HHS Public Access

Author manuscript

J Stat Plan Inference. Author manuscript; available in PMC 2024 September 01.

Published in final edited form as:

J Stat Plan Inference. 2024 March ; 229: . doi:10.1016/j.jspi.2023.07.004.

Regression-Assisted Bayesian Record Linkage for Causal Inference in Observational Studies with Covariates Spread Over Two Files

Sharmistha Guha[a], **Jerome P. Reiter**[b]

[a]Department of Statistics, Texas A&M University, College Station, 77843, TX, USA

[b]Department of Statistical Science, Duke University, Durham, 27708, NC, USA

Abstract

We consider causal inference for 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 causally-relevant covariates measured on a partially overlapping set of individuals. By linking records in the two databases, the analyst can control for more covariates, thereby reducing the risk of bias compared to using only one file alone. When analysts do not have access to a unique identifier that enables perfect, error-free linkages, they typically rely on probabilistic record linkage to construct a single linked data set, and estimate causal effects using these linked data. This typical practice does not propagate uncertainty from imperfect linkages to the causal inferences. Further, it does not take advantage of relationships among the variables to improve the linkage quality. We address these shortcomings by fusing regression-assisted, Bayesian probabilistic record linkage with causal inference. The Markov chain Monte Carlo sampler generates multiple plausible linked data files as byproducts that analysts can use for multiple imputation inferences. Here, we show results for two causal estimators based on propensity score overlap weights. Using simulations and data from the Italy Survey on Household Income and Wealth, we show that our approach can improve the accuracy of estimated treatment effects.

Keywords

Data fusion; Entity resolution; Overlap; Propensity score; Treatment

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. Analysts seek to link the two databases to control for more causally-relevant covariates and thereby reduce the risk of bias from unmeasured confounding, relative to using only one file alone.

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. Often, however, 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 based on variants of the framework developed by Fellegi and Sunter [1].

Typically, researchers perform causal inference with linked files in a two-stage process. They use probabilistic record linkage to construct a single file comprising linked records, and then carry out causal inference on the linked file [e.g., 2]. 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., 3, 4, 5, 6] have shown that leveraging these relationships in fact can improve the quality of the linkages. Second, estimation with a single linked file does not propagate uncertainty arising from imperfect linkages to the causal inferences.

In this article, we address these shortcomings by proposing regression-assisted, Bayesian probabilistic record linkage with causal inference, henceforth abbreviated as RegBRLC. 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 [7]. For the sake of illustrating our modeling approach, we estimate a weighted average treatment effect [WATE, 8] using the propensity score overlap weights of [9]. Analysts could replace the overlap weights estimators with any other causal estimator.

Our work contributes to existing methods for statistical inference with probabilistic record linkage [e.g., 3, 4, 6, 10, 11, 12, 13, 14, 15, 16, 17, 18], though none of these works consider causal inference as the analysis goal. A version of simultaneous causal inference and record linkage is presented in [19]. 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-Sunter [1] algorithm. They do not use relationships among the

variables to determine the record pairs to consider as possible links in the first place, nor do they propagate uncertainty from imperfect linkages; our approach does both. A model for Bayesian causal inference and record linkage was proposed by [20] for when the treatment and all covariates reside in one file and the outcome in another. Because our data setting differs—the causally-relevant covariates are spread over two files rather than all in one file—we estimate a propensity score model and a covariate model simultaneously with the outcome and linkage models. Additionally, [20] 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 do not impute counterfactual outcomes. Instead, we estimate causal effects based on balancing scores like the overlap weights, which reduces sensitivity to the fit of the outcome model. To our knowledge, embedding outcome, propensity score, and covariate models in a Bayesian probabilistic record linkage model while enabling causal inference based on balancing scores has not been implemented previously.

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 our methodology. In Section 3, we present an illustrative specification of the RegBRLC model. We also describe a regression-adjusted causal estimator exploiting overlap weights. We believe this estimator has not appeared previously in the literature, even for settings where probabilistic record linkage is not needed, e.g., all the data are in one file. In Section 4, we present results of simulation studies comparing the illustrative RegBRLC model to a corresponding two-stage approach. Results from additional simulation studies are included in the supplementary material. 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. The sets of simulation results demonstrate the potential of RegBRLC models to improve on the two-stage approach in terms of both record linkage quality and causal inference accuracy. Finally, in Section 6, we conclude with a discussion.

2. Causal Inference and Record Linkage: An Overview

We first review a few key concepts and assumptions related to causal inference in Section 2.1. For ease of exposition, we review causal inference for data where record linkage is not needed, i.e., all relevant outcomes, covariates, and treatments are in the same file. We then review the Bayesian probabilistic record linkage model that we utilize in Section 2.2.

2.1. Causal Inference

For causal inference, we work in the potential outcomes framework [21]. Let $z = 1$ and $z = 0$ indicate assignment to the treatment and control conditions, respectively. Each unit has an outcome under treatment, $y(1)$, and an outcome under control, $y(0)$. For any unit, we observe only one of $y(1)$ and $y(0)$. The observed outcome for any unit can be written as $y = zy(1) + (1 - z)y(0)$. In this article, we work with binary outcomes $y(0), y(1) \in \{0, 1\}$. Typical of causal studies, we assume the stable unit treatment value assumption [SUTVA,

21] and strong ignorability [22]; that is, for some $p \times 1$ vector of covariates $\mathbf{x}$, we have $0 < P(z = 1 \mid \boldsymbol{x}) < 1$ and $(y(0), y(1)) \perp z \mid \mathbf{x}$.

Many causal inference procedures utilize propensity scores, defined as $e(\boldsymbol{x}) = P(z = 1 \mid \boldsymbol{x})$, i.e., the probability of being assigned a treatment given $\boldsymbol{x}$. As shown in [22], 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, define the conditional average controlled difference for a given $\boldsymbol{x}$, $\tau(\boldsymbol{x}) = E[y \mid z = 1, \boldsymbol{x}] - E[y \mid z = 0, \boldsymbol{x}]$. Under strong ignorability, $E[y(z) \mid \boldsymbol{x}] = E[y \mid \boldsymbol{x}, z]$, so that $\tau(\boldsymbol{x})$ becomes the average treatment effect conditional on $\boldsymbol{x}$, i.e., $\tau(\boldsymbol{x}) = E[y(1) - y(0) \mid \boldsymbol{x}]$. To complete the definition of the causal estimand, one averages $\tau(\boldsymbol{x})$ over some distribution of $\boldsymbol{x}$. The choice of distribution corresponds to the region of covariate space for the target population of interest. For example, to estimate the effect of the treatment on the treated, the relevant covariate distribution is for treated cases.

