbol{x}) = P(z = 1|\boldsymbol{x})$, i.e., the probability of being assigned a treatment given the covariate $\boldsymbol{x}$. As shown by Rosenbaum and Rubin (1983), the treatment assignment is independent of $\boldsymbol{x}$ given $e(\boldsymbol{x})$ under SUTVA and strong ignorability. Propensity scores are used in a variety of causal estimators, including matching, stratification, inverse probability weighting, and overlap weighting, as we do here.

To compare outcomes under treatment and control, we first define the conditional average controlled difference for a given $\boldsymbol{x}$,

$$\tau(\boldsymbol{x}) = E[y|z = 1, \boldsymbol{x}] - E[y|z = 0, \boldsymbol{x}]. \tag{1}$$

Under strong ignorability, we have $E[y(z)|\boldsymbol{x}] = E[y|\boldsymbol{x}, z]$, so that $\tau(\boldsymbol{x})$ in (1) becomes the average treatment effect conditional on $\boldsymbol{x}$, i.e., $\tau(\boldsymbol{x}) = E[y(1) - y(0)|\boldsymbol{x}]$.

To complete the definition of the causal estimand, we average $\tau(\boldsymbol{x})$ over some distribution of $\boldsymbol{x}$. The choice of the distribution corresponds to the region of the covariate space for the target population of interest. For example, if one seeks to estimate the effect of the treatment on the treated, the relevant covariate distribution is for treated cases. In this article, we follow Li *et al.* (2018) and consider the overlap population, which is the population with the most overlap in covariate values for the treatment and control groups.

Let $f(\boldsymbol{x})$ be the marginal density of the covariates, defined with respect to a base measure $\Delta(\boldsymbol{x})$. Li *et al.* (2018) show that, for many populations typically of interest in causal inference, the distribution of the covariates in the target population can be represented as $g(\boldsymbol{x}) = f(\boldsymbol{x})t(\boldsymbol{x})$. For example, $t(\boldsymbol{x}) = e(\boldsymbol{x})$ when the target population comprises the treated subjects, and $t(\boldsymbol{x}) = 1$ when the target population is the entire study. Using this expression, causal estimands for different target populations can be expressed as special cases of the WATE,

$$\tau = \frac{\int \tau(\boldsymbol{x})t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x})}{\int t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x})}. \tag{2}$$

We use τ_O to represent the WATE for the overlap population.

For any unit i in a study with n units, let $w_{1i} = t(\boldsymbol{x}_i)/e(\boldsymbol{x}_i)$, and let $w_{0i} = t(\boldsymbol{x}_i)/(1 - e(\boldsymbol{x}_i))$. A consistent estimator of τ for any target population is

$$\hat{\tau} = \frac{\sum_{i=1}^n w_{1i}z_iy_i}{\sum_{i=1}^n w_{1i}z_i} - \frac{\sum_{i=1}^n w_{0i}(1 - z_i)y_i}{\sum_{i=1}^n w_{0i}(1 - z_i)}. \tag{3}$$

For the overlap population, we set $t(\boldsymbol{x}) = e(\boldsymbol{x})(1 - e(\boldsymbol{x}))$. The resulting estimator for τ_O is the estimated average treatment effect for the overlap population, given by

$$\hat{\tau}_O = \frac{\sum_{i=1}^n (1 - e(\boldsymbol{x}_i))z_iy_i}{\sum_{i=1}^n (1 - e(\boldsymbol{x}_i))z_i} - \frac{\sum_{i=1}^n e(\boldsymbol{x}_i)(1 - z_i)y_i}{\sum_{i=1}^n e(\boldsymbol{x}_i)(1 - z_i)}. \tag{4}$$

The overlap weights are attractive in causal studies. They are bounded, as $0 < e(\boldsymbol{x}_i) < 1$),

and thus $\hat{\tau}_O$ is not affected by extreme weights. Compared to the common practice of truncating weights or discarding units, the overlap weights are continuously defined and avoid arbitrary choices of cutoffs for inclusion in the analysis. Under mild conditions, the overlap weights leading to $\hat{\tau}_O$ minimize the asymptotic variance of the estimators of the form in (3) within the class of balancing weights (Li *et al.*, 2018).

Li *et al.* (2019) derive a closed form variance estimator of $\hat{\tau}_O$ using the empirical sandwich method. Let

$$\hat{\tau}_{O,1} = \frac{\sum_{i=1}^{n}(1-e(\boldsymbol{x}_i))z_i y_i}{\sum_{i=1}^{n}(1-e(\boldsymbol{x}_i))z_i} \tag{5}$$

$$\hat{\tau}_{O,0} = \frac{\sum_{i=1}^{n} e(\boldsymbol{x}_i)(1-z_i)y_i}{\sum_{i=1}^{n} e(\boldsymbol{x}_i)(1-z_i)}. \tag{6}$$

The variance estimator is given by $(n\hat{\theta})^{-2}\sum_{i=1}^{n}\hat{I}_i^2$, where $\hat{\theta} = \sum_{i=1}^{n} e(\boldsymbol{x}_i)(1-e(\boldsymbol{x}_i))/n$ and

$$\hat{I}_i = z_i(y_i - \hat{\tau}_{O,1})(1-e(\boldsymbol{x}_i)) - (1-z_i)(y_i-\hat{\tau}_{O,0})e(\boldsymbol{x}_i) - (z_i - e(\boldsymbol{x}_i))\hat{\boldsymbol{H}}'\hat{\boldsymbol{E}}^{-1}\boldsymbol{x}_i \tag{7}$$

$$\hat{\boldsymbol{H}} = \sum_{i=1}^{n}\left[z_i(y_i-\hat{\tau}_{O,1}) + (1-z_i)(y_i-\hat{\tau}_{O,0})\right] e(\boldsymbol{x}_i)(1-e(\boldsymbol{x}_i))\boldsymbol{x}_i/n \tag{8}$$

$$\hat{\boldsymbol{E}} = \sum_{i=1}^{n} e(\boldsymbol{x}_i)(1-e(\boldsymbol{x}_i))\boldsymbol{x}_i\boldsymbol{x}_i'/n. \tag{9}$$

We use $\hat{\tau}_O$ as a representative causal estimator to show the advantages of joint causal inference and probabilistic record linkage framework. We also employ a regression-adjusted causal estimator based on overlap weights, which we describe in Section 3.2.

2.2 Record Linkage

We develop methodology for bipartite record linkage scenarios (Sadinle, 2017). Under this setting, each individual is recorded at most once within each file, i.e., no file contains any duplicates. Let File B comprise n_B records, for which we measure the outcome, treatment status and p_B causally-relevant covariates. Let File A comprise n_A records, for which we measure only a set of p_A additional causally-relevant covariates not in File B. We assume that

some of the same individuals are in File A and File B. Both files include a set of imperfect linking variables that can be used to link records from File A and File B. Without loss of generality, we assume that $n_A \geq n_B$. Finally, let $p = p_A + p_B$.

For any individual i, let $\boldsymbol{x}_i^{(A)} = (x_{i,1}^{(A)}, \ldots, x_{i,p_A}^{(A)})'$ and $\boldsymbol{x}_i^{(B)} = (x_{i,1}^{(B)}, \ldots, x_{i,p_B}^{(B)})'$ be the values of the covariates that are present in File A and File B, respectively; and, let y_i be the observed outcome and z_i be the treatment status. We directly observe $\boldsymbol{x}_i^{(A)}$ for all records in File A, but not $(\boldsymbol{x}_i^{(B)}, y_i, z_i)$. Likewise, we directly observe $(\boldsymbol{x}_i^{(B)}, y_i, z_i)$ for all records in File B, but not $\boldsymbol{x}_i^{(A)}$.

Following Sadinle (2017), we introduce $\boldsymbol{d} = (d_1, \ldots, d_{n_B})'$ for the records in File B to encode a particular linkage of the two files. Specifically, for any record j in File B, let

$$d_j = \begin{cases} i, & \text{if record } i \text{ in File A and record } j \text{ in File B is a match} \\ n_A + j, & \text{if record } j \text{ in File B has no match in File A.} \end{cases}$$

In the context of bipartite matching, we enforce $d_j \neq d_{j'}$ for any $j \neq j'$.

Suppose we have F imperfect linking variables, also referred to as fields. For each pair of records (i, j) in File A$\times$File B, we define an F-dimensional vector $\boldsymbol{\gamma}_{ij} = (\gamma_{1,ij}, \ldots, \gamma_{F,ij})'$, where $\gamma_{f,ij}$ is the score reflecting the similarity in the field f for the record pair. Here, we use binary comparisons, i.e., $\gamma_{f,ij} = 1$ when the records i and j have the same value of field f, and $\gamma_{f,ij} = 0$ otherwise. One can also use ordered comparisons with multiple levels to capture the strength of agreement in the fields, which can be especially useful for string fields like names.

