Published in final edited form as:

Stat Med. 2016 January 15; 35(1): 115–129. doi:10.1002/sim.6586.

Error Adjustments for File Linking Methods Using Encrypted Unique Client Identifier (eUCI) with Application to Recently Released HIV+ Prisoners

R. Gutman^{a,*}, C.J. Sammartino^b, T.C. Green^c, and B.T. Montague^d

^aDepartment of Biostatistics, Brown University ^bDepartment of Health Services, Policy and Practice, Brown University ^cDepartments of Emergency Medicine and Epidemiology, the Warren Alpert School of Medicine of Brown University, Rhode Island Hospital ^dBrown University and the Miriam Hospital, in Providence, Rhode Island

Abstract

Incarceration provides an opportunity to test for HIV, provide treatment such as highly active anti-retroviral therapy, as well as link infected persons to comprehensive HIV care upon their release. A key factor in assessing the success of a program that links released individuals to care is the time from release to receiving care in the community (linkage time). To estimate the linkage time, records from correction systems are linked to Ryan White Clinic data using encrypted Unique Client Identifier (eUCI). Most of the records linked using eUCI belong to the same individual; however, in some cases it may link records incorrectly, or not identify records that should have been linked. We propose a Bayesian procedure that relies on the relationships between variables that appear in either of the data sources, as well as variables that exists in both to identify correctly linked records among all linked records. The procedure generates K datasets in which each pair of linked records is identified as a true link or a false link. The K datasets are analyzed independently and the results are combined using Rubin's multiple imputation rules. A small validation dataset is used to examine different statistical models, and to inform the prior distributions of the parameters. In comparison to previously proposed methods, the proposed method utilizes all of the available data and is both flexible and computationally efficient. In addition, this approach can be applied in other file linking applications.

Keywords

File Linking; Multiple Imputation; eUCI; Mixture Models

1. Introduction

1.1. Background

Incarcerated populations have been disproportionately impacted by the HIV epidemic. The prevalence of HIV in this community is amongst the highest of tracked subpopulations and

*Correspondence to: 121 S. Main St., Providence, RI 02912 roee_gutman@brown.edu.

consistently higher than in the general community. Prior to incarceration, many inmates lack health insurance or a regular source of medical care, and periods of incarceration present the opportunity to diagnose, initiate treatment, and provide regular care for other comorbid conditions. The health benefits for those identified and treated during incarceration are frequently lost as infected individuals transition back into the community, with many experiencing substantial interruptions in treatment and care [1, 2]. Although programs to support linkage and retention of HIV infected persons upon reentry to the community exist, they are often limited in scope and systematic frameworks are needed to support the development of programs on a larger scale [3, 4].

Key metrics that reflect the adequacy of linkage to care upon reentry to the community include the time from release until the first service in the community (linkage time), the clinical status of the patient at the time of first service and the retention in care following the initial service. Among these metrics, linkage time is central because prolonged linkage time is generally associated with a decline in clinical status at the time of first service. In order to generate these metrics, release records for HIV+ inmates as well as records of their care once released to the community are required. Although full linkage of these datasets would facilitate estimation of the distribution of linkage time as well as the associations of linkage time with various demographic characteristics, it would require access to identifying information that is not publicly available. Thus, confidential linking of limited datasets is needed to assess the adequacy of linkage to care.

The Linkage Into Care Study (LINCS), proposed a framework for linking corrections release data to clinical data from Ryan White service providers using the confidential encrypted Unique Client Identifier (eUCI) [5]. The eUCI was developed by the Sphere Institute under contract from the Health Resources Services Administration (HRSA) to support confidential reporting of client level data regarding HIV care [6]. The eUCI is a unique identifier that is based on individuals' first and third characters of their first and last names, their full birth date, and their gender. The combination of these elements is encrypted using an encryption algorithm with the important property that once encrypted these elements cannot be reconstructed from the final identifier [7]. The eUCI is the identifier used for client-level data reporting for Ryan White care programs to the HRSA as well as the eHARS system used by the state departments of health for viral load surveillance. As key safety providers, Ryan White care programs are likely to be the first source of care for persons upon reentry to the community in many jurisdictions. The national correctional datasets, which are managed by the Bureau of Justice Statistics do not provide eUCI; however, they include the components required to calculating eUCIs. Using publicly available software [8] eUCI is created for every record in the correctional dataset, and deterministic eUCI matching is used to link it to the Ryan White dataset.