Let $f(\boldsymbol{x})$ be the marginal density of $\boldsymbol{x}$, defined with respect to a base measure $\Delta(\boldsymbol{x})$. 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,

$$\tau = \frac{\int \tau(\boldsymbol{x}) t(\boldsymbol{x}) f(\boldsymbol{x}) \Delta(d\boldsymbol{x})}{\int t(\boldsymbol{x}) f(\boldsymbol{x}) \Delta(d\boldsymbol{x})}. \tag{1}$$

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

$$\hat{\tau} = \frac{\sum_{i=1}^{n} w_{1i} z_i y_i}{\sum_{i=1}^{n} w_{1i} z_i} - \frac{\sum_{i=1}^{n} w_{0i}(1 - z_i) y_i}{\sum_{i=1}^{n} w_{0i}(1 - z_i)}. \tag{2}$$

In our example of RegBRLC, we estimate τ for the overlap population, described in [9] as the target population with the most overlap in covariate values for the treatment and control groups. Formally, the overlap population is implicitly defined by the overlap weights [9], which set $t(\boldsymbol{x}) = e(\boldsymbol{x})(1 - e(\boldsymbol{x}))$ for the covariate distribution defined by $g(\boldsymbol{x})$. The resulting estimator for the WATE for the overlap population, which we write as τ_O, is

$$\hat{\tau}_O = \frac{\sum_{i=1}^{n} (1 - e(\boldsymbol{x}_i)) z_i y_i}{\sum_{i=1}^{n} (1 - e(\boldsymbol{x}_i)) z_i} - \frac{\sum_{i=1}^{n} e(\boldsymbol{x}_i)(1 - z_i) y_i}{\sum_{i=1}^{n} e(\boldsymbol{x}_i)(1 - z_i)}. \tag{3}$$

The overlap weights have appealing features for causal inference. They are bounded, as $0 < e(\boldsymbol{x}_i) < 1$, and thus $\hat{\tau}_O$ is not affected by extreme weights. Unlike truncating or setting some weights to zero, the overlap weights are continuously defined and avoid arbitrary choices of cutoffs. Under mild conditions, the overlap weights leading to $\hat{\tau}_O$ minimize the asymptotic variance of the estimators of the form in (2) within the class of balancing weights [9].

A closed form variance estimator of $\hat{\tau}_O$ is provided in [23]. Let

$$\hat{\tau}_{O,1} = \frac{\sum_{i=1}^{n}(1-e(\boldsymbol{x}_i))z_i y_i}{\sum_{i=1}^{n}(1-e(\boldsymbol{x}_i))z_i}, \qquad \hat{\tau}_{O,0} = \frac{\sum_{i=1}^{n} e(\boldsymbol{x}_i)(1-z_i)y_i}{\sum_{i=1}^{n} e(\boldsymbol{x}_i)(1-z_i)}. \tag{4}$$

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

$$\hat{I}_i = z_i(y_i - \hat{\tau}_{O,1})(1-e(\boldsymbol{x}_i)) - (1-z_i)(y_i - \hat{\tau}_{O,0})e(\boldsymbol{x}_i) - (z_i - e(\boldsymbol{x}_i))\hat{\boldsymbol{H}}'\hat{\boldsymbol{E}}^{-1}\boldsymbol{x}_i \tag{5}$$

$$\hat{\boldsymbol{H}} = \sum_{i=1}^{n}[z_i(y_i - \hat{\tau}_{O,1}) + (1-z_i)(y_i - \hat{\tau}_{O,0})]e(\boldsymbol{x}_i)(1-e(\boldsymbol{x}_i))\boldsymbol{x}_i/n \tag{6}$$

$$\hat{\boldsymbol{E}} = \sum_{i=1}^{n} e(\boldsymbol{x}_i)(1-e(\boldsymbol{x}_i))\boldsymbol{x}_i\boldsymbol{x}_i'/n\,. \tag{7}$$

2.2. Record Linkage

We develop methodology for bipartite record linkage scenarios [1, 24]. Under this setting, each individual is recorded at most once within each file. 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 $n_A \geq n_B$ records, for which we measure a set of p_A additional causally-relevant covariates not in File B. 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. Finally, let $p = p_A + p_B$.

For any individual i, let $\boldsymbol{x}_i^{(A)} = \left(x_{i,1}^{(A)}, \ldots, x_{i,pA}^{(A)}\right)'$ and $\boldsymbol{x}_i^{(B)} = \left(x_{i,1}^{(B)}, \ldots, x_{i,pB}^{(B)}\right)'$ be the values of the covariates present in File A and File B, respectively; and, let y_i be the 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 [24], 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. For any record j in File B, let $d_j = i$ if record i in File A and record j in File B is a match, and $d_j = n_A + j$ if record j in File B has no match in File A. 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 now, assume that none of the covariates in $\boldsymbol{x}^{(A)}$ or $\boldsymbol{x}^{(B)}$ are used as linking variables. We discuss using a set of variables as both covariates and linking variables in Section 5 and Section 6. For each pair of records (i, j) in File A×File B, we define a 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 records i and j have the same value of field f, and $\gamma_{f,ij} = 0$ otherwise. One also can 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.

Following [1] and related work, we assume that $\boldsymbol{\gamma}_{ij}$ is a realization from a mixture of two distributions, one for true links and one for nonlinks. We have

$$\boldsymbol{\gamma}_{ij} \mid (d_j = i) \overset{iid}{\sim} g(\boldsymbol{\theta}_m), \qquad \boldsymbol{\gamma}_{ij} \mid (d_j \neq i) \overset{iid}{\sim} g(\boldsymbol{\theta}_u), \tag{8}$$

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. With binary fields, we compute

$$g(\boldsymbol{\theta}_m) = P(\boldsymbol{\gamma}_{ij} \mid d_j = i) = \prod_{f=1}^{F} P(\gamma_{f,ij} \mid d_j = i) = \prod_{f=1}^{F} \theta_{f,m}^{\gamma_{f,ij}} (1 - \theta_{f,m})^{1 - \gamma_{f,ij}} \tag{9}$$

$$g(\boldsymbol{\theta}_u) = P(\boldsymbol{\gamma}_{ij} \mid d_j \neq i) = \prod_{f=1}^{F} P(\gamma_{f,ij} \mid d_j \neq i) = \prod_{f=1}^{F} \theta_{f,u}^{\gamma_{f,ij}} (1 - \theta_{f,u})^{1 - \gamma_{f,ij}}. \tag{10}$$

This model implies that the linking fields are independent of the outcomes, treatments, and covariates. This is commonly assumed in record linkage settings, although it is possible to make the distributions depend on some variables [25].

To specify a prior distribution on $\boldsymbol{d}$ with the constraint $d_j \neq d_{j'}$ for any $j \neq j'$, we follow an approach described in, for example, [24, 26, 27]. Let $I(\mathscr{E})$ represent the indicator for an event $\mathscr{E}$. We assume $I(d_j \leq n_A) \sim \text{Bernoulli}(\pi)$, where π represents the proportion of matches expected *a priori* as a fraction of the smaller file. We assume $\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi)$. The hyperparameters α_π and β_π provide prior information on the number of intersecting records in

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 specific choices of α_π, β_π, a and b in Section 3.1.