Probabilistic record linkage is most effective when (i) records that refer to the same entity have similar values for most linking variables, and (ii) records that refer to different entities have very different values for most linking variables. When these are not the case, for example, the amount of recording error in the files is large, the record linkage task may be practically infeasible.

Following Fellegi and Sunter (1969) and related literature, we assume that $\boldsymbol{\gamma}_{ij}$ is a random realization from a mixture of two distributions, one for true links and the other for nonlinks.

We have

$$\boldsymbol{\gamma}_{ij}|(d_j = i) \overset{iid}{\sim} g(\boldsymbol{\theta}_m), \quad \boldsymbol{\gamma}_{ij}|(d_j \neq i) \overset{iid}{\sim} g(\boldsymbol{\theta}_u), \tag{10}$$

where $\boldsymbol{\theta}_m = (\theta_{1,m}, \ldots, \theta_{F,m})'$ and $\boldsymbol{\theta}_u = (\theta_{1,u}, \ldots, \theta_{F,u})'$ are parameters specific to each mixture component. Following common practice in probabilistic record linkage, for computational convenience we posit conditional independence across fields to compute,

$$g(\boldsymbol{\theta}_m) = P(\boldsymbol{\gamma}_{ij}|d_j = i) = \prod_{f=1}^{F} P(\gamma_{f,ij}|d_j = i) = \prod_{f=1}^{F} \theta_{f,m}^{\gamma_{f,ij}}(1-\theta_{f,m})^{1-\gamma_{f,ij}} \tag{11}$$

$$g(\boldsymbol{\theta}_u) = P(\boldsymbol{\gamma}_{ij}|d_j \neq i) = \prod_{f=1}^{F} P(\gamma_{f,ij}|d_j \neq i) = \prod_{f=1}^{F} \theta_{f,u}^{\gamma_{f,ij}}(1-\theta_{f,u})^{1-\gamma_{f,ij}}. \tag{12}$$

To specify a prior distribution on $\boldsymbol{d}$ with the constraint $d_j \neq d_{j'}$ for any $j \neq j'$, we follow a construct used in the bipartite record linkage literature (e.g., Fortini *et al.*, 2002; Larsen, 2010; Sadinle, 2017). Let $I(\mathcal{E})$ represent the indicator for an event $\mathcal{E}$. We assume $I(d_j \leq n_A) \sim Ber(\pi)$, where π represents the proportion of matches expected a priori as a fraction of the smaller file. We assume $\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi)$. Marginalizing over π, the total number of matches between File A and File B, given by $o_{AB}(\boldsymbol{d}) = \sum_{j=1}^{n_B} I(d_j \leq n_A)$, is distributed according to a Beta-binomial $(n_B, \alpha_\pi, \beta_\pi)$ distribution.

Conditional on the knowledge of which records in File B have a match, we assume all possible bipartite matchings are equally likely. The final form of the prior distribution of $\boldsymbol{d}$, marginalizing over π, is given by

$$P(\boldsymbol{d}|\alpha_\pi, \beta_\pi) = \frac{(n_A - o_{AB}(\boldsymbol{d}))!}{n_A!} \frac{B(o_{AB}(\boldsymbol{d}) + \alpha_\pi, n_B - o_{AB}(\boldsymbol{d}) + \beta_\pi)}{B(\alpha_\pi, \beta_\pi)}, \tag{13}$$

where $B(\cdot)$ denotes the Beta function, and $B(\alpha_\pi, \beta_\pi) = \frac{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)}{\Gamma(\alpha_\pi + \beta_\pi)}$. The choice of the hyper-parameters α_π and β_π provides prior information on the number of intersections between the two files. Finally, the parameters $\theta_{f,m}$ and $\theta_{f,u}$ follow i.i.d. Beta(a, b) distributions for all $f = 1, \ldots, F$. We discuss the specific choices of α_π, β_π, a and b in Section 3.1.

3 Bayesian Joint Model for Causal Inference and Record Linkage

We now present a Bayesian joint modeling framework for simultaneous causal inference and probabilistic record linkage. We begin by presenting the model in general form, followed by an illustrative model specification for normally distributed data.

The joint model requires sub-models relating the outcomes, treatment indicator, and covariates in File B to the covariates in File A. The contribution to the likelihood of a record in File B changes depending on whether it is linked to a record in File A, or not. For any record j in File B linked to a record i in File A, we specify the joint distribution of $(y_j, z_j, \boldsymbol{x}_j^{(B)} | \boldsymbol{x}_i^{(A)})$ through an outcome model, a propensity score model and a covariate model, while for a record j in File B not linked to any record i in File A, we specify the joint distribution of $(y_j, z_j, \boldsymbol{x}_j^{(B)})$. For the outcomes, for any record j in File B linked to record i in File A, we specify the conditional distribution, $y_j | (\boldsymbol{x}_i^{(A)}, z_j, \boldsymbol{x}_j^{(B)})$ denoted as $f_1(y_j \mid \boldsymbol{x}_i^{(A)}, z_j, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{ym})$. For any record j in File B that does not have a link in File A and hence missing $\boldsymbol{x}_i^{(A)}$, we write $y_j | (\boldsymbol{x}_j^{(B)}, z_j)$ as $f_2(y_j \mid \boldsymbol{x}_j^{(B)}, z_j, \boldsymbol{\theta}_{yu})$. Similarly, for the treatment indicator, for any record j linked to some record i, we model the propensity score with $g_1(z_j \mid \boldsymbol{x}_j^{(B)}, \boldsymbol{x}_i^{(A)}, \boldsymbol{\theta}_{zm})$. We model the propensity score for any non-linked record j with $g_2(z_j | \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zu})$. In typical applications, $g_1(\cdot)$ and $g_2(\cdot)$ are logistic or probit regressions. Finally, we use $h_1(\boldsymbol{x}_j^{(B)} | \boldsymbol{x}_i^{(A)}, \boldsymbol{\theta}_{xm})$ to represent the conditional distribution of $\boldsymbol{x}_j^{(B)} | \boldsymbol{x}_i^{(A)}$ when record j links to record i, and $h_2(\boldsymbol{x}_j^{(B)} | \boldsymbol{\theta}_{xu})$ to represent the marginal distribution of $\boldsymbol{x}_j^{(B)}$ when record j is not linked to any record in File A.

Let $\boldsymbol{y} = (y_1, \ldots, y_{n_B})'$ and $\boldsymbol{z} = (z_1, \ldots, z_{n_B})'$ be the $n_B \times 1$ vectors of outcomes and treatment indicators for the records in File B. Let $\boldsymbol{X}^{(A)} = [\boldsymbol{x}_1^{(A)\prime} : \cdots : \boldsymbol{x}_{n_A}^{(A)\prime}]'$ be a $n_A \times p_A$ dimensional matrix of covariates in File A, and $\boldsymbol{X}^{(B)} = [\boldsymbol{x}_1^{(B)\prime} : \cdots : \boldsymbol{x}_{n_B}^{(B)\prime}]'$ be a $n_B \times p_B$ dimensional matrix of covariates in File B. For any record j in File B, the contribution to

the joint likelihood function is given by

$$L_j^{AB} = \begin{cases} f_1(y_j|\boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, z_j, \boldsymbol{\theta}_{ym})\ g_1(z_j|\boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zm})\ h_1(\boldsymbol{x}_j^{(B)}|\boldsymbol{x}_i^{(A)}, \boldsymbol{\theta}_{xm}), & \text{when } d_j = i \\ f_2(y_j|\boldsymbol{x}_j^{(B)}, z_j, \boldsymbol{\theta}_{yu})\ g_2(z_j|\boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zu})\ h_2(\boldsymbol{x}_j^{(B)}|\boldsymbol{\theta}_{xu}), & \text{when } d_j \neq i, \text{ for all i.} \end{cases} \tag{14}$$

Thus, the joint likelihood incorporating the contributions from (14) and the linkage model in (10) and (11) is

$$L(\boldsymbol{\theta}_{ym}, \boldsymbol{\theta}_{zm}, \boldsymbol{\theta}_{xm}, \boldsymbol{\theta}_{yu}, \boldsymbol{\theta}_{zu}, \boldsymbol{\theta}_{xu}, \boldsymbol{\theta}_m, \boldsymbol{\theta}_u, \boldsymbol{d}|\{\boldsymbol{\gamma}_{ij} : 1 \leq i \leq n_A, 1 \leq j \leq n_B\}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X}^{(A)}, \boldsymbol{X}^{(B)}) \propto$$
$$\prod_{\substack{(i,j):\\ d_j=i}} L_j^{AB} \prod_{\substack{j:d_j\neq i\\ \forall i}} L_j^{AB} \prod_{i,j} \left[\left\{ \prod_{f=1}^{F} {\theta_{f,m}}^{\gamma_{f,ij}} (1-\theta_{f,m})^{1-\gamma_{f,ij}} \right\}^{I(d_j=i)} \times \left\{ \prod_{f=1}^{F} {\theta_{f,u}}^{\gamma_{f,ij}} (1-\theta_{f,u})^{1-\gamma_{f,ij}} \right\}^{I(d_j\neq i)} \right]$$
$$I(d_j \neq d_{j'}, \text{ whenever } j \neq j'). \tag{15}$$

