The Annals of Applied Statistics
2021, Vol. 15, No. 1, 412–436
https://doi.org/10.1214/20-AOAS1397

A MULTIPLE IMPUTATION PROCEDURE FOR RECORD LINKAGE AND CAUSAL INFERENCE TO ESTIMATE THE EFFECTS OF HOME-DELIVERED MEALS

BY MINGYANG SHAN[1,*], KALI S. THOMAS[2] AND ROEE GUTMAN[1,†]

[1]*Department of Biostatistics, Brown University School of Public Health,* [*]*mingyang_shan@brown.edu;* [†]*roee_gutman@brown.edu*

[2]*Department of Health Services, Policy and Practice, Brown University School of Public Health, kali_thomas@brown.edu*

Causal analysis of observational studies requires data that comprise a set of covariates, a treatment assignment indicator and the observed outcomes. However, data confidentiality restrictions or the nature of data collection may distribute these variables across two or more datasets. In the absence of unique identifiers to link records across files, probabilistic record linkage algorithms can be leveraged to merge the datasets. Current applications of record linkage are concerned with estimation of associations between variables that are exclusive to one file and not causal relationships. We propose a Bayesian framework for record linkage and causal inference where one file comprises all the covariate and observed outcome information, and the second file consists of a list of all individuals who receive the active treatment. Under certain ignorability assumptions, the procedure properly propagates the error in the record linkage process, resulting in valid statistical inferences. To estimate the causal effects, we devise a two-stage procedure. The first stage of the procedure performs Bayesian record linkage to multiply-impute the treatment assignment for all individuals in the first file, while adjustments for covariates' imbalance and imputation of missing potential outcomes are performed in the second stage. This procedure is used to evaluate the effect of Meals on Wheels services on mortality and healthcare utilization among homebound older adults in Rhode Island. In addition, an interpretable sensitivity analysis is developed to assess potential violations of the ignorability assumptions.

1. Introduction. Meals on Wheels (MOW) programs are community-based social-service organizations that provide home-delivered meals to homebound older adults in order to reduce hunger and food insecurity, promote socialization and encourage community independence. Providing home-delivered meals to these populations is associated with beneficial nutritional outcomes, decreased risk of falls and improved mental health (Thomas and Mor (2013), Campbell et al. (2015), Thomas, Akobundu and Dosa (2015), Thomas et al. (2018a, 2018b)). However, healthcare payers, providers and policy makers are also interested in the effects of community based organizations, like MOW, on premature mortality and healthcare utilization, such as hospitalizations, emergency department visits and nursing home placement.

One major challenge to performing such research is that MOW, by the nature of the services provided, does not submit medical claims or maintain clients' health records. Medicare enrollment records and claims data contain comprehensive information on patients' demographics, diagnoses, healthcare utilization and long-term health outcomes but exclude information about receipt of social services, such as MOW. In order to estimate the effects of MOW services on the healthcare utilization of its clients, linkage of Medicare claims data

Received December 2019; revised August 2020.
Key words and phrases. Record linkage, missing data, causal inference, multiple imputation, Bayesian data analysis.

to MOW client data is required. However, data confidentiality restrictions prevent unique identifiers to be available for linking.

Many studies seek to draw causal conclusions about the impact of an intervention. Randomized experiments are the gold standard for inferring causality; however, they are sometimes infeasible because of logistical, ethical or financial considerations. In these instances, researchers are limited to use nonrandomized observational studies to estimate the causal effects. Causal analysis of observational studies requires data that comprise a set of covariates, a treatment assignment indicator and the observed outcome (Rubin (1973a, 1973b, 1978, 1979)). In some studies, because of confidentiality restrictions or the nature of the data collection, the covariates, treatment assignment and outcome information are distributed across two or more data sources without unique identifiers to link records that belong to the same entity. One way to overcome this limitation is to incorporate an initial record linkage step in the design phase of the study.

Record linkage is a set of statistical procedures that identifies individuals or entities that are shared across datasets (Jaro (1989), Winkler (1993)). Record linkage techniques can be classified into two broad classes, deterministic and probabilistic. Deterministic record linkage methods identify records that represent the same entity based on deterministic agreement functions between data elements common to both records. Probabilistic methods calculate probabilities or weights that a pair of records represents the same entity. The Fellegi–Sunter procedure estimates the probabilities of observed agreement patterns of data elements between a pair of records if these records were true links or false links (Fellegi and Sunter (1969)). These probabilities are commonly used in an iterative algorithm that classifies the pair of records with the highest probability as a link and then removes these records from the pool of possible links. This process continues until a certain threshold is reached (Larsen and Rubin (2001)). The remaining records are either sent for clerical review or are declared nonlinks. Deterministic methods are used widely in practice and are reported to yield a higher proportion of true links than probabilistic methods (Gomatam et al. (2002), Campbell, Deck and Krupski (2008)). However, when the data elements are subject to variations in spelling, data entry inaccuracies or incompleteness, deterministic methods may miss more true links than probabilistic linking methods (Gomatam et al. (2002), Campbell, Deck and Krupski (2008)). In applications involving large public health datasets, these missed links may limit the practicality of deterministic linking methods (Newman et al. (2009)). Probabilistic linkage methods are less sensitive to errors among the identifying fields, and some of these methods, such as hit-miss models, can account for measurement error or missingness within records (Copas and Hilton (1990)).

Both probabilistic and deterministic methods may suffer from incorrectly linked entities. Neter, Maynes and Ramanathan (1965) noted that, in finite population sampling, a small amount of incorrectly linked records can lead to biased regression estimates and inflated variances. Multiple methods have been developed to reduce bias and propagate linkage errors using linear and generalized linear models (Scheuren and Winkler (1993, 1997), Lahiri and Larsen (2005), Chambers et al. (2009), Hof and Zwinderman (2012)). However, these methods are model specific and assume that the linkage probabilities are known or estimable. Furthermore, all of these methods rely on the noninformative linkage assumption, which states that the outcome of interest is conditionally independent from the linkage process, given the variables that appear in both files.

Bayesian file linkage procedures are probabilistic record linkage techniques that were proposed as possible solutions to overcome some of these limitations (Fortini, Liseo and Nuccitelli (2001), Larsen (2005), Tancredi and Liseo (2011), Gutman, Afendulis and Zaslavsky (2013), Steorts (2015), Steorts, Hall and Fienberg (2016), Sadinle (2017)). These methods introduce a latent structure indicating the pairing of records from one file with records from

another file. By generating multiple samples from the posterior distribution of the latent linking structure, Bayesian record linkage procedures account for the uncertainty in the linkage process.

The objective of the previously mentioned Bayesian and Frequentist methods is to estimate marginal and conditional associations and not causal effects. Wortman and Reiter (2018) proposed a method for estimating causal effects using linked observational data sources and described the possible effects of errors in linkage when estimating causal effects. The proposed method incorporates record pair selection strategies that can improve treatment effect estimation when using propensity score subclassification. However, this method does not adjust for error in the linkage process, is limited to the application of propensity score subclassification and only applies to settings in which the treatment assignment and covariates are in one file and the outcomes are in another.

Based on the potential outcome framework originally proposed by Neyman (1923) in the context of randomization-based inference in randomized experiments and generalized to other settings in Rubin (1978), we propose a joint Bayesian framework for record linkage and causal inference, where one file comprises all of the covariates and observed outcome information and the second file consists of a list of all individuals who receive the active treatment. To estimate the causal effects, we propose a computationally efficient two-stage procedure that accounts for the uncertainty in the linkage process and the unobserved potential outcomes under certain ignorability assumptions. Bayesian record linkage is performed in the first stage to inform the treatment assignment for all units in the first file. Adjustments for covariates' imbalance and imputation of the unobserved potential outcomes are performed in the second stage. In addition, we develop a procedure to examine the sensitivity of our results to the ignorability assumptions. We apply this procedure to estimate the effect of MOW services on mortality and healthcare utilization among homebound older adults in Rhode Island and find that MOW receipt does not have a significant effect on 30 day hospitalization rates or nursing home stays.

2. Framework.

2.1. *Notation.* The potential outcomes framework posits that, for a population of size $\mathcal{N}$, where $\mathcal{N}$ can be infinite, the effect of a binary treatment W on outcome Y for unit i ($i = 1, \ldots, \mathcal{N}$) is the comparison of two "potential" outcomes, $Y_i(0)$ and $Y_i(1)$, which correspond to the two possible levels of W: $W_i = 1$ indicates the receipt of the active level of the treatment, and $W_i = 0$ indicates the receipt of the control level. We assume the stable unit treatment value assumption (SUTVA) (Rubin (1980, 1990)) so that this notation is functionally well defined. For each unit i, there is also a vector of P covariates that are unaffected by W_i, $\mathbf{X}_i = (X_{i1}, \ldots, X_{iP})$.

We assume that the observed data is distributed across two files, $\mathbf{A}$ and $\mathbf{B}$, which comprise n_A and n_B records, respectively. The P covariates are partitioned into P_1 covariates that only appear in file $\mathbf{A}$, $\mathbf{X}_A = (\mathbf{X}_1, \ldots, \mathbf{X}_{P_1})$, P_2 covariates that only appear in file $\mathbf{B}$, $\mathbf{X}_B = (\mathbf{X}_{P_1+1}, \ldots, \mathbf{X}_{P_1+P_2})$ and P_3 covariates that appear in both files and can be used as semi-identifying information $\mathbf{Z}_A = \mathbf{Z}_B = (\mathbf{X}_{P_1+P_2+1}, \ldots, \mathbf{X}_P)$, where $P_1 + P_2 + P_3 = P$. Record $l \in \{1, \ldots, n_A\}$ in file $\mathbf{A}$ comprise $\mathbf{X}_{Al}$, $\mathbf{Z}_{Al}$ and the observed outcome Y_l^{obs}. Record $j \in \{1, \ldots, n_B\}$ in file $\mathbf{B}$ comprise $\mathbf{X}_{Bj}$ and $\mathbf{Z}_{Bj}$. We further assume that file $\mathbf{B}$ represents a list of records that all receive the active treatment, such that $W_j = 1, \forall j = 1 \ldots, n_B$, and that all records receiving the active treatment in file $\mathbf{A}$ also appear in file $\mathbf{B}$, or $\{l \in \mathbf{A} : W_l = 1\} \in \mathbf{B}$.

We introduce a latent structure $\mathbf{C} = (C_1, \ldots, C_{n_A})$, which represents the link designations in file $\mathbf{B}$ for each record in file $\mathbf{A}$,

$$C_l = \begin{cases} j & \text{if record } l \in \mathbf{A} \text{ is linked with record } j \in \mathbf{B}; \\ 0 & \text{if record } l \in \mathbf{A} \text{ is not linked with any record from file } \mathbf{B}. \end{cases} \tag{1}$$

An immediate consequence of this definition is that all records in $\mathbf{A}$ with $C_l > 0$ receive the active treatment ($W_l = 1$), and the remaining records with $C_l = 0$ receive the control treatment ($W_l = 0$). The observed potential outcomes are defined as $\mathbf{Y}^{\text{obs}} = \{Y_l^{\text{obs}}\}$, where $Y_l^{\text{obs}} = \mathbb{1}(C_l > 0)Y_l(1) + \mathbb{1}(C_l = 0)Y_l(0)$. Note that the information in $\mathbf{X}_B$ for the units with $C_l = 0$ is missing. We define $\mathbf{X}_B = (\mathbf{X}_B^{\text{obs}}, \mathbf{X}_B^{\text{mis}})$, where $\mathbf{X}_B^{\text{obs}}$ represents the observed information for records in $\mathbf{A}$ with $C_l > 0$ and the records in $\mathbf{B}$ with $j \notin \mathbf{C}$ that do not link with any record in $\mathbf{A}$, and $\mathbf{X}_B^{\text{mis}}$ represents the unobserved information for records in $\mathbf{A}$ with $C_l = 0$ that do not link with any record in $\mathbf{B}$.

