

# Bayesian Causal Inference with Bipartite Record Linkage

Sharmistha Guha<sup>\*</sup>, Jerome P. Reiter<sup>†</sup>, and Andrea Mercatanti<sup>‡</sup>

**Abstract.** In some scenarios, the observational data needed for causal inferences are spread over two data files. In particular, we consider scenarios where one file includes covariates and the treatment measured on a set of individuals, and a second file includes responses measured on another, partially overlapping set of individuals. In the absence of error-free direct identifiers like social security numbers, straightforward merging of separate files is not feasible, so that records must be linked using error-prone variables such as names, birth dates, and demographic characteristics. Typical practice in such situations generally follows a two-stage procedure: first link the two files using a probabilistic linkage technique, then make causal inferences with the linked dataset. This does not propagate uncertainty due to imperfect linkages to the causal inference, nor does it leverage relationships among the study variables to improve the quality of the linkages. We propose a joint model for simultaneous Bayesian inference on probabilistic linkage and causal effects that addresses these deficiencies. Using simulation studies and theoretical arguments, we show that the joint model can improve the accuracy of estimated treatment effects, as well as the record linkages, compared to the two-stage modeling option. We illustrate the joint model using a constructed causal study of the effects of debit card possession on household spending.

**Keywords:** treatment, matching, observational, fusion, propensity.

## 1 Introduction

In some scenarios, researchers seek to make causal inferences from variables spread over two datasets. For example, a social scientist seeks to link records from a survey and an administrative database to assess the effect of some policy on economic outcomes. Similarly, a health researcher seeks to link patients' electronic health records and Medicare claims data to assess the effect of some medical intervention. As a final example, a researcher seeks to link records from a study done in the past to records in a current database to make inferences about long-term effects of a treatment, without having to incur the substantial costs of collecting new primary data.

When perfectly measured, unique identifiers like social security numbers or Medicare patient IDs are available in the two files, it is reasonably straightforward to link individuals across the files (based on these identifiers). However, in some circumstances, direct identifiers may be missing from one or more files, or may not be made available

---

<sup>\*</sup>Department of Statistics, Texas A&M University, College Station, TX 77843, [sharmistha@tamu.edu](mailto:sharmistha@tamu.edu)

<sup>†</sup>Department of Statistical Science, Duke University, Durham, NC 27708, [jreiter@duke.edu](mailto:jreiter@duke.edu)

<sup>‡</sup>Department of Economics, University of Verona, Via Cantarane 24, 37129 Verona, Italy, [andrea.mercatanti@univr.it](mailto:andrea.mercatanti@univr.it)

due to privacy restrictions. In such situations, data files have to be linked based on indirect identifiers, such as individuals' names, birth dates, addresses and demographic information. These are inherently imperfect, e.g., they could be recorded differently on the files. This introduces uncertainty in linkages that should be propagated to the causal inferences.

Historically, record linkage and causal inference have been carried out as a two-stage process. The researcher first links records using a probabilistic record linkage model based on indirect identifiers, not taking into account available information on the outcome, covariate or treatment status. Subsequently, the researcher uses the set of linked records in a causal inference procedure. This two-stage approach suffers from two drawbacks. First, it does not propagate uncertainty from imperfect linkages. Second, it does not take advantage of relationships among the study variables that could enhance the accuracy of the linkages.

In this article, we propose a Bayesian joint modeling framework for simultaneous causal inference and record linkage in observational studies. In particular, we consider scenarios where one file includes the treatment indicator and causally relevant covariates measured on a set of individuals, and the other file includes outcomes measured on a partially overlapping set of individuals. We follow the Bayesian paradigm for causal inference and posit models for the missing potential outcomes, conditional on the linking status and known covariates. We couple these outcome models with a probabilistic model for the unknown linkage statuses, i.e., which record pairs are links and which are not. For the outcome models, we consider both parametric and semi-parametric forms, with the latter based on a regression of the outcome on a flexible function of the propensity scores (Rosenbaum and Rubin, 1983). For the record linkage model, we use the Bayesian version of the Fellegi and Sunter (1969) model proposed by Sadinle (2017). As part of the model estimation, we generate plausible values of the missing potential outcomes, which we then use to estimate posterior distributions of causal effects.

Our work adds to a body of literature that uses Bayesian methods for simultaneous record linkage and statistical inference, including regression modeling (Gutman et al., 2013; Dalzell and Reiter, 2018) and population size estimation (Domingo-Ferrer, 2011; Tancredi and Liseo, 2011; Sadinle et al., 2018; Tancredi et al., 2018). It also adds to the literature on non-Bayesian methods for simultaneous record linkage and estimation (e.g., Scheuren and Winkler, 1991; Lahiri and Larsen, 2005; Chipperfield et al., 2011; Solomon and O'Brien, 2019). None of these Bayesian and non-Bayesian works consider causal inference as the analysis goal. Wortman and Reiter (2018) introduced the concept of allowing the causal model to inform the linkage model. Their (non-Bayesian) approach uses point estimates of average causal effects to determine the thresholds at which record pairs are declared links in a Fellegi and Sunter (1969) algorithm. It does not use the causal estimates to determine the record pairs to consider as possible links in the first place, which our Bayesian approach does. Further, their approach does not provide uncertainty quantification.

The remainder of the article proceeds as follows. In Section 2 we discuss the background, notations and the formulation of the Bayesian joint model. We also present theoretical results arguing for improved inference on record linkage from a joint model

compared to a two-stage model. In Section 3 we describe posterior computation for the model. In Section 4, we provide results from simulation studies used to assess the effectiveness of the joint model, both for causal inference and for linkage quality. In Section 5 we apply the joint model to data from an Italian household survey to link records between different files and assess the effect of debit card possession on household spending. Finally, in Section 6 we conclude with an eye towards future work.

## 2 Model and Prior Formulation

We define a few key concepts and assumptions related to causal inference in Section 2.1, and describe probabilistic record linkage in Section 2.2. We propose the Bayesian joint modeling approach in Section 2.3.

### 2.1 Background and Notation for Bayesian Causal Inference

We assume a binary treatment,  $w_i \in \{0, 1\}$ , with  $w_i = 1$  and  $w_i = 0$  indicating treatment and control assignment to individual  $i$ , respectively. Let  $\mathbf{x}_i$  be the  $p \times 1$  covariate vector and  $y_i$  be the (continuous) outcome for individual  $i$ . Each individual is assumed to have two potential outcomes (Rubin, 1974), one under each value of the treatment. We denote  $y_i(1)$  and  $y_i(0)$  as the potential outcomes for individual  $i$  when  $w_i = 1$  or  $w_i = 0$ , respectively. The treatment effect for the  $i$ th individual is given by  $T_i = y_i(1) - y_i(0)$ . Other treatment effects can be defined as well, such as  $y_i(1)/y_i(0)$ , although here we consider effects in the form of  $T_i$ .

In reality, for each individual  $i$ , we can observe only one of  $y_i(1)$  and  $y_i(0)$ ; that is, we can observe  $y_i = w_i y_i(1) + (1 - w_i) y_i(0)$ . Bayesian approaches to causal inference essentially treat the unobserved potential outcomes as missing data (Rubin, 2005; Hill, 2011; Ding et al., 2018). One can impute the missing values repeatedly by sampling from posterior predictive distributions, and use the resulting draws of each  $T_i$  to make statements about causal effects. For example, one can compute the posterior distribution of the average of the causal effects for the  $n$  individuals in the study,  $\bar{T} = \sum_i T_i/n$ .

Following convention, we make the following assumptions to facilitate causal inferences.

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 any treatment (Rubin, 1974).
2. *Strong ignorability*: Strong ignorability stipulates that  $(y_i(0), y_i(1)) \perp w_i | \mathbf{x}_i$  for all  $i$ , which means that there is no unobserved confounding, and that  $0 < P(w_i = 1 | \mathbf{x}_i) < 1$ .

We also make use of propensity scores (Rosenbaum and Rubin, 1983). For any individual  $i$ , let the propensity score  $e(\mathbf{x}_i) = P(w_i = 1 | \mathbf{x}_i)$ , i.e., the probability of being assigned to treatment given the covariate  $\mathbf{x}_i$ . Rosenbaum and Rubin (1983) show that

the treatment assignment is independent of  $\mathbf{x}_i$  given  $e(\mathbf{x}_i)$  under strong ignorability. Typically, propensity scores are estimated using binary regressions of the treatment on causally-relevant covariates (Rosenbaum and Rubin, 1983). Analysts can use the resulting estimate in a variety of ways, for example, to create subsets of matched treated and control records (Stuart, 2010).

## 2.2 Background and Notation for Probabilistic Record Linkage

We consider the scenario where we seek to link two files, File A and File B, comprising  $n_A$  and  $n_B$  records, respectively. Without loss of generality, we assume that  $n_A \geq n_B$ . We suppose each individual or entity is recorded at most once within each file, i.e., each file contains no duplicates. Under this setting, the goal of record linkage is to identify which records in File A and File B refer to the same subject. This setting is known as bipartite record linkage (Sadinle, 2017).

A corollary to the no-duplicates assumption comes in the form of a maximum one-to-one restriction in the linkage, i.e., a record in one file can be linked with a maximum of one record in the other file. Most commonly, the one-to-one linkage is enforced as a post-processing step after identifying a set of potentially many-to-one links (e.g., Fellegi and Sunter, 1969; Jaro, 1989; Winkler, 1993; Belin and Rubin, 1995; Larsen and Rubin, 2001; Herzog et al., 2007). Alternatively, one can embed the bipartite matching constraint into a Bayesian model (Fortini et al., 2002; Tancredi and Liseo, 2011; Larsen, 2010; Gutman et al., 2013; Sadinle, 2017; Dalzell and Reiter, 2018), as we do here.