The posterior distribution of the parameters can be obtained from

$$L(\boldsymbol{\theta}_{ym}, \boldsymbol{\theta}_{zm}, \boldsymbol{\theta}_{xm}, \boldsymbol{\theta}_{yu}, \boldsymbol{\theta}_{zu}, \boldsymbol{\theta}_{xu}, \boldsymbol{\theta}_m, \boldsymbol{\theta}_u, \boldsymbol{d}|\{\boldsymbol{\gamma}_{ij} : 1 \leq i \leq n_A, 1 \leq j \leq n_B\}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X}^{(A)}, \boldsymbol{X}^{(B)}) \times$$
$$P(\boldsymbol{d}|\alpha_\pi, \beta_\pi) \times \prod_{f=1}^{F} {\theta_{f,m}}^{a-1} (1-\theta_{f,m})^{b-1} \times \prod_{f=1}^{F} {\theta_{f,u}}^{a-1} (1-\theta_{f,u})^{b-1} \times$$
$$\Pi(\boldsymbol{\theta}_{ym}, \boldsymbol{\theta}_{zm}, \boldsymbol{\theta}_{xm}, \boldsymbol{\theta}_{yu}, \boldsymbol{\theta}_{zu}, \boldsymbol{\theta}_{xu}). \tag{16}$$

This modeling strategy is sufficiently general to incorporate several choices of f_1, f_2, g_1, g_2, h_1, and h_2. For the sake of illustration, we present a specific choice of these distributions in the sections below. We also use this model in the empirical investigations.

3.1 Illustrative Specification

We now illustrate the general modeling strategy with specific choices of f_1, f_2, g_1, g_2, h_1, and h_2. We also discuss the prior distributions and the choices of hyper-parameters that we use throughout the simulations.

3.1.1 Outcome Models, Propensity Score Models and Covariate Models

For the outcomes, we assume linear regressions for f_1 and f_2, such that

$$y_j = \alpha_{ym}^{(0)} + z_j \alpha_{ym}^{(1)} + \boldsymbol{x}_i^{(A)'} \boldsymbol{\alpha}_{ym}^{(2)} + \boldsymbol{x}_j^{(B)'} \boldsymbol{\alpha}_{ym}^{(3)} + \epsilon_j, \quad \epsilon_j \sim N(0, \sigma_m^2) \tag{17}$$

for records with links, and

$$y_j = \alpha_{yu}^{(0)} + z_j \alpha_{yu}^{(1)} + \boldsymbol{x}_j^{(B)'} \boldsymbol{\alpha}_{yu}^{(2)} + \epsilon_j, \;\; \epsilon_j \sim N(0, \sigma_u^2) \tag{18}$$

for records without links. As noted previously, we do not have the $\boldsymbol{x}_i^{(A)}$ for the non-links.

For the propensity score model, we assume logistic regressions for both g_1 and g_2 with

$$Pr(z_j = 1 | \boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zm}) = \frac{\exp(\alpha_{zm}^{(0)} + \boldsymbol{x}_i^{(A)'} \boldsymbol{\alpha}_{zm}^{(1)} + \boldsymbol{x}_j^{(B)'} \boldsymbol{\alpha}_{zm}^{(2)})}{1 + \exp(\alpha_{zm}^{(0)} + \boldsymbol{x}_i^{(A)'} \boldsymbol{\alpha}_{zm}^{(1)} + \boldsymbol{x}_j^{(B)'} \boldsymbol{\alpha}_{zm}^{(2)})}, \;\; \boldsymbol{\theta}_{zm} = (\alpha_{zm}^{(0)}, \boldsymbol{\alpha}_{zm}^{(1)'}, \boldsymbol{\alpha}_{zm}^{(2)'})' \tag{19}$$

for records with links, and

$$Pr(z_j = 1 | \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zu}) = \frac{\exp(\alpha_{zu}^{(0)} + \boldsymbol{x}_j^{(B)'} \boldsymbol{\alpha}_{zu}^{(1)})}{1 + \exp(\alpha_{zu}^{(0)} + \boldsymbol{x}_j^{(B)'} \boldsymbol{\alpha}_{zu}^{(1)})}, \;\; \boldsymbol{\theta}_{zu} = (\alpha_{zu}^{(0)}, \boldsymbol{\alpha}_{zu}^{(1)'})' \tag{20}$$

for records without links.

Finally, we use a multivariate normal regression for the conditional distribution of $\boldsymbol{x}_j^{(B)} | \boldsymbol{x}_i^{(A)}$ when $d_j = i$, i.e., the covariate model, so that

$$\boldsymbol{x}_j^{(B)'} = \boldsymbol{\eta}_{xm}' + \boldsymbol{x}_i^{(A)'} \boldsymbol{B}_{xm} + \boldsymbol{\epsilon}_{ij}, \boldsymbol{\epsilon}_{ij} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{xm}), \tag{21}$$

where $\boldsymbol{B}_{xm}$ is a $p_A \times p_B$ dimensional matrix, $\boldsymbol{\eta}_{xm}$ is a p_B-dimensional vector and $\boldsymbol{\Sigma}_{xm}$ is a $p_B \times p_B$ covariance matrix. For records without links, we assume $\boldsymbol{x}_j^{(B)}$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}_{xu}$ (a p_B-dimensional vector) and covariance matrix $\boldsymbol{\Sigma}_{xu}$ (of

dimension $p_B \times p_B$).

3.1.2 Prior distributions and Choice of Hyper-parameters

In this illustrative model specification and in our simulation studies, we assign all regression coefficients in the outcome model and in the propensity score model i.i.d. N(0,1) prior distributions. We assign σ^2_{xm} and σ^2_{xu} i.i.d. Inverse-Gamma (a_σ, b_σ) priors. For the covariate model, we set a priori $\Pi(\boldsymbol{B}_{xm}, \boldsymbol{\Sigma}_{xm}) = \Pi_1(\boldsymbol{B}_{xm}|\boldsymbol{\Sigma}_{xm})\Pi_2(\boldsymbol{\Sigma}_{xm})$, where $\boldsymbol{B}_{xm}|\boldsymbol{\Sigma}_{xm}$ follows a matrix normal distribution $\mathcal{MN}_{p_A,p_B}(\boldsymbol{0}, \boldsymbol{I}, \boldsymbol{\Sigma}_{xm})$ and $\boldsymbol{\Sigma}_{xm}$ follows an IW$(\nu, \boldsymbol{I})$ prior, where IW$(\nu, \boldsymbol{I})$ denotes an Inverse-Wishart prior with parameters ν and the identity matrix. The prior specification is completed by assigning an IW$(\nu, \boldsymbol{I})$ prior on $\boldsymbol{\Sigma}_{xu}$. We set $a = b = 1$, $a_\sigma = b_\sigma = 1$, $\alpha_\pi = \beta_\pi = 1$, $\nu = 10$. The choice of $a_\sigma = b_\sigma = 1$ leads to Inverse-Gamma prior distributions which are sufficiently non-informative, while $\alpha_\pi = \beta_\pi = 1$ ensures equal prior probabilities for a pair of records being a link or a non-link. The value of $\nu = 10$ implies that the prior distributions on $\boldsymbol{\Sigma}_{xm}$ and $\boldsymbol{\Sigma}_{xu}$ are sufficiently concentrated around the identity matrix. Moderate perturbations of these hyperparameters lead to practically indistinguishable results in our simulation studies.

Summaries of the posterior distribution cannot be computed in closed form. However, the full conditional distributions for all the parameters are available. For the illustrative model, they correspond to standard families. Thus, posterior computation can proceed through a MCMC algorithm. Details of the full conditional distributions are provided in the supplementary material.

The MCMC sampling also offers inferences on the record linkages. For $j = 1, \ldots, n_B$, let $(d_j^{(1)}, \ldots, d_j^{(L)})$ be the L post burn-in MCMC iterates of d_j. For each j, we empirically estimate $P(d_j = q|-)$ using the proportion of post burn-in samples where d_j takes the value q, i.e., $\hat{P}(d_j = q|-) = \#\{l : d_j^{(l)} = q\}/L$, for $q \in \mathcal{J}_j = \{1, \ldots, n_A, n_A + j\}$. When $1 \leq q^* = arg\max_{q \in \mathcal{J}_j} \hat{P}(d_j = q|-) \leq n_A$, we conclude that the record q^* in File A is the most likely match for the record j in File B; denote this $\hat{d}_j = q^*$. On the other hand, when $q^* = n_A + j$, we conclude that most likely record j in File B does not match to any record in

File A. In addition to posterior modes, our framework can estimate the posterior probability of linkage between any record pair. See Sadinle (2017) for further discussion of using the posterior probabilities to determine links.