To summarize, $\mathbf{X}_A$, $\mathbf{Z}_A$ and $\mathbf{Y}^{\text{obs}}$ are observed in $\mathbf{A}$, and $\mathbf{X}_B$, $\mathbf{Z}_B$ and $\mathbf{W}$ are observed in $\mathbf{B}$. The additional unobserved variables are $\mathbf{C}$ and $\mathbf{Y}^{\text{mis}} = \{Y_l^{\text{mis}}\}$, where $Y_l^{\text{mis}} = \mathbb{1}(C_l = 0)Y_l(1) + \mathbb{1}(C_l > 0)Y_l(0)$. The joint distribution of the observed and unobserved data across both files is

$$
\begin{aligned}
&f\big(\mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{C}, \mathbf{W}\big) \\
&\quad = f(\mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B) f\big(\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}) | \mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B\big) \\
&\qquad \times f\big(\mathbf{C} | \mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})\big) f\big(\mathbf{W} | \mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{C}\big).
\end{aligned} \tag{2}
$$

2.2. *Causal estimand.* Causal treatment effects are commonly summarized by estimands, $\tau = \tau(\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{W}) = \tau(\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{W})$, which are functions of the unit level potential outcomes of all units on a common subset of $\mathcal{N}$ units (Rubin (1978)). A Bayesian inference for the effects of an exposure using linked data source considers the observed values $\mathbf{X}_A$, $\mathbf{X}_B^{\text{obs}}$, $\mathbf{Z}_B$, $\mathbf{Z}_A$ and $\mathbf{Y}^{\text{obs}}$ as a realization of random variables and $\mathbf{X}_B^{\text{mis}}$, $\mathbf{Y}^{\text{mis}}$, $\mathbf{C}$ and $\mathbf{W}$ to be unobserved random variables. This perspective explicitly confronts the observed and missing random variables by conditioning on the observed variables in $\mathbf{A}$ and sampling from the posterior distribution of τ,

$$
\begin{aligned}
&f\big(\tau | \mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}^{\text{obs}}\big) \\
&\quad = \int f\big(\tau | \mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{C}, \mathbf{W}\big) \\
&\qquad \times f\big(\mathbf{X}_B^{\text{mis}}, \mathbf{Y}^{\text{mis}}, \mathbf{C}, \mathbf{W} | \mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}^{\text{obs}}\big) d\big(\mathbf{X}_B^{\text{mis}}, \mathbf{Y}^{\text{mis}}, \mathbf{C}, \mathbf{W}\big).
\end{aligned} \tag{3}
$$

Equation (3) shows that obtaining the posterior distribution of τ involves integrating over

$$
\begin{aligned}
&f\big(\mathbf{X}_B^{\text{mis}}, \mathbf{Y}^{\text{mis}}, \mathbf{C}, \mathbf{W} | \mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B\big) \\
&\quad = \frac{f(\mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{C}, \mathbf{W})}{\int f(\mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{C}, \mathbf{W}) \, d(\mathbf{X}_B^{\text{mis}}, \mathbf{Y}^{\text{mis}}, \mathbf{C}, \mathbf{W})}.
\end{aligned} \tag{4}
$$

2.3. *Simplifying assumptions.* To perform the integration in equation (3), we make the following simplifying assumptions. Some of these assumptions are made explicitly in many applications and some are made implicitly.

ASSUMPTION 1. The covariates that are unique to file $\mathbf{B}$ are independent from the potential outcomes given $\mathbf{X}_A$ and $\mathbf{Z}_A$. Formally,

$$
f\big(\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{X}_B, \mathbf{Z}_B | \mathbf{X}_A, \mathbf{Z}_A\big) = f\big(\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}) | \mathbf{X}_A, \mathbf{Z}_A\big) f(\mathbf{X}_B, \mathbf{Z}_B | \mathbf{X}_A, \mathbf{Z}_A). \tag{5}
$$

This assumption implies that only the covariates in $\mathbf{A}$ are required to predict the potential outcomes and that $\mathbf{X}_B$ and $\mathbf{Z}_B$ may only be informative for the linkage process. This assumption is valid in our application, because $\mathbf{Z}_B$ represents the same covariates as $\mathbf{Z}_A$, and $\mathbf{X}_B$ is only composed of administrative variables that are not influencing the health outcomes given

the clinical and demographic covariates defined in $\mathbf{X}_A$. It is possible to incorporate additional covariate information in $\mathbf{X}_B$ by using values of $\mathbf{X}_B^{\text{obs}}$ for linked records in $\mathbf{A}$ with $C_l > 0$ and imputing $\mathbf{X}_B^{\text{mis}}$ for records with $C_l = 0$.

ASSUMPTION 2. The treatment assignment mechanism is a deterministic function of the linkage structure.

Because file $\mathbf{B}$ comprises all units that received the active treatment, the treatment assignment for units $l \in \mathbf{A}$ can be derived as a deterministic function of their linkage status, $W_l = g(C_l)$, where

$$W_l = \begin{cases} 1 & \text{if } C_l > 0; \\ 0 & \text{if } C_l = 0. \end{cases} \tag{6}$$

This implies that $f(\mathbf{W}|\mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{C})$ is a degenerate distribution that is completely defined by $\mathbf{C}$ and that $\tau = \tau(\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{W}) = \tau(\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{C})$.

ASSUMPTION 3. The linkage is strongly noninformative.

This assumption states that $\mathbf{C}$ is conditionally independent from the potential outcomes and any unobserved data components. Formally,

$$f\big(\mathbf{C}|\mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})\big) = f\big(\mathbf{C}|\mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B\big). \tag{7}$$

This is a modified version of the noninformative linkage assumption commonly made when estimating noncausal associations using linked data. The implicit noninformative linkage assumption in the Fellegi–Sunter record linkage model implies that the linkage structure is conditionally independent from $\mathbf{Y}^{\text{obs}}$, $\mathbf{X}_A$ and $\mathbf{X}_B$, given $\mathbf{Z}_A$ and $\mathbf{Z}_B$ (Harron, Goldstein and Dibben (2015)). Equation (7) also implies that the treatment assignment is unconfounded (Rubin (1990)).

By combining Assumptions 1–3, equation (2) can be expressed as

$$\begin{aligned} &f(\mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{C}, \mathbf{W}) \\ &\quad = f(\mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B) f\big(\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})|\mathbf{X}_A, \mathbf{Z}_A\big) f\big(\mathbf{C}|\mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B\big). \end{aligned} \tag{8}$$

These assumptions simplify equation (3) to

$$\begin{aligned} &f\big(\tau|\mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}^{\text{obs}}\big) \\ &\quad = \int f\big(\tau|\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{C}\big) f\big(\mathbf{Y}^{\text{mis}}, \mathbf{C}|\mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}^{\text{obs}}\big)\, d\big(\mathbf{Y}^{mis}, \mathbf{C}\big). \end{aligned} \tag{9}$$

2.4. *Parametric models.* Given the linkage structure $\mathbf{C}$, we can assume the distributions of $\mathbf{X}_{Al}$, $\mathbf{X}_{Bl}$, $\mathbf{Z}_{Al}$, $\mathbf{Z}_{Bl}$, $Y_l(0)$ and $Y_l(1)$ are row exchangeable. Using de-Finetti's theorem, equation (8) can be written as a product of independent random variables, given the parameters $\boldsymbol{\Theta} = (\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y\cdot X}, \boldsymbol{\theta}_C)$, where $\boldsymbol{\theta}_X = (\boldsymbol{\theta}_{XA}, \boldsymbol{\theta}_{XB}, \boldsymbol{\theta}_{ZA}, \boldsymbol{\theta}_{ZB})$, $\boldsymbol{\theta}_{Y\cdot X} = (\boldsymbol{\theta}_{Y1\cdot X}, \boldsymbol{\theta}_{Y0\cdot X})$ and $\boldsymbol{\theta}_C = (\boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU})$. When the parameters $\boldsymbol{\theta}_X$, $\boldsymbol{\theta}_{Y\cdot X}$, and $\boldsymbol{\theta}_C$ are a priori independent with prior distribution $p(\boldsymbol{\Theta}) = p(\boldsymbol{\theta}_X)p(\boldsymbol{\theta}_{Y\cdot X})p(\boldsymbol{\theta}_C)$, equation (8), for linked and unlinked data, can be expressed as

$$\begin{aligned} &f(\mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{C}, \mathbf{W}) \\ &\quad = \int \Bigg[\prod_{l:C_l>0} f\big(\mathbf{X}_{Al}, \mathbf{X}_{Bl}^{\text{obs}}, \mathbf{Z}_{Al}, \mathbf{Z}_{Bl}|\boldsymbol{\theta}_X\big) f\big(Y_l(0), Y_l(1)|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\theta}_{Y\cdot X}\big) \Bigg] \end{aligned}$$

$$
\begin{aligned}
(10)\qquad &\times \Bigg[\prod_{l:C_l=0} f\big(\mathbf{X}_{Al}, \mathbf{X}_{Bl}^{\mathrm{mis}}, \mathbf{Z}_{Al}, \mathbf{Z}_{Bl}|\boldsymbol{\theta}_X\big) f\big(Y_l(0), Y_l(1)|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\theta}_{Y\cdot X}\big)\Bigg] \\
&\times \Bigg[\prod_{j:j\notin\mathbf{C}} f\big(\mathbf{X}_{Bj}^{\mathrm{obs}}, \mathbf{Z}_{Bj}|\boldsymbol{\theta}_{XB}, \boldsymbol{\theta}_{ZB}\big)\Bigg] \\
&\times f\big(\mathbf{C}, \boldsymbol{\theta}_C|\mathbf{X}_A, \mathbf{X}_B^{\mathrm{obs}}, \mathbf{Z}_A, \mathbf{Z}_B\big) p(\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y\cdot X})\, d\boldsymbol{\Theta}.
\end{aligned}
$$

The information in the second line in equation (10) represents the data components for linked records in $\mathbf{A}$, the third line reflects the information for unlinked records in $\mathbf{A}$ and the last product represents the unlinked records in $\mathbf{B}$. The integrals over $\boldsymbol{\theta}_X$ pass through the distributions for $Y_l(0)$, $Y_l(1)$ and $\mathbf{C}$ such that the marginal distribution $f(\mathbf{X}_A, \mathbf{X}_B, \mathbf{Z}_A, \mathbf{Z}_B)$ is irrelevant for $\boldsymbol{\theta}_{Y\cdot X}$ and $\boldsymbol{\theta}_C$.

We will make the assumption of no contamination of imputation across treatments (Rubin (2008)).

ASSUMPTION 4. The conditional distribution of potential outcomes for the exposed and unexposed units are independent, given baseline covariates, and the parameters governing their distributions are a priori independent:

$$
\begin{aligned}
(11)\qquad f\big(Y_l(0), Y_l(1)|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\theta}_{Y\cdot X}\big) &= f\big(Y_l(0)|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\theta}_{Y0\cdot X}\big) f\big(Y_l(1)|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\theta}_{Y1\cdot X}\big), \\
p(\boldsymbol{\theta}_{Y\cdot X}) &= p(\boldsymbol{\theta}_{Y0\cdot X}) p(\boldsymbol{\theta}_{Y1\cdot X}).
\end{aligned}
$$

Based on this assumption, the conditional distribution for the potential outcomes is

$$
\begin{aligned}
&f\big(\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})|\mathbf{X}_A, \mathbf{Z}_A\big) \\
(12a)\qquad &= \int \prod_{l=1}^{n_A} f\big(Y_l(0), Y_l(1)|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\theta}_{Y\cdot X}\big) p(\boldsymbol{\theta}_{Y\cdot X})\, d\boldsymbol{\theta}_{Y\cdot X} \\
(12b)\qquad &= \int \prod_{l:C_l>0} f\big(Y_l(0)|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\theta}_{Y0\cdot X}\big) \\
&\quad \times \prod_{l:C_l=0} f\big(Y_l(0)|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\theta}_{Y0\cdot X}\big) p(\boldsymbol{\theta}_{Y0\cdot X})\, d\boldsymbol{\theta}_{Y0\cdot X} \\
(12c)\qquad &\quad \times \int \prod_{l:C_l=0} f\big(Y_l(1)|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\theta}_{Y1\cdot X}\big) \\
&\quad \times \prod_{l:C_l>0} f\big(Y_l(1)|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\theta}_{Y1\cdot X}\big) p(\boldsymbol{\theta}_{Y1\cdot X})\, d\boldsymbol{\theta}_{Y1\cdot X} \\
(12d)\qquad &\propto \int f\big(\mathbf{Y}(\mathbf{0})^{\mathrm{mis}}|\mathbf{X}_A, \mathbf{Z}_A, \boldsymbol{\theta}_{Y0\cdot X}\big) p\big(\boldsymbol{\theta}_{Y0\cdot X}|\mathbf{X}_A, \mathbf{Z}_A, \mathbf{Y}(\mathbf{0})^{\mathrm{obs}}\big)\, d\boldsymbol{\theta}_{Y0\cdot X} \\
(12e)\qquad &\quad \times \int f\big(\mathbf{Y}(\mathbf{1})^{\mathrm{mis}}|\mathbf{X}_A, \mathbf{Z}_A, \boldsymbol{\theta}_{Y1\cdot X}\big) f\big(\boldsymbol{\theta}_{Y1\cdot X}|\mathbf{X}_A, \mathbf{Z}_A, \mathbf{Y}(\mathbf{1})^{\mathrm{obs}}\big)\, d\boldsymbol{\theta}_{Y1\cdot X}.
\end{aligned}
$$

The first factor in equation (12d) and the first factor in equation (12e) are the posterior predictive distributions of the unobserved potential outcomes and the remaining terms in each line are the posterior distributions of $\boldsymbol{\theta}_{Y0\cdot X}$ and $\boldsymbol{\theta}_{Y1\cdot X}$, respectively.