Following Sadinle (2017), we introduce  $\mathbf{z} = (z_1, \dots, z_{n_B})'$  for the records in File B to encode a particular linking status between the two files. Specifically, let

$$z_j = \begin{cases} i, & \text{if record } i \text{ in File A and record } j \text{ in File B belong to same individual,} \\ n_A + j, & \text{if record } j \text{ in File B has no link in File A.} \end{cases}$$

In the context of bipartite matching, one enforces  $z_j \neq z_{j'}$  whenever  $j \neq j'$ .

Suppose the two files include  $F$  variables in common that can be used to link records across the files. We call these the linking variables or fields. For each pair of records  $(i, j)$  in File A  $\times$  File B, we define a  $F$ -dimensional vector  $\boldsymbol{\gamma}_{ij} = (\gamma_{1,ij}, \dots, \gamma_{F,ij})'$ , where  $\gamma_{f,ij}$  is a score reflecting the similarity in field  $f$  for the record pair. In this article, for categorical variables (e.g., birth year, birth date), we set  $\gamma_{f,ij} = 1$  when the values of field  $f$  for records  $i$  and  $j$  are equal, and set  $\gamma_{f,ij} = 0$  otherwise. For string fields (e.g., names) we take into account partial agreement using the normalized Levenshtein Similarity (LS) metric (Winkler, 1990). This metric ranges between 0 (no agreement) and 1 (full agreement). We obtain it using the “levenshteinSim” function in the `RecordLinkage` package in R. We convert the LS metric into a binary  $\gamma_{f,ij}$  by setting  $\gamma_{f,ij} = 1$  when the LS metric exceeds a predetermined threshold – we use 0.95 in the simulations – and  $\gamma_{f,ij} = 0$  otherwise. One can convert the LS metric into a multinomial variable for more refined comparisons (Sadinle et al., 2018; Wortman and Reiter, 2018).

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

$$\gamma_{ij}|(z_j = i) \stackrel{iid}{\sim} g(\boldsymbol{\theta}_m), \quad \gamma_{ij}|(z_j \neq i) \stackrel{iid}{\sim} g(\boldsymbol{\theta}_u), \tag{2.1}$$

where  $\boldsymbol{\theta}_m = (\theta_{1,m}, \dots, \theta_{F,m})'$  and  $\boldsymbol{\theta}_u = (\theta_{1,u}, \dots, \theta_{F,u})'$  comprise probabilities of agreement for each field specific to each mixture component. For computational convenience we assume conditional independence across fields, so that

$$\begin{aligned} g(\boldsymbol{\theta}_m) &= P(\gamma_{ij}|z_j = i) = \prod_{f=1}^F P(\gamma_{f,ij}|z_j = i) = \prod_{f=1}^F \theta_{f,m}^{\gamma_{f,ij}} (1 - \theta_{f,m})^{1-\gamma_{f,ij}}, \\ g(\boldsymbol{\theta}_u) &= P(\gamma_{ij}|z_j \neq i) = \prod_{f=1}^F P(\gamma_{f,ij}|z_j \neq i) = \prod_{f=1}^F \theta_{f,u}^{\gamma_{f,ij}} (1 - \theta_{f,u})^{1-\gamma_{f,ij}}. \end{aligned} \tag{2.2}$$

As a prior distribution on the set of  $z_j$ , such that  $z_j \neq z_{j'}$  for any  $j \neq j'$ , we follow a construct used in the bipartite record linkage literature, including Fortini et al. (2002), Larsen (2010) and Sadinle (2017). Specifically, let  $I(z_j \leq n_A) \sim Ber(\pi)$ , where  $\pi$  represents the proportion of matches expected a priori as a fraction of the smaller file. Here and throughout,  $I(\mathcal{E}) = 1$  when its argument  $\mathcal{E}$  is true, and  $I(\mathcal{E}) = 0$  otherwise. We assume  $\pi$  is distributed according to a  $Beta(\alpha_\pi, \beta_\pi)$  a priori. Marginalizing over  $\pi$ , the total number of links between File A and File B, given by  $n_{AB}(\mathbf{z}) = \sum_{j=1}^{n_B} I(z_j \leq n_A)$ , is distributed according to a Beta-binomial  $(n_B, \alpha_\pi, \beta_\pi)$  distribution. Conditioning on the number of records in File B with a link, all possible bipartite pairings are taken as equally likely. The final form of the prior distribution of  $\mathbf{z}$ , marginalizing over  $\pi$ , is given by

$$P(\mathbf{z}|\alpha_\pi, \beta_\pi) = \frac{(n_A - n_{AB}(\mathbf{z}))! B(n_{AB}(\mathbf{z}) + \alpha_\pi, n_B - n_{AB}(\mathbf{z}) + \beta_\pi)}{n_A! B(\alpha_\pi, \beta_\pi)}. \tag{2.3}$$

The choice of the hyper-parameters  $\alpha_\pi$  and  $\beta_\pi$  provides prior information on the number of overlapping records between the two files. We discuss the specific choices of  $\alpha_\pi$  and  $\beta_\pi$  in Section 3. Finally, the parameters  $\theta_{f,m}$  and  $\theta_{f,u}$  follow i.i.d.  $Beta(a = 1, b = 1)$  distributions for all  $f = 1, \dots, F$ .

### 2.3 Joint Model for Bayesian Causal Inference and Record Linkage

We now present the joint model for Bayesian causal inference and record linkage for the setting where the outcomes  $y$  are in File A, and the covariates  $\mathbf{x}$  and the treatment status  $w$  are in File B. We presume that the sets of covariates and linking variables are disjoint. This is the case, for example, when the linking variables comprise only string variables like names, addresses, or phone numbers. We discuss relaxing this assumption after presenting the model.

With this setting, we must specify the distribution of  $y_i$  in File A depending on whether or not it is linked to any record's covariates and treatment in File B. For linked records, we specify the conditional distribution of  $y_i|(\mathbf{x}_j, w_j)$  using a regression of our choice. For records without a link, we specify a model for the marginal distribution of  $y_i$ . We couple these with the model for record linkage in (2.1)–(2.3).

More specifically, the contribution to the likelihood function from the  $i$ th record in File A is  $f_1(y_i | \mathbf{x}_j, w_j, \boldsymbol{\theta}_c)$  when  $z_j = i$ , and is  $f_2(y_i | \boldsymbol{\theta}_d)$  when  $z_j \neq i$ , for any  $j$ . Here,  $\boldsymbol{\theta}_c$  and  $\boldsymbol{\theta}_d$  represent parameters in the regression and in the marginal model for outcomes, respectively. Let  $\mathbf{y} = (y_1, \dots, y_{n_A})'$  and  $\mathbf{w} = (w_1, \dots, w_{n_B})'$  be the  $n_A \times 1$  vector of outcomes in File A and  $n_B \times 1$  vector of treatment statuses in File B, respectively, and  $\mathbf{X} = [\mathbf{x}'_1 : \dots : \mathbf{x}'_{n_B}]'$  be a  $n_B \times p$  dimensional matrix of covariates obtained from File B. The joint likelihood is given by

$$\begin{aligned} & L(\boldsymbol{\theta}_d, \boldsymbol{\theta}_c, \boldsymbol{\theta}_m, \boldsymbol{\theta}_u, \mathbf{z} | \{\gamma_{ij} : 1 \leq i \leq n_A, 1 \leq j \leq n_B\}, \mathbf{y}, \mathbf{w}, \mathbf{X}) \\ & \propto \prod_{\substack{(i,j): \\ z_j=i}} f_1(y_i | \mathbf{x}_j, w_j, \boldsymbol{\theta}_c) \times \prod_{\substack{i: z_j \neq i \\ \forall j}} f_2(y_i | \boldsymbol{\theta}_d) \\ & \times \prod_{i,j} \left\{ \prod_{f=1}^F \theta_{f,m}^{\gamma_{f,ij}} (1 - \theta_{f,m})^{1-\gamma_{f,ij}} \right\}^{I(z_j=i)} \times \left\{ \prod_{f=1}^F \theta_{f,u}^{\gamma_{f,ij}} (1 - \theta_{f,u})^{1-\gamma_{f,ij}} \right\}^{I(z_j \neq i)} \\ & \times I(z_j \neq z_{j'}, \text{ whenever } j \neq j'). \end{aligned} \quad (2.4)$$

Here, as frequently done in record linkage models, we assume that  $\mathbf{x}$  does not affect the distribution for  $\mathbf{z}$ . When this is not plausible, analysts may want to condition on features of  $\mathbf{x}$  when modeling  $\mathbf{z}$  in (2.2). Additionally, analysts may want to include a model for  $\mathbf{x}$  in the likelihood when the covariates suffer from missing values; otherwise, it is not necessary to specify a model for  $\mathbf{x}$ . More broadly, we assume that the distributions in (2.2) are the same across individuals, e.g., mismatches in the linking variables across files occur randomly. This is typically done for record linkage algorithms (Christen, 2012). When this is not the case, the joint model may not perform as effectively.

We can adapt the model to settings where analysts use some variables as both linkage variables and covariates. When such variables are recorded identically across the two files, this presents no issue for the modeling framework. In such cases we can view them as blocking variables—i.e., records are required to have the same values of these variables, or else they are not linked—rather than treat them as linkage variables having some  $\gamma_{f,ij}$ . We then replace  $f_2$  with a regression of  $y$  on this subset of variables. When these variables are not recorded identically across files, one can use the values in both files as linking variables and select one set, presumably those in File B with the rest of  $\mathbf{x}$ , to include among the covariates in the model for  $y$ . We illustrate this approach in the debit card analysis as well as in a simulation in the supplementary material (Guha et al., 2022). For the remainder of this section, we assume that the covariates and linking variables are disjoint.