1.2. File Linking - Previous Work

The statistical literature includes methods both for exact matching record linkage and for analysis of linked datasets. The exact file linking process involves two files, A and B , where each file has n_a and n_b records, respectively. Record $i \in \{1, \dots, n_a\}$ includes the variables $(\mathbf{Y}_{Ai}, \mathbf{X}_{Ai})$ and record $j \in \{1, \dots, n_b\}$ includes the variables $(\mathbf{Y}_{Bj}, \mathbf{X}_{Bj})$, where $\mathbf{Y}_{Ai}$ and $\mathbf{Y}_{Bj}$

respectively represent variables available exclusively in file A and file B , and $\mathbf{X}_{Ai}$ and $\mathbf{X}_{Bj}$ represent variables that are assumed to be reported in both files, sometimes with errors. The linking process attempts to identify a subset of the set of pairs of records $A \times B = (i, j)$, $i \in A$, $j \in B$ which include information on the same individuals.

Record-linkage methods fall into two general categories: deterministic and probabilistic. Deterministic methods link records based on agreement, usually exact, between data elements common to both records, while probabilistic methods link records based on probabilities (or likelihoods under a probability model) that pairs of records from the two files are exact links. These probabilities are commonly estimated from the distribution of records' characteristics agreement in the general population, and are sometimes based on preliminary study or on a manually identified subset of records.

Deterministic methods are widely used in routine practice and can be as simple as establishing record linkages based on exact characters correspondence between one or more common data elements, such as the first and last names. However, very little has been written on them in peer-reviewed journals [9]. Probabilistic record linkage relies on a model proposed by Fellegi and Sunter [10], which formalized ideas of Newcombe et al. [11]. The Fellegi and Sunter [10] model assumes that the set $A \times B$ is composed of two disjoint subsets, the set of true links, M , and the set of true non-links, U . For each pair $\{i, j\}$ $i \in A$ and $j \in B$ let $\gamma_{ij} = (\gamma_{ij1}, \dots, \gamma_{ijP})$ be an observed agreement vector on P covariates that appear in both files, $(\mathbf{X}_{Ai}, \mathbf{X}_{Bj})$. The Fellegi-Sunter algorithm assigns a score to each pair based on the conditional probability that γ_{ij} is a true link, $m(\gamma_{ij}) = P(\gamma_{ij}|(i, j) \in M)$, and the conditional probability that γ_{ij} is a non-link, $u(\gamma_{ij}) = P(\gamma_{ij}|(i, j) \in U)$. These scores are then used in probabilistic record linking algorithms, often of the "greedy" type, to iteratively link and remove from the matching pool the pair with the highest probability of a match. Possible links with probabilities that are below a cutoff value are either clerically reviewed or declared non-links.

Deterministic linking based on perfect agreements was reported to have higher rate of true links among those reported than probabilistic linking [9, 12]. However, when the underlying data elements are subject to variation in spelling, data-entry inaccuracies, completeness of data, etc., deterministic linking may have a high level of missed true links in comparison to probabilistic linking [9, 12]. These drawbacks may limit the usefulness of simple deterministic methods in practical record-linking applications that use large public health datasets [13]. Both deterministic and probabilistic linking suffer from the presence of incorrectly linked records, which may become more prevalent when only a small proportion of the records in the two files correspond to the same subjects. Mixture models have been proposed to estimate the probabilities of incorrect links [14, 15]. Let π_m be the probability that a pair of records is a true link, then the probability of observing pattern γ_{ij} can be formally defined as:

$$P(\gamma_{ij}|\pi_m, \theta) = P(\gamma_{ij}|(i, j) \in M, \theta_M)\pi_m + P(\gamma_{ij}|(i, j) \in U, \theta_U)(1 - \pi_m) \quad (1)$$