By combining equation (9) with equations (10–12), the causal estimand is

$$
\begin{aligned}
&f(\tau|\mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B, \mathbf{Y}^{\text{obs}}) \\
&\quad = \int f(\tau|\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{C}) f(\mathbf{Y}(\mathbf{0})^{\text{mis}}|\mathbf{X}_A, \mathbf{Z}_A, \boldsymbol{\theta}_{Y0\cdot X}) p(\boldsymbol{\theta}_{Y0\cdot X}|\mathbf{X}_A, \mathbf{Z}_A, \mathbf{Y}(\mathbf{0})^{\text{obs}}) \\
&\qquad \times f(\mathbf{Y}(\mathbf{1})^{\text{mis}}|\mathbf{X}_A, \mathbf{Z}_A, \boldsymbol{\theta}_{Y1\cdot X}) p(\boldsymbol{\theta}_{Y1\cdot X}|\mathbf{X}_A, \mathbf{Z}_A, \mathbf{Y}(\mathbf{1})^{\text{obs}}) \\
&\qquad \times f(\mathbf{C}, \boldsymbol{\theta}_C|\mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B) d(\boldsymbol{\theta}_C, \mathbf{C}, \boldsymbol{\theta}_{Y0\cdot X}, \boldsymbol{\theta}_{Y1\cdot X}, \mathbf{Y}(\mathbf{0})^{\text{mis}}, \mathbf{Y}(\mathbf{1})^{\text{mis}}).
\end{aligned} \tag{13}
$$

2.5. *Record linkage models.* To model $f(\mathbf{C}, \boldsymbol{\theta}_C|\mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B)$, we will rely on the record linkage framework initially proposed by Fellegi and Sunter (1969). This framework considers the set of all possible $n_A \times n_B$ pairs of records from $\mathbf{A}$ and $\mathbf{B}$ as the union of two disjoint sets of links $\mathbf{M} = \{(l, j) : l \in \mathbf{A}, j \in \mathbf{B}, C_l = j\}$ and non-links $\mathbf{U} = \{(l, j) : l \in \mathbf{A}, j \in \mathbf{B}, C_l \neq j\}$. Without a loss of generality, assume $n_A \geq n_B$. To ensure that each record in file $\mathbf{A}$ is linked to, at most, one record in file $\mathbf{B}$ and vice versa, we introduce the following constraint on $\mathbf{C}$: $\sum_{l=1}^{n_A} \mathbb{1}\{C_l = j\} \leq 1, \forall j = 1, \ldots, n_B$ (Larsen (2005), Sadinle (2017)).

For records $l \in \mathbf{A}$ and $j \in \mathbf{B}$, let $\boldsymbol{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj}) = (\gamma_{lj1}, \ldots, \gamma_{ljP_3})$ be an agreement vector for the $k = 1, \ldots, P_3$ identifying variables that exist in both files. The agreement of field k between two values can be evaluated on an ordinal scale with $r_k = 1, \ldots, R_k$ levels, where 1 represents complete disagreement and R_k represents complete agreement (Winkler (1990), Sadinle (2017)). Let $\boldsymbol{\theta}_{CM} = \{\boldsymbol{\theta}_{CMk}\}$ and $\boldsymbol{\theta}_{CU} = \{\boldsymbol{\theta}_{CUk}\}$ represent the parameters governing the distributions of the comparison functions for record pairs in $\mathbf{M}$ and $\mathbf{U}$, respectively, such that $\boldsymbol{\theta}_{CMk} = \{\theta_{CMkr}\}$, where $\theta_{CMkr} = \Pr(\gamma_{ljk} = r|C_l = j)$ for $k = 1, \ldots, P_3$ and $r = 1, \ldots, R_k$ and, similarly, $\boldsymbol{\theta}_{CUk} = \{\theta_{CUkr}\}$ and $\theta_{CUkr} = \Pr(\gamma_{ljk} = r|C_l \neq j)$.

Mixture models have been proposed to estimate $\boldsymbol{\theta}_{CM}$, $\boldsymbol{\theta}_{CU}$ and $\mathbf{C}$ based on $\boldsymbol{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj})$ (Jaro (1989), Larsen and Rubin (2001)),

$$
\begin{aligned}
\boldsymbol{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj})|C_l = j &\sim f(\boldsymbol{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj})|\boldsymbol{\theta}_{CM}), \\
\boldsymbol{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj})|C_l \neq j &\sim f(\boldsymbol{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj})|\boldsymbol{\theta}_{CU}), \\
\mathbf{C} &\sim p(\mathbf{C}, n_m),
\end{aligned} \tag{14}
$$

where $n_m = \sum_{l=1}^{n_A} \sum_{j=1}^{n_B} \mathbb{1}(C_l = j)$ represents the number of true links.

Let π represent the expected proportion of records that represent the same entities in both files, such that $n_m \sim \text{Binomial}(n_B, \pi)$ and $\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi)$ a priori. Sadinle (2017) proposes a Beta-Binomial prior for $\mathbf{C}$ and n_m that marginalizes over π,

$$
p(\mathbf{C}, n_m|\alpha_\pi, \beta_\pi) = \frac{(n_A - n_m)!}{n_A!} \frac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)} \frac{\Gamma(n_m + \alpha_\pi)\Gamma(n_B - n_m + \beta_\pi)}{\Gamma(n_m + \alpha_\pi)}. \tag{15}
$$

A simplifying assumption that is frequently made in the Fellegi and Sunter record linkage model is that each of the comparison functions are conditionally independent given $\mathbf{C}$ (Winkler (1988), Jaro (1989)). Under this assumption, the likelihood for $\mathbf{C}$, $\boldsymbol{\theta}_{CM}$, and $\boldsymbol{\theta}_{CU}$ is

$$
\mathcal{L}(\mathbf{C}, \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}|\mathbf{Z}_A, \mathbf{Z}_B) = \prod_{l=1}^{n_A} \prod_{j=1}^{n_B} \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} [\theta_{CMkr}^{\mathbb{1}(\gamma_{ljk}=r)}]^{\mathbb{1}(C_l=j)} [\theta_{CUkr}^{\mathbb{1}(\gamma_{ljk}\neq r)}]^{\mathbb{1}(C_l\neq j)}. \tag{16}
$$

Independent conjugate priors $\boldsymbol{\theta}_{CMk} \sim \text{Dirichlet}(\alpha_{Mk1}, \ldots, \alpha_{MkR_k})$ and $\boldsymbol{\theta}_{CUk} \sim \text{Dirichlet}(\alpha_{Uk1}, \ldots, \alpha_{UkR_k})$ for $k = 1, \ldots, K$ can be specified to complete the Bayesian model.

To sample from $f(\mathbf{C}, \boldsymbol{\theta}_C|\mathbf{Z}_A, \mathbf{Z}_B) \propto p(\mathbf{C})p(\boldsymbol{\theta}_C)\mathcal{L}(\mathbf{C}, \boldsymbol{\theta}_C|\mathbf{Z})$, we will use the data augmentation algorithm (Tanner and Wong (1987)). The I-Step involves drawing $\mathbf{C}^{[t+1]}$ from $f(\mathbf{C}|\mathbf{Z}_A, \mathbf{Z}_B, \boldsymbol{\theta}_C^{[t]})$ and the P-step will update values of $\boldsymbol{\theta}_C^{[t+1]}$ from $f(\boldsymbol{\theta}_C|\mathbf{Z}_A, \mathbf{Z}_B, \mathbf{C}^{[t+1]})$ (see Appendix A for a detailed description).

2.6. *Causal treatment effect estimation*. The record linkage facilitates the identification of the exposed and unexposed units, but it does not guarantee that units are similar across treatment groups. When the distribution of covariates between the treatment and control groups are different, simple comparison of the two groups may result in biased estimates of the treatment effect (Rubin (1973a, 1973b)). Several types of procedures have been proposed to address this issue when the treatment effect is unconfounded (Imbens and Rubin (2015), Gutman and Rubin (2017)).

Matching is a design phase causal estimation technique that reduces bias by identifying units with similar covariate values between the two treatment groups (Stuart (2010)). With a single covariate it is often easy to identify a match. This task is more complicated with multiple covariates (Stuart (2010)). Matching on the propensity score (Rosenbaum and Rubin (1983)), which is the probability of $W_l = 1$ given the covariates, was proposed as a possible solution. Formally, the propensity score for unit l is defined as $e(\mathbf{X}_{Al}, \mathbf{Z}_{Al}) \equiv f(W_l|\mathbf{X}_{Al}, \mathbf{Z}_{Al}, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ are the parameters governing this distribution. Point estimates of τ, using matching, have been shown to be consistent but may underestimate its sampling variance when ignoring the variability in the matching procedure (Abadie and Imbens (2011), Gutman and Rubin (2017)). In addition, because matching on $e(\mathbf{X}_{Al}, \mathbf{Z}_{Al})$ is not exact, some covariates may still suffer from minor imbalances, which is often addressed using regression adjustments (Imbens and Rubin (2015)).

A different approach to estimate τ is to combine matching with a Bayesian imputation framework (Rubin (2008), Gutman and Rubin (2013, 2015)). This combination reduces the bias resulting from minor covariate imbalances and increases precision by using modeling to impute $\mathbf{Y}^{\text{mis}}$. Under the Bayesian causal inference framework, the missing potential outcomes are taken to be unobserved random variables that can be sampled from their posterior predictive distribution. Because sampling from a posterior distribution is complex, we use a multiple imputation procedure as an approximation of the posterior distribution of $\mathbf{Y}^{\text{mis}}$ (Gutman and Rubin (2013, 2015)).

2.7. *Two-stage multiple imputation estimation procedure*. Equation (13) suggests a two-step estimation procedure. In the first step the record linkage structure is sampled. Using this linkage structure, the potential outcomes are imputed in the second step.

We now explicate and summarize this two-stage multiple imputation approach (Shen (2000), Rubin (2003)) to estimate τ:

1. Sample $\mathbf{C}^{(m)}$ from $f(\mathbf{C}, \boldsymbol{\theta}_C|\mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B)$ for $m = 1, \ldots, M$ random draws (Appendix A).
2. For $\mathbf{C}^{(m)}, m = 1, \ldots, M$:

   (a) Perform nearest neighbor matching using the estimated the propensity score $\hat{e}(\mathbf{X}_A, \mathbf{Z}_A)^{(m)}$ to obtain a sample of exposed and unexposed units with similar covariate distributions. Let $\mathbf{G}^{(m)} = 1$ represent the units in this matched sample.

   (b) For the units with $\mathbf{G}_l^{(m)} = 1$, partition $\mathbf{Y}(\mathbf{0})$ and $\mathbf{Y}(\mathbf{1})$ into $\mathbf{Y}^{\text{obs}(m)}$ and $\mathbf{Y}^{\text{mis}(m)}$.

   (c) Sample $\theta_{Y0\cdot X}^{(m,q)}$, $q = 1, \ldots, Q$, from $p(\theta_{Y0\cdot X}|\mathbf{X}_A, \mathbf{Z}_A, \mathbf{Y}(\mathbf{0})^{\text{obs}(m)}, \mathbf{G}^{(m)} = 1)$ and $\theta_{Y1\cdot X}^{(m,q)}$, from $p(\theta_{Y1\cdot X}|\mathbf{X}_A, \mathbf{Z}_A, \mathbf{Y}(\mathbf{1})^{\text{obs}(m)}, \mathbf{G}^{(m)} = 1)$.

   (d) For each $q = 1, \ldots, Q$, use $\theta_{Y0\cdot X}^{(m,q)}$ and $\theta_{Y1\cdot X}^{(m,q)}$ to independently impute the missing potential outcomes for each unit in $\mathbf{G}^{(m)} = 1$ from the posterior predictive distributions $f(\mathbf{Y}(\mathbf{0})^{\text{mis}}|\mathbf{X}_A, \mathbf{Z}_A, \mathbf{G}^{(m)} = 1, \theta_{Y0\cdot X}^{(m,q)})$ and $f(\mathbf{Y}(\mathbf{1})^{\text{mis}}|\mathbf{X}_A, \mathbf{Z}_A, \mathbf{G}^{(m)} = 1, \theta_{Y1\cdot X}^{(m,q)})$. This will result in Q datasets with imputed $\mathbf{Y}(\mathbf{0})^{\text{mis}}$ and $\mathbf{Y}(\mathbf{1})^{\text{mis}}$ for the matched units.

3. For each of the $M \times Q$ imputations, estimate the treatment effect $\hat{\tau}^{(m,q)}$ and the within imputation sampling variance, $U^{(m,q)}$.