To illustrate the potential benefit of joint modeling over two-stage modeling, we examine the likelihood ratio that any pair of records is linked versus not linked. Under the joint model, the likelihood ratio of  $i \sim j$  (i.e., record  $i$  is linked to record  $j$ ) and  $i \not\sim j$  is given by

$$\text{Ratio}_{\text{Joint}} = \frac{L(\boldsymbol{\theta}_d, \boldsymbol{\theta}_c, \boldsymbol{\theta}_m, \boldsymbol{\theta}_u, \mathbf{z} | \{\gamma_{ij} : 1 \leq i \leq n_A, 1 \leq j \leq n_B\}, \mathbf{y}, \mathbf{w}, \mathbf{X}, i \sim j)}{L(\boldsymbol{\theta}_d, \boldsymbol{\theta}_c, \boldsymbol{\theta}_m, \boldsymbol{\theta}_u, \mathbf{z} | \{\gamma_{ij} : 1 \leq i \leq n_A, 1 \leq j \leq n_B\}, \mathbf{y}, \mathbf{w}, \mathbf{X}, i \not\sim j)}. \quad (2.5)$$

Notably, (2.5) depends on a contribution to the likelihood from the outcome model. In contrast, the likelihood ratio for linking records in the traditional two-stage model only involves the likelihood from the assumed probabilistic record linkage model. For this model, we have

$$\text{Ratio}_{2\text{Stage}} = \prod_{f=1}^F \left( \frac{\theta_{f,m}}{\theta_{f,u}} \right)^{\gamma_{f,ij}} \left( \frac{1 - \theta_{f,m}}{1 - \theta_{f,u}} \right)^{1 - \gamma_{f,ij}}. \quad (2.6)$$

Theorem 2.1 offers insight into the behavior of  $\text{Ratio}_{\text{Joint}}$  and  $\text{Ratio}_{2\text{Stage}}$ .

**Theorem 2.1.** *Assuming  $\frac{f_1(y_i|x_j, w_j, \theta_c)}{f_2(y_i|\theta_d)}$  is bounded away from 0 and  $\infty$  in its support, we have*

- (a)  $E_{i \sim j}[\text{Ratio}_{\text{Joint}}] \geq E_{i \sim j}[\text{Ratio}_{2\text{Stage}}]$ ,
- (b)  $E_{i \not\sim j}[\text{Ratio}_{\text{Joint}}] \leq E_{i \not\sim j}[\text{Ratio}_{2\text{Stage}}]$ .

*Proof.* The likelihood ratio under the joint model,  $\text{Ratio}_{\text{Joint}}$ , can be expressed as

$$\begin{aligned} & \frac{\prod_{\substack{(k,l): z_k=l, k \neq j \\ l \neq i}} f_1(y_l|x_k, w_k, \theta_c) \times f_1(y_i|x_j, w_j, \theta_c) \times \prod_{l: z_k \neq l} f_2(y_l|\theta_d)}{\prod_{\substack{(k,l): z_k=l, k \neq j \\ l \neq i}} f_1(y_l|x_k, w_k, \theta_c) \times f_2(y_i|\theta_d) \times \prod_{l: z_k \neq l} f_2(y_l|\theta_d)} \\ & \quad \times \frac{\prod_{f=1}^F \theta_{f,m}^{\gamma_{f,ij}} (1 - \theta_{f,m})^{1 - \gamma_{f,ij}}}{\prod_{f=1}^F \theta_{f,u}^{\gamma_{f,ij}} (1 - \theta_{f,u})^{1 - \gamma_{f,ij}}} \\ & = \frac{f_1(y_i|x_j, w_j, \theta_c)}{f_2(y_i|\theta_d)} \prod_{f=1}^F \left( \frac{\theta_{f,m}}{\theta_{f,u}} \right)^{\gamma_{f,ij}} \left( \frac{1 - \theta_{f,m}}{1 - \theta_{f,u}} \right)^{1 - \gamma_{f,ij}}. \end{aligned} \quad (2.7)$$

The likelihood ratio under the two-stage model,  $\text{Ratio}_{2\text{Stage}}$ , is given by (2.6), which we abbreviate as  $h(\theta_{f,m}, \theta_{f,u})$ . Thus,  $\log(\text{Ratio}_{\text{Joint}}) = \log(h(\theta_{f,m}, \theta_{f,u})) + \log\left[\frac{f_1(y_i|x_j, w_j, \theta_c)}{f_2(y_i|\theta_d)}\right]$ , and  $\log(\text{Ratio}_{2\text{Stage}}) = \log(h(\theta_{f,m}, \theta_{f,u}))$ . Therefore, we have

$$\begin{aligned} & E_{i \sim j}[\log(\text{Ratio}_{\text{Joint}}) - \log(\text{Ratio}_{2\text{Stage}})] \\ & = \int \int \left[ \prod_{f=1}^F \theta_{f,m}^{\gamma_{f,ij}} (1 - \theta_{f,m})^{1 - \gamma_{f,ij}} \right] f_1(y_i|x_j, w_j, \theta_c) \log \left[ \frac{f_1(y_i|x_j, w_j, \theta_c)}{f_2(y_i|\theta_d)} \right] \geq 0 \end{aligned}$$

as a consequence of this expression being a Kullback-Leibler divergence. And, we have

$$\begin{aligned} & E_{i \not\sim j}[\log(\text{Ratio}_{\text{Joint}}) - \log(\text{Ratio}_{2\text{Stage}})] \\ & = \int \int \left[ \prod_{f=1}^F \theta_{f,u}^{\gamma_{f,ij}} (1 - \theta_{f,u})^{1 - \gamma_{f,ij}} \right] f_2(y_i|\theta_d) \log \left[ \frac{f_1(y_i|x_j, w_j, \theta_c)}{f_2(y_i|\theta_d)} \right] \leq 0, \end{aligned}$$

where the last inequality follows from the fact that the expression is  $(-1)$  times the Kullback-Leibler divergence between the two densities  $f_1$  and  $f_2$ .  $\square$

Theorem 2.1 indicates that the likelihood ratio for the joint model is more extreme than the likelihood ratio for the two stage model, which facilitates more accurate identification of a link or no link between records  $i$  and  $j$ .

### Outcome Models

Naturally, one should specify  $f_1(y_i|w_j, \mathbf{x}_j, \boldsymbol{\theta}_c)$  and  $f_2(y_i|\boldsymbol{\theta}_d)$  to describe the distribution of outcomes as faithfully as possible. In this article, we specify models for  $y_i \in \mathcal{R}$  and assume  $f_2(y_i|\boldsymbol{\theta}_d) = N(y_i|\mu_2, \sigma_2^2)$ ; setting a more complicated distributional form for  $f_2$  or extending to a categorical  $y_i$  is relatively straightforward. For  $f_1(y_i|w_j, \mathbf{x}_j, \boldsymbol{\theta}_c)$ , we use a general mean-zero additive error form,

$$y_i = m(\mathbf{x}_j, w_j) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2). \quad (2.8)$$

The specification for  $m(\mathbf{x}_j, w_j)$  could be a simple linear form, although often in observational studies it is advantageous to use more flexible modeling (Hill, 2011).

We use a computationally favorable yet flexible specification for  $m(\mathbf{x}_j, w_j)$ . In particular, we assume

$$m(\mathbf{x}_j, w_j) = m_1(\hat{e}(\mathbf{x}_j)) + m_2(\hat{e}(\mathbf{x}_j))w_j, \quad (2.9)$$

where  $\hat{e}(\mathbf{x}_j)$  is the estimated propensity score. Here, one can determine  $\hat{e}(\mathbf{x}_j)$  using the covariates and treatment statuses for all the records in File B. Alternatively, one can determine  $\hat{e}(\mathbf{x}_j)$  for just linked records, so as to facilitate modeling over the specific region of the propensity score distribution relevant for (2.8). We illustrate the second strategy in the simulations of Section 4, with  $\hat{e}(\mathbf{x}_j) = \rho^{-1}(\mathbf{x}_j' \hat{\boldsymbol{\eta}})$ , where  $\rho(\cdot)$  is the logit link function and  $\hat{\boldsymbol{\eta}}$  is the maximum likelihood estimate of  $\boldsymbol{\eta}$  obtained by fitting a logistic regression of  $w_j$  on  $\mathbf{x}_j$  for all  $j \in \mathcal{B} = \{j : z_j = i, \text{ for some } i, 1 \leq i \leq n_A\}$ . We note that, in our simulations and illustration with the debit card study, estimating propensity scores with all of File B did not change results meaningfully.

To afford model flexibility, we propose a semi-parametric choice for  $m_1(\cdot)$  and  $m_2(\cdot)$  using penalized splines (Ruppert et al., 2003). Let  $\kappa_1 < \kappa_2 < \dots < \kappa_m$  be a set of  $m$  fixed knot points in  $(0, 1)$ . The functions  $m_1(\cdot)$  and  $m_2(\cdot)$  are represented using spline basis functions,

$$\begin{aligned} m_1(\hat{e}(\mathbf{x}_j)) &= \beta_0 + \sum_{l_1=1}^s \beta_{l_1} \hat{e}(\mathbf{x}_j)^{l_1} + \sum_{l_2=1}^m \beta_{s+l_2} (\hat{e}(\mathbf{x}_j) - \kappa_{l_2})_+^{l_2}, \\ m_2(\hat{e}(\mathbf{x}_j)) &= \sum_{l_1=1}^s \gamma_{l_1} \hat{e}(\mathbf{x}_j)^{l_1} + \sum_{l_2=1}^m \gamma_{s+l_2} (\hat{e}(\mathbf{x}_j) - \kappa_{l_2})_+^{l_2}. \end{aligned} \quad (2.10)$$

So, the parameters are  $\boldsymbol{\theta}_c = (\beta_0, \beta_1, \dots, \beta_s, \beta_{s+1}, \dots, \beta_{s+m}, \gamma_1, \dots, \gamma_s, \gamma_{s+1}, \dots, \gamma_{s+m}, \sigma^2)'$ . This modeling framework is motivated by the penalized spline regression approaches in the Bayesian survey sampling literature (Zheng and Little, 2003, 2005), with survey weights replaced by propensity scores. Penalized splines are also used in



causal inference (Myers and Louis, 2012; Gutman and Rubin, 2013, 2015; Zhou et al., 2019). Some of these articles also recommend adjusting for covariates along with the propensity score to remove any residual imbalance. This is straightforward to implement in the joint modeling framework; we add the desired covariates to  $f_1$ . We illustrate this approach in the supplementary material.