3. The RegBRLC Model

RegBRLC requires 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 depends 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 $(y_j, z_j, \boldsymbol{x}_j^{(B)} \mid \boldsymbol{x}_i^{(A)})$, whereas for a record j in File B not linked to any record in File A, we specify the joint distribution of $(y_j, z_j, \boldsymbol{x}_j^{(B)})$. More precisely, when any record j in File B is linked to record i in File A, we specify the conditional distribution of the outcome denoted as $f_1(y_j \mid \boldsymbol{x}_i^{(A)}, z_j, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{ym})$, the propensity score denoted as $g_1(z_j \mid \boldsymbol{x}_j^{(B)}, \boldsymbol{x}_i^{(A)}, \boldsymbol{\theta}_{zm})$, and the covariates denoted as $h_1(\boldsymbol{x}_j^{(B)} \mid \boldsymbol{x}_i^{(A)}, \boldsymbol{\theta}_{xm})$. For any record j in File B that does not have a link in File A and hence is missing $\boldsymbol{x}_i^{(A)}$, we specify an outcome model $f_2(y_j \mid \boldsymbol{x}_j^{(B)}, z_j, \boldsymbol{\theta}_{yu})$, a propensity score model $g_2(z_j \mid \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zu})$, and and a marginal distribution $h_2(\boldsymbol{x}_j^{(B)} \mid \boldsymbol{\theta}_{xu})$. Throughout, we assume that the covariates in File A are fixed quantities and thus do not require specified probability distributions.

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

$$L_j^{AB} = \begin{cases} f_1(y_j \mid \boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, z_j, \boldsymbol{\theta}_{ym}) g_1(z_j \mid \boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zm}) h_1(\boldsymbol{x}_j^{(B)} \mid \boldsymbol{x}_i^{(A)}, \boldsymbol{\theta}_{xm}), & \text{for } d_j = i \le n_A \\ f_2(y_j \mid \boldsymbol{x}_j^{(B)}, z_j, \boldsymbol{\theta}_{yu}) g_2(z_j \mid \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zu}) h_2(\boldsymbol{x}_j^{(B)} \mid \boldsymbol{\theta}_{xu}), & \text{for } d_j = n_A + j. \end{cases} \tag{11}$$

The likelihood including the contributions from (11) and the linkage model in (8)–(10) is

$$L\left(\boldsymbol{\theta}_{ym}, \boldsymbol{\theta}_{zm}, \boldsymbol{\theta}_{xm}, \boldsymbol{\theta}_{yu}, \boldsymbol{\theta}_{zu}, \boldsymbol{\theta}_{xu}, \boldsymbol{\theta}_m, \boldsymbol{\theta}_u, \boldsymbol{d} \mid \{\boldsymbol{\gamma}_{ij} : 1 \le i \le n_A, 1 \le j \le n_A\}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X}^{(A)}, \boldsymbol{X}^{(B)}\right) \propto$$
$$\prod_{\substack{(i,j): \\ d_j = i}} L_j^{AB} \prod_{\substack{j: d_j \ne i \\ \forall i}} L_{jj}^{AB} \prod_{i,j} \left\{ \prod_{f=1}^{F} \theta_{f,m}^{\gamma_{f,ij}} (1 - \theta_{f,m})^{1 - \gamma_{f,ij}} \right\}^{I(d_j = i)} \left\{ \prod_{f=1}^{F} \theta_{f,u}^{\gamma_{f,ij}} (1 - \theta_{f,u})^{1 - \gamma_{f,ij}} \right\}^{I(d_j \ne i)}$$
$$\times I(d_j \ne d_{j'}, \text{ whenever } j \ne j'). \tag{12}$$

This modeling strategy is sufficiently general to incorporate many choices of f_1, f_2, g_1, g_2, h_1, and h_2. To illustrate and to facilitate simulation studies, we present a specific choice of these distributions in Section 3.1. We use outcome and covariate models based on normal distributions, and a propensity score model based on a logistic regression. In Section 3.3, we discuss some strategies for simplifying model specifications and for model checking.

Before we present the illustrative model specification, it is worth emphasizing the purpose of the outcome and covariate models: they are intended to improve the quality of the linkages.

By leveraging relationships among variables across the files, we hope to find more reliable links. Once we have plausibly linked records, we use a causal estimator that can be specified independently of the models used in (12), as discussed in Section 3.2.

3.1. Illustrative Specification

For the illustrative outcome model, we use linear regressions for f_1 and f_2 so that

$$y_j = \alpha_{ym}^{(0)} + z_j\alpha_{ym}^{(1)} + \boldsymbol{x}_i^{(A)'}\boldsymbol{\alpha}_{ym}^{(2)} + \boldsymbol{x}_j^{(B)'}\boldsymbol{\alpha}_{ym}^{(3)} + \epsilon_j, \qquad \epsilon_j \sim N\left(0, \sigma_m^2\right) \tag{13}$$

for records with links, and

$$y_j = \alpha_{yu}^{(0)} + z_j\alpha_{yu}^{(1)} + \boldsymbol{x}_j^{(B)'}\boldsymbol{\alpha}_{yu}^{(2)} + \epsilon_j, \qquad \epsilon_j \sim N\left(0, \sigma_u^2\right) \tag{14}$$

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

For the illustrative propensity score model, we use logistic regressions for g_1 and g_2 with

$$Pr\left(z_j = 1 \mid \boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zm}\right) = \frac{\exp\left(\alpha_{zm}^{(0)} + \boldsymbol{x}_i^{(A)'}\boldsymbol{\alpha}_{zm}^{(1)} + \boldsymbol{x}_j^{(B)'}\boldsymbol{\alpha}_{zm}^{(2)}\right)}{1 + \exp\left(\boldsymbol{\alpha}_{zm}^{(0)} + \boldsymbol{x}_i^{(A)'}\boldsymbol{\alpha}_{zm}^{(1)} + \boldsymbol{x}_j^{(B)'}\boldsymbol{\alpha}_{zm}^{(2)}\right)} \tag{15}$$

for records with links, where $\boldsymbol{\theta}_{zm} = \left(\alpha_{zm}^{(0)}, \boldsymbol{\alpha}_{zm}^{(1)'}, \boldsymbol{\alpha}_{zm}^{(2)'}\right)'$. We use

$$Pr\left(z_j = 1 \mid \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zu}\right) = \frac{\exp\left(\alpha_{zu}^{(0)} + \boldsymbol{x}_j^{(B)'}\boldsymbol{\alpha}_{zu}^{(1)}\right)}{1 + \exp\left(\alpha_{zu}^{(0)} + \boldsymbol{x}_j^{(B)'}\boldsymbol{\alpha}_{zu}^{(1)}\right)} \tag{16}$$

for records without links, where $\boldsymbol{\theta}_{zu} = \left(\alpha_{zu}^{(0)}, \boldsymbol{\alpha}_{zu}^{(1)'}\right)'$.