3.2 Incorporating the Overlap Weights Estimator

The plausibly linked files also provide means to estimate a WATE. When each record in File B has a link in File A, the WATE is defined over the full study population in File B. When some records in File B do not have links in File A, the WATE is defined over a subset of the study population in File B, which becomes the target population. Specifically, we define the target population as the overlap population among records that can be linked. Using the notation in Section 2.1, the WATE for this target population, which we denote as $\tau_{O,linked}$, can be obtained by letting $t(\boldsymbol{x}) = e(\boldsymbol{x})(1-e(\boldsymbol{x}))a(\boldsymbol{x})$ in (2), where $a(\boldsymbol{x}) = 1$ when the record corresponding to covariate $\boldsymbol{x}$ is linkable and $a(\boldsymbol{x}) = 0$ when it is not linkable. This can be a reasonable target population for causal inferences based on File A and File B, as it is the only set of individuals for which we could observe their full set of outcomes, treatments, and covariates.

An important question, however, is when we can generalize $\tau_{O,linked}$ to treatment effects for broader populations. Here, we focus on generalizing to τ_O, which is based on the full overlap population based on File A. This is the subset of records in File A that results from applying the overlap weights defined in Section 2.1, but computed with the full $\boldsymbol{x}$ for all individuals. Of course, in our setting we do not observe the full overlap population, as we can know $\boldsymbol{x}$ only for linkable records. However, we can generalize $\tau_{O,linked} = \tau_O$ when the distribution of the full set of $\boldsymbol{x}$ is the same for linkable and non-linkable records; that is, $g(\boldsymbol{x})$ when $a(\boldsymbol{x}) = 1$ is the same as $g(\boldsymbol{x})$ for the full overlap population. A special case of this scenario arises when all records in the full overlap population are linkable. We also can generalize $\tau_{O,linked} = \tau_O$ in the case where $\tau(\boldsymbol{x}) = \tau$ for all $\boldsymbol{x}$ in File A. Of course, as with any observational study, generalizing treatment effects beyond the study population requires additional assumptions, such as constant treatment effects for all individuals (Hill, 2011).

We focus our attention here on scenarios where $\tau_{O,linked} = \tau_O$ and discuss estimators for $\tau_{O,linked}$. For the l-th MCMC iterate after burn-in, let $\mathcal{M}^{(l)}$ indicate the indices of record pairs in File A and File B that are linked, i.e., $\mathcal{M}^{(l)} = \{(i,j) : d_j^{(l)} = i, i \leq n_A\}$. Let $(\boldsymbol{\theta}_{ym}^{(l)}, \boldsymbol{\theta}_{zm}^{(l)}, \boldsymbol{\theta}_{yu}^{(l)}, \boldsymbol{\theta}_{zu}^{(l)})$ be the l-th post burn-in iterate of $(\boldsymbol{\theta}_{ym}, \boldsymbol{\theta}_{zm}, \boldsymbol{\theta}_{yu}, \boldsymbol{\theta}_{zu})$. For the l-th iteration, we first compute an estimate of the propensity score for all observations in File B that are matched with some observation in File A. Specifically, if (i,j) is a matched pair, then the estimated propensity score for the (i,j)th pair is given by $\hat{e}_{i,j}^{(l)} = e(\boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zm}^{(l)})$. Following (4), define the lth post burn-in iterate of $\hat{\tau}_O$ as

$$\hat{\tau}_O^{(l)} = \left[\left(\frac{\sum_{(i,j)\in\mathcal{M}^{(l)}} (1 - \hat{e}_{i,j}^{(l)}) z_j y_j}{\sum_{(i,j)\in\mathcal{M}^{(l)}} (1 - \hat{e}_{i,j}^{(l)}) z_j} \right) - \left(\frac{\sum_{(i,j)\in\mathcal{M}^{(l)}} \hat{e}_{i,j}^{(l)} (1 - z_j) y_j}{\sum_{(i,j)\in\mathcal{M}^{(l)}} \hat{e}_{i,j}^{(l)} (1 - z_j)} \right) \right]. \tag{22}$$

We compute $(\hat{\tau}_O^{(1)}, \ldots, \hat{\tau}_O^{(L)})$ and use $\overline{\tau}_O = \sum_{l=1}^{L} \hat{\tau}_O^{(l)}/L$ as the point estimator of τ_O. To estimate the variance of $\overline{\tau}_O$, we use multiple imputation formulae with all L iterates (Hu *et al.*, 2013), computing

$$\widehat{\text{Var}(\overline{\tau}_O)} = \frac{1}{L} \sum_{l=1}^{L} U_O^{(l)} + \left(1 + \frac{1}{L}\right) \frac{1}{L-1} \sum_{l=1}^{L} (\hat{\tau}_O^{(l)} - \overline{\tau}_O)^2. \tag{23}$$

Here, each $U_O^{(l)}$ is computed using (7), plugging in the values from the l-th iterate in the expression. Assuming large L, inferences are based on a normal distribution with mean $\overline{\tau}_O$ and variance $\widehat{\text{Var}(\overline{\tau}_O)}$.

As noted previously, the generality of the joint modeling framework allows analysts to use a causal estimator of their choice with the plausibly linked data files. For the purposes of illustrating this flexibility, we now present a regression-adjusted estimator based on the overlap weights. As this estimator has not been discussed previously in the literature, we discuss some of its properties in the supplementary material.

Suppose we have a model for the outcome; for illustrative purposes, we use the model in

(17). Let the mean function of the outcome under the model, evaluated at the l-th MCMC iterate of the parameters, be $\hat{\mu}_{i,j}^{(l)}(\zeta) = \mu(z_j = \zeta, \boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{ym}^{(l)})$, where $\zeta = 0, 1$, represent control and treatment, respectively. For example, with a linear regression as the outcome model, the mean function is the predicted value of the outcome using the linked data and parameter estimates in iteration l. For any linked record pair (i, j) at the l-th iteration, the residual for the fitted outcome model is $\hat{R}_{i,j}^{(l)} = y_j - \hat{\mu}_{i,j}^{(l)}(z_j, \boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{ym}^{(l)})$. The regression-adjusted estimator for the l-th iteration is defined as

$$\hat{\tau}_{O,r}^{(l)} = \left\{ \frac{\sum_{(i,j)\in\mathcal{M}^{(l)}} (1 - \hat{e}_{i,j}^{(l)}) z_j \hat{R}_{i,j}^{(l)}}{\sum_{(i,j)\in\mathcal{M}^{(l)}} (1 - \hat{e}_{i,j}^{(l)}) z_j} - \frac{\sum_{(i,j)\in\mathcal{M}^{(l)}} \hat{e}_{i,j}^{(l)} (1 - z_j) \hat{R}_{i,j}^{(l)}}{\sum_{(i,j)\in\mathcal{M}^{(l)}} \hat{e}_{i,j}^{(l)} (1 - z_j)} \right\} + \frac{\sum_{(i,j)\in\mathcal{M}^{(l)}} \left(\hat{\mu}_{i,j}^{(l)}(1) - \hat{\mu}_{i,j}^{(l)}(0) \right) \hat{e}_{i,j}^{(l)} (1 - \hat{e}_{i,j}^{(l)})}{\sum_{(i,j)\in\mathcal{M}^{(l)}} \hat{e}_{i,j}^{(l)} (1 - \hat{e}_{i,j}^{(l)})}. \tag{24}$$

We compute $(\hat{\tau}_{O,r}^{(1)}, \ldots, \hat{\tau}_{O,r}^{(L)})$ and use $\overline{\tau}_{O,r} = \sum_{l=1}^{L} \hat{\tau}_{O,r}^{(l)} / L$ as the new estimator of τ_O.

To estimate the variance of $\overline{\tau}_{O,r}$, we use multiple imputation and compute

$$\widehat{\text{Var}(\overline{\tau}_{O,r})} = \frac{1}{L} \sum_{l=1}^{L} U_{O,r}^{(l)} + \left(1 + \frac{1}{L}\right) \frac{1}{L-1} \sum_{l=1}^{L} (\hat{\tau}_{O,r}^{(l)} - \overline{\tau}_{O,r})^2. \tag{25}$$

We derive $U_{O,r}^{(l)}$ as an empirical sandwich variance estimator based on the theory of M-estimation. To save space, we present the expression for $U_{O,r}^{(l)}$ and its derivation in the supplementary material. We use normal-based inferences for τ_O with mean $\overline{\tau}_{O,r}$ and variance $\widehat{\text{Var}(\overline{\tau}_{O,r})}$.