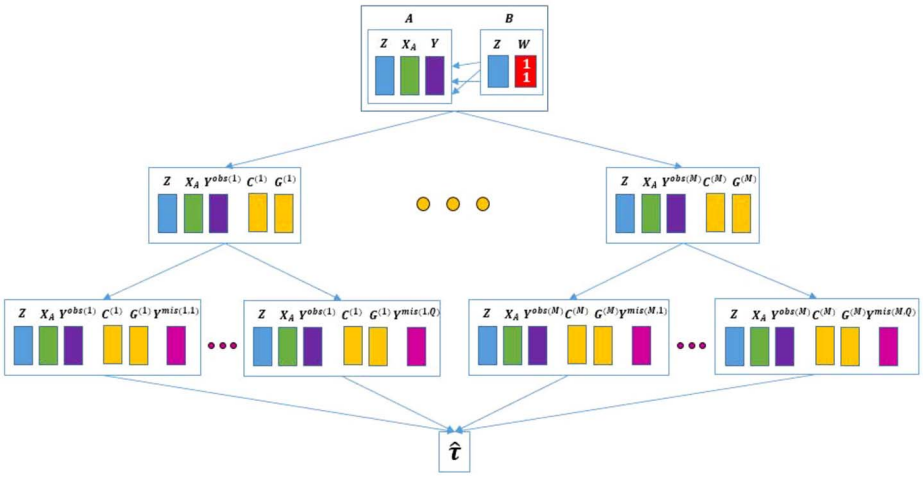

FIG. 1. *Two-Stage Multiple Imputation Procedure.*

4. The point estimate of the treatment effect τ is calculated as $\hat{\tau} = \frac{1}{MQ}\sum_{m=1}^{M}\sum_{q=1}^{Q}\hat{\tau}^{(m,q)}$. The total variance can be derived, according to Shen (2000), as $T = \hat{U} + (1 + M^{-1})B + (1 - Q^{-1})W$, where $\hat{U} = \frac{1}{MQ}\sum_{m=1}^{M}\sum_{q=1}^{Q}U^{(m,q)}$, $B = \frac{1}{M-1}\sum_{m=1}^{M}(\bar{\tau}^{(m.)} - \hat{\tau})^2$ and $W = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{Q-1}\sum_{q=1}^{Q}(\hat{\tau}^{(m,q)} - \bar{\tau}^{(m.)})^2$. Inference for $\hat{\tau}$ is based on a t-distribution $(\tau - \hat{\tau})/\sqrt{T} \sim t_\nu$ with $\nu^{-1} = \frac{1}{M-1}(\frac{(1+1/M)B}{T})^2 + \frac{1}{M(Q-1)}(\frac{(1-1/Q)W}{T})^2$.

An illustration of this two-stage multiple imputation procedure is presented in Figure 1.

3. Application to meals on wheels data. We applied the proposed procedure to estimate the effects of Meals on Wheels (MOW) programs on mortality and healthcare utilization among Medicare beneficiaries. We compared the difference in mortality rate after 30 days of initiating meal delivery service and the difference in the frequency of acute inpatient, emergency department (ED) and nursing home (NH) events between MOW recipients and nonrecipients who were alive after 30 days of enrollment in both the observed treatment and predicted control arms. Because we do not expect MOW receipt to influence mortality, identification of such an effect may indicate potential violations of the unconfoundedness assumption.

3.1. *Data description.* A comprehensive list of all clients who received home-delivered meals between January 1, 2010 and December 31, 2013 was submitted by Meals on Wheels Rhode Island (MOWRI), which serves the entire State of Rhode Island. This list contained information on each client's sex, date of birth, start and end date of service and the nine-digit ZIP code corresponding to the address where their meal was delivered. While MOW serves clients of a variety of ages, we restricted the analysis only to individuals older than 65 at enrollment in MOW, because only those individuals are expected to be enrolled in Medicare. This resulted in a total of $n_B = 3916$ MOW recipients.

The Medicare Master Beneficiary Summary File (MBSF) is a comprehensive database that contains demographic information on all Medicare enrollees including gender, date of birth, date of death and the nine-digit ZIP code corresponding to their mailing address. In addition, the MBSF identifies preexisting chronic medical conditions including congestive heart failure, kidney failure, diabetes, pelvic fracture, stroke, dementia or Alzheimer's disease, chronic obstructive pulmonary disease, coronary artery disease and various types of cancer. Using

unique identifiers, the MBSF is linked to Medicare inpatient, outpatient, skilled nursing facility and home health claims as well as the nursing home Minimum Data Set (MDS) between calendar years 2009 and 2014. These claims and assessment data were used to calculate the frequency and cost of inpatient events, emergency department visits, nursing home stays and home health utilization in the 30, 90 and 180 days prior to enrollment in MOW as well as the frequency of acute inpatient, ED and NH events 30 days following MOW enrollment. A total of $n_A = 179{,}269$ Medicare beneficiaries over the age of 65 resided in the five-digit ZIP codes serviced by MOWRI.

3.2. *Record linkage of MOW and Medicare data.* Comparison of all possible record pairs, where one record appears in the MOW file and another record in the Medicare file would result in over 700 million possible record pairs. "Blocking" is a common record linkage to reduce the computational complexity by only considering record pairs that agree on specific blocking fields (Newcombe et al. (1959), Newcombe (1988)). We generated blocks based on five-digit ZIP codes and gender (Herzog, Scheuren and Winkler (2007)). We assumed that $\boldsymbol{\theta}_{CM}$ and $\boldsymbol{\theta}_{CU}$ do not differ across blocks, and we restricted record pairs such that the MOW enrollment date precedes the date of death in the Medicare file. These criteria reduced the total number of possible record pairs to 13,786,172. A sensitivity analysis of our results to this blocking criteria is provided in Appendix D.

A recipient's date of birth (DOB) and nine-digit ZIP code were used as linking variables to classify record pairs into links and non-links. Ordinal agreement patterns were used for both linking variables, whose levels of agreement are described in Table 1. In addition, we modeled the interaction between agreement on DOB and ZIP code (Winkler (1989)). The resulting record linkage likelihood is

$$
\begin{aligned}
&\mathcal{L}\big(\mathbf{C}, \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU} | \boldsymbol{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B)\big) \\
&\quad = \prod_{l=1}^{n_A} \prod_{j=1}^{n_B} \prod_{r_D=1}^{4} \prod_{r_Z=1}^{5} \big[\theta_{CMDr_D}^{\mathbb{1}(\gamma_{ljD}=r_D)} \theta_{CMZr_Z|r_D}^{\mathbb{1}(\gamma_{ljZ}=r_Z, \gamma_{ljD}=r_D)}\big]^{\mathbb{1}(C_l=j)\mathbb{1}(B_{lj}=1)} \\
&\qquad \times \big[\theta_{CUDr_D}^{\mathbb{1}(\gamma_{ljD}=r_D)} \theta_{CUZr_Z|r_D}^{\mathbb{1}(\gamma_{ljZ}=r_Z, \gamma_{ljD}=r_D)}\big]^{\mathbb{1}(C_l\neq j)\mathbb{1}(B_{lj}=1)},
\end{aligned} \tag{17}
$$

where B_{lj} is an indicator that record pair (l, j) was successfully blocked and met the MOW enrollment date constraint, and $\boldsymbol{\theta}_{CM} = \{\boldsymbol{\theta}_{CMD}, \boldsymbol{\theta}_{CMZ}\}$ and $\boldsymbol{\theta}_{CU} = \{\boldsymbol{\theta}_{CUD}, \boldsymbol{\theta}_{CUZ}\}$ are the parameters governing the distribution of agreement functions, such that $\boldsymbol{\theta}_{CMD} = \{\theta_{CMDr_D}\}$, $\boldsymbol{\theta}_{CMZ} = \{\theta_{CMZr_Z|r_D}\}$, $\boldsymbol{\theta}_{CUD} = \{\theta_{CUDr_D}\}$, $\boldsymbol{\theta}_{CUZ} = \{\theta_{CUZr_Z|r_D}\}$. Each $\theta_{CMDr_D} = \Pr(\gamma_{ljD} = r_D | C_l = j, B_{lj} = 1)$, $\theta_{CMZr_Z|r_D} = \Pr(\gamma_{ljZ} = r_Z | \gamma_{ljD} = r_D, C_l = j, B_{lj} = 1)$, $\theta_{CUDr_D} =$

TABLE 1
Linking variable description and agreement level

Agreement type	Level
Disagreement on DOB	$r_D = 1$
Agree on DOB Year only	$r_D = 2$
Agree on DOB Year and Month only	$r_D = 3$
Agree on DOB Year, Month, and Day	$r_D = 4$
Agree on first 5 digits of ZIP code only	$r_Z = 1$
Agree on first 6 digits of ZIP code only	$r_Z = 2$
Agree on first 7 digits of ZIP code only	$r_Z = 3$
Agree on first 8 digits of ZIP code only	$r_Z = 4$
Agree on all 9 digits of ZIP code	$r_Z = 5$

$\Pr(\gamma_{ljD} = r_D | C_l \neq j, B_{lj} = 1)$ and $\theta_{CUZr_Z|r_D} = \Pr(\gamma_{ljZ} = r_Z | \gamma_{ljD} = r_D, C_l \neq j, B_{lj} = 1)$ for $r_D = 1, \ldots, 4$ and $r_Z = 1, \ldots, 5$. A total of $M = 100$ different linkage structures were imputed from the linkage algorithm.

3.3. *Propensity score matching.* Each of the $m = 1, \ldots, M$ linked datasets identifies Medicare beneficiaries who received MOW and those who did not. Prior research suggested that MOW programs target older adults who have higher social and economic needs and are at higher risk for institutional care (Lloyd and Wellman (2015), Lee, Shannon and Brown (2015)). Thus, enrollment in MOW programs may be confounded with preexisting health conditions or prior healthcare utilization. To reduce covariates' imbalances, matching on the estimated propensity score was performed.

Prior to matching, all linked individuals who were enrolled in a Medicare Advantage (MA) program in the six months prior to receiving MOW or were enrolled in MA during the month they began MOW were removed. This truncation was implemented because MA plans are not required to submit claims, and it was not possible to fully observe the prior history of chronic conditions or healthcare utilization for these individuals. A start date for individuals who are not enrolled in MOW programs is not available. Instead, for individuals who were not linked to MOW records, we calculated the medical history and prior healthcare utilization at the start of each quarter for every year in our study period. This resulted in 16 sets of pretreatment covariates calculated at different potential enrollment dates for each unlinked individual.

Matching was implemented by enforcing exact agreement on patients' gender, race, age categories and whether the patients had any inpatient, ER or SNF claim in the 90 days prior to enrollment. The remaining covariates were balanced using propensity score models that included preexisting medical conditions, prior healthcare utilization frequency and prior healthcare costs. We selected nearest pair matches without replacement based on the propensity score within each quarter (Stuart (2010)). This process was replicated on each of the $M = 100$ linked datasets to identify beneficiaries that resemble MOW recipients but did not enroll in the program.

3.4. *Imputation of unobserved outcomes.* To assess the impact of MOW on mortality, we examine the average treatment effect on the treated (ATT) among linked MOW clients who are matched to a control individual. Let $D_l(1)$ and $D_l(0)$ represent the potential 30-day mortality for individual l had they received meals or not, respectively. The estimand of interest is $\tau_{\text{ATT}} = E(D(1) - D(0) | W = 1, G = 1)$, which can be estimated within each imputation as

$$\hat{\tau}_{\text{ATT}}^{(m,q)} = \frac{1}{n_G^{(m)}} \sum_{l: C_l^{(m)} > 0, G_l^{(m)} = 1} \big(D_l(1)^{(m)} - D_l(0)^{(m,q)}\big), \tag{18}$$

where $n_G^{(m)}$ represents the number of linked MOW clients matched with a control. To predict the unobserved 30-day mortality for MOW clients had they not received meals, we used a Bayesian logistic regression model that included the covariates used in matching, $X_{Al1}, \ldots, X_{AlP1}$, and all of their two-way interactions,

$$\operatorname{logit}\big(P\big(D_l(0) = 1\big)\big) = \beta_0 + \sum_{p=1}^{P_1} \beta_p X_{Alp} + \sum_{p=1}^{P_1} \sum_{s=1}^{p-1} \delta_{ps} X_{Alp} X_{Als}. \tag{19}$$

A one-level hierarchical normal-gamma shrinkage prior (Griffin and Brown (2017)) was constructed for the coefficients of the main effect and interaction terms such that $\beta_p \overset{\text{i.i.d.}}{\sim}$

$N(0, \Phi_1)$, $\Phi_1 \sim \text{Gamma}(1, 1)$, $\delta_{ps} \overset{\text{i.i.d.}}{\sim} N(0, \Phi_2)$, and $\Phi_2 \sim \text{Gamma}(1, 2)$. These prior distributions attenuate interaction terms more aggressively. We also assumed that $\beta_0 \sim N(0, 10{,}000)$.