We suggest placing a large number of knots to estimate the semi-parametric functions accurately. We recommend and use a fixed set of knots chosen on a regular grid that remains unchanged for all iterations of the Markov chain Monte Carlo (MCMC) sampler. Even a moderately large choice of  $m$  may result in model over-fitting. We therefore regularize the spline coefficients,  $\beta_{s+1}, \dots, \beta_{s+m}$  and  $\gamma_{s+1}, \dots, \gamma_{s+m}$ , using Bayesian Lasso shrinkage priors. Following Park and Casella (2008), a scale-mixture representation of the Bayesian Lasso shrinkage prior is given by

$$\begin{aligned} \beta_{s+k} | \tau_{1,k}^2 &\sim N(0, \sigma^2 \tau_{1,k}^2), & \gamma_{s+k} | \tau_{2,k}^2 &\sim N(0, \sigma^2 \tau_{2,k}^2), \\ \tau_{1,k}^2 &\stackrel{iid}{\sim} \text{Exp}(\lambda_1^2), & \tau_{2,k}^2 &\stackrel{iid}{\sim} \text{Exp}(\lambda_2^2), \quad k = 1, \dots, m, \\ \lambda_1^2 &\sim \text{Gamma}(r_1, \delta_1), & \lambda_2^2 &\sim \text{Gamma}(r_2, \delta_2). \end{aligned} \tag{2.11}$$

We assign  $\beta_1, \dots, \beta_s \stackrel{i.i.d.}{\sim} N(0, 1)$  and  $\gamma_1, \dots, \gamma_s \stackrel{i.i.d.}{\sim} N(0, 1)$  priors. We also assign  $\beta_0 \sim N(0, 1)$  and  $\sigma^2 \sim IG(a_\sigma, b_\sigma)$  priors. The prior specification is completed by setting prior distributions on  $\theta_d = (\mu_d, \sigma_d^2)'$  as  $\mu_d \sim N(0, 1)$  and  $\sigma_d^2 \sim IG(a_{\sigma_d}, b_{\sigma_d})$  a priori. We discuss the choice of hyper-parameters further in Section 3.

For comparisons, we also consider a parametric outcome regression. In this model, we set  $m_1(\hat{e}(\mathbf{x}_j)) = \beta_0 + \hat{e}(\mathbf{x}_j)\beta_1$  and  $m_2(\hat{e}(\mathbf{x}_j)) = \alpha$ , so that  $\theta_c = (\beta_0, \beta_1, \alpha, \sigma^2)'$ . We assign  $\beta = (\beta_0, \beta_1)'$  and  $\alpha$  a multivariate normal prior distribution. We let  $(\beta, \alpha)' \sim N(\mathbf{0}, \mathbf{I})$ , and let  $\sigma^2$  follow an  $IG(a_\sigma, b_\sigma)$  prior.

### Estimation of the Causal Effect

As the causal effect of interest, we define the average treatment effect for the linked records, which we abbreviate as ATEL.

$$\text{ATEL} = \frac{\sum_{i \in \mathcal{A}} (y_i(1) - y_i(0))}{n_{AB}} = \frac{\sum_{i \in \mathcal{A}} T_i}{n_{AB}}, \tag{2.12}$$

where  $\mathcal{A} = \{i : j\text{th record in File B is linked to the } i\text{th record in File A, for some } j \leq n_B\} = \{i : z_j = i, \text{ for some } j \leq n_A\}$ , and  $n_{AB}$  denotes the cardinality of  $\mathcal{A}$ . When each record in File B has a link in File A, the ATEL is defined over and is an average treatment effect for the full study population in File B. When some records in File B do not have links in File A, the ATEL is defined over a subset of the study population in File B. If the cases that have links can be viewed as a random sample of the study population, in expectation the ATEL equals the average treatment effect for the study population in File B. When this is not the case, the study population that comprises the ATEL does not readily map onto populations typically used to define average treatment effects. This can complicate the task of generalizing the ATEL to the study or broader populations,

although the latter can be difficult in general with observational studies (Hill, 2011). When seeking to generalize results beyond the linked records, analysts should examine the distributions of  $\mathbf{x}$  in the linked and not linked cases. When substantial differences exist, it is prudent to be cautious in generalizing the ATEL to the study population in File B. One special case where the ATEL readily generalizes is when the individual treatment effect  $y_i(1) - y_i(0) = T$  for all  $i$  in File B, i.e., a constant treatment effect. Then, the ATEL is an average treatment effect for the study population in File B.

As an extension of the Bayesian joint model, analysts could impute plausible values of  $(y_i(0), y_i(1))$  for cases in File B that do not have links. This would allow analysts to compute average treatment effects for the population of records in File B. We leave comparison of this estimator to the ATEL for future research.

### 3 Posterior Computation

Incorporating the prior information, the full posterior for the model with the semi-parametric outcome regression is proportional to

$$\begin{aligned}
& L(\boldsymbol{\theta}_m, \boldsymbol{\theta}_u, \boldsymbol{\theta}_d, \boldsymbol{\theta}_c, \mathbf{z} | \{\gamma_{ij} : 1 \leq i \leq n_A, 1 \leq j \leq n_B\}, \mathbf{y}, \mathbf{w}, \mathbf{X}) \times P(\mathbf{z} | \alpha_\pi, \beta_\pi) \\
& \times \prod_{f=1}^F \theta_{f,m}^{a-1} (1 - \theta_{f,m})^{b-1} \times \prod_{f=1}^F \theta_{f,u}^{a-1} (1 - \theta_{f,u})^{b-1} \times IG(\sigma^2 | a_\sigma, b_\sigma) \times N(\beta_0 | 0, 1) \\
& \times \prod_{k=1}^s N((\beta_k, \gamma_k)' | 0, I) \times \prod_{k=1}^m [N(\beta_{k+s} | 0, \sigma^2 \tau_{1,k}^2) \times N(\gamma_{k+s} | 0, \sigma^2 \tau_{2,k}^2)] \\
& \times \prod_{k=1}^m [Exp(\tau_{1,k}^2 | \lambda_1^2) \times Exp(\tau_{2,k}^2 | \lambda_2^2)] \times Gamma(\lambda_1^2 | r_1, \delta_1) \times Gamma(\lambda_2^2 | r_2, \delta_2) \\
& \times N(\mu_2 | 0, 1) \times IG(\sigma_2^2 | a_{\sigma_2}, b_{\sigma_2}). \tag{3.1}
\end{aligned}$$

Similarly, the full posterior for the model with the parametric outcome regression is proportional to

$$\begin{aligned}
& L(\boldsymbol{\theta}_m, \boldsymbol{\theta}_u, \boldsymbol{\theta}_d, \boldsymbol{\theta}_c, \mathbf{z} | \{\gamma_{ij} : 1 \leq i \leq n_A, 1 \leq j \leq n_B\}, \mathbf{y}, \mathbf{w}, \mathbf{X}) \times P(\mathbf{z} | \alpha_\pi, \beta_\pi) \\
& \times \prod_{f=1}^F \theta_{f,m}^{a-1} (1 - \theta_{f,m})^{b-1} \times \prod_{f=1}^F \theta_{f,u}^{a-1} (1 - \theta_{f,u})^{b-1} \times IG(\sigma^2 | a_\sigma, b_\sigma) \times N((\boldsymbol{\beta}, \boldsymbol{\alpha})' | \mathbf{0}, \mathbf{I}) \\
& \times N(\mu_2 | 0, 1) \times IG(\sigma_2^2 | a_{\sigma_2}, b_{\sigma_2}). \tag{3.2}
\end{aligned}$$

Summaries of these posterior distributions cannot be computed in closed form. Thus, posterior computation proceeds through MCMC algorithms. In each iteration, we update the outcome regression parameters using the current set of model-determined links. We also re-estimate propensity scores based only on those records in File B that have been linked to File A in that iteration. The full posterior conditionals can be found in the supplementary material.

For either model, we let the MCMC chain run until apparent convergence (2000 iterations in our simulations) and discard an appropriate burn-in. Let  $z_j^{(1)}, \dots, z_j^{(L)}$  be the  $L$  post burn-in MCMC iterates of  $z_j$ , where  $j = 1, \dots, n_B$ . For each  $j$ , we empirically estimate  $P(z_j = q| -)$  using the proportion of post burn-in samples where  $z_j$  takes the value  $q$ , i.e.,  $\hat{P}(z_j = q| -) = \#\{l : z_j^{(l)} = q\} / L$ , for  $q \in \mathcal{J}_j = \{1, \dots, n_A, n_A + j\}$ . The most likely link for record  $j$  in File B is the record  $q$  satisfying  $1 \leq q^* = \arg \max_{q \in \mathcal{J}_j} \hat{P}(z_j = q| -) \leq n_A$ . We denote this record as  $\hat{z}_j = q$ . When  $q^* = n_A + j$ , we declare it most likely that record  $j$  does not have a link in File A. The posterior distributions of each  $z_j$  characterize the uncertainties associated with the links.