3.3 Useful Modeling Simplifications

Using all the conditional distributions in (14) offers a path to take advantage of as much information as possible from File A. However, it may be convenient to assume that variables in File B are independent of subsets of variables in File A to simplify model specification and reduce computational overhead. The goal of modeling the relationships among the study variables in File B and File A is to enhance the quality of the probabilistic record linkage.

Once we obtain links, these models are largely irrelevant, as we apply a causal estimator on each plausibly linked file. Thus, it is possible for the conditional distributions to be mis-specified yet still useful, as we now describe.

One simplification is to set the outcome y to be conditionally independent of $\boldsymbol{x}^{(A)}$. Effectively, this eliminates the contribution of the model for $y|\boldsymbol{x}$ from (14). Thus, analysts who make this assumption need not specify a model for y when obtaining draws of $(d_j^{(1)}, \ldots, d_j^{(L)})$, for $j = 1, \ldots, n_B$. This accords with the "design-first" philosophy of causal inference (Rubin, 2008), which argues that one should avoid using the outcomes when manipulating the covariates, such as when computing propensity scores or linking records. Using the framework with this simplification still can improve linkage quality. For example, if one can find covariates in File B that are highly correlated with some function of the variables in File A, the joint model will be able to use that information to improve linkage accuracy.

Another simplification is to assume $\boldsymbol{x}^{(B)}$ is independent of $\boldsymbol{x}^{(A)}$. This eliminates the contribution from the model for $\boldsymbol{x}^{(B)}|\boldsymbol{x}^{(A)}$ from (14) and hence eliminates the need to model this conditional distribution. When p_B or p_A is large, or when the covariates in File B have complicated distributions, this simplification can reduce modeling and computational effort substantially. Alternatively, analysts may be able to posit covariate models for fewer than p_B and p_A variables. Again, as the goal of the joint model is solely to augment the probabilistic record linkage model with information to assist in linking records, such simplifications still can provide benefits, even if they are based on faulty assumptions.

As with any model specification, it is good practice to check the quality of model fit. This can be challenging, particularly for relationships of variables across the two files. One possibility is to use pairs known a priori to be certain links, when such pairs are available. For example, one can estimate the posited outcome, propensity score, and covariate models on these certain links, and perform the usual model checking procedures to arrive at reasonable models. These certain links also could be used to identify variables across the two files that have strong relationships, so as to suitably discard variables in File A that offer little information about the variables in File B. When an adequate number of certain links are

not available for model checking, one can use record pairs that have very high probability of being links according to standard probabilistic record linkage algorithms such as that of Fellegi and Sunter (1969).

Another model checking tool is to generate replicate datasets from the joint model using draws of model parameters (Fosdick *et al.*, 2016). Analysts can compare results from these replicates to those in the observed data, akin to posterior predictive model checking. For example, one could examine the replicated and observed distributions of the outcome variable; if they are dissimilar in appearance, it suggests the outcome model might be improved.

4 Simulation Studies

We illustrate the performance of the joint modeling strategy using repeated sampling simulations. We also compare the performance of the joint model to the performance of estimators based on a two stage approach. For simplicity, we assume that both files have the same number of covariates and that all covariates are important in the outcome and the propensity score models. We present additional simulations in the supplementary material where data are generated assuming that the two files have different number of predictors (i.e., $p_A \neq p_B$) and both the propensity score and outcome models include unimportant predictors.

4.1 Simulated Data Generation

We work with the RLdata10000 data from the R package `RecordLinkage` (Sariyar and Borg, 2010). These data comprise an artificial population of 10000 records with birth years, birth months, birth dates, first names and last names. Among these, there are 1000 individuals for whom the values of these variables have been duplicated and then randomly perturbed, introducing errors into these potential linking variables.

The RLdata10000 data do not include covariates, treatments, or outcomes. We generate values of these for each of the 9000 unique individuals in the RLdata10000 file. In particular, for each individual k, we generate $p = 4$ covariates, $(x_{1,k}, x_{2,k}, x_{3,k}, x_{4,k})$ as follows. We sample

$(x_{1,k}, x_{2,k})'$ from a bivariate normal distribution with mean zero, marginal variance 1 for each component, and covariance $\rho^{(0)} = 0.2$. Here and for all parameters to follow, the superscript 0 emphasizes that the parameter value is from the true data generating mechanism. For each simulated $x_{1,k}$ and $x_{2,k}$, we generate $(x_{3,k}, x_{4,k})'$ from a bivariate normal distribution with mean $(x_{1,k}, x_{2,k})$, marginal variance 1 for each component, and correlation also equal to $\rho^{(0)}$. This represents a modest amount of correlation among the predictors.

We simulate each individual's binary treatment assignment z_k from a Bernoulli distribution such that

$$P(z_k = 1|\boldsymbol{x}_k) = \frac{e^{\alpha_0^{(0)} + \sum_{l=1}^{p} \alpha_l^{(0)} x_{l,k}}}{\left(1 + e^{\alpha_0^{(0)} + \sum_{l=1}^{p} \alpha_l^{(0)} x_{l,k}}\right)}, \tag{26}$$

where $(\alpha_0^{(0)}, \alpha_1^{(0)}, \alpha_2^{(0)}, \alpha_3^{(0)}, \alpha_4^{(0)}) = (1, 1.5, -1, 2, -3)$. We generate each individual's outcome y_k from

$$y_k = \beta_0^{(0)} + \sum_{l=1}^{p} \beta_l^{(0)} x_{l,k} + \beta_C^{(0)} z_k + \epsilon_k, \qquad \epsilon_k \overset{i.i.d.}{\sim} N(0, \sigma^{(0)2}), \tag{27}$$

where $(\beta_0^{(0)}, \beta_1^{(0)}, \beta_2^{(0)}, \beta_3^{(0)}, \beta_4^{(0)}) = (1, -1, 2, -3, -2)$. We consider $\sigma^{(0)2} \in \{1, 4, 16\}$. These correspond to R^2 values of $(.95, .82, .55)$, respectively. Thus, we can evaluate the performance of the methods under differing strength of association among the outcomes and the remaining variables. Since (27) implies $\tau(x_{1,k}, x_{2,k}, x_{3,k}, x_{4,k}) = \beta_C^{(0)}$, we have $\tau_{O,linked} = \tau_O = 5$.

We construct File A and File B by putting subsets of records into two files. Every record in File A has measured (x_1, x_2), and every record in File B has measured the outcome, treatment, and (x_3, x_4). Both files include three imperfect linking variables: birth year, birth month and birth date. We do not use the first names and last names in these simulations, reflecting the common setting where names are unavailable. When string fields like names are used for linking, one can construct comparison vectors from metrics like the Jaro-Winkler or the Levenshtein similarity measure (Jaro, 1989). For ease of simulation, we set the sizes of File A and File B to be $n_A = n_B = 1000$, although the method does not require $n_A = n_B$.

In any simulation, we randomly sample a subset of the 1000 individuals with duplicates.

We put these records in File A and their duplicates in File B. The number of these intersecting individuals is denoted by n_{AB}, which is varied to be 200, 500, or 800. In this way, we can evaluate the performance of the methods under different amounts of intersecting records. For the remaining $(n_A - n_{AB})$ records in File A, we randomly choose $(n_A - n_{AB})$ records from the 8000 individuals without duplicates, discarding their treatments, outcomes, and $\boldsymbol{x}^{(B)}$, and keeping only $\boldsymbol{x}^{(A)}$ and the linking variables. To ensure that the non-intersecting records of File A and File B correspond to different individuals, we set aside these $(n_A - n_{AB})$ records from the 8000 records. To add the remaining $(n_B - n_{AB})$ records to File B, we randomly choose $(n_B - n_{AB})$ records from the remaining $(8000 - n_A + n_{AB})$ records, discarding $\boldsymbol{x}^{(A)}$, and keeping the treatments, outcomes and $\boldsymbol{x}^{(B)}$, along with the linking variables.

When estimating the models, we let the MCMC chains run for 2000 iterations. We discard the first 1500 as burn-in, and draw inference on both the causal effects and record linkages based on the post burn-in iterates. We assess convergence of the Markov chains by observing the trace-plots of 10 randomly chosen parameters from the outcome and the propensity score models for the linked and unlinked data, which show satisfactory mixing. The average effective sample size for all parameters of the outcome model is 307 (out of 500 iterates).