To examine the impact of MOW on healthcare utilization, we estimate the survivor average treatment effect on the treated (SATT), which compares the effect of MOW receipt among MOWRI clients who would be alive 30 days after their enrollment date, irrespective of whether they received services from MOW or not (Frangakis and Rubin (2002), Rubin (2006), Frangakis et al. (2007)). The SATT is defined as $\tau_{\text{SATT}} = E(H(1) - H(0)|W = 1, D(0) = D(1) = 0, G = 1)$, where $H(1)$ and $H(0)$ denote the potential utilization frequency among MOW clients and controls, respectively. The SATT is estimated within each imputation as

$$\hat{\tau}_{\text{SATT}}^{(m,q)} = \frac{1}{n_S^{(m,q)}} \sum_{\substack{l: C_l^{(m)} > 0, G_l^{(m)} = 1, \\ D_l(1)^{(m)} = D_l(0)^{(m,q)} = 0}} \big(H_l(1)^{(m)} - H_l(0)^{(m,q)}\big), \tag{20}$$

where $n_S^{(m,q)}$ is the number of linked individuals who were matched to a control and are alive after 30 days following enrollment, according to their observed and predicted mortality status.

Bayesian zero-inflated negative binomial models were fitted to impute the frequency of inpatient admissions, emergency department visits and nursing home stays among MOW clients had they not received meals within each imputed linkage structure. The zero-inflated negative binomial distribution for count response $H_l(0)$ is given by

$$\begin{aligned} &P\big(H_l(0) = h\big) \\ &= \begin{cases} \pi_l + (1 - \pi_l)\left(\dfrac{\alpha}{\mu_l + \alpha}\right)^{\alpha}, & H_l(0) = 0, \\ (1 - \pi_l)\dfrac{\Gamma(H_l(0) + \alpha)}{\Gamma(H_l(0) + 1)\Gamma(\alpha)}\left(\dfrac{\mu_l}{\mu_l + \alpha}\right)^{H_l(0)}\left(\dfrac{\tau}{\mu_l + \alpha}\right)^{\alpha}, & H_l(0) > 0, \end{cases} \end{aligned} \tag{21}$$

where α represents the shape parameter

$$\log(\mu_l) = \zeta_0 + \sum_{p=1}^{P_1} \zeta_p X_{Alp} + \sum_{p=1}^{P_1} \sum_{s=1}^{p-1} \eta_{ps} X_{Alp} X_{Als} \tag{22}$$

and

$$\text{logit}(\pi_l) = \psi_0 + \sum_{p=1}^{P_1} \psi_p X_{Alp} + \sum_{p=1}^{P_1} \sum_{s=1}^{p-1} \xi_{ps} X_{Alp} X_{Als}. \tag{23}$$

The negative binomial component is modeled in equation (22) and equation (23) models the zero inflation. To complete the Bayesian model, we assumed that $\zeta_p \overset{\text{i.i.d.}}{\sim} N(0, \Phi_3)$, $\Phi_3 \sim \text{Gamma}(1, 1)$, $\eta_{ps} \overset{\text{i.i.d.}}{\sim} N(0, \Phi_4)$, $\Phi_4 \sim \text{Gamma}(1, 2)$, $\psi_p \overset{\text{i.i.d.}}{\sim} N(0, \Phi_5)$, $\Phi_5 \sim \text{Gamma}(1, 1)$, $\xi_{ps} \overset{\text{i.i.d.}}{\sim} N(0, \Phi_6)$ and $\Phi_6 \sim \text{Gamma}(1, 2)$. Lastly, $\tau \sim U(0, 1000)$, $\zeta_0 \sim N(0, 10{,}000)$ and $\psi_0 \sim N(0, 10{,}000)$. All models were fit using Rstan version 2.17.2 (Stan Development Team (2018)). All outcomes were imputed for 100 datasets within each of the 100 linked datasets, resulting in $M \times Q = 10{,}000$ complete datasets.

4. Results. Of the $n_B = 3916$ MOW clients eligible for linkage to Medicare data, an average of $\bar{n}_m = 3608.02$ records (95% CI: 3570.91, 3645.14) were linked over $M = 100$ imputations. Among the Medicare beneficiaries who were linked to MOW clients, an average of 1748.35 (95% CI: 1735.28, 1761.44) individuals had at least one month of MA coverage in the six months prior to enrollment and were excluded from our analysis. Of the remaining linked individuals, an average of 1859.67 (95% CI: 1835.63, 1883.70) treated units were matched to a control unit that did not receive meals.

Figure 2 displays the median and range of absolute standardized differences for each of the pretreatment variables between the MOW and control samples before and after matching for the $M = 100$ linked datasets. The absolute standardized difference exceeded 0.25 in 22 out of the 40 covariates prior to matching. After matching, all absolute standardized differences are less than 0.25 for all covariates which suggests that the covariates are adequately balanced (Rosenbaum and Rubin (1985, 2015)).

The observed and imputed potential outcomes for MOW clients and the estimated treatment effects are provided in Table 2. The estimated ATT on mortality using our two-stage multiple imputation procedure is 0.008 (95% CI: −0.067, 0.083). This indicates that there is no significant difference between the observed and predicted mortality rate among the linked and matched MOW clients. An average of 51.21 individuals died within 30 days of their MOW enrollment or were predicted to die within 30 days without MOW services. This results in an average of 1808.46 individuals who are alive whether they received MOW or not across imputations. Among these individuals the estimated SATTs are 0.010 (95% CI: −0.174, 0.194) on acute inpatient admissions, −0.013 (95% CI: −0.236, 0.209) on ED visits, and 0.003 (95% CI: −0.268, 0.274) for NH stays. This suggests that among MOW recipients,

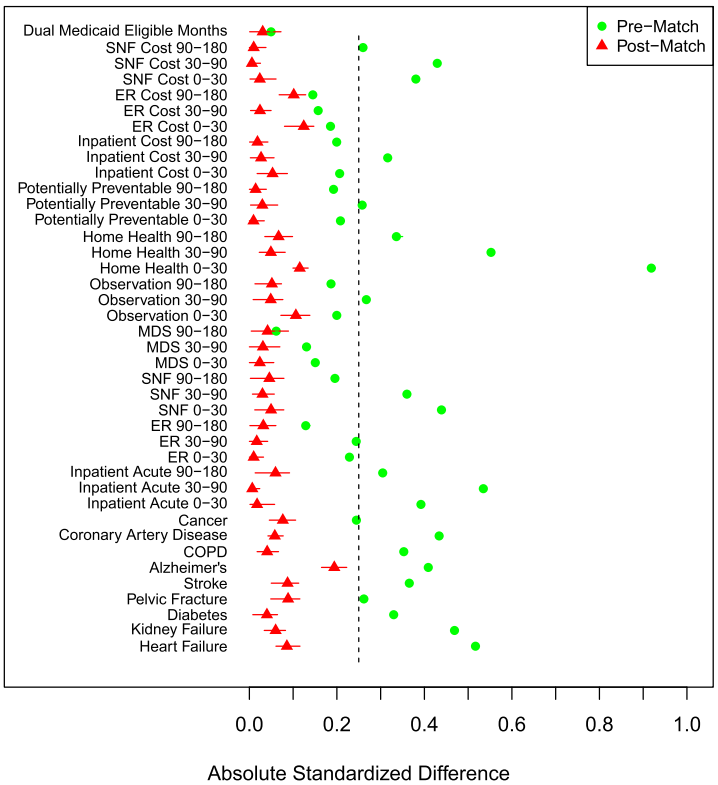

FIG. 2. *Covariate balance before and after propensity score matching for $M = 100$ linked datasets. The points represent the median absolute standardized differences, while the horizontal lines represent the range between the minimum and maximum absolute standardized difference for each covariate.*

TABLE 2
Estimated causal treatment effects on MOW recipients

30 day outcome	$\bar{n}_G$	$\bar{\mathbf{D}}(\mathbf{1})$	$\bar{\mathbf{D}}(\mathbf{0})$	$\hat{\tau}_{\text{ATT}}$	95% CI
Mortality	1859.67	0.023 (0.001)	0.015(0.004)	0.008	$(-0.067, 0.083)$
30 day outcome	$\bar{n}_S$	$\bar{\mathbf{H}}(\mathbf{1})$	$\bar{\mathbf{H}}(\mathbf{0})$	$\hat{\tau}_{\text{SATT}}$	95% CI
Acute Inpatient	1808.46	0.085 (0.002)	0.075 (0.010)	0.010	$(-0.174, 0.194)$
Emergency Care	1808.46	0.073 (0.002)	0.088 (0.012)	-0.013	$(-0.236, 0.209)$
Nursing Home	1808.46	0.078 (0.003)	0.075 (0.015)	0.003	$(-0.268, 0.274)$

$\bar{n}_G$ is the average sample size of MOW individuals linked and matched to a control individual across $M = 100$ imputations. $\bar{n}_S$ is the average number of linked and matched MOW individuals who are observed and predicted to be alive irrespective of their treatment, across $M \times Q = 10{,}000$ imputations. $\bar{\mathbf{D}}(\mathbf{1})$ and $\bar{\mathbf{D}}(\mathbf{0})$ represent the mortality rates for linked and matched MOW individuals if they received MOW or not, respectively. $\bar{\mathbf{H}}(\mathbf{1})$ and $\bar{\mathbf{H}}(\mathbf{0})$ are the estimated healthcare utilization rates for linked and matched individuals who are alive for 30 days irrespective of whether they received MOW or not, respectively. $\hat{\tau}_{\text{ATT}}$ is the estimated average treatment effect on the treated and $\hat{\tau}_{\text{SATT}}$ is the survivor average treatment effect on the treated.

who would be alive 30 days after enrollment irrespective of whether they received MOW or not, no significant differences in the number of acute inpatient admissions, ED visits or NH stays are detected.

5. Sensitivity analysis.

5.1. *Sensitivity of the strongly noninformative linkage assumption.* We examine the sensitivity of our results to the strongly noninformative linkage assumption (Assumption 3). Under Assumption 3, errors in the linkage only depend on comparisons of semi-identifying information that exists in both files. Thus, the probability of record $l \in \mathbf{A}$ forming a link with record $j \in \mathbf{B}$ given $\mathbf{C}_{(-l)} = \{C_{l'} : l' \neq l\}$ is (see Appendix B for additional details)

$$P\big(C_l = j | \mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj}), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big) \propto \frac{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj}) | \boldsymbol{\theta}_{CM})}{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj}) | \boldsymbol{\theta}_{CU})} \mathbb{1}(j \notin \mathbf{C}_{(-l)}). \tag{24}$$

To examine the impact of potential violations of the strongly noninformative linkage assumption on the estimation of the treatment effect, we assume that the errors in the linkage model depend on 30-day mortality status. Let D_{lj}^{obs} be an indicator that is equal to 1 if individual $l \in \mathbf{A}$ died within 30 days of the start date indicated by record $j \in \mathbf{B}$. Let λ_M and λ_U represent parameters governing the distribution of D_{lj}^{obs} for links and non-links, respectively. We assume that the distribution of D_{lj}^{obs} is

$$f\big(D_{lj}^{\text{obs}} | C_l, \lambda_M, \lambda_U\big) = \begin{cases} \lambda_M^{D_{lj}^{\text{obs}}} & \text{if } C_l = j, \\ \lambda_U^{D_{lj}^{\text{obs}}} & \text{if } C_l \neq j. \end{cases}$$

The posterior probability of individual $l \in \mathbf{A}$ linking with individual $j \in \mathbf{B}$ given $\mathbf{C}_{(-l)}$ will then take the form

$$\begin{aligned} &P\big(C_l = j | \mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj}), D_{lj}^{\text{obs}}, \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \lambda_M, \lambda_U, \mathbf{C}_{(-l)}\big) \\ &\quad \propto \frac{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj}) | \boldsymbol{\theta}_{CM})}{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj}) | \boldsymbol{\theta}_{CU})} \frac{f(D_{lj}^{\text{obs}} | \lambda_M)}{f(D_{lj}^{\text{obs}} | \lambda_U)} \mathbb{1}(j \notin \mathbf{C}_{(-l)}). \end{aligned} \tag{25}$$

A violation of the strongly noninformative linkage assumption influences the likelihood of record pairs by a sensitivity factor of $\Delta = (\frac{\lambda_M}{\lambda_U})^{D_{lj}^{\text{obs}}}$. When $\Delta = 1$, Assumption 3 is valid. Values of $\Delta > 1$ suggest that MOWRI was more likely to enroll individuals who may die shortly after enrollment. Values of $\Delta < 1$ represent scenarios where MOWRI may have selectively avoided enrolling individuals who may die shortly after enrollment.

To estimate the effect of the sensitivity parameter on the outcomes, we implement the algorithm in Section 2.7 after replacing $f(\mathbf{C}, \boldsymbol{\theta}_C | \mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B)$ in Step 1 with equation (25).