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

$$\boldsymbol{x}_j^{(B)'} = \boldsymbol{\eta}_{xm}^{'} + \boldsymbol{x}_i^{(A)'}\boldsymbol{B}_{xm} + \boldsymbol{\epsilon}_{ij}, \boldsymbol{\epsilon}_{ij} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{xm}), \tag{17}$$

where $\boldsymbol{B}_{xm}$ is a $p_A \times p_B$ matrix, $\boldsymbol{\eta}_{xm}$ is a $p_B \times 1$ 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 $\boldsymbol{\mu}_{xu}$ and covariance matrix $\boldsymbol{\Sigma}_{xu}$.

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_{xm}^2 and σ_{xu}^2 i.i.d. Inverse-Gamma (a_σ, b_σ) priors. For the covariate

model, we set $\Pi(\boldsymbol{B}_{xm}, \boldsymbol{\Sigma}_{xm}) = \Pi_1(\boldsymbol{B}_{xm} \mid \boldsymbol{\Sigma}_{xm})\Pi_2(\boldsymbol{\Sigma}_{xm})$, where $\boldsymbol{B}_{xm} \mid \boldsymbol{\Sigma}_{xm}$ follows a matrix normal distribution $\mathcal{MN}_{p_A, p_B}(\mathbf{0}, \boldsymbol{I}, \boldsymbol{\Sigma}_{xm})$ and $\boldsymbol{\Sigma}_{xm}$ follows an IW(ν, $\boldsymbol{I}$) prior, where IW(ν, $\boldsymbol{I}$) denotes an Inverse-Wishart prior with parameters ν and the identity matrix. The prior specification is completed by assigning an IW(ν, 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. Thus, posterior computation can proceed through a MCMC algorithm. Details of the full conditional distributions for the illustrative model are provided in the supplementary material.

The MCMC sampling also offers inferences on the record linkages. For $j = 1, \ldots, n_B$, let $\left(d_j^{(1)}, \ldots, d_j^{(L)}\right)$ be the L post burn-in MCMC iterates of d_j. For each j, we empirically estimate $P(d_j = q \mid -)$ using the proportion of samples where $d_j = q$, i.e., $\widehat{P}(d_j = q \mid -) = \#\left\{l : d_j^{(l)} = q\right\}/L$, for $q \in \mathcal{J}_j = \{1, \ldots, n_A, n_A + j\}$. When $1 \leq q^* = arg\ \max_{q \in \mathcal{J}_j} \widehat{P}(d_j = q \mid -) \leq n_A$, we conclude that the record q^* in File A is the most likely match for the record j in File B; denote this $\widehat{d}_j = q^*$. 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, one can estimate the posterior probability of linkage between any record pair. See Sadinle [24] for further discussion of using the posterior probabilities to determine links.

3.2. Defining the Causal Estimand and Overlap Weights Estimator

The plausibly linked files also provide means to estimate a WATE. Let $m_j = 1$ when the identity of record j in File B with values $(y_j, z_j, \boldsymbol{x}_j^{(B)})$ matches the identity of some record in File A, and let $m_j = 0$ when record j does not have a true match in File A. Thus, when $m_j = 1$, the record's full data vector $(y_j, z_j, \boldsymbol{x}_j^{(B)}, \boldsymbol{x}_j^{(A)})$ is potentially observable with accurate record linkage. To define the WATE, we assume SUTVA and a version of strong ignorability for all records in File B, namely $(y(0), y(1)) \perp z \mid \boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)}, m$ for any generic record in File B with a link in File A. Under SUTVA and this version of strong ignorability, we have $E\left[y(z) \mid \boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)}, m = 1\right] = E\left[y \mid \boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)}, z, m = 1\right]$, so that $\tau\left(\boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)}\right)$ becomes the average treatment effect conditional on the covariates $(\boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)})$ for linked records, i.e., $\tau\left(\boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)}\right) = E\left[y(1) - y(0) \mid \boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)}, m = 1\right]$. In actuality, we require strong ignorability only for records with true links ($m = 1$). However, in practical settings it seems a stretch to claim that strong ignorability holds only for the subset of linkable records but not for others.

Following Section 2.1, to define the WATE we need to average over a target population. We use the overlap population among records that can be linked, and denote the WATE

for this target population as $\tau_{O,linked}$. When each record in File B has a link in File A, $\tau_{O,linked}$ is defined over the full study population in File B. When some records in File B do not have links in File A, $\tau_{O,linked}$ is defined over a subset of the study population in File B. 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. The expression for $\tau_{O,linked}$ can be obtained by letting $t(\boldsymbol{x}) = e\left(\boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)}\right)\left(1 - e\left(\boldsymbol{x}^{(A)}, \boldsymbol{x}^{(B)}\right)\right)m\left(\boldsymbol{x}^{(B)}\right)$ in (1), where $m\left(\boldsymbol{x}^{(B)}\right) = 1$ when the record corresponding to covariate $\boldsymbol{x}^{(B)}$ in File B is linkable and $m\left(\boldsymbol{x}^{(B)}\right) = 0$ when it is not linkable. Substituting this new $t(\boldsymbol{x})$ into (2), we estimate $\tau_{O,linked}$ by

$$\hat{\tau}_{O,linked} = \left(\frac{\sum(i,j) \in \mathcal{M}\left(1 - e\left(\boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}\right)\right)z_j y_j}{\sum(i,j) \in \mathcal{M}\left(1 - e\left(\boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}\right)\right)z_j}\right) - \left(\frac{\sum(i,j) \in \mathcal{M}\, e\left(\boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}\right)(1 - z_j)y_j}{\sum(i,j) \in \mathcal{M}\, e\left(\boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}\right)(1 - z_j)}\right), \tag{18}$$

where $\mathcal{M}$ is the set of all linked records in File A and File B.

An important question is when we can generalize $\tau_{O,linked}$ to treatment effects for broader populations. Here, we focus on generalizing to τ_O. As described in Section 2.1, τ_O is defined on the overlap population among the records in File B and computed with the complete $\boldsymbol{x}$ for all individuals. Of course, in our setting we do not observe this 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 $\boldsymbol{x}$ is the same for linkable and non-linkable records; that is, when $g(\boldsymbol{x} \mid m(\boldsymbol{x}) = 1) = g(\boldsymbol{x})$ for the full overlap population. A special case of this scenario arises when all records in the full overlap population are linkable. Another arises when true linkage status is independent of covariates. This can arise, for example, when the linking variables comprise string metrics (like names) that are erroneously recorded at random. We also can generalize $\tau_{O,linked} = \tau_O$ in the case where $\tau(\boldsymbol{x}) = \tau$ for all $\boldsymbol{x}$ in File B. 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 [28].