where $\theta = \{\theta_M, \theta_U\}$ are the parameters governing these distributions. Using Bayes Rule, we obtain that the probability of a linked pair to be a true link given γ_{ij} is

$$P((i, j) \in M | \gamma_{ij}, \pi_m, \theta) = \frac{P(\gamma_{ij} | (i, j) \in M, \theta_M) \pi_m}{P(\gamma_{ij} | (i, j) \in M, \theta_M) \pi_m + P(\gamma_{ij} | (i, j) \in U, \theta_U) (1 - \pi_m)} \quad (2)$$

At times, statistical analysis may be adversely affected by incorrectly linked records [16]. Neter et al. [17] recognized that in finite population sampling, a relatively small number of incorrectly linked records could lead to a substantial bias in estimating regression coefficients of an outcome Y_A on $(\mathbf{X}_B, \mathbf{Y}_B)$. Scheuren and Winkler [18] also investigated the effects of incorrectly linked records on the bias of ordinary least squares estimators of regression coefficients and proposed a method that adjusts for the resulting bias. Scheuren and Winkler [19] advanced the work further by proposing an iterative procedure that modified the regression estimates and the linking results for apparent outliers.

Lahiri and Larsen [16] considered an alternative adjustment by providing an unbiased estimator for a transformed linear regression model. Chambers et al. [20] generalized this method by solving a set of estimating equations. Recently, Hof and Zwinderman [21] have extended the method by Lahiri and Larsen [16] to other generalized linear models using weighted least squares estimators. All of these adjustments assume that either $P((i, j) \in M)$ or $P((i, j) \in M | \gamma_{ij}, \pi_m, \theta)$ are known or can be estimated from the data. Using those probabilities and assuming that the linkage is non-informative [20]:

$$P((i, j) \in M | \mathbf{Y}_A, \mathbf{X}_A, \mathbf{X}_B, \mathbf{Y}_B, \pi_m, \theta) = P((i, j) \in M | \mathbf{X}_A, \mathbf{X}_B, \mathbf{Y}_B, \pi_m, \theta), \quad (3)$$

unbiased estimates of the regression coefficients can be obtained. Gutman et al. [22] proposed an algorithm that utilizes variables that are available in both files as well as variables that are available in only one of the files to provide multiply imputed linked datasets, that can be analyzed separately and combined using Rubin's Rules for multiple imputation (MI) [23]. Their approach assumes that both files include the same individuals, and combines record linkage and estimation within a single procedure by relying on the correlations between all of the variables in files A and B .

1.3. Current Work

Linking records of released inmates who are HIV+ with Ryan White clinical records introduces several complications that require extensions of the previously proposed methods. First, only a small subset of the records in both files represents the same individuals. The other records include inmates who are HIV−, or patients who are HIV+ that have not been incarcerated. This scenario has been identified as difficult for record linking [19]. Second, the number of available linking covariates that appear in both files is small, and the requirement for using eUCI reduces the number of covariates even further. The eUCI linked pairs, generally have $\mathbf{X}_{Ai} = \mathbf{X}_{Bj}$, which complicates the estimation of the probabilities of true and false links using agreement vectors. As a result, the assumption of non-informative linking is less plausible because the observed covariates do not contain information on the classification of eUCI identified links as true links or false links. Third, the adoption of eUCI, leads to using a deterministic linking algorithm, and one cannot determine the underlying covariates responsible for differences between two eUCI values, which precludes the application of algorithms like Gutman et al. [22]. Fourth, although the

number of linked records is not large and most of these are true links, manual review of the linked records entails a significant amount of work, because it requires a specialist to compare prison medical files and Ryan White records. A significant amount of work was invested to generate a small validation dataset in Rhode Island (RI), and repeating this procedure annually, and in larger states, is not sustainable. Fifth, a validation dataset was used to examine the performance of deterministic and probabilistic linking on such datasets, and it would be useful to utilize this dataset to inform the analysis of other linked files in which the classification of links is unknown.