5.2. *Sensitivity analysis results*. The average number of records linked for different values of Δ is presented in Table 3. Because a small proportion of the linked individuals die within 30 days after enrollment when $\Delta = 1$, decreasing the sensitivity parameter does not yield significantly lower amounts of linked records. However, increasing the sensitivity parameter significantly increases the number of linked records, specifically record pairs that indicate death within 30 days of enrollment. At $\Delta = 100$, the narrow confidence interval indicates the linkage algorithm links almost all of the MOW records to the Medicare enrollment file.

The difference in 30-day mortality rate and inpatient acute utilization rate for the various sensitivity levels are presented in Table 4. The estimated mortality difference is similar for values of $\Delta \leq 1$. While the difference in mortality between MOW clients and controls is negative when $\Delta = 1/100$, this effect is small and insignificant. When $\Delta \geq 50$, the mortality difference is significant and MOW is estimated to increase mortality among their clients. This results from clients who die within 30 days being added to the linked sample and, subsequently, being removed from the potential control cohort to be matched. When $\Delta = 100$, the sensitivity parameter trumps the linkage likelihood such that many MOW clients are linked to Medicare records indicating 30 day mortality despite major disagreements between the linking information in $\mathbf{Z}_A$ and $\mathbf{Z}_B$. Assuming that MOW beneficiaries are 50–100 times more likely to die within 30 days of enrollment is highly implausible. The SATT of MOW on inpatient acute admission frequency does not significantly differ across the sensitivity scenarios. These results imply that our analysis is robust to the strongly noninformative linkage assumption with regards to mortality. Similar results were observed for inpatient acute hospitalization (data not shown).

TABLE 3
Sensitivity analysis linkage results

Δ	$\bar{n}_m$	95% CI
1/100	3566.74	(3521.15, 3612.33)
1/50	3571.50	(3529.74, 3613.26)
1/10	3580.75	(3537.22, 3624.29)
1/5	3589.74	(3549.78, 3629.70)
1/2	3607.69	(3563.14, 3652.24)
1	3608.02	(3570.91, 3645.14)
2	3610.76	(3575.98, 3645.54)
5	3662.55	(3621.70, 3703.40)
10	3714.30	(3677.96, 3750.64)
50	3844.21	(3821.23, 3867.19)
100	3874.40	(3860.50, 3888.31)

TABLE 4
Sensitivity analysis causal treatment effect estimates

	30 day mortality		30 day acute inpatient	
Δ	$\hat{\tau}_{\text{ATT}}$	95% CI	$\hat{\tau}_{\text{SATT}}$	95% CI
1/100	−0.006	(−0.055, 0.044)	0.006	(−0.125, 0.137)
1/50	0.004	(−0.070, 0.078)	0.009	(−0.178, 0.197)
1/10	0.005	(−0.070, 0.081)	0.009	(−0.178, 0.195)
1/5	0.006	(−0.068, 0.081)	0.009	(−0.177, 0.196)
1/2	0.007	(−0.067, 0.082)	0.010	(−0.178, 0.197)
1	0.008	(−0.067, 0.083)	0.010	(−0.176, 0.196)
2	0.010	(−0.066, 0.086)	0.010	(−0.174, 0.194)
5	0.016	(−0.059, 0.092)	0.010	(−0.173, 0.193)
10	0.025	(−0.052, 0.102)	0.010	(−0.175, 0.194)
50	0.086	(0.002, 0.170)	0.014	(−0.170, 0.198)
100	0.549	(0.412, 0.686)	0.079	(−0.120, 0.277)

6. Discussion. We have proposed a novel Bayesian framework to estimate causal treatment effects using linked data sources. We examine a linkage scenario that combines covariate information and outcome information from one file with the treatment assignment defined by a second file. Under a series of conditional independence and ignorability assumptions, we provide a two-stage multiple imputation procedure to obtain statistically valid treatment effect point and interval estimates. This procedure accounts for both the errors in the linkage and the unobserved outcomes. The first stage of the procedure imputes the linkage structure, and the missing potential outcomes are imputed in the second stage. Because the strong noninformative linkage assumption cannot be examined using observed data, we developed a sensitivity analysis to assess its violations.

In the linkage setting that we considered, all of the records in file **B** receive the active treatment. This allows us to derive the treatment assignment as a deterministic function of the linkage status for records in file **A**. More research and, possibly, stronger assumptions are required to estimate treatment effects in different linkage scenarios, such as when one file contains the treatment assignment and some of the covariates, while the other file includes the rest of the covariates and the observed outcomes. In addition, the proposed linkage algorithm is based on the Fellegi–Sunter framework which does not account for relationships between variables that are exclusive to one file. A possible extension would be to incorporate such relationships as described in Gutman, Afendulis and Zaslavsky (2013). The modularity of our procedure allows for adjustment of the record linkage algorithm without the need to adjust the causal inference component.

We applied our framework to estimate the effect of receiving services from a MOW program in Rhode Island on mortality and healthcare utilization for its clients. Our analysis suggested that MOW does not have a significant impact on reducing 30 day mortality among its clients. Furthermore, among clients who would be alive after 30 days, irrespective of MOW services, no significant differences in the frequency of acute inpatient admissions, ED visits or NH stays are observed. A sensitivity analysis that examined the strong noninformative linkage assumption showed that our analysis is robust against potential violations of this assumption. However, this assumption may not be valid in different applications, and designing procedures that relax this assumption is an area for future research.

A major limitation of Bayesian record linkage procedures is that they are computationally intensive, and commonly require specialized programming that most nontechnical researchers cannot implement. We have addressed these issues by utilizing two-stage multiple

imputation and blocking. Both of these techniques reduce the computational complexity and enable nontechnical researchers to perform the causal inference analyses while scaling down the record linkage complexity. Multiple imputation has been widely used in other missing data applications and was shown to provide valid inferences (Rubin (1996)). We have examined the performance of our proposed two-stage imputation in simulation analyses (Appendix C). The results show that the procedure provides valid statistical inference but that the coverage is higher than nominal. This implies that the proposed procedure may overestimate the sampling variance. Increased efficiency of such procedures is a future area of research.

Using blocking increases the computational efficiency and scalability of the record linkage procedure. However, it may exclude true links and influence subsequent inferences (Murray (2015)). We have examined the performance of our two-stage procedure with stricter and looser blocking criteria (Appendix D). Using a single CPU, the current runtime of our linkage algorithm was, approximately, 18 days. Loosening the blocking criteria such that the number of record pairs was more than four times larger resulted in a runtime of approximately 30 days. The point estimates were relatively similar for the stricter and loosened blocking criteria, but the stricter criteria had larger sampling variance because less record pairs were linked. This shows that our algorithm is relatively efficient and that inferences were robust to blocking. Use of parallel computing, more efficient programming languages and improvement of the MCMC sampling procedure based on ideas proposed by Zanella (2020) may improve the scaling of the proposed algorithm to even larger settings.

In conclusion, this manuscript describes a statistical framework to estimate causal effects using linked datasets where one file contains the covariates and the observed outcome and the second file contains the treatment assignment. Under the strongly noninformative linkage assumption, we develop a two stage multiple imputation procedure that provides statistically valid treatment effect estimates, and we describe a sensitivity analysis for this assumption.

APPENDIX A: RECORD LINKAGE GIBBS SAMPLING ALGORITHM

A Gibbs sampling algorithm proposed by Sadinle (2017) can be used to iterate between sampling the posterior distributions of the linking parameters and the linking configuration given the observed data. Starting with initial values for $\mathbf{C}$, we sample from $f(\mathbf{C}, \boldsymbol{\theta}_C | \mathbf{X}_A, \mathbf{X}_B^{\text{obs}}, \mathbf{Z}_A, \mathbf{Z}_B)$ using the following procedure:

1. Sample new values of $\boldsymbol{\theta}_{CMk}^{[t+1]}$ from

$$\begin{aligned}\boldsymbol{\theta}_{CMk}^{[t+1]} | \mathbf{C}^{[t]}, \boldsymbol{\Gamma}(\mathbf{Z}_{Alk}, \mathbf{Z}_{Bjk}) \sim \text{Dirichlet}\Bigg(&\alpha_{Mk1} + \sum_{l=1}^{n_A}\sum_{j=1}^{n_B} \mathbb{1}(C_l = j)\mathbb{1}(\gamma_{ljk} = 1), \ldots, \\ &\alpha_{MkR_k} + \sum_{l=1}^{n_A}\sum_{j=1}^{n_B} \mathbb{1}(C_l = j)\mathbb{1}(\gamma_{ljk} = R_k)\Bigg)\end{aligned} \tag{26}$$

and $\boldsymbol{\theta}_{CUk}^{[t+1]}$ from

$$\begin{aligned}\boldsymbol{\theta}_{CUk}^{[t+1]} | \mathbf{C}^{[t]}, \boldsymbol{\Gamma}(\mathbf{Z}_{Alk}, \mathbf{Z}_{Bjk}) \sim \text{Dirichlet}\Bigg(&\alpha_{Uk1} + \sum_{l=1}^{n_A}\sum_{j=1}^{n_B} \mathbb{1}(C_l \neq j)\mathbb{1}(\gamma_{ljk} = 1), \ldots, \\ &\alpha_{UkL_k} + \sum_{l=1}^{n_A}\sum_{j=1}^{n_B} \mathbb{1}(C_l \neq j)\mathbb{1}(\gamma_{ljk} = R_k)\Bigg).\end{aligned} \tag{27}$$

2. Sample a new state of $\mathbf{C}^{[t+1]}$ by iterating through each entry and proposing updates one entry at a time. Define $\mathbf{C}_{(-l)}^{[t+1]} = (C_1^{[t+1]}, \ldots, C_{l-1}^{[t+1]}, C_{l+1}^{[t+1]}, \ldots, C_{n_A}^{[t+1]})$ as the collection of

link designations without the lth entry, and let $n_{m(-l)}^{[t+1]} = \sum_{j=1}^{n_B} \mathbb{1}(j \in \mathbf{C}_{(-l)}^{[t+1]})$ be the number of designated links at iteration $[t+1]$, excluding the lth entry. The posterior distribution of $\mathbf{C}_l$ given $\mathbf{C}_{(-l)}$ is a multinomial distribution where the labels are $\{j : j \notin \mathbf{C}, 0\}$. The probability for $l \in \mathbf{A}$ to pair with the unlinked record $j \in \mathbf{B}$ and increase the total number of links by 1 is

$$
\begin{aligned}
&P\big(C_l^{[t+1]} = j | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}^{[t+1]}, \boldsymbol{\theta}_{CU}^{[t+1]}, \mathbf{C}_{(-l)}\big) \\
&= \frac{\frac{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj})|\boldsymbol{\theta}_{CM}^{[t+1]})}{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj})|\boldsymbol{\theta}_{CU}^{[t+1]})} \mathbb{1}(j \notin \mathbf{C}_{(-l)}^{[t+1]})}{\sum_{j'=1}^{n_B} \frac{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj'})|\boldsymbol{\theta}_{CM}^{[t+1]})}{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj'})|\boldsymbol{\theta}_{CU}^{[t+1]})} \mathbb{1}(j' \notin \mathbf{C}_{(-l)}^{[t+1]}) + \frac{(n_A - n_{m(-l)})(n_B - n_{m(-l)} + \beta_\pi - 1)}{n_{m(-l)} + \alpha_\pi}}.
\end{aligned} \tag{28}
$$

Similarly, the probability for $l \in \mathbf{A}$ not pairing with any record from $\mathbf{B}$ and $n_m^{[t+1]} = n_{m(-l)}^{[t+1]}$ is

$$
\begin{aligned}
&P\big(C_l^{[t+1]} = 0 | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}^{[t+1]}, \boldsymbol{\theta}_{CU}^{[t+1]}, \mathbf{C}_{(-l)}\big) \\
&= \frac{\frac{(n_A - n_{m(-l)})(n_B - n_{m(-l)} + \beta_\pi - 1)}{n_{m(-l)} + \alpha_\pi}}{\sum_{j'=1}^{n_B} \frac{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj'})|\boldsymbol{\theta}_{CM}^{[t+1]})}{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj'})|\boldsymbol{\theta}_{CU}^{[t+1]})} \mathbb{1}(j' \notin \mathbf{C}_{-l}^{[t+1]}) + \frac{(n_A - n_{m(-l)})(n_B - n_{m(-l)} + \beta_\pi - 1)}{n_{m(-l)} + \alpha_\pi}}.
\end{aligned} \tag{29}
$$

APPENDIX B: DERIVATION OF POSTERIOR LINKAGE PROBABILITIES

The posterior distribution of $\mathbf{C}_l$ and n_m, given the remaining link designations $\mathbf{C}_{(-l)}$, is

$$
\begin{aligned}
&f\big(C_l, n_m | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big) \\
&\quad \propto p(\mathbf{C}, n_m) \prod_{j=1}^{n_B} f\big(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj}) | \boldsymbol{\theta}_{CM}\big)^{\mathbb{1}(C_l = j)} \mathbb{1}(\mathbf{C}_{(-l)} \neq j) \\
&\qquad \times f\big(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj}) | \boldsymbol{\theta}_{CU}\big)^{\mathbb{1}(C_l \neq j)},
\end{aligned} \tag{30}
$$

where $p(\mathbf{C}, n_m)$ takes the form of equation (15).