In practice we do not know and must estimate which records have links. We consider MCMC iterations (suitably thinned) as draws of plausible linkages and therefore providing estimates of $\tau_{O,linked}$. 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., $\mathcal{M}^{(l)} = \left\{(i,j): d_j^{(l)} = i, i \leq n_A\right\}$. 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 $\hat{e}_{i,j}^{(l)} = e\left(\boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{zm}^{(l)}\right)$ Following (3), define the l-th post burn-in iterate of $\hat{\tau}_{O,linked}$ as

$$\hat{\tau}_{O,linked}^{(l)} = \left[\left(\frac{\sum(i,j) \in \mathcal{M}^{(l)}\left(1 - \hat{e}_{i,j}^{(l)}\right)z_j y_j}{\sum(i,j) \in \mathcal{M}^{(l)}\left(1 - \hat{e}_{i,j}^{(l)}\right)z_j}\right) - \left(\frac{\sum(i,j) \in \mathcal{M}^{(l)}\, \hat{e}_{i,j}^{(l)}(1 - z_j)y_j}{\sum(i,j) \in \mathcal{M}^{(l)}\, \hat{e}_{i,j}^{(l)}(1 - z_j)}\right)\right]. \tag{19}$$

We compute $\left(\hat{\tau}_{O,linked}^{(1)}, \ldots, \hat{\tau}_{O,linked}^{(L)}\right)$ and use $\bar{\tau}_{O,linked} = \sum_{l=1}^{L} \hat{\tau}_{O,linked}^{(l)}/L$ as the point estimator of $\tau_{O,linked}$.

Draws of $\hat{\tau}_{O,linked}$ vary only because of different linkages across MCMC iterations; the variation in these drawn values does not reflect inherent sampling variability. For example, even if all true links were known perfectly, we still have uncertainty in the estimated WATE. We account for this sampling variability and the variability from estimating the missing true links using multiple imputation. More specifically, to estimate the variance of $\bar{\tau}_{O,linked}$, we use multiple imputation formulae with all L iterates [29], computing

$$\widehat{\mathrm{Var}}(\bar{\tau}_{O,linked}) = \frac{\sum_{l=1}^{L} U_{O,linked}^{(l)}}{L} + \left(1 + \frac{1}{L}\right)\frac{\sum_{l=1}^{L}\left(\hat{\tau}_{O,linked}^{(l)} - \bar{\tau}_{O,linked}\right)^2}{L-1}. \tag{20}$$

Each $U_{O,linked}^{(l)}$ is computed using (5), plugging in the values from the linked data in the l-th iterate. Assuming large L, inferences are based on a normal distribution with mean $\bar{\tau}_{O,linked}$ and variance $\widehat{\mathrm{Var}}(\bar{\tau}_{O,linked})$

The modeling framework allows analysts to use other causal estimators with the plausibly linked data files. To illustrate this flexibility in the simulation studies, we now present a regression-adjusted estimator of $\tau_{O,linked}$ based on the overlap weights. As a regression-adjusted, overlap weights estimator has not been discussed previously in the literature, we discuss some of its properties with perfectly linked data in the supplementary material.

Suppose we have a model for the outcome; for illustrative purposes, we use the model in (13). Let the mean function of the outcome under the model, evaluated at the l-th MCMC iterate of the parameters, be $\hat{\mu}_{i,j}^{(l)}(\zeta) = \mu\left(z_j = \zeta, \boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{ym}^{(l)}\right)$, 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 $\hat{R}_{i,j}^{(l)} = y_j - \hat{\mu}_{i,j}^{(l)}\left(z_j, \boldsymbol{x}_i^{(A)}, \boldsymbol{x}_j^{(B)}, \boldsymbol{\theta}_{ym}^{(l)}\right)$. The regression-adjusted estimator for the l-th iteration is defined as

$$\hat{\tau}_{O,linked,reg}^{(l)} = \left\{\frac{\sum_{(i,j)\in\mathcal{M}^{(l)}}\left(1-\hat{e}_{i,j}^{(l)}\right)z_j\hat{R}_{l,j}^{(l)}}{\sum_{(i,j)\in\mathcal{M}^{(l)}}\left(1-\hat{e}_{i,j}^{(l)}\right)z_j} - \frac{\sum_{(i,j)\in\mathcal{M}^{(l)}}\hat{e}_{l,j}^{(l)}(1-z_j)\hat{R}_{i,j}^{(l)}}{\sum_{(i,j)\in\mathcal{M}^{(l)}}\hat{e}_{i,j}^{(l)}(1-z_j)}\right\} + \frac{\sum_{(i,j)\in\mathcal{M}^{(l)}}\left(\hat{\mu}_{i,j}^{(l)}(1) - \hat{\mu}_{i,j}^{(l)}(0)\right)\hat{e}_{i,j}^{(l)}\left(1-\hat{e}_{i,j}^{(l)}\right)}{\sum_{(i,j)\in\mathcal{M}^{(l)}}\hat{e}_{i,j}^{(l)}\left(1-\hat{e}_{i,j}^{(l)}\right)}. \tag{21}$$

We compute $\left(\hat{\tau}_{O,linked,\ reg}^{(1)}, \ldots, \tilde{\tau}_{O,linked,\ reg}^{(L)}\right)$ and use $\bar{\tau}_{O,linked,reg} = \sum_{l=1}^{L}\hat{\tau}_{O,linked,reg}^{(l)}/L$ as the new estimator of $\tau_{O,linked}$.

To estimate the variance of $\bar{\tau}_{O,linked,reg}$, we use multiple imputation and compute

$$\widehat{\mathrm{Var}}(\bar{\tau}_{O,linked,reg}) = \frac{\sum_{l=1}^{L} U_{O,linked,reg}^{(l)}}{L} + \left(1 + \frac{1}{L}\right) \frac{\sum_{l=1}^{L} \left(\hat{\tau}_{O,linked,reg}^{(l)} - \bar{\tau}_{O,linked,reg}\right)^2}{L-1}. \quad (22)$$

We present the expression for $U_{O,linked,reg}^{(l)}$ and its derivation in the supplementary material. We use normal-based inferences for $\tau_{O,linked}$ with the mean from (21) and variance from (22).

3.3. Useful Modeling Simplifications

Using all the conditional distributions in (11) 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 $y \mid \boldsymbol{x}$ from (11). Thus, analysts who make this assumption need not specify a model for y when obtaining draws of $\left(d_j^{(1)}, \ldots, d_j^{(L)}\right)$, for $j = 1, \ldots, n_B$. This accords with the "design-first" philosophy of causal inference [30], 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 covariates in File A, the proposed 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 $\boldsymbol{x}^{(B)} \mid \boldsymbol{x}^{(A)}$ from (11) 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 covariate 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 to be certain links, when such pairs are available. For example, one can estimate the posited outcome, propensity score, and covariate models using 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 with 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, one can use record pairs that have high probability of being links according to off-the-shelf probabilistic record linkage algorithms like implementations of [1].

Another model checking tool is to generate replicate datasets from the posited RegBRLC model using draws of model parameters [31]. 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, the outcome model might be improved.