We compare the performance of the joint model with estimators from a two-stage model as follows. First, we fit the bipartite Bayesian record linkage model from Section 2.2 without using the covariates, treatments, or outcomes. Each of the L post burn-in samples of $\boldsymbol{d}$ corresponds to a plausibly linked database. In each plausibly linked database, we compute the maximum likelihood estimates (MLEs) of the coefficients in the outcome and propensity score models, which we substitute into (22) and (24). As the two-stage point estimates, we compute $\overline{\tau}_O = \sum_{l=1}^{L} \hat{\tau}_O^{(l)}/L$ and $\overline{\tau}_{O,r} = \sum_{l=1}^{L} \hat{\tau}_{O,r}^{(l)}/L$. We also estimate their variances based on (23) and (25). Since this model links the files without using information on the outcomes, treatments, and covariates, comparisons with it reveal if the sharing of information between the record linkage and study variable models offers benefits.

We compare the performances of the joint and two-stage models in terms of both causal

inference and record linkage using 100 replications. For linkage quality, we compute the precision and the recall in each of the 100 replications. Following the notation in Section 2.2 and Section 3, in any replication, let $\widehat{\boldsymbol{d}}$ be the point estimate of $\boldsymbol{d} = (d_1, \ldots, d_{n_B})'$. The precision is the proportion of links that are actual matches. Let $\mathcal{A}_j = \{\widehat{d}_j = d_j, \widehat{d}_j \leq n_A\}$. The precision is defined as $\sum_{j=1}^{n_B} I(\mathcal{A}_j) / \sum_{j=1}^{n_B} I(\widehat{d}_j \leq n_A)$. The recall is the proportion of actual matches that are determined as links, $\sum_{j=1}^{n_B} I(\mathcal{A}_j) / \sum_{j=1}^{n_B} I(d_j \leq n_A)$. A perfect record linkage procedure would result in precision and recall equal to one.

To assess the quality of the causal inferences, we report the averages and empirical standard deviations of $\overline{\tau}_O$ and $\overline{\tau}_{O,r}$ over the 100 replications for both the joint and the two-stage models. We also present the empirical coverage rates of multiple imputation 95% confidence intervals (based on 100 replications) for each of these estimators. Finally, we present the results for the causal estimators applied to the subsets of records that are true links, i.e., when we have perfect record linkage. This provides baseline results to assess how much accuracy is lost from imperfect linkages. As an extra benefit, it also allows us to assess the properties of the regression-adjusted overlap weights estimator and its variance estimator in settings where record linkage is not needed.

4.2 Simulation Results

The first three rows of Table 1 display the averages of the precision and recall over 100 replications of each of the three intersection scenarios with $\sigma^{(0)2} = 1$. In these three scenarios, we observe a modest increase in precision and a sharp increase in recall as the number of intersecting records increases for both the joint and the two-stage models. The joint model dominates the two-stage model, with higher average precision and recall in all three simulation scenarios. The differences in average recall are substantial and grow with the number of intersecting records in the two files. Evidently, the joint model uses the relationships among the variables in the two files to learn more accurately which records should be paired, as the linkage variables are not sufficient by themselves to identify pairs as accurately.

The improved performance of the joint model over the two-stage model in terms of record

Percentage of Intersection	$\sigma^{(0)2}$	Precision (Joint)	Recall (Joint)	Precision (Two-stage)	Recall (Two-stage)
20	1	0.78	0.69	0.75	0.56
50	1	0.87	0.81	0.77	0.62
80	1	0.87	0.94	0.77	0.78
80	4	0.77	0.87	0.77	0.78
80	16	0.76	0.86	0.77	0.78

Table 1: Simulated average precision and recall values for the joint and the two-stage models over 100 replications of each scenario. Scenarios vary the number of intersecting records in the two files or the outcome model variance $\sigma^{(0)2}$. All Monte Carlo standard errors are 0.008 or less.

linkage has a positive impact on the estimation of the causal effect. The first three rows of Table 2 display properties of $\overline{\tau}_O$ and $\overline{\tau}_{O,r}$ over the 100 replications of the three scenarios for both joint and two-stage models. Table 2 also displays properties of these estimators when applied to the perfectly linked records. For both estimators, the joint model accurately estimates the true causal effect $\tau_O = 5$ in all scenarios, with greatest deviation for the scenario with only 20% intersection between two files. In contrast, the two-stage model significantly underestimates the causal effect in all three scenarios. The joint model has smaller empirical standard deviations than the two-stage model. The empirical standard deviations also reveal the cost of imperfect linkages. They are higher for the joint model and two-stage model than for the analysis with the perfectly linked data. As expected, the empirical standard deviations decrease as the percentage of intersection between two files increases. Finally, the empirical standard deviations are consistently higher for $\overline{\tau}_O$ compared to $\overline{\tau}_{O,r}$, suggesting benefits to using the regression-adjusted estimator.

We next vary the signal to noise ratios for the outcome model. Specifically, we consider $\sigma^{(0)2} \in \{4, 16\}$ in (27). Here, we perform simulation studies only for the 80% intersecting records scenario, as this scenario gives each model its best chance to perform effectively. The last two rows of Table 1 present the average precision and recall values corresponding to the higher outcome model variances. Comparing the third row of Table 1 with the last two rows, we find that the precision and recall values decline for the joint model as the regression

Percentage of Intersection	$\sigma^{(0)2}$	$\overline{\tau}_O$			$\overline{\tau}_{O,r}$		
		Joint	Two-stage	Perfect	Joint	Two-stage	Perfect
20	1	4.58 (0.58)	3.42 (0.61)	4.86 (0.49)	4.78 (0.36)	3.51 (0.48)	4.95 (0.32)
50	1	4.92 (0.43)	3.84 (0.49)	5.05 (0.34)	4.92 (0.20)	3.81 (0.28)	4.96 (0.17)
80	1	4.93 (0.23)	3.97 (0.29)	5.07 (0.27)	4.96 (0.17)	3.99 (0.24)	5.02 (0.09)
80	4	4.89 (0.36)	3.91 (0.40)	4.93 (0.37)	4.88 (0.30)	3.88 (0.35)	4.98 (0.27)
80	16	4.64(0.39)	3.64(0.48)	4.78(0.46)	4.68(0.35)	3.67(0.40)	4.76(0.34)

Table 2: Simulated averages and standard deviations (in parentheses) of $\overline{\tau}_O$ and $\overline{\tau}_{O,r}$ for the joint and the two-stage models, as well as the causal inferences based on the perfectly linked data. Scenarios vary the numbers of intersecting records in the files or the outcome model variance $\sigma^{(0)2}$. Results in each scenario are based on 100 replications. Monte Carlo standard errors, obtained by dividing each empirical standard deviation by 10, are all less than .08.

variance increases. As the predictive power of the covariates weakens, the outcome model offers increasingly less information about the correct linkages. For the two stage model, the average precision and recall values are unchanged (other than by small Monte Carlo errors) when changing the outcome model variance. This is expected, since the record linkage in the two-stage model is done independently of the outcomes, treatments, and covariates. Overall, under both variance values, the joint model exhibits better performance than the two-stage model in terms of recall and similar performance in terms of precision.

The last two rows of Table 2 present the simulation results for $\overline{\tau}_O$ and $\overline{\tau}_{O,r}$ in these scenarios with larger outcome model variances. The joint model continues to estimate the causal effect accurately, although with increased standard deviations, as expected. In comparison, the two-stage model continues to underestimate the causal effect. For these two larger values of $\sigma^{(0)2}$, the empirical standard deviations for $\overline{\tau}_{O,r}$ trend smaller than those for $\overline{\tau}_O$.

We next turn to the coverage rates for the multiple imputation 95% confidence intervals. For the joint model, in all but the 20% intersection scenario, the intervals based on $\overline{\tau}_O$ cover in 100% of the replications; the 20% intersection scenario has a coverage of 99%. The consistent over-coverage occurs because, in these simulations, the distribution of $\hat{\tau}_O$ across the 100 replications is platykurtic rather than normally distributed. The coverage rates for the intervals based on $\overline{\tau}_{O,r}$ for the five scenarios are $(91\%, 96\%, 98\%, 99\%, 99\%)$ for

the scenarios with, respectively, 20% intersection, 50% intersection, 80% intersection and $\sigma^{(0)2} = 1$, 80% intersection and $\sigma^{(0)2} = 4$, and 80% intersection and $\sigma^{(0)2} = 16$ scenarios. In contrast, the intervals for the two-stage approach demonstrate substantial under-coverage, never rising above 41%. For both estimators, the intervals based on $\overline{\tau}_O$ tend to be wider than those based on $\overline{\tau}_{O,r}$. Finally, for both estimators, the lengths of the intervals decrease steadily as overlap between the two files increases, reflecting reduced uncertainty in linkages.