The marginal posterior probability for record $l \in \mathbf{A}$ to form a true link with $j \in \mathbf{B}$, given $\mathbf{C}_{(-l)}$, is

$$
\begin{aligned}
&P\big(C_l = j | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big) \\
&\quad = \big(f\big(C_l = j, n_m = n_{m(-l)} + 1 | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big)\big) \\
&\qquad \Big/ \Bigg(\sum_{j=1}^{n_B} f\big(C_l = j, n_m = n_{m(-l)} + 1 | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big) \\
&\qquad\quad + f\big(C_l = 0, n_m = n_{m(-l)} | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big) \Bigg),
\end{aligned} \tag{31}
$$

and the marginal posterior probability for record $l \in \mathbf{A}$ to remain unlinked, given $\mathbf{C}_{(-l)}$, is

$$
\begin{aligned}
&P\big(C_l = 0 | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big) \\
&\quad = \big(f\big(C_l = 0, n_m = n_{m(-l)} | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big)\big) \\
&\qquad \Big/ \Bigg(\sum_{j=1}^{n_B} f\big(C_l = j, n_m = n_{m(-l)} + 1 | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big) \\
&\qquad\quad + f\big(C_l = 0, n_m = n_{m(-l)} | \mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big) \Bigg).
\end{aligned} \tag{32}
$$

Dividing the numerator and denominator of equation (31) by

$$p(\mathbf{C}, n_m = n_{m(-i)} + 1) \prod_{j=1}^{n_B} f\big(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj})|\boldsymbol{\theta}_{CU}\big) \tag{33}$$

results in equation (28). Similarly, dividing the numerator and denominator of equation (32) by equation (33) results in equation (29). Therefore, we see that

$$P\big(C_l = j|\mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big) \propto \frac{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj})|\boldsymbol{\theta}_{CM})}{f(\mathbf{\Gamma}(\mathbf{Z}_{Al}, \mathbf{Z}_{Bj})|\boldsymbol{\theta}_{CU})} \mathbb{1}(j \notin \mathbf{C}_{(-l)}),$$

and

$$P\big(C_l = 0|\mathbf{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B), \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU}, \mathbf{C}_{(-l)}\big) \propto \frac{(n_A - n_{m(-l)})(n_B - n_{m(-l)} + \beta_\pi - 1)}{n_{m(-l)} + \alpha_\pi}.$$

APPENDIX C: SIMULATION STUDY FOR TWO-STAGE MULTIPLE IMPUTATION PROCEDURE

We examine the operating characteristics of a t-distribution approximation for inference of the treatment effect in our two-stage multiple imputation procedure, described in Section 2.7, using a simulation study. We consider a simulation setting with $n_A = 20{,}000$ records in file $\mathbf{A}$, $n_B = 2000$ records in file $\mathbf{B}$, and $n_m = 1600$ true links between both files. The treatment effect is simulated for continuous, count and binary outcomes. Appendix Table 5 depicts the linking variables, covariates and response surfaces used to generate the simulations. Linking variables for true-links were simulated according to $f(\mathbf{Z}_A)$. No errors were simulated among the linking variables, such that $\mathbf{Z}_A$ and $\mathbf{Z}_B$ took the same values among true-links in both files. When the intervention has no effect, we assumed that $f(\mathbf{Y}(\mathbf{0})|\mathbf{X}_A, \mathbf{Z}_A) = f(\mathbf{Y}(\mathbf{1})|\mathbf{X}_A, \mathbf{Z}_A)$. When the treatment effect exists, we assumed that it is constant on the linear, logarithm or logistic scale for continuous, binary and count outcomes. The response surfaces were calibrated such that $E(\mathbf{Y}(\mathbf{0}))$ is equal to 10 for the continuous outcome, 2 for the count outcome and 0.3 for the binary outcome. A total of 100 simulated sets of data were generated from this simulation configuration.

To conduct record linkage of the pairs of simulated files, an individual's continuous age values were converted to a date of birth (DOB) with a year, month and day value. Four levels of similarity are used to compare the elements of DOB: no agreement on DOB year, agreement on DOB year only, agreement on DOB year and month only and agreement on all elements of DOB. Five levels of similarity are used to compare ZIP codes: disagreement on the first ZIP digit, agreement on the first ZIP digit only, agreement on the first and second ZIP digits only, agreement on the first through third ZIP digits only, and agreement on all ZIP digits. Exact agreement is used to compare values of gender. Conditional independence is assumed between agreement on the three linking variables, and equation (16) is used as the record linkage likelihood. Independent Dirichlet$(1, \ldots, 1)$ prior distributions are used for each $\boldsymbol{\theta}_{CM}$ and $\boldsymbol{\theta}_{CU}$. $M = 100$ imputations of the linkage structure was taken for each simulated dataset.

Propensity score models were fitted on linked and unlinked records in $\mathbf{A}$ using patients' age, gender, prior hospitalization and prior log-healthcare cost. Nearest neighbor matching without replacement was performed based on the propensity score to identify a set of controls with similar covariate distributions as the linked records. We calculated the average treatment effect on the treated for the continuous, count and binary outcomes by fitting Bayesian linear, Poisson and logistic regression models using the set of matched records. Each Bayesian model contained the covariates used in matching and all of their two-way interactions similar to the model specified in equation (19). This results in all of the imputation models being misspecified, but there is no unmeasured confounding. Noninformative prior distributions were placed on all of the parameters in each model.

TABLE 5
Simulated linking variables, covariates and outcomes

Linking variables	$f(\mathbf{Z}_A)$	$f(\mathbf{Z}_B)$
Age	$N(50, 5^2)$	$N(45, 10^2)$
Gender	Bernoulli(0.5)	Bernoulli(0.5)
ZIP code	1st Digit $\sim$ Discrete Uniform(6)	1st Digit $\sim$ Discrete Uniform(6)
	2nd Digit $\sim$ Discrete Uniform(7)	2nd Digit $\sim$ Discrete Uniform(7)
	3rd Digit $\sim$ Discrete Uniform(7)	3rd Digit $\sim$ Discrete Uniform(7)
	4th Digit $\sim$ Discrete Uniform(7)	4th Digit $\sim$ Discrete Uniform(7)

$f(\mathbf{X}_A)$	$l : C_l > 0$	$l : C_l = 0$
Prior hospitalization	Poisson(1)	Poisson(0.75)
Prior log-healthcare cost	$N(10, 3^2)$	$N(6, 4^2)$

Outcome type	τ_{ATT}	$f(\mathbf{Y}(\mathbf{1})\|\mathbf{X}_A, \mathbf{Z}_A)$
Continuous	0	$N(7.95 - 0.7 * \text{Gender} + 0.4 * \text{PriorHosp} + 0.2 * \text{PriorCost}, 0.1)$
Continuous	0.05	$N(8.00 - 0.7 * \text{Gender} + 0.4 * \text{PriorHosp} + 0.2 * \text{PriorCost}, 0.1)$
Continuous	0.10	$N(8.05 - 0.7 * \text{Gender} + 0.4 * \text{PriorHosp} + 0.2 * \text{PriorCost}, 0.1)$
Count	0	$\text{Poisson}(\exp(0.3431 - 0.7 * \text{Gender} + 0.2 * \text{PriorHosp} + 0.05 * \text{PriorCost}))$
Count	0.15	$\text{Poisson}(\exp(0.4155 - 0.7 * \text{Gender} + 0.2 * \text{PriorHosp} + 0.05 * \text{PriorCost}))$
Count	0.20	$\text{Poisson}(\exp(0.4384 - 0.7 * \text{Gender} + 0.2 * \text{PriorHosp} + 0.05 * \text{PriorCost}))$
Binary	0	$\text{Bernoulli}(\text{expit}(-1.1973 - 0.7 * \text{Gender} + 0.2 * \text{PriorHosp} + 0.05 * \text{PriorCost}))$
Binary	0.04	$\text{Bernoulli}(\text{expit}(-1.0132 - 0.7 * \text{Gender} + 0.2 * \text{PriorHosp} + 0.05 * \text{PriorCost}))$
Binary	0.08	$\text{Bernoulli}(\text{expit}(-0.8395 - 0.7 * \text{Gender} + 0.2 * \text{PriorHosp} + 0.05 * \text{PriorCost}))$

Appendix Table 6 displays the $\overline{\tau}$, $\overline{\text{Bias}}$, $\overline{SE}$ and coverage over the 100 simulated datasets for the different types of outcomes and treatment effect sizes. In settings where $\bar{n}_m$ and $\bar{n}_G$ are approximately 1600, our proposed two-stage procedure can accurately estimate linear, count and binary treatment effects with minimal bias. Interval estimates according to a t-distribution approximation provides nominal type 1 error and valid statistical inference for true treatment effect for all three types of outcomes and across different treatment effect sizes. However,

TABLE 6
Average bias, SE, coverage and type I error or power for different outcome distributions and treatment effect sizes

Outcome type	τ_{ATT}	$\overline{Estimate}$	$\overline{SE}$	$\overline{\text{Bias}}$	Coverage
Linear	0	0.0006	0.0041	0.0006	1
	0.05	0.0695	0.0102	0.0195	1
	0.10	0.1187	0.0102	0.0187	1
Count	0	0.0296	0.0608	0.0296	1
	0.15	0.1688	0.0606	0.0188	1
	0.20	0.2608	0.0611	0.0608	1
Binary	0	0.0005	0.0182	0.0005	1
	0.04	0.0410	0.0182	0.0010	1
	0.08	0.0784	0.0182	−0.0016	1

these intervals seem to be too wide, because the coverage probabilities are close to 1 and are not around the expected nominal coverage of 0.95.

APPENDIX D: SENSITIVITY ANALYSIS OF BLOCKING CRITERIA

In our linkage of MOW clients to Medicare enrollment records, blocks were generated based on clients' gender and the first five digits of ZIP code to reduce the number of possible record pairs and increase the efficiency of the linkage procedure. Sadinle and Fienberg (2013) demonstrate that blocking can significantly increase the accuracy of the linkage and subsequent inference, even when record linkage is computationally feasible without blocking, especially when the linking variables are limited or prone to error. However, Murray (2015) notes that blocking on variables that may be recorded with error can exclude true matches and influence the subsequent inference on the linked data. We examine the sensitivity of our results to different blocking criteria based on ZIP code digits.

Let n_Z represent the number of ZIP code digits used in the blocking criteria, where $n_Z = 5$ in our application. To examine the potential impact of different blocking restraints on the estimation of the causal treatment effect, we consider different values of $n_Z = (4, 6, 7)$. Altering the blocking criteria shifts the number of ZIP code digits available as linking variables. The record linkage likelihood in equation (17) can be reexpressed in terms of n_Z as

$$\begin{aligned}
&\mathcal{L}\big(\mathbf{C}, \boldsymbol{\theta}_{CM}, \boldsymbol{\theta}_{CU} | \boldsymbol{\Gamma}(\mathbf{Z}_A, \mathbf{Z}_B)\big) \\
&\quad = \prod_{l=1}^{n_A} \prod_{j=1}^{n_B} \prod_{r_D=1}^{4} \prod_{r_Z=1}^{9-n_Z+1} \big[\theta_{CMDr_D}^{\mathbb{1}(\gamma_{ljD}=r_D)} \theta_{CMZr_Z|r_D}^{\mathbb{1}(\gamma_{ljZ}=r_Z, \gamma_{ljD}=r_D)}\big]^{\mathbb{1}(C_l=j)\mathbb{1}(B_{lj}=1)} \\
&\qquad \times \big[\theta_{CUDr_D}^{\mathbb{1}(\gamma_{ljD}=r_D)} \theta_{CUZr_Z|r_D}^{\mathbb{1}(\gamma_{ljZ}=r_Z, \gamma_{ljD}=r_D)}\big]^{\mathbb{1}(C_l\neq j)\mathbb{1}(B_{lj}=1)}.
\end{aligned} \tag{34}$$

Treatment effects were estimated according to the algorithm in Section 2.7 after replacing the record linkage likelihood under each blocking scenario with equation (34).