4. Simulation Studies

We illustrate the performance of RegBRLC modeling using repeated sampling simulations with the model in Section 3.1. For simplicity, we assume that both File A and File B 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 in which $p_A \neq p_B$ and both the propensity score and outcome models include unimportant predictors, as well as a simulation where the fitted outcome model is poorly specified, including an illustrative comparison with the approach from [20].

4.1. Simulated Data Generation

We work with the RLdata10000 data from the R package RecordLinkage [32]. These data comprise an artificial population of 10000 records with birth years, birth months, birth dates, first names and last names. Among these 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. 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 equal to 0.2. 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 covariance also equal to 0.2. This represents a modest amount of correlation among the predictors. We note that this generates covariates independently of the linkage variables.

We simulate each z_k from a Bernoulli distribution such that

$$P(z_k = 1 \mid \boldsymbol{x}_k) = \frac{e^{\alpha_0^{(0)}} + \sum_{l=1}^{p} \alpha_l^{(0)} x_{l,k}}{\left(1 + e^{\alpha_0^{(0)}} + \sum_{l=1}^{p} \alpha_l^{(0)} x_{l,k}\right)}, \tag{23}$$

where $\left(\alpha_0^{(0)}, \alpha_1^{(0)}, \alpha_2^{(0)}, \alpha_3^{(0)}, \alpha_4^{(0)}\right) = (1, 1.5, -1, 2, -3)$. The superscript 0 emphasizes that the parameter value is from the true data generating mechanism. We generate each y_k from

$$y_k = \beta_0^{(0)} + \sum_{l=1}^{p} \beta_l^{(0)} x_{l,k} + \beta_C^{(0)} z_k + \epsilon_k, \qquad \epsilon_k \overset{i.i.d}{\sim} N\left(0, \sigma^{(0)2}\right), \tag{24}$$

where $\left(\beta_0^{(0)}, \beta_1^{(0)}, \beta_2^{(0)}, \beta_3^{(0)}, \beta_4^{(0)}\right) = (1, -1, 2, -3, -2)$. 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 (24) 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 [33]. For ease of simulation, we set the sizes of File A and File B to be $n_A = n_B = 1000$, although RegBRLC in general 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.

We let the MCMC chains run for 2000 iterations. We discard the first 1500 as burn-in, and make inferences 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 models and the propensity score models for the linked and unlinked data, which suggest 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 RegBRLC with estimators from a two-stage approach 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 (19) and (21). As the point estimates, we compute $\bar{\tau}_{O,linked}$ and $\bar{\tau}_{O,linked,reg}$ using the L linked databases. We also estimate their variances based on (20) and (22). Since this model links the files without using information on the outcomes, treatments, and covariates, comparisons, it can reveal if the sharing of information between the record linkage and study variable models offers benefits. We refer to this model as BRLC.

We use 100 independent simulations runs to compare the performances of RegBRLC and BRLC in terms of both causal inference and record linkage. 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 $\hat{\boldsymbol{d}}$ 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 $\mathscr{A}_j = \left\{\hat{d}_j = d_j, \hat{d}_j \leq n_A\right\}$. The precision is defined as $\sum_{j=1}^{n_B} I(\mathscr{A}_j) / \sum_{j=1}^{n_B} I\left(\hat{d}_j \leq n_A\right)$. The recall is the proportion of actual matches that are determined as links, $\sum_{j=1}^{n_B} I(\mathscr{A}_j) / \sum_{j=1}^{n_B} I(d_j \leq n_A)$. 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 $\bar{\tau}_{O,linked}$ and $\bar{\tau}_{O,linked,reg}$ over the 100 replications for both RegBRLC and BRLC. We also present the empirical coverage rates of multiple imputation 95% confidence intervals (based on 100 replications) for $\tau_{O,linked}$. 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 a baseline to assess the accuracy lost due to imperfect linkages. As a side 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

Figure 1 displays averages of the precision and recall for various simulation scenarios. For the three scenarios with $\sigma^{(0)2} = 1$, we observe a modest increase in precision and a sharp increase in recall as the number of intersecting records increases for both RegBRLC and BRLC. In all three scenarios, RegBRLC dominates BRLC, with higher average precision and recall. The differences in average recall are substantial and grow with the number of intersecting records in the two files. Evidently, RegBRLC 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 RegBRLC over BRLC in terms of record linkage has a positive impact on the estimation of the causal effect. The first three rows of Table 1 display properties of $\bar{\tau}_{O,linked}$ and $\bar{\tau}_{O,linked,reg}$ as applied for RegBRLC, BRLC, and the perfectly linked records when $\sigma^{(0)2} = 1$. For both estimators, RegBRLC accurately estimates the true $\tau_{O,linked} = 5$ in all scenarios, with greatest deviation for the scenario with only 20% intersection between two files. In contrast, BRLC significantly underestimates $\tau_{O,linked}$ in all three scenarios. RegBRLC has smaller empirical standard deviations than BRLC. The empirical standard deviations also reveal the cost of imperfect linkages. They are higher

for RegBRLC and BRLC 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 $\bar{\tau}_{O,linked}$ compared to $\bar{\tau}_{O,linked,reg}$, 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 (24). Here, we perform simulation studies only for the 80% intersecting records scenario, as this scenario gives each model its best chance to perform effectively. Comparing all 80% intersection scenarios in Figure 1, we find that the precision and recall values decline for RegBRLC as $\sigma^{(0)2}$ increases. As the predictive power of the covariates weakens, the outcome model offers increasingly less information about the correct linkages. For BRLC, the average precision and recall values are unchanged (other than by small Monte Carlo errors) when changing $\sigma^{(0)2}$. This is expected, since the record linkage in BRLC is done independently of the outcomes, treatments, and covariates. Overall, RegBRLC exhibits better performance than BRLC on recall and similar performance on precision.

The last two rows of Table 1 display the simulation results for $\bar{\tau}_{O,linked}$ and $\bar{\tau}_{O,linked,reg}$ in these scenarios with larger $\sigma^{(0)2}$. RegBRLC continues to estimate the causal effect accurately, although with increased standard deviations, as expected. In comparison, BRLC continues to underestimate the causal effect. For these two larger values of $\sigma^{(0)2}$, the empirical standard deviations for $\bar{\tau}_{O,linked,reg}$ tend to be smaller than those for $\bar{\tau}_{O,linked}$.