To address the issues described above, we propose a four-stage procedure that is based on a missing data perspective and builds upon concepts from previously suggested methodologies. First, a latent indicator variable defining whether a linked record is a true link is introduced. Second, using a Bayesian mixture model we estimate the probability that a pair of records is a true link and using this probability we multiply impute a latent indicator for each pair of linked records. The outcome of this stage comprise K imputed linked datasets in which each pair of linked records is classified as a true link or a false link. Third, for each imputed dataset we calculate the estimand of interest based on the true linked records. Fourth, we combine the estimands using Rubin's Rule for multiple imputation (MI) [23] when K is small, or using posterior quantiles when K is large. This method enables the investigators to identify true links once, using an extensive model, and later on perform statistical analysis using different models of interest. In addition, this method utilizes the validation study as a prior distribution when estimating the probabilities of true links.

2. Validation Study

The error rates for record linkage using the eUCI were assessed by the Sphere Institute using simulation. The estimated false negative linkage rate was 3.8% and false positive linkage rate was 5.0% [6]. However, the performance of this method when applied to larger datasets as well as datasets from outside the Ryan White care system such as correctional institutions have not been investigated. We have constructed a validation dataset that enables comparison of eUCI linking and commonly used deterministic and probabilistic linking procedures. Varying combinations of covariates were included in the linking algorithms to assess their linking rates (see Table 1 for more details). eUCI matching methods included both simple and alias-inclusive algorithms. Alias-inclusive algorithms are similar to simple eUCI matching algorithm, but for each released inmate for whom an alias is recorded, additional eUCIs are created for each alias name using the same birth date. Exact deterministic matching is performed for all eUCIs, and a pair of records are declared a link if at least one of those eUCIs is matched to a record in the other file. The deterministic linking algorithm involved direct matching based on combinations of elements from the datasets including first name, last name, aliases, age and gender. The probabilistic matching method was implemented using Link King Software (version 6.5.1, Tacoma Seattle, WA).

2.1. Validation Data

The RI ACI validation dataset consisted of sentenced inmates who were released between January 1, 2010 and December 31, 2010. Sentenced inmates were those who had served out their prison sentence. This dataset includes 3,957 unique individuals, of whom 3,708 were

released to the community with an opportunity to link to care, provided they were HIV+. Of those not released to the community four were never released, four passed away while incarcerated, and 241 were either discharged to immigration, to the US Marshall, to a mental health institution or to the military.

The Miriam Hospital's Immunology Clinic is the main provider of Ryan White-funded HIV care in RI, providing over 75% of the HIV care within RI, with roughly 1,383 HIV/AIDS patients in 2010. More than 90% of persons with HIV released from corrections who receive care are seen at the Miriam Hospital Clinic. The clinical dataset from The Miriam Hospital Immunology Center include all current and previously active patients who received care services during the validation study interval. Clinic utilization data was evaluated for 1,432 people with HIV-related clinic or social work visits within the period January 1, 2010 to June 30, 2011, to allow for a six-month window for possible linkage to care following release.

Both the RI ACI and the HIV clinic datasets include individuals' first name, last name, date of birth and gender. The RI ACI file also includes aliases, dates of admission, dates of release and release types, while the HIV clinic database include dates of care and care type (clinic or social work visit).

To construct the validation dataset we followed the following three-stage procedure: 1) electronically matching all released inmates to the clinic data to determine maximum possible links; 2) manually reviewing all candidate links in the clinic data system to confirm the accuracy of the link; and 3) for those who did not link, reviewing the clinical records from the corrections HIV care program to confirm HIV status or challenge the link status. This manual procedure identified 102 individuals infected with HIV that were incarcerated. However, only 79 of these individuals linked to care at the HIV clinic. Thus, only 79 individuals can be identified from linking the two datasets.

2.2. Validation Results

Table 1 depicts the number of links detected by each linking method as well as the number of true links (true positives (TP)), false links (false positives (FP)), and false non-links (false negatives (FN)). Linking by eUCI inclusive of aliases identified the largest number of true links (93%). Moreover, eUCI inclusive of aliases also had the smallest percentage of FP among all links identified. eUCI linking without aliases identified only 84% of the true links, but with relatively small percentage of FP in comparison to the deterministic methods. Probabilistic linking was the second best method identifying 90% of the TP and only 13% of the total links identified were FP.