To draw posterior inferences on the ATEL, define  $y_{miss,i} = (1 - w_i)y_i(1) + w_i y_i(0)$  as the counterfactual outcome for the  $i$ th record in File A, where  $i = 1, \dots, n_A$ . At the  $l$ -th post burn-in iteration, we impute the counterfactual outcomes  $y_{miss,i}^{(l)}$  for all linked individuals, i.e., all  $i \in \mathcal{A}^{(l)} = \{i : z_j^{(l)} = i, \text{ for some } j \leq n_A\}$ , from their posterior predictive distributions,

$$p(y_{miss,i} | y_1, \dots, y_{n_A}, \mathbf{x}_j, w_i, z_j = i) = \int f_1(y_{miss,i} | w_i = 1 - w_j, \mathbf{x}_j, \boldsymbol{\theta}_c) p(\boldsymbol{\theta}_c | y_1, \dots, y_{n_B}) d\boldsymbol{\theta}_c. \quad (3.3)$$

In (3.3), we sample  $y_{miss,i}^{(l)}$  using a treatment indicator that is opposite of what is observed for its linked record, i.e., we set  $w_i = (1 - w_j)$ . We obtain the  $l$ -th post burn-in iterate for the ATEL using (2.12) with  $(y_i, y_{miss,i}^{(l)})$  for all  $i \in \mathcal{A}^{(l)}$ .

In the simulations in Section 4 and analyses in Section 5, we choose the values of the hyperparameters as  $a_\sigma = 1$ ,  $b_\sigma = 1$ ,  $\alpha_\pi = 1$ ,  $\beta_\pi = 1$ ,  $a_{\sigma_2} = 1$ ,  $b_{\sigma_2} = 1$ ,  $r_1 = r_2 = \delta_1 = \delta_2 = 1$ . Moderate perturbations of the hyperparameter values lead to practically indistinguishable results.

## 4 Simulation Studies

We assess the performance of the Bayesian joint model using simulation studies. We consider scenarios in which we vary (a) the proportion of records in common between the two files and (b) the data generation model for the outcomes. Within these, we consider scenarios with the correctly specified and a mis-specified outcome regression model. Finally, we present results from a simulation with missing outcome values.

### 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 first names, last names, birth years and birth dates, which we use as linking variables. Among these records, there are 1000 individuals whose values of these variables have been duplicated and then randomly perturbed, introducing errors into these potential

linking variables; see Sariyar and Borg (2010) for details. The `RLdata10000` dataset is widely used as benchmark data for record linkage methods (Binette and Steorts, 2020).

The `RLdata10000` data do not include covariates, treatments, or outcomes. Thus, we generate values of these for each of the 9000 unique individuals in the `RLdata10000` file. For each individual  $j$ , we generate  $p = 2$  covariates,  $x_{1,j}$  and  $x_{2,j}$ , sampled i.i.d. from standard normal distributions. We generate each individual's binary treatment assignment  $w_j$  from a Bernoulli distribution with probability given by

$$e(\mathbf{x}_j) = P(w_j = 1 | \mathbf{x}_j) = \frac{e^{\alpha_0 + \sum_{l=1}^p \alpha_l x_{l,j}}}{(1 + e^{\alpha_0 + \sum_{l=1}^p \alpha_l x_{l,j}})}, \quad (4.1)$$

where  $(\alpha_0, \alpha_1, \alpha_2) = (1, 1.5, -1)$ . We generate each individual's outcome  $y_j$  from

$$y_j = m_1^0(\mathbf{x}_j) + m_2^0(\mathbf{x}_j)w_j + \epsilon_j, \quad \epsilon_j \sim N(0, 1), \quad (4.2)$$

where the superscript 0 indicates the true data generating mechanism. We examine results for two choices of  $(m_1^0, m_2^0)$ . The first uses linear functions in  $e(\mathbf{x}_j)$ :  $m_1^0(\mathbf{x}_j) = 1 + 2e(\mathbf{x}_j)$  and  $m_2^0(\mathbf{x}_j) = 4$ . We call this Scheme L. The second uses nonlinear functions in  $e(\mathbf{x}_j)$ :  $m_1^0(\mathbf{x}_j) = 0.58 - 1.5 e(\mathbf{x}_j)$  and  $m_2^0(\mathbf{x}_j) = \exp(-0.8 + 2.6 e(\mathbf{x}_j))$ . We call this Scheme N. In both schemes, the true regression function can be viewed as a nonlinear function of the covariates.

We construct File A and File B by putting subsets of these records into two files. For any record, File A includes the outcome information, while File B includes the covariate and treatment information; both files include the imperfect linking variables. For ease of simulation, we set the sizes of File A and File B to be  $n_A = n_B = 1000$ .

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 overlapping individuals is denoted by  $O_{AB}$ , which is varied to be 100, 500, or 900. For the remaining  $(n_A - O_{AB})$  records in File A, we randomly choose  $(n_A - O_{AB})$  records from the 8000 individuals without duplicates, discarding their treatments and covariates and keeping their outcomes and the linking variables. To ensure that the non-overlapping records of File A and File B correspond to different individuals, we set aside these  $(n_A - O_{AB})$  records from the 8000 records. To add the remaining  $(n_B - O_{AB})$  records to File B, we randomly choose  $(n_B - O_{AB})$  records from the remaining  $(8000 - n_A + O_{AB})$  records, discarding their outcomes and keeping the treatments, covariates, and linking variables. This method of selection results in the same distributions of all variables for linked and not linked records. This matches an implied assumption of the model and is a condition for the validity of typical estimation tasks with linked data (Christen, 2012).

For comparisons, we consider the performance of two alternatives. The first is a two-stage model in which we fit the bipartite Bayesian record linkage method as described in Section 2.2, without using the covariates, treatments, or outcomes. For each of the  $L$  post burn-in MCMC samples of  $\mathbf{z}$ , we estimate the outcome model using a Bayesian regression. Using this model, we take a sample from the posterior predictive distribution of the missing potential outcomes. Thus, we have  $L$  draws of each missing potential

outcome, which we use for inferences about the ATEL. Comparisons with this approach reveal if the sharing of information between the record linkage and outcome models offers any inferential advantages. We also consider using the known links, that is, we make causal inferences with the true links. Although this approach is not feasible in practice, as one does not know the true links in genuine scenarios, we consider it a benchmark for the best we can do in these simulation scenarios.

We compare the performances of the joint model and two-stage model in terms of record linkage as well as causal inference. For the former, we examine the positive predictive value (PPV) and the negative predictive value (NPV), defined as follows. After running the record linkage procedure on the two files, e.g., as we do for each simulation replication, let  $\hat{z}$  be the posterior mode of  $\mathbf{z}$ . Let  $\mathcal{A}_{1,j} = \{\hat{z}_j = z_j, z_j \leq n_A\}$  and  $\mathcal{A}_{2,j} = \{\hat{z}_j = z_j, z_j = n_A + j\}$ . Let  $I(\mathcal{A}_{k,j})$  be the indicator function corresponding to set  $\mathcal{A}_{k,j}$ , where  $k = 1, 2$  and  $j = 1, \dots, n_B$ . The PPV is the proportion of links that are actual matches, that is,  $\sum_{j=1}^{n_B} I(\mathcal{A}_{1,j}) / \sum_{j=1}^{n_B} I(z_j \leq n_A)$ . The NPV is the proportion of non-links that are actual non-matches; that is,  $\sum_{j=1}^{n_B} I(\mathcal{A}_{2,j}) / \sum_{j=1}^{n_B} I(z_j = n_A + j)$ . A perfect record linkage procedure would result in  $\text{PPV} = \text{NPV} = 1$ .

To assess the quality of causal inference for all three methods in any simulation replication, we use the mean squared error (MSE) of the  $L$  post burn-in causal effects, i.e.,  $\text{MSE} = \sum_{l=1}^L (\text{ATEL}^{(l)} - \text{ATEL}_0)^2 / L$ , where  $\text{ATEL}_0$  is the true causal effect and  $\text{ATEL}^{(l)}$  is the  $l$ th post burn-in estimate of ATEL.

## 4.2 Results

We begin with results with no missing outcomes and with correct outcome model specifications. That is, we specify parametric or semi-parametric outcome regressions that match the choices of  $m_1(\cdot)$  and  $m_2(\cdot)$  in the data generation models in Section 4.1. We estimate the joint model and the two-stage model using four linking variables: first name, last name, birth date and birth year. We treat birth dates and years as categorical variables, setting each  $\gamma_{f,ij} = 1$  when they match across files and each  $\gamma_{f,ij} = 0$  otherwise. We compute LS metrics for first and last names, and for each set  $\gamma_{f,ij} = 1$  when the LS metrics are at least 0.95 and  $\gamma_{f,ij} = 0$  otherwise. We note that, in these and all other simulation studies, the effective sample sizes for all regression parameters are greater than 400, among the  $L = 500$  post burn-in samples. Convergence happens quickly for both models, producing practically uncorrelated post burn-in samples.

Table 1 summarizes the average PPVs and NPVs for the joint model and two-stage model over 100 independent replications—enough to generate sufficiently small Monte Carlo errors—when using the correct outcome model specifications. For both data generation schemes, the PPVs and the NPVs of the joint model decrease as the percentage of overlap between File A and File B decrease. The two-stage model follows a similar pattern. Comparing the two models, we see that the joint model tends to have a higher PPV than the two-stage model. The improved performance of the joint model becomes increasingly apparent as the amount of overlap decreases. The joint model also tends to have a higher NPV than the two-stage model, although the differences are negligible in the scenario with 10% overlap under Scheme N.