We next turn to the coverage rates for the multiple imputation 95% confidence intervals. For RegBRLC, in all but the 20% intersection scenario, the intervals based on $\bar{\tau}_{O,linked}$ cover in 100% of the replications; the 20% intersection scenario has a coverage rate of 99%. The consistent over-coverage occurs because, in these simulations, the distribution of $\hat{\tau}_{O,linked}$ is platykurtic rather than normally distributed. The coverage rates for the intervals based on $\bar{\tau}_{O,linked,reg}$ 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 $\sigma^{(0)2} = 16$. In contrast, the intervals for BRLC demonstrate substantial under-coverage, never rising above 41%. For both models, the intervals based on $\bar{\tau}_{O,linked}$ tend to be wider than those based on $\bar{\tau}_{O,linked,reg}$. 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. Let $\hat{\tau}_O$ and $\hat{\tau}_{O,\ reg}$ be the unadjusted and regression-adjusted estimators of τ_O based on perfectly linked data. For the five scenarios, the coverage rates when using $\hat{\tau}_O$ based on the perfectly linked data are (97%, 96%, 96%, 98%, 98%), respectively. And, the coverage rates when using $\hat{\tau}_{O,\ reg}$ based on the perfectly linked data are (95%, 96%, 96%, 98%, 98%), respectively. For the perfectly linked data, the intervals based on $\hat{\tau}_O$ again tend to be wider than those based on $\hat{\tau}_{O,\ reg}$.

5. Illustration with Causal Study of Debit Cards

To illustrate regression-assisted Bayesian record linkage with causal inference further, we follow Guha et al. [20] 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 [20], our description of the data follows theirs.

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. The SHIW collects information on Italian households' economic and financial behavior. We link two files with data collected during 1995 and 1998. Some households participated in both years and some did not. Our study 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 negative values of the outcome, or 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 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. [20] 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.

We implement RegBRLC using a simplification from Section 3.3, namely that $\boldsymbol{x}^{(B)}$ is independent of $\boldsymbol{x}^{(A)}$. 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 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 them lead to practically indistinguishable results. We let the MCMC chains run for 2000 iterations and discard the first 1500 as burn-in. We also examine results for BRLC and results using the perfectly linked data for comparisons.

Table 2 presents the precision and recall values, along with the multiple imputation means and 95% confidence intervals using $\bar{\tau}_{O,linked}$ and $\bar{\tau}_{O,linked,reg}$ (in thousand Italian Liras) for all methods. Consistent with the simulation results, RegBRLC offers better precision and recall than BRLC. Using results from the perfect-links model as a benchmark, we find that RegBRLC more closely tracks the mean treatment effect estimates from the perfectlinks model than BRLC does. This also holds for the 95% confidence intervals, particularly for $\bar{\tau}_{O,linked,reg}$, although the differences arguably are modest. The estimated variance of $\bar{\tau}_{O,linked,reg}$ is smaller than the estimated variance of $\bar{\tau}_{O,linked}$ across all three methods, reflecting the benefits of using the regression-adjusted estimator. It should be noted that the point estimates from both RegBRLC and BRLC models differ from those for the perfect-links model, reflecting the effects of inevitably imperfect linkages.

These results suggest that, on average, possession of a single debit card for a household leads to more monthly consumption than not possessing any debit card, during the study period. Similar findings are presented by [34].

6. Conclusion and Future Work

The empirical studies suggest that regression-assisted Bayesian record linkage with causal inference modeling strategies 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. Although we assumed $n_A \geq n_B$, we conjecture that these findings will hold when $n_A < n_B$, albeit with a different joint likelihood function tailored to the data setting.

The RegBRLC 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 sampling. 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.

RegBRLC models can be 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 [25]. By allowing duplicate matchings, the linkage steps can be done for each observation in parallel, thereby speeding computation significantly.

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 RegBRLC framework. In such cases it may make sense to view these as blocking variables—the analyst requires record pairs to match exactly on the blocking variables [35, 36]—rather than use them as linkage variables. When these variables are not recorded identically across files, the path forward to using RegBRLC models is less clear, as the true value of the covariate is unobserved. In the SHIW application, we used the values in one of the files, File A, as the 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

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors wish to thank Dr. Andrea Mercatanti for generously sharing the SHIW debit card data utilized in this manuscript.

References

[1]. Fellegi IP, Sunter AB, A theory for record linkage, Journal of the American Statistical Association 64 (1969) 1183–1210.

[2]. Hernandez AF, Greiner MA, Fonarow GC, Hammill BG, Heidenreich PA, Yancy CW, Peterson ED, Curtis LH, Relationship between early physician follow-up and 30-day readmission among Medicare beneficiaries hospitalized for heart failure, Journal of the American Medical Association 303 (2010) 1716–1722. [PubMed: 20442387]

[3]. Gutman R, Afendulis CC, Zaslavsky AM, A Bayesian procedure for file linking to analyze end-of-life medical costs, Journal of the American Statistical Association 108 (2013) 34–47. [PubMed: 23645944]

[4]. Dalzell NM, Reiter JP, Regression modeling and file matching using possibly erroneous matching variables, Journal of Computational and Graphical Statistics 27 (2018) 728–738.

[5]. Steorts RC, Tancredi A, Liseo B, Generalized Bayesian record linkage and regression with exact error propagation, in: Domingo-Ferrer J, Montes F. (Eds.), Privacy in Statistical Databases, Springer, 2018, pp. 297–313.

[6]. Tang J, Reiter JP, Steorts RC, Bayesian modeling for simultaneous regression and record linkage, in: International Conference on Privacy in Statistical Databases, Springer, 2020, pp. 209–223.

[7]. Rubin DB, Multiple Imputation for Survey Nonresponse, New York: Wiley, 1987.

[8]. Hirano K, Imbens GW, Ridder G, Efficient estimation of average treatment effects using the estimated propensity score, Econometrica 71 (2003) 1161–1189.

[9]. Li F, Morgan KL, Zaslavsky AM, Balancing covariates via propensity score weighting, Journal of the American Statistical Association 113 (2018) 390–400.

[10]. Scheuren F, Winkler WE, Regression analysis of data files that are computer matched, Survey Methodology 19 (1993) 39–58.

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

[12]. Lahiri P, Larsen MD, Regression analysis with linked data, Journal of the American Statistical Association 100 (2005) 222–230.

[13]. Chipperfield JO, Bishop G, Campbell PD, et al. , Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data, Survey Methodology 37 (2011) 13–24.

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

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

[16]. Ventura SL, Nugent R, Hierarchical linkage clustering with distributions of distances for large-scale record linkage, in: Domingo-Ferrer J, Montes F. (Eds.), Privacy in Statistical Databases, Springer, 2014, pp. 283–298.

[17]. Solomon NC, A Framework for Decision Threshold Selection in Record Linkage, Ph.D. thesis, Duke University, 2019.

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

[19]. Heck Wortman J, Reiter JP, Simultaneous record linkage and causal inference with propensity score subclassification, Statistics in Medicine 37 (2018) 3533–3546. [PubMed: 30069901]

[20]. Guha S, Reiter JP, Mercatanti A, Bayesian causal inference with bipartite record linkage, Bayesian Analysis 17 (2022) 1275–1299.

[21]. Rubin DB, Estimating causal effects of treatments in randomized and nonrandomized studies., Journal of Educational Psychology 66 (1974) 688.

[22]. Rosenbaum PR, Rubin DB, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1983) 41–55.