The simulation results for the perfectly linked data also offer insight into the accuracy of the variance estimators for $\hat{\tau}_O$ and $\hat{\tau}_{O,r}$. For the five scenarios, the coverage rates when using $\overline{\tau}_O$ based on the perfectly linked data are $(97\%, 97\%, 96\%, 98\%, 98\%)$, respectively. And, the coverage rates when using $\hat{\tau}_{O,r}$ based on the perfectly linked data are $(95\%, 96\%, 96\%, 97\%, 98\%)$, respectively. As shown in the supplemental material, the variance estimators for $\hat{\tau}_O$ and $\hat{\tau}_{O,r}$ for the perfectly linked data offer reasonable estimates of the true variances, which results in close-to-nominal coverage rates. For the perfectly linked data, the intervals based on $\hat{\tau}_O$ again tend to be wider than those based on $\hat{\tau}_{O,r}$.

4.3 Illustrative Performances Under Model Simplifications

We assess the performance of the joint model under the two modeling simplifications suggested in Section 3.3.

1. **Strategy I:** The fitted covariate model assumes that $\boldsymbol{x}_j^{(B)}$ is independent of $\boldsymbol{x}_i^{(A)}$ for every linked pair of records i and j.

2. **Strategy II:** The fitted outcome model assumes that y_j is independent of $\boldsymbol{x}_i^{(A)}$ given $\boldsymbol{x}_j^{(B)}$ for every linked pair of records i and j.

We continue to generate the simulated data from the full model without any simplifications, using the scenarios with $\sigma^{(0)2} = 1$ described in Section 4.1.

Table 3 summarizes the properties of the record linkages under each simplification. The joint model generally maintains its advantage over the two-stage model on both precision and recall, especially for scenarios with a higher intersection between the two files. However, the

Strategy	Percentage of Intersection	Precision (Joint)	Recall (Joint)	Precision (Two-stage)	Recall (Two-stage)
I	20	0.74	0.63	0.75	0.56
	50	0.83	0.68	0.77	0.60
	80	0.88	0.91	0.77	0.78
II	20	0.76	0.64	0.75	0.55
	50	0.77	0.72	0.76	0.61
	80	0.82	0.84	0.77	0.78

Table 3: Summary of record linkage properties for simulations with simplifications for outcome and covariate models. Results are simulated average precision and recall over 100 replications. All Monte Carlo standard errors are 0.014 or less.

precision and recall values for the joint model tend to be lower than those in the first three rows of Table 1, reflecting the loss in accuracy for using incorrect, simplifying assumptions. As evident in Table 4, this results in increased bias for estimating the causal effect, whether using $\overline{\tau}_O$ or $\overline{\tau}_{O,r}$. Notably, $\overline{\tau}_{O,r}$ again tends to estimate τ_O more accurately than does $\overline{\tau}_O$.

Under the joint model, the multiple imputation 95% confidence intervals using $\overline{\tau}_O$ cover 100% of the replications in all three simulation scenarios when using Strategy I. The coverage rates when using $\overline{\tau}_O$ and Strategy II are $(100\%, 93\%, 89\%)$ corresponding to the $(20\%, 50\%, 80\%)$ intersection scenarios. The coverage rates when using $\overline{\tau}_{O,r}$ and Strategy I are $(82\%, 84\%, 84\%)$ corresponding to the $(20\%, 50\%, 80\%)$ intersection scenarios. For Strategy II, these coverage rates are $(99\%, 89\%, 90\%)$. Apparently, the bias induced by the model simplifications is substantial enough to produce less than nominal coverage rates. However, these coverage rates are still much higher than those for the two-stage models.

5 Illustration with Constructed Causal Study of Debit Cards

To illustrate the Bayesian joint model further, we follow the approach used by Guha *et al.* (2020) and generate a record linkage scenario for an observational study of the causal effect of possession of debit cards on household consumption. As we use the same survey as Guha

Strategy	Percentage of Intersection	$\overline{\tau}_O$			$\overline{\tau}_{O,r}$		
		Joint	Two-stage	Perfect	Joint	Two-stage	Perfect
I	20	4.51 (0.98)	3.38 (0.94)	5.39 (0.68)	4.62 (0.42)	3.48 (0.80)	5.23 (0.27)
	50	4.60 (0.46)	3.62 (0.52)	4.94 (0.31)	4.72 (0.27)	3.71 (0.45)	4.93 (0.09)
	80	4.84 (0.30)	3.89 (0.31)	5.14 (0.24)	4.83 (0.20)	3.84 (0.28)	5.09 (0.07)
II	20	5.27 (0.61)	3.46 (0.69)	5.24 (0.35)	5.14 (0.56)	3.46 (0.67)	5.12 (0.15)
	50	5.30 (0.19)	3.73 (0.38)	4.91 (0.14)	5.15 (0.16)	3.79 (0.34)	4.97 (0.11)
	80	5.26 (0.18)	3.74 (0.19)	5.15 (0.14)	5.14 (0.13)	3.75 (0.20)	5.04 (0.05)

Table 4: Summary of causal inference properties for simulations with simplifications for outcome and covariate models. Average treatment effects are computed over 100 replications. Empirical standard deviations of the 100 estimated treatment effects are in parentheses.

et al. (2020), our description of the data closely follows theirs.

5.1 Data Description and Background

We use data from the Italy Survey on Household Income and Wealth (SHIW), which is a nationally representative survey run by the Bank of Italy once in every two years since 1965, with the only exception being that the 1997 survey was delayed to 1998. This survey collects information on various aspects of Italian households' economic and financial behavior.

We link two files with data collected during the years 1995 and 1998. Some households participated in both years and some did not. Our target population is the set of households possessing at least one current bank account but no debit cards before 1995. The treatment $z = 1$ if the household (all members combined) possesses one and only one debit card at 1998, and $z = 0$ if the household does not possess any debit cards at 1998. Households with more than one debit card are excluded from our sample. As the SHIW data have information on debit card ownership only at the household level, we assume that the owner of the debit card is the household head.

The outcome is the monthly average spending of the household on all consumer goods, measured in the 1998 survey. For data quality control, we delete the observations that have either negative values of the outcome (monthly spending), unusually high values of monthly income or ratios of monthly spending to monthly income. The final data file corresponding to 1998 contains 3088 observations with information on the outcome, the treatment, and

several covariates, and the final data file corresponding to 1995 contains 582 observations with information on additional covariates.

Both files contain a common set of variables that we can use as imperfect linking variables. For this illustration, we use the household head's gender, birth year, marital status, and highest educational qualification, the geographical area of residence of the household, the region and the province in which the household is located, and the number of inhabitants in the town in which the household is located. The data values for these variables are collected in each survey year from questionnaires completed by the participants. Hence, linking on these variables is imperfect, as participants can and do enter different values in the two surveys. Fortunately, we also have a unique identifier (ID) that we can use to perfectly link households across years. We use this ID to assess how well the models link observations in the two files based on the imperfect linking variables described above. Based on the unique ID, among the intersecting individuals in the two files, there are 190 individuals in the treatment group (who possess a debit card) and 392 individuals in the control group.

We consider covariates in this study measured in the 1995 survey and the 1998 survey. The covariates in the 1995 data consist of the monthly average spending of the household on consumer goods, the net wealth of the household, the household net disposable income, the monthly average cash inventory held by the household, the average interest rate and the number of banks in the municipality where the household is located; all values are measured in 1995. Guha *et al.* (2020) provide a detailed justification for inclusion of the covariates in the 1995 survey. The covariates in the 1998 data consist of the number of household income earners and the age of the head of the household, measured in 1998.

5.2 Results

We implement the joint model following Strategy I described in Section 4.3. In fitting the model, we let the data from 1995 comprise File B and the data from 1998 comprise File A, as the data file from 1995 has smaller sample size. This means that the outcome and treatment are in File A. Although this allocation of variables differs from the presentation

Method	Precision	Recall	$\hat{\tau}_O$			$\hat{\tau}_{O,r}$		
			Mean (SD)	2.5%	97.5%	Mean (SD)	2.5%	97.5%
Perfect	–	–	140.38 (3.36)	74.47	174.84	181.61 (2.81)	165.87	198.33
Joint	0.897	0.876	184.34 (4.66)	50.05	340.31	192.44 (3.67)	162.34	246.19
Two-stage	0.849	0.842	201.18 (4.82)	101.87	364.67	221.36 (3.96)	183.29	272.06

Table 5: Results of the analysis of the SHIW data. Entries include the precision and recall for linking the 1995 and 1998 files, and the means and multiple imputation 95% confidence intervals using $\hat{\tau}_O$ and $\hat{\tau}_{O,r}$ (in thousand Italian Liras) for all methods. In the parentheses are the standard deviations (SDs) corresponding to $\hat{\tau}_O$ and $\hat{\tau}_{O,r}$.

in Section 3, practically it makes no difference to the model specification. We include both covariates in 1998 in $\boldsymbol{x}^{(A)}$ and all six covariates in 1995 in $\boldsymbol{x}^{(B)}$. In addition, because gender, marital status and highest educational qualification of the head of the household could be important predictors of the outcome, we also include their 1995 values in $\boldsymbol{x}^{(B)}$.