True Model	Percentage of Overlap	PPV ( <i>Joint</i> )	NPV ( <i>Joint</i> )	PPV ( <i>Two-Stage</i> )	NPV ( <i>Two-Stage</i> )
Scheme L	90	0.99 (0.001)	0.98 (0.001)	0.97 (0.002)	0.94 (0.002)
	50	0.99 (0.003)	0.97 (0.002)	0.94 (0.003)	0.91 (0.003)
	10	0.92 (0.003)	0.93 (0.003)	0.85 (0.003)	0.90 (0.003)
Scheme N	90	0.99 (0.001)	0.96 (0.001)	0.97 (0.001)	0.94 (0.001)
	50	0.97 (0.002)	0.96 (0.001)	0.94 (0.002)	0.91 (0.002)
	10	0.96 (0.002)	0.91 (0.002)	0.86 (0.002)	0.91 (0.003)

Table 1: Positive predictive values (PPV) and negative predictive values (NPV) for the joint model and the two-stage model for different overlap levels when using correct outcome model specifications. The known link model uses the true links and hence is not included. Monte Carlo standard errors are in parentheses.

The improvements in the linkages when using the joint model have benefits for the estimation of the causal effect. As evident in Table 2, the joint model performs significantly better on MSE than the two-stage model. Because the two-stage method makes fewer correct linkages than the joint model, the posterior distributions of its parameter estimates tend to be centered away from the true values, which results in less accurate imputations of the missing potential outcomes. Additionally, the variability of the estimated mean regression function tends to be larger for the two-stage method than the joint method. Among these two reasons, the bias tends to represent a larger portion of the MSE. For example, under the two-stage model under scheme L and with 10% overlap, the squared bias of the posterior mean is around 10.9, whereas the variance is around 0.7. The corresponding squared bias and variance for the joint model are 1.53 and 0.19. We also observe that the performance gap in MSE becomes more substantial as the percentage of overlap decreases, especially under Scheme L. Notably, the results from the joint model are similar to those from the gold-standard Known Link model in the 50% and 90% overlap scenarios.

We next examine performance when the outcome model does not exactly match the data generating model. In particular, we fit an outcome regression that is linear in the propensity score even though the outcomes are generated using Scheme N; and, we fit an outcome regression that uses the penalized splines even though the outcomes are generated using Scheme L. Here, we only consider the scenario with 90% overlap, which gives both methods the best chance to perform well. As evident in Table 3, not surprisingly performances of both models deteriorate substantially compared to the results in Table 1 and Table 2. We are imputing potential outcomes from misspecified models, after all. When fitting the semi-parametric model to data generated under Scheme L, we observe higher PPV and NPV, as well as a lower MSE, for the

True Model	Percentage of Overlap	True Causal Effect ( $ATEL_0$ )	Joint Model	Two-Stage Model	Known Link Model
Scheme L	90	10.01	0.02 (0.001)	0.12 (0.002)	0.01 (0.001)
	50	9.98	0.11 (0.001)	7.41 (0.254)	0.09 (0.001)
	10	10.02	1.74 (0.042)	11.62 (0.271)	0.68 (0.024)
Scheme N	90	3.12	0.01 (0.001)	0.03 (0.001)	0.01 (0.001)
	50	3.09	0.02 (0.001)	0.56 (0.032)	0.02 (0.001)
	10	3.11	0.46 (0.031)	1.58 (0.059)	0.16 (0.002)

Table 2: MSE of estimating the true causal effect ( $ATEL_0$ ) for the joint model, the two-stage model, and using the known links for different overlap levels when using correct outcome model specifications. Monte Carlo standard errors are in parentheses. The true causal effect is averaged over 100 replications.

True Model	Fitted Model	PPV		NPV		MSE	
		Joint	Two-Stage	Joint	Two-Stage	Joint	Two-Stage
Scheme L	Splines	0.99 (0.004)	0.97 (0.003)	0.96 (0.003)	0.94 (0.003)	0.83 (0.031)	1.23 (0.048)
Scheme N	Linear	0.98 (0.003)	0.97 (0.003)	0.95 (0.002)	0.94 (0.003)	0.56 (0.015)	0.61 (0.013)

Table 3: PPV, NPV, and MSE for the joint model and the two-stage models under model mis-specification. Monte Carlo standard errors for PPV, NPV and MSE values are presented in parentheses. Results based on 90% overlap of records between File A and File B. The true causal effects averaged over 100 replications for Scheme L and Scheme N for 90% overlap are given in Table 2.

joint model compared to the two-stage model. This is also the case when fitting the parametric model to data generated under Scheme N; however, in this scenario the differences are practically modest. Taken together, these results suggest that, even with some model mis-specification, it may be advantageous to use the joint model over the two-stage model.

Finally, we examine the performance of the joint model and two-stage model in the presence of missing outcomes in File A. We blank either 5% or 10% of the values of  $y_i$  in File A using a missing completely at random mechanism. We examine the cases of correct model specifications with 90% overlap of records between File A and File B. To handle the missing values in the joint model, we sample the missing observations from their posterior predictive distributions in each MCMC iteration. For the two-stage model, based on the posterior modes of  $\mathbf{z}$ , we impute the missing values from their

True Model	Missing %	PPV		NPV		MSE	
		Joint	Two-Stage	Joint	Two-Stage	Joint	Two-Stage
Scheme L	5%	0.99	0.97	0.97	0.94	0.11	0.16
		(0.001)	(0.002)	(0.001)	(0.002)	(0.003)	(0.003)
Scheme L	10%	0.98	0.97	0.96	0.94	0.16	0.19
		(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)
Scheme N	5%	0.99	0.97	0.97	0.95	0.02	0.05
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Scheme N	10%	0.99	0.97	0.95	0.94	0.05	0.07
		(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)

Table 4: PPV, NPV, and MSE for the joint and the two-stage models with 5% and 10% missing outcomes in File A. Results based on 90% overlap of records between File A and File B. The true causal effects averaged over 100 replications for Scheme L and Scheme N for 90% overlap are given in Table 2.

posterior predictive distributions after fitting the outcome model on the linked dataset.

Table 4 summarizes results over 100 independent simulation runs. The performance of the joint model worsens as the percentage of missing data increases, although not by much in these scenarios.

A similar trend is observed for the two-stage model. We continue to see advantages of the joint model over the two-stage model.

In the supplementary material, we describe results from additional simulation scenarios. In particular, we find that the relative performances of the joint model and two-stage model remain qualitatively unchanged when using correlated (rather than independent) covariates (Supplement A). We lower the signal to noise ratio by increasing the regression variance (Supplement B). Not surprisingly, the performance gap between the joint model and the two-stage model closes as the variance increases. We study the performance of the two models when a subset of the predictors are also used as linking variables (Supplement C). The joint model continues to offer improved performance over the two-stage model. Finally, we present a simulation study when the true regression function is not entirely a function of  $e(\mathbf{x}_j)$  and the fitted spline model also adjusts for covariates along with the propensity score (Supplement D). Using the covariate adjustments, the joint model offers lower MSEs than the two-stage model.

We next turn to demonstrating the joint model with data from an observational study, which concerns the causal effect of possession of debit cards on household consumption. As we describe, we have true links in these data. Thus, the analysis serves primarily to illustrate the methodology.

## 5 Causal Study of Debit Cards

As context, we begin with background on the motivation for the causal question of interest. The past few decades have seen a steadily increasing global trend in the use of



non cash payment instruments like credit, debit and prepaid cards. Thaler (1985) and Thaler (1999) argue that the form of payment instruments can have a significant impact on consumer decisions via mental accounting, a set of cognitive operations used by individuals and households to keep track of financial activities. Indeed, there is evidence that consumers who have cards would spend more than ones who do not (Cole, 1998). A comprehensive causal study carried out by Mercatanti et al. (2014) in this regard focuses on the effect of debit cards on spending. Mercatanti et al. (2014) argue that debit cards, unlike credit cards, do not allow consumers to incorporate additional long-term sources of funds in their spending decisions, thus eliminating any confounding intertemporal reallocations of wealth from the psychological effects on spending (Soman, 2001), and hence are more appropriate to look at for this kind of a causal study.

## 5.1 Data Description and Background

We use data from the Italy Survey on Household Income and Wealth (SHIW). The SHIW 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. The purpose of this survey is to collect information on several aspects of Italian households' economic and financial behavior. Since the data contain information related to household characteristics, spending and payment instruments, the SHIW can provide a useful opportunity to evaluate the causal effect of debit card possession on spending in Italian households.

We link two files comprising data collected during the years 1995 and 1998. A number of the same households participated in both years. In particular, our target population is the set of households having at least one current bank account but no debit cards before 1995. The treatment  $w = 1$  if the household (all members combined) possesses one and only one debit card at 1998, and  $w = 0$  if the household does not possess any debit cards at 1998. Households with more than one debit card are excluded from our sample. Ideally, we would perform the analysis with units being individuals that possess debit cards, because debit cards are typically issued to individuals. However, the SHIW survey only has this information at the household level. Our strategy to limit the sample of treated units to households possessing only one debit card ensures that a possible effect on household spending will be due to a certain individual possessing this card. Though we do not have exact information on the ownership of the card, we make the (reasonable) assumption that the head of the household has possession of the sole debit card.

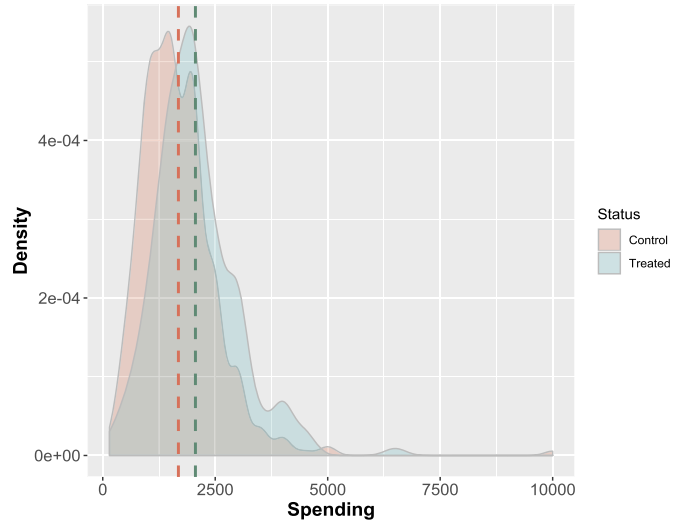
The outcome on which we evaluate the treatment effect is the monthly average spending of the household on all consumer goods, measured in the latter survey (1998). For data quality control, we delete 15 observations which have either negative values of the outcome (monthly spending) or unusually high ratios (greater than 5 and going up to 900) of monthly spending to monthly income. Upon implementing such data quality control measures, the data file corresponding to 1995 contains 589 observations with information on the treatment (debit card possession) and covariates, while the data file corresponding to 1998 (3919 observations) contains information on the outcome (monthly average household spending).