D.1. Sensitivity analysis results. A comparison of the computational complexity of the linkage for different blocking scenarios and the linkage results are shown in Appendix Table 7. While increasing the number of ZIP digits used for blocking significantly reduces the computational complexity, there is also a significant decrease in the number of records that are linked. This is likely due to potential errors or discrepancies in how the six to nine ZIP code digits are recorded across both files. When blocking on these error-prone ZIP code digits, many true links are classified as non-links. Increasing the blocking constraints also reduces the number of available linking variables, which increases the efficiency of the computation

TABLE 7
Blocking sensitivity analysis linkage complexity and results

Blocking criteria	n_A	$n_A n_B$	$\bar{n}_m$	95% CI	Run time
4 digit ZIP	251,285	56,706,359	3829.67	(3814.40, 3844.94)	30 days
5 digit ZIP	247,724	13,786,172	3608.02	(3570.91, 3645.14)	18 days
6 digit ZIP	234,331	2,666,450	2807.21	(2728.63, 2885.79)	2 days
7 digit ZIP	144,230	436,049	2767.80	(2543.67, 2991.93)	<1 day

n_A represents the number of unique records in the Medicare data, $n_A n_B$ represents the total number of possible record pairs that are partitioned into a gender and ZIP code block, $\bar{n}_m$ represents the average number of linked records over $M = 100$ imputations and run time reflects the approximate time in days our Bayesian record linkage algorithm required to complete 500 iterations using a single CPU core on a Linux system.

TABLE 8
Blocking sensitivity analysis causal treatment effect estimates

	30 day mortality		30 day acute inpatient	
Blocking criteria	$\hat{\tau}_{\text{ATT}}$	95% CI	$\hat{\tau}_{\text{SATT}}$	95% CI
First 4 digits	0.005	(−0.067, 0.077)	0.007	(−0.168, 0.182)
First 5 digits	0.008	(−0.067, 0.083)	0.010	(−0.174, 0.194)
First 6 digits	0.010	(−0.074, 0.094)	0.016	(−0.180, 0.213)
First 7 digits	0.007	(−0.083, 0.096)	0.009	(−0.203, 0.222)

as well. Loosening the blocking criteria to the first four ZIP code digits results in more than a four-fold increase in the number of possible record pairs as well as an increase in the number of records linked.

The estimates of the causal treatment effects for mortality and acute inpatient admissions for different blocking criteria are presented in Appendix Table 8. Overall, we see that the use of blocking on the first five ZIP code digits in our application provides similar results compared to less strict blocking criteria. Gender and the first five digits of ZIP code are well defined and unlikely to be reported with errors which would not result in biased point estimates or suboptimal interval estimates if used as blocking criteria. Using stricter blocking criteria generally results in fewer false links but, possibly, more true links missed. The point estimates for the stricter criteria were practically the same, but the interval estimates were larger because of the smaller number of true links that were identified.

Acknowledgments. This work was supported in part by a grant from the Gary and Mary West Foundation and a grant from the Patient-Centered Outcomes Research Institute (PCORI/ME-2017C3-9241).

SUPPLEMENTARY MATERIAL

Supplemental code: A multiple imputation procedure for record linkage and causal inference to estimate the effects of home-delivered meals (DOI: 10.1214/20-AOAS1397 SUPP; .zip). We provide supplemental code (Shan, Thomas and Gutman (2021)) to demonstrate the implementation of the Bayesian Record Linkage algorithm, propensity score matching, imputation of the missing potential outcomes, and calculation of the two-stage multiple imputation treatment effects as proposed in this manuscript.

REFERENCES

ABADIE, A. and IMBENS, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *J. Bus. Econom. Statist.* **29** 1–11. MR2789386 https://doi.org/10.1198/jbes.2009.07333

CAMPBELL, K. M., DECK, D. and KRUPSKI, A. (2008). Record linkage software in the public domain: A comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. *Health Inform. J.* **14** 5–15.

CAMPBELL, A. D., GODFRYD, A., BUYS, D. R. and LOCHER, J. L. (2015). Does participation in home-delivered meals programs improve outcomes for older adults? Results of a systematic review. *J. Nutr. Gerontol. Geriatr.* **34** 124–167. https://doi.org/10.1080/21551197.2015.1038463

CHAMBERS, R., CHIPPERFIELD, J., DAVIS, W. and KOVACEVIC, M. (2009). Inference based on estimating equations and probability-linked data. Centre for Statistical and Survey Methodology, University of Wollongong Working Paper Series 18-09.

COPAS, J. B. and HILTON, F. J. (1990). Record linkage: Statistical models for matching computer records. *J. Roy. Statist. Soc. Ser. A* **153** 287–320.

FELLEGI, I. P. and SUNTER, A. B. (1969). A theory for record linkage. *J. Amer. Statist. Assoc.* **64** 1183–1210.

FORTINI, M., LISEO, B. and NUCCITELLI, A. (2001). On Bayesian record linkage **4** 185–198.

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039 https://doi.org/10.1111/j.0006-341X.2002.00021.x

FRANGAKIS, C. E., RUBIN, D. B., AN, M.-W. and MACKENZIE, E. (2007). Principal stratification designs to estimate input data missing due to death. *Biometrics* **63** 641–649. MR2395697 https://doi.org/10.1111/j.1541-0420.2007.00847_1.x

GOMATAM, S., CARTER, R., ARIET, M. and MITCHELL, G. (2002). An empirical comparison of record linkage procedures. *Stat. Med.* **21** 1485–1496.

GRIFFIN, J. and BROWN, P. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Anal.* **12** 135–159. MR3597570 https://doi.org/10.1214/15-BA990

GUTMAN, R., AFENDULIS, C. C. and ZASLAVSKY, A. M. (2013). A Bayesian procedure for file linking to analyze end-of-life medical costs. *J. Amer. Statist. Assoc.* **108** 34–47. MR3174601 https://doi.org/10.1080/01621459.2012.726889

GUTMAN, R. and RUBIN, D. B. (2013). Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Stat. Med.* **32** 1795–1814. MR3067363 https://doi.org/10.1002/sim.5627

GUTMAN, R. and RUBIN, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Stat. Med.* **34** 3381–3398. MR3412639 https://doi.org/10.1002/sim.6532

GUTMAN, R. and RUBIN, D. B. (2017). Estimation of causal effects of binary treatments in unconfounded studies with one continuous covariate. *Stat. Methods Med. Res.* **26** 1199–1215. MR3660989 https://doi.org/10.1177/0962280215570722

HARRON, K., GOLDSTEIN, H. and DIBBEN, C. (2015). *Methodological Developments in Data Linkage*. *Wiley Series in Probability and Statistics*. Wiley, New York.

HERZOG, T., SCHEUREN, F. and WINKLER, W. (2007). *Data Quality and Record Linkage Techniques*. Springer, New York.

HOF, M. H. P. and ZWINDERMAN, A. H. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Stat. Med.* **31** 4231–4242. MR3040077 https://doi.org/10.1002/sim.5498

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics*, *Social*, *and Biomedical Sciences*: *An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

JARO, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *J. Amer. Statist. Assoc.* **84** 414–420.

LAHIRI, P. and LARSEN, M. D. (2005). Regression analysis with linked data. *J. Amer. Statist. Assoc.* **100** 222–230. MR2156832 https://doi.org/10.1198/016214504000001277

LARSEN, M. D. (2005). Hierarchical Bayesian record linkage theory.

LARSEN, M. D. and RUBIN, D. B. (2001). Iterative automated record linkage using mixture models. *J. Amer. Statist. Assoc.* **96** 32–41. MR1973781 https://doi.org/10.1198/016214501750332956

LEE, J. S., SHANNON, J. and BROWN, A. (2015). Characteristics of older Georgians receiving older Americans act nutrition program services and other home- and community-based services: Findings from the Georgia aging information management system (GA AIMS). *J. Nutr. Gerontol. Geriatr.* **34** 168–188.

LLOYD, J. L. and WELLMAN, N. S. (2015). Older americans act nutrition programs: A community-based nutrition program helping older adults remain at home. *J. Nutr. Gerontol. Geriatr.* **34** 90–109. https://doi.org/10.1080/21551197.2015.1031592

MURRAY, J. S. (2015). Probabilistic record linkage and deduplication after indexing, blocking, and filtering. *J. Priv. Confid.* **7**.

NETER, J., MAYNES, E. S. and RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *J. Amer. Statist. Assoc.* **60** 1005–1027. MR0193729

NEWCOMBE, H. B. (1988). *Handbook of Record Linkage*: *Methods for Health and Statistical Studies*, *Administration*, *and Buisiness*. Oxford Univ. Press, Oxford.

NEWCOMBE, H. B., KENNEDY, J. M., AXFORD, S. J. and JAMES, A. P. (1959). Automatic linkage of vital records. *Science* **130** 954–959.

NEWMAN, L. M., SAMUEL, M. C., STENGER, M. R., GERBER, T. M., MACOMBER, K., STOVER, J. A. and WISE, W. (2009). Practical considerations for matching STD and HIV surveillance data with data from other sources. *Public Health Rep.* **124** 7–17. https://doi.org/10.1177/00333549091240S203

NEYMAN, J. (1923). Sur les applications de la thar des probabilities aux experiences agaricales: Essay de principle. English translation of excerpts by Dabrowska, D. and Speed, T. (1990). *Statist. Sci.* **5** 465–472.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

RUBIN, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics* **29** 159–183.

RUBIN, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29** 185–203.
RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152
RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.
RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.
RUBIN, D. B. (1990). Formal mode of statistical inference for causal effects. *J. Statist. Plann. Inference* **25** 279–292.
RUBIN, D. B. (1996). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91** 473–489.
RUBIN, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Stat. Neerl.* **57** 3–18. MR2055518 https://doi.org/10.1111/1467-9574.00217
RUBIN, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statist. Sci.* **21** 299–309. MR2339125 https://doi.org/10.1214/088342306000000114
RUBIN, D. B. (2008). Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. In *Epidemiology and Medical Statistics. Handbook of Statist.* **27** 28–63. Elsevier/North-Holland, Amsterdam. MR2500431 https://doi.org/10.1016/S0169-7161(07)27002-6
SADINLE, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *J. Amer. Statist. Assoc.* **112** 600–612. MR3671755 https://doi.org/10.1080/01621459.2016.1148612
SADINLE, M. and FIENBERG, S. E. (2013). A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. *J. Amer. Statist. Assoc.* **108** 385–397. MR3174628 https://doi.org/10.1080/01621459.2012.757231
SCHEUREN, F. and WINKLER, W. (1993). Regression analysis of data files that are computer matched—Part I. *Surv. Methodol.* **19**.
SCHEUREN, F. and WINKLER, W. (1997). Regression analysis of data files that are computer matched—Part II. *Surv. Methodol.* **23** 157–165.
SHAN, M., THOMAS, K. S. and GUTMAN, R. (2021). Supplement to "A multiple imputation procedure for record linkage and causal inference to estimate the effects of home-delivered meals." https://doi.org/10.1214/20-AOAS1397SUPP
SHEN, Z. (2000). Nested multiple imputations. Thesis (Ph.D.)—Harvard University. MR2700720
STAN DEVELOPMENT TEAM (2018). RStan: The R interface to Stan. R package version 2.17.3.
STEORTS, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Anal.* **10** 849–875. MR3432242 https://doi.org/10.1214/15-BA965SI
STEORTS, R. C., HALL, R. and FIENBERG, S. E. (2016). A Bayesian approach to graphical record linkage and deduplication. *J. Amer. Statist. Assoc.* **111** 1660–1672. MR3601725 https://doi.org/10.1080/01621459.2015.1105807
STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313
TANCREDI, A. and LISEO, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Ann. Appl. Stat.* **5** 1553–1585. MR2849786 https://doi.org/10.1214/10-AOAS447
TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357
THOMAS, K. S., AKOBUNDU, U. and DOSA, D. (2015). More than a meal? A randomized control trial comparing the effects of home-delivered meals programs on participants' feelings of loneliness. *J. Gerontol. Ser. B* **71** 1049–1058.
THOMAS, K. S. and MOR, V. (2013). Providing more home-delivered meals is one way to keep older adults with low care needs out of nursing homes. *Health Aff.* **32** 1796–1802.
THOMAS, K. S., GADBOIS, E. A., SHIELD, R. R., AKOBUNDU, U., MORRIS, A. M. and DOSA, D. M. (2018a). "It's not just a simple meal. It's so much more": Interactions between meals on wheels clients and drivers. *J. Appl. Gerontol.* **39** 151–158.
THOMAS, K. S., PARIKH, R. B., ZULLO, A. R. and DOSA, D. (2018b). Home-delivered meals and risk of self-reported falls: Results from a randomized trial. *J. Appl. Gerontol.* **37** 41–57. https://doi.org/10.1177/0733464816675421
WINKLER, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 667–671.
WINKLER, W. (1989). Near automatic weight computation in the Fellegi–Sunter model of record linkage.
WINKLER, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 354–359.

WINKLER, W. E. (1993). Improved decision rules in the Fellegi–Sunter model of record linkage.
WORTMAN, J. H. and REITER, J. P. (2018). Simultaneous record linkage and causal inference with propensity score subclassification. *Stat. Med.* **37** 3533–3546. MR3862902 https://doi.org/10.1002/sim.7911
ZANELLA, G. (2020). Informed proposals for local MCMC in discrete spaces. *J. Amer. Statist. Assoc.* **115** 852–865. MR4107684 https://doi.org/10.1080/01621459.2019.1585255