For the outcome model, we use a linear regression of 1998 monthly average spending of the household on all consumer goods on linear functions of $(\boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)})$. For the propensity score model, we use a logistic regression of z on linear functions of $(\boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)})$. We do not specify a covariate model. We use the prior hyperparameter values described in the simulation studies, moderate perturbations of which lead to practically indistinguishable results. We let the MCMC chains run for 2000 iterations and discard the first 1500 as burn-in, drawing inferences on both the treatment effect and record linkage based on the post burn-in iterates. We also include results for the two-stage model and results using the perfectly linked data for comparisons.

Table 5 presents the precision and recall values, along with the multiple imputation means and 95% confidence intervals using $\hat{\tau}_O$ and $\hat{\tau}_{O,r}$ (in thousand Italian Liras) for all methods. Consistent with the simulation results, the joint model offers better precision and recall than the two-stage model. Using results from the perfect-links model as a benchmark, we find that the joint model more closely tracks the mean treatment effect estimates from the perfect-links model than the two-stage model does. This also holds for the 95% confidence intervals, particularly for $\hat{\tau}_{O,r}$, although the differences arguably are modest. The estimated variance of $\hat{\tau}_{O,r}$ is smaller than the estimated variance of $\hat{\tau}_O$ across all three methods, reflecting

the benefits of using the regression-adjusted estimator. It should be noted that the point estimates from both the joint and the two-stage models differ from those for the perfect-links model, reflecting the effects of inevitably imperfect linkages.

These results suggest that, on average, the effect of possession of a single debit card for a household leads to more monthly consumption than households that do not possess any debit card during the study period. Similar results are presented by Mercatanti and Li (2014) who show that the possession of debit cards in a household is generally accompanied with higher levels of income, wealth and education of the members in comparison with households without debit cards.

6 Conclusion and Future Work

The empirical studies suggest that the Bayesian joint modeling strategy for causal inference and record linkage can improve the quality of the linkages and the accuracy of the causal inferences. They also suggest potential benefits of using a regression-adjusted estimator when applying overlap weights approaches to propensity score inference.

The modeling framework has other advantages. First, it can accommodate missing outcomes, treatment status or linking variables in the two files. These values can be imputed from predictive distributions as part of the MCMC algorithms. In such cases, using the full modeling strategy can be preferable to using a simplification, so as to preserve relationships across variables during imputation. Second, the modeling framework accommodates any causal estimator, such as those based on inverse probability weighting or matching using propensity scores. Third, it can accommodate prior information, such as estimates of relationships among the study variables from other studies or domain knowledge, via specification of informative prior distributions.

The joint model is computationally intensive, as is generally the case with Bayesian versions of bipartite probabilistic record linkage in general. In addition to simplifying the models as discussed in Section 3.3, it may be possible to speed computation by modifying the estimation algorithms. For example, in large samples, one can approximate the distributions

of coefficients of binary or other categorical regression models using normal distributions, thereby simplifying some MCMC steps. Another approach is not to enforce bipartite matching in the Bayesian record linkage model. By allowing duplicate matchings, the linkage steps can be done for each observation in parallel, thereby speeding computation significantly. Further, it may be possible to adapt some of the strategies in recent work on scalable record linkage (McVeigh *et al.*, 2019; Marchant *et al.*, 2021).

In some contexts, analysts may desire to use some variables as linkage variables and as covariates, as we do in the SHIW analysis. When these variables are recorded identically across files, this presents no issue for the joint modeling framework. In such cases it makes sense to view these as blocking variables rather than use them as linkage variables. When these variables are not recorded identically across files, the path forward to using the joint model is less clear. We treated the values in one of the files, File A in our SHIW application, as covariates while using the values in both files as linking variables. Evaluating this approach as a general strategy in probabilistic record linkage is a topic for future research.

Supplementary Material

Section 1: Introduction to the supplementary material.

Section 2: This section provides full conditional distributions for the joint model described in Section 3 of the main article.

Section 3: This section states theorems about $\hat{\tau}_{O,r}$ as a causal estimator in complete-data contexts, i.e., a single database has all relevant variables.

Section 4: This section provides proofs of the theorems in Section 3 of the supplementary material.

Section 5: This section presents the derivation of the asymptotic variance estimator of $\overline{\tau}_{O,r}$.

Section 6: This section demonstrates performance of the joint and two-stage models in additional simulations with unequal number of predictors in two files and with unimportant predictors in the outcome and propensity score models.

References

Chipperfield, J. O., Bishop, G., Campbell, P. D., *et al.* (2011). Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data. *Survey Methodology*.

Dalzell, N. M. and Reiter, J. P. (2018). Regression modeling and file matching using possibly erroneous matching variables. *Journal of Computational and Graphical Statistics*, **27**(4), 728–738.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64**(328), 1183–1210.

Fortini, M., Nuccitelli, A., Liseo, B., and Scanu, M. (2002). Modelling issues in record linkage: a Bayesian perspective. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, pages 1008–1013.

Fosdick, B. K., Yoreo, M. D., and Reiter, J. P. (2016). Categorical data fusion using auxiliary information. *Annals of Applied Statistics*, **10**, 1907—-1929.

Guha, S., Reiter, J. P., and Mercatanti, A. (2020). A joint Bayesian framework for causal inference and bipartite matching for record linkage. *arXiv preprint arXiv:2002.09119*.

Gutman, R., Afendulis, C. C., and Zaslavsky, A. M. (2013). A Bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association*, **108**(501), 34–47.

Heck Wortman, J. and Reiter, J. P. (2018). Simultaneous record linkage and causal inference with propensity score subclassification. *Statistics in Medicine*, **37**(24), 3533–3546.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, **20**(1), 217–240.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, **71**(4), 1161–1189.

Hu, J., Mitra, R., and Reiter, J. P. (2013). Are independent parameter draws necessary for multiple imputation? *The American Statistician*, **67**, 143–149.

Jaro, M. A. (1989). 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.

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

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

Larsen, M. D. (2010). Record linkage modeling in federal statistical databases. In *Proceedings of the 2009 FCSM Research Conference*.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, **113**(521), 390–400.

Li, F., Thomas, L. E., and Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, **188**(1), 250–257.

Marchant, N. G., Kaplan, A., Elazar, D. N., Rubinstein, B. I., and Steorts, R. C. (2021). d-blink: Distributed end-to-end Bayesian entity resolution. *Journal of Computational and Graphical Statistics*, **30**(2), 406–421.

McVeigh, B. S., Spahn, B. T., and Murray, J. S. (2019). Scaling Bayesian probabilistic record linkage with post-hoc blocking: An application to the California Great Registers. *arXiv preprint arXiv:1905.05337*.

Mercatanti, A. and Li, F. (2014). Do debit cards increase household spending? Evidence from a semiparametric causal analysis of a survey. *Annals of Applied Statistics*, **8**(4), 2485–2508.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**(5), 688.

Rubin, D. B. (1987). *Multiple Imputation for Survey Nonresponse*. New York: Wiley.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, **2**(3), 808–840.

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

Sadinle, M. (2018). Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *Annals of Applied Statistics*, **12**(2), 1013–1038.

Sariyar, M. and Borg, A. (2010). The RecordLinkage package: Detecting errors in data. *The R Journal*, **2**(2), 61–67.

Scheuren, F. and Winkler, W. E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, **19**(1), 39–58.

Solomon, N. C. (2019). *A Framework for Decision Threshold Selection in Record Linkage*. Ph.D. thesis, Duke University.

Steorts, R. C., Tancredi, A., and Liseo, B. (2018). Generalized Bayesian record linkage and regression with exact error propagation. In J. Domingo-Ferrer and F. Montes, editors, *Privacy in Statistical Databases*, pages 297–313. Springer.

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

Tancredi, A., Steorts, R., Liseo, B., *et al.* (2020). A unified framework for de-duplication and population size estimation (with discussion). *Bayesian Analysis*, **15**(2), 633–682.

Tang, J., Reiter, J. P., and Steorts, R. C. (2020). Bayesian modeling for simultaneous regression and record linkage. In J. Domingo-Ferrer and K. Muralidhar, editors, *Privacy in Statistical Databases*, pages 209–223. Springer.

Ventura, S. L. and Nugent, R. (2014). Hierarchical linkage clustering with distributions of distances for large-scale record linkage. In J. Domingo-Ferrer and F. Montes, editors, *Privacy in Statistical Databases*, pages 283–298. Springer.