Both files contain a common set of imperfect linking variables, including the geographical area of residence of the household, the number of inhabitants in the town of the household, the number of members in the household, and the gender, birth year, marital status, region of birth and highest educational qualification of the head of the household. Fortunately, we also have a unique ID that we can use to perfectly link households across years. We use this ID variable to assess how well our model has linked observations in the two files, based on the other imperfect linking variables noted above. Using the unique matching ID, we observe that the file contains 191 observations in the treatment group (who possess a debit card) and 398 observations in the control group. An initial check on the spending distribution for the treatment and control groups (see Figure 1(a)) hints at a positive effect of acquiring a debit card on household spending.

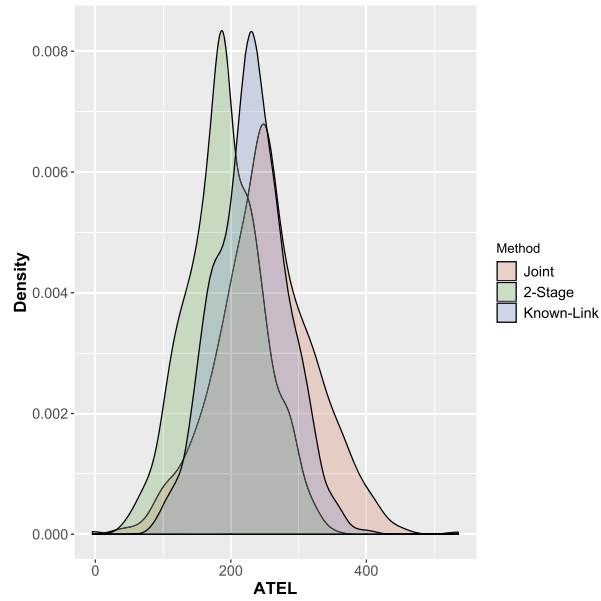
The covariates (possible confounders) we consider in this study are all measured in the initial survey (1995), and consist of the monthly average spending of the household on consumer goods in the initial survey year (lagged outcome), the net wealth of the household, the household net disposable income, the monthly average cash inventory held by the household, the number of members in the household, the marital status of the head of the household, the average interest rate and the number of banks in the municipality where the household is located. The choice of these confounders appears to satisfy the strong ignorability condition, as we discuss in the online supplement (Supplement F). The lagged outcome is generally indicated in the economic literature as a fundamental confounder (Angrist and Pischke, 2009; Frölich and Sperlich, 2019). The cash inventory held by the household was introduced in the specific context by Mercatanti et al. (2014). The net wealth and the net disposable income are important indicators of the household economic condition. The last two covariates have been suggested by Attanasio et al. (2002), who have shown in non-causal contexts that the interest rate and the number of banks in the municipality where the household lives had a significant contribution to the probability of acquiring debit card in Italy. Moreover, the number of banks is a good indicator of the size of the municipality. We note that the number of members in the household and the marital status of the household head are used both as covariates and linking variables.

## 5.2 Results

We implement the joint model with the semi-parametric outcome regression specification discussed in Section 2.3. We also include the two-stage model and the results using the known-links for comparisons. We use the same prior hyperparameter values as in the simulation studies; moderate perturbation of them leads to practically indistinguishable results. For the outcome models, we re-estimate the propensity scores using the linked records at each MCMC iteration. Results based on outcome models with propensity scores computed from all records in the 1995 data are presented in Supplement I; they are essentially identical to what we present here. We let the MCMC chains run for 2000 iterations and discard the first 1500 as burn-in, and draw inferences on both the ATEL and record linkage based on the post burn-in iterates. The effective sample sizes averaged over all parameters are 411 and 428, for the 500 post burn-in iterates corresponding to the joint and two-stage models, respectively. The post burn-in iterates



(a)



(b)

Figure 1: Figure 1(a) shows the average per month spending distribution of the treatment and control groups. The vertical lines indicate the means of the two distributions. Figure 1(b) shows the distribution of the ATEL for the joint, two-stage and the known-link models. The numbers are in per thousand Italian Liras.

Fitted Model	PPV	NPV	ATEL		
			2.5%	50%	97.5%
Known-Link	–	–	125.49	230.99	329.66
Joint	0.874	0.978	101.57	248.84	393.44
Two-Stage	0.841	0.885	82.08	188.99	297.16

Table 5: PPV and NPV for linking the 1995 and 1998 files in the SHIW causal study. Also included are the 2.5%, 50% and 97.5% quantiles of the posterior distribution of the ATEL (in thousand Italian Liras) for all methods.

are reasonably uncorrelated for both models. Additionally, we present trace-plots for a few representative parameters in Supplement G that show rapid convergence of the MCMC chains for both models. Finally, as shown in Supplement H, the covariate means for linked and non-linked records in the 1995 data, as determined by the joint model, are reasonably similar. This offers some support to interpreting the ATEL as approximately the same as a treatment effect for the study population.

Table 5 presents the PPV and NPV values, along with the posterior median and 95% credible intervals of the estimated ATEL (in thousand Italian Liras) for all models. Consistent with the simulation results, the joint model offers a noticeably better PPV and NPV than the two-stage model. Using the results from the known links as a benchmark, we find that the posterior inferences for the joint model seem more plausible than those from the two-stage model. First, the posterior medians for the joint model and the known link model are more similar to one another than are the posterior medians for the known-link model and two-stage model. Second, the 95% credible interval for the joint model is wider than the interval for the known-links model, which is sensible in that it reflects additional uncertainty from imperfect linkages. On the other hand, the 95% credible interval from the two-stage model actually is practically the same length as the interval for the known-links model, effectively portraying no propagation of uncertainty from imprecise linkages.

Figure 1(b) displays the posterior distributions of the ATEL. The 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. Our analysis largely eliminates any potential confounding effect of intertemporal reallocation of wealth, since debit cards do not allow for long-term fund sources (Soman and Cheema, 2002). Hence, the significant estimated effects of debit card possession on spending may be attributed to psychological reasons (increased perceived amount of money) (Soman, 2001) and easier accessibility to financial resources (Morewedge et al., 2007). The estimated ATEL is higher than the average treatment effect on the treated ( $\sim 200$  thousand Italian Liras) in Mercatanti et al. (2014). This result is interpretable in the light of some recent economic models for the use of debit cards (e.g., see Kim and Lee, 2010 and references therein), which imply that people with less economic resources adopt debit cards later than the rest of the population. This is confirmed by Mercatanti et al. (2014) who show that households with debit cards generally have members with higher levels of income, wealth and education compared to

households without debit cards. Therefore, our estimated ATEL values indicate larger psychological effects on spending for people in disadvantageous social and economic conditions.

## 6 Discussion and Future Work

The Bayesian approach to causal inference and record linkage offers interesting future directions. For example, some data applications have predictors and treatment status residing in different files. This requires significant modifications of the approach presented here, as one needs a model for the covariates as well as the outcomes. Another important future direction is to extend this approach to other flexible outcome models. In fact, to address outcome model mis-specification, it may be possible to adopt a doubly robust approach to causal inference (Graham et al., 2016; Saarela et al., 2016).

When one or more files have large numbers of records, Bayesian probabilistic record linkage techniques, including our joint modeling approaches, are likely to face computational challenges. These models may require comparing very large numbers of record pairs in each MCMC iteration (Binette and Steorts, 2020). In such cases, it may be possible to leverage techniques such as blocking, filtering and indexing (Murray, 2016; Steorts et al., 2014) to reduce computational burdens. Nonetheless, efficient computation with joint models is an important topic for research.

We conclude with a connection to the philosophy of causal inference. Performing causal inference and record linkage simultaneously allows the values of the outcome variables to influence which records are used in the causal estimator. This is in conflict with the often followed advice that the design of the observational study should proceed separately from the analysis (Imbens and Rubin, 2015). As suggested by Wortman and Reiter (2018), if one seeks the potential gains in accuracy from using the relationships among the variables, this is the price to pay for working with imperfect linkages.

## Supplementary Material

Supplementary Material: Bayesian Causal Inference with Bipartite Record Linkage (DOI: [10.1214/21-BA1297SUPP](https://doi.org/10.1214/21-BA1297SUPP); .pdf). Supplement A includes additional simulations where the predictors are correlated. Supplement B includes additional simulations by varying the signal to noise ratio. Supplement C includes additional simulations that incorporate overlap between linking variables and covariates. Supplement D includes additional simulations with a different data generation scheme, and with the spline model adjusting for covariates along with the propensity score. Supplement E contains full conditional distributions for implementing the Gibbs sampler. Supplement F examines the plausibility of the strong ignorability assumption for the debit card usage data. Supplement G looks at convergence diagnostics for the MCMC chains for the competing methods. Supplement H provides additional descriptive analysis for the debit card study. Supplement I provides additional analysis for the debit card study, where the propensity score is computed only once from the full 1995 file, rather than in every

iteration. Supplement J re-analyzes the real data using the ratio of spending to income as the outcome.