[23]. Li F, Thomas LE, Li F, Addressing extreme propensity scores via the overlap weights, American Journal of Epidemiology 188 (2019) 250–257. [PubMed: 30189042]

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

[25]. Heck Wortman J, Record Linkage Methods with Applications to Causal Inference and Elections Voting Data, Ph.D. thesis, 2018.

[26]. Fortini M, Nuccitelli A, Liseo B, Scanu M, Modelling issues in record linkage: a Bayesian perspective, in: Proceedings of the American Statistical Association, Survey Research Methods Section, 2002, pp. 1008–1013.

[27]. Larsen MD, Record linkage modeling in federal statistical databases, in: Proceedings of the 2009 FCSM Research Conference, 2010.

[28]. Hill JL, Bayesian nonparametric modeling for causal inference, Journal of Computational and Graphical Statistics 20 (2011) 217–240.

[29]. Hu J, Mitra R, Reiter JP, Are independent parameter draws necessary for multiple imputation?, The American Statistician 67 (2013) 143–149.

[30]. Rubin DB, For objective causal inference, design trumps analysis, Annals of Applied Statistics 2 (2008) 808–840.

[31]. Fosdick BK, De Yoreo M, Reiter JP, Categorical data fusion using auxiliary information, Annals of Applied Statistics 10 (2016) 1907–1929.

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

[33]. Jaro MA, Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, Journal of the American Statistical Association 84 (1989) 414–420.

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

[35]. Christen P, A survey of indexing techniques for scalable record linkage and deduplication, IEEE Transactions on Knowledge and Data Engineering 24 (2011) 1537–1555.

[36]. Herzog TN, Scheuren FJ, Winkler WE, Data Quality and Record Linkage Techniques, Springer Science & Business Media, 2007.

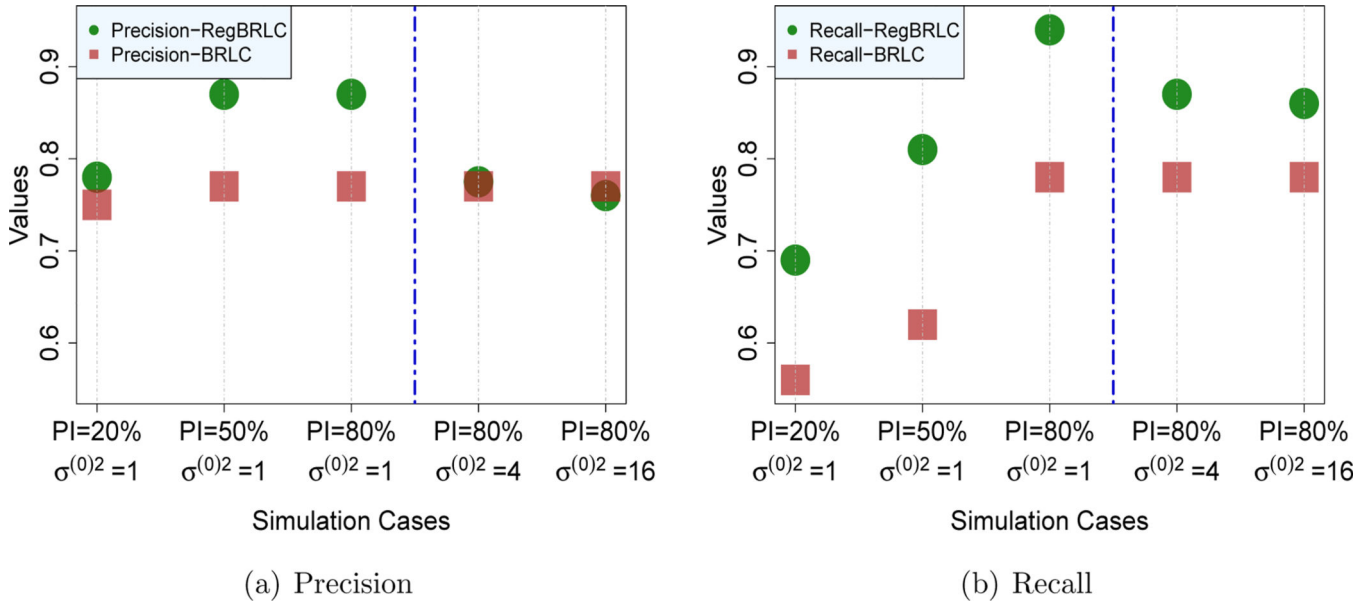

Figure 1:

Simulated average precision and recall values for RegBRLC and BRLC over 100 replications of each scenario. Scenarios vary the number of intersecting records in the two files or the outcome model variance $\sigma^{(0)2}$. The dotted blue vertical lines separate scenarios where $\sigma^{(0)2} = 1$ and $\sigma^{(0)2} > 1$. All Monte Carlo standard errors are 0.008 or less.

Table 1:

Simulated averages and standard deviations (in parentheses) of $\bar{\tau}_{O,linked}$ and $\bar{\tau}_{O,linked,reg}$ for RegBRLC and BRLC, as well as estimates based on the perfectly linked data. Scenarios vary the numbers of intersecting records in the files or the outcome model variance $\sigma^{(0)2}$. All scenarios have $\tau_{O,\ inked} = 5$. Results in each scenario based on 100 replications. Monte Carlo standard errors are all less than .08.

		$\bar{\tau}_{O,linked}$			$\bar{\tau}_{O,linked,reg}$		
Percentage of Intersection	$\sigma^{(0)2}$	**RegBRLC**	**BRLC**	**Perfect**	**RegBRLC**	**BRLC**	**Perfect**
20	1	4.58(0.58)	3.42(0.61)	4.94(0.36)	4.78(0.36)	3.51(0.48)	4.94(0.23)
50	1	4.92(0.43)	3.84(0.49)	4.98(0.27)	4.92(0.20)	3.81(0.28)	4.99(0.08)
80	1	4.93(0.23)	3.97(0.29)	5.01(0.16)	4.96(0.17)	3.99(0.24)	5.02(0.03)
80	4	4.89(0.36)	3.91(0.40)	4.98(0.23)	4.88(0.30)	3.88(0.35)	4.99(0.11)
80	16	4.64(0.39)	3.64(0.48)	4.94(0.29)	4.68(0.35)	3.67(0.40)	4.96(0.24)

Table 2:

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 $\bar{\tau}_{O,linked}$ and $\bar{\tau}_{O,linked,reg}$ (in thousand Italian Liras) for all methods. In the parentheses are the standard deviations (SDs) corresponding to $\bar{\tau}_{O,linked}$ and $\bar{\tau}_{O,linked,reg}$.

	Precision	Recall	$\bar{\tau}_{O,linked}$			$\bar{\tau}_{O,linked,reg}$		
Method			**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
RegBRLC	0.897	0.876	184.34 (4.66)	50.05	340.31	192.44 (3.67)	162.34	246.19
BRLC	0.849	0.842	201.18 (4.82)	101.87	364.67	221.36 (3.96)	183.29	272.06