These results indicate that using eUCI based linking can provide results that are comparable to probabilistic linking when linking Ryan White Clinic data with correction system records. However, both methods suffer from FP and FN errors. In order to obtain valid inferences, adjustments for such errors are required.

3. Methods

3.1. Notations and Model Structure

We replace mixture model (1) with a mixture model that includes $(\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj})$, s.t $i \in A, j \in B$, instead of agreement vector γ_{ij} :

$$P(\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj} | \pi_M^*, \theta^*) = \pi_M^* P((\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj}) | (i, j) \in M, \theta_M^*) + (1 - \pi_M^*) P((\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj}) | (i, j) \in U, \theta_U^*). \quad (4)$$

where π_M^* is the probability that a pair of records is a true link, and $\theta^* = \{\theta_M^*, \theta_U^*\}$ are the parameters governing the two conditional distributions. It is important to note that these parameters might be different than the ones in (1). This replacement may seem technical, but using variables that do not appear in both files as well as the actual values of the covariates could improve the identification of true links among all links, in comparison to agreement vector γ_{ij} .

The eUCI linking algorithm relies on exact matching of unique identifiers, that are based on $\mathbf{X}_A$ and $\mathbf{X}_B$, to generate n pairs of records that are declared as links. Let g_{ij} be an indicator variable that is equal to one if eUCI algorithm identifies the pair (i, j) as a link and 0 otherwise, and let $\pi = \{\pi_{mg}, m \in \{M, U\}, g \in \{0, 1\}\}$, such that $\pi_{mg} = P((i, j) \in m \cap g_{ij} = g)$ then

$$P(\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj} | \pi, \theta^{**}) = \pi_{M1} P((\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj}) | (i, j) \in M \cap g_{ij} = 1, \theta_{M1}) + \pi_{U1} P((\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj}) | (i, j) \in U \cap g_{ij} = 1, \theta_{U1}) + \pi_{M0} P((\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj}) | (i, j) \in M \cap g_{ij} = 0, \theta_{M0}) + \pi_{U0} P((\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj}) | (i, j) \in U \cap g_{ij} = 0, \theta_{U0}). \quad (5)$$

where $\pi_{M1} + \pi_{U1} + \pi_{M0} + \pi_{U0} = 1$ and $\theta^{**} = \{\theta_{mg}, m \in \{M, U\}, g \in \{0, 1\}\}$ are the parameters governing those distributions. The fourth line of Equation (5) is commonly not of interest in most applications, as it represents the distribution of non-links that were not identified as links by the linking algorithm. The third line of Equation (5) represents the distribution of false non-links. Estimating the proportion and distribution of false non-links is still an open problem, even in situations where a representative sample is available [24]. The difficulty arises because this analysis requires manual sorting through all possible unlinked pairs to identify individuals whose characteristics are recorded differently in the two files. This problem is exacerbated here, because there are many individuals who are HIV+ that were never incarcerated, and many released individuals that are HIV-. A simplifying assumption that will be made is that $P((\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj}) | (i, j) \in M \cap g_{ij} = 1, \theta_{M1}) = P((\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj}) | (i, j) \in M \cap g_{ij} = 0, \theta_{M0})$. Based on this assumption, results that are observed for records that were linked correctly by the linking algorithm can be generalized to the population of individuals that were missed by the linking algorithm. This is a strong assumption, but the validation data does not provide any evidence to contradict it. The first line of Equation (5) comprises the distribution of all the pairs that were identified as links by the eUCI linking algorithm. Thus, $P(\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj} | g_{ij} = 1, \theta^{**})$ is also a

mixture of correctly linked and falsely linked individuals. Estimating this mixture will enable the identification of true links among the entire cohort of linked records.

3.2. Method

For probabilistic linking, several algorithms that rely on a mixture distribution have been proposed to estimate the probability that a declared link is false [14, 15, 