## References

- Angrist, J. D. and Pischke, J. (2009). “Instrumental variables in action: sometimes you get what you need. Mostly harmless econometrics: an empiricist’s companion.” 1292
- Attanasio, O. P., Guiso, L., and Jappelli, T. (2002). “The demand for money, financial innovation, and the welfare cost of inflation: An analysis with household data.” *Journal of Political Economy*, 110(2): 317–351. 1292
- Belin, T. R. and Rubin, D. B. (1995). “A method for calibrating false-match rates in record linkage.” *Journal of the American Statistical Association*, 90(430): 694–707. 1278
- Binette, O. and Steorts, R. C. (2020). “(Almost) All of Entity Resolution.” *arXiv preprint arXiv: 2008.04443*. 1286, 1295
- Chipperfield, J. O., Bishop, G., Campbell, P. D., et al. (2011). “Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data.” 1276
- Christen, P. (2012). “The data matching process.” In *Data matching*, 23–35. Springer. 1280, 1286
- Cole, C. (1998). “Identifying interventions to reduce credit card misuse through consumer behavior research.” In *Proceedings of the Marketing and Public Policy Conference*, 11–13. Washington, DC: Georgetown University Press. 1291
- 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. MR3890865. doi: <https://doi.org/10.1080/10618600.2018.1458624>. 1276, 1278
- Ding, P., Li, F., et al. (2018). “Causal inference: A missing data perspective.” *Statistical Science*, 33(2): 214–237. MR3797711. doi: <https://doi.org/10.1214/18-STS645>. 1277
- Domingo-Ferrer, J. (2011). *Privacy in statistical databases*. Springer. 1276
- Fellegi, I. P. and Sunter, A. B. (1969). “A theory for record linkage.” *Journal of the American Statistical Association*, 64(328): 1183–1210. 1276, 1278
- 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*, 1008–1013. 1278, 1279
- Frölich, M. and Sperlich, S. (2019). *Impact evaluation*. Cambridge University Press. 1292
- Graham, D. J., McCoy, E. J., Stephens, D. A., et al. (2016). “Approximate Bayesian in-

- ference for doubly robust estimation.” *Bayesian Analysis*, 11(1): 47–69. MR3447091. doi: <https://doi.org/10.1214/14-BA928>. 1295
- Guha, S., Reiter, J. P., and Mercatanti, A. (2022). “Supplementary Material: Bayesian Causal Inference with Bipartite Record Linkage.” *Bayesian Analysis*. doi: <https://doi.org/10.1214/21-BA1297SUPP>. 1280
- 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. MR3174601. doi: <https://doi.org/10.1080/01621459.2012.726889>. 1276, 1278
- Gutman, R. and Rubin, D. B. (2013). “Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes.” *Statistics in Medicine*, 32(11): 1795–1814. MR3067363. doi: <https://doi.org/10.1002/sim.5627>. 1283
- Gutman, R. and Rubin, D. B. (2015). “Estimation of causal effects of binary treatments in unconfounded studies.” *Statistics in medicine*, 34(26): 3381–3398. MR3412639. doi: <https://doi.org/10.1002/sim.6532>. 1283
- Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer Science & Business Media. MR2654644. doi: [https://doi.org/10.1016/S0169-7161\(08\)00014-X](https://doi.org/10.1016/S0169-7161(08)00014-X). 1278
- Hill, J. L. (2011). “Bayesian nonparametric modeling for causal inference.” *Journal of Computational and Graphical Statistics*, 20(1): 217–240. MR2816546. doi: <https://doi.org/10.1198/jcgs.2010.08162>. 1277, 1282, 1284
- Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press. MR3309951. doi: <https://doi.org/10.1017/CB09781139025751>. 1295
- 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. 1278
- Kim, Y. and Lee, M. (2010). “A model of debit card as a means of payment.” *Journal of Economic Dynamics & Control*, 34: 1359–1368. MR2652371. doi: <https://doi.org/10.1016/j.jedc.2010.03.003>. 1294
- Lahiri, P. and Larsen, M. D. (2005). “Regression analysis with linked data.” *Journal of the American statistical association*, 100(469): 222–230. MR2156832. doi: <https://doi.org/10.1198/016214504000001277>. 1276
- Larsen, M. D. (2010). “Record linkage modeling in federal statistical databases.” In *FCSM Research Conference*. 1278, 1279
- Larsen, M. D. and Rubin, D. B. (2001). “Iterative automated record linkage using mixture models.” *Journal of the American Statistical Association*, 96(453): 32–41. MR1973781. doi: <https://doi.org/10.1198/016214501750332956>. 1278

- Mercatanti, A., Li, F., et al. (2014). “Do debit cards increase household spending? Evidence from a semiparametric causal analysis of a survey.” *The Annals of Applied Statistics*, 8(4): 2485–2508. MR3292506. doi: <https://doi.org/10.1214/14-AOAS784>. 1291, 1292, 1294
- Morewedge, C. K., Holtzman, L., and Epley, N. (2007). “Unfixed resources: Perceived costs, consumption, and the accessible account effect.” *Journal of Consumer Research*, 34(4): 459–467. 1294
- Murray, J. S. (2016). “Probabilistic Record Linkage and De-duplication after Indexing, Blocking, and Filtering.” *Journal of Privacy and Confidentiality*, 7. 1295
- Myers, J. A. and Louis, T. A. (2012). “Comparing treatments via the propensity score: stratification or modeling?” *Health Services and Outcomes Research Methodology*, 12(1): 29–43. 1283
- Park, T. and Casella, G. (2008). “The Bayesian Lasso.” *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: <https://doi.org/10.1198/016214508000000337>. 1283
- 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. MR0742974. doi: <https://doi.org/10.1093/biomet/70.1.41>. 1276, 1277, 1278
- Rubin, D. B. (1974). “Estimating causal effects of treatments in randomized and non-randomized studies.” *Journal of educational Psychology*, 66(5): 688. 1277
- Rubin, D. B. (2005). “Bayesian inference for causal effects.” *Handbook of statistics*, 25: 1–16. MR2490521. doi: [https://doi.org/10.1016/S0169-7161\(05\)25001-0](https://doi.org/10.1016/S0169-7161(05)25001-0). 1277
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. 12. Cambridge university press. MR1998720. doi: <https://doi.org/10.1017/CB09780511755453>. 1282
- Saarela, O., Belzile, L. R., and Stephens, D. A. (2016). “A Bayesian view of doubly robust causal inference.” *Biometrika*, 103(3): 667–681. MR3551791. doi: <https://doi.org/10.1093/biomet/asw025>. 1295
- Sadinle, M. (2017). “Bayesian estimation of bipartite matchings for record linkage.” *Journal of the American Statistical Association*, 112(518): 600–612. MR3671755. doi: <https://doi.org/10.1080/01621459.2016.1148612>. 1276, 1278, 1279
- Sadinle, M. et al. (2018). “Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations.” *The Annals of Applied Statistics*, 12(2): 1013–1038. MR3834293. doi: <https://doi.org/10.1214/18-AOAS1178>. 1276, 1278
- Sariyar, M. and Borg, A. (2010). “The RecordLinkage package: Detecting errors in data.” *The R Journal*, 2(2): 61–67. 1285, 1286
- Scheuren, F. and Winkler, W. E. (1991). “Regression analysis of data files that are computer matched.” 1276



- Solomon, N. C. and O'Brien, S. M. (2019). "A Framework for Decision Threshold Selection in Record Linkage." 1276
- Soman, D. (2001). "Effects of payment mechanism on spending behavior: The role of rehearsal and immediacy of payments." *Journal of Consumer Research*, 27(4): 460–474. 1291, 1294
- Soman, D. and Cheema, A. (2002). "The effect of credit on spending decisions: The role of the credit limit and credibility." *Marketing Science*, 21(1): 32–53. 1294
- Steorts, R. C., Ventura, S. L., Sadinle, M., and Fienberg, S. E. (2014). "A Comparison of Blocking Methods for Record Linkage." In *International conference on privacy in statistical databases*, 253–268. Springer. 1295
- Stuart, E. A. (2010). "Matching methods for causal inference: A review and a look forward." *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1. MR2741812. doi: <https://doi.org/10.1214/09-STS313>. 1278
- Tancredi, A. and Liseo, B. (2011). "A hierarchical Bayesian approach to record linkage and population size problems." *The Annals of Applied Statistics*, 5(2B): 1553–1585. MR2849786. doi: <https://doi.org/10.1214/10-AOAS447>. 1276, 1278
- Tancredi, A., Steorts, R., Liseo, B., et al. (2018). "A Unified Framework for De-Duplication and Population Size Estimation." *Bayesian Analysis*. MR4122517. doi: <https://doi.org/10.1214/19-BA1146>. 1276
- Thaler, R. (1985). "Mental accounting and consumer choice." *Marketing science*, 4(3): 199–214. 1291
- Thaler, R. H. (1999). "Mental accounting matters." *Journal of Behavioral decision making*, 12(3): 183–206. 1291
- Winkler, W. E. (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." 1278
- Winkler, W. E. (1993). *Improved decision rules in the Fellegi-Sunter model of record linkage*. Citeseer. 1278
- Wortman, J. H. and Reiter, J. P. (2018). "Simultaneous record linkage and causal inference with propensity score subclassification." *Statistics in Medicine*, 37(24): 3533–3546. MR3862902. doi: <https://doi.org/10.1002/sim.7911>. 1276, 1278, 1295
- Zheng, H. and Little, J. (2005). "Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model." *Journal of Official Statistics*, 21(1): 1. 1282
- Zheng, H. and Little, R. J. (2003). "Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples." *Journal of Official Statistics*, 19(2): 99. 1282
- Zhou, T., Elliott, M. R., and Little, R. J. (2019). "Penalized spline of propensity methods for treatment comparison." *Journal of the American Statistical Association*, 114(525): 1–19. MR3941229. doi: <https://doi.org/10.1080/01621459.2018.1518234>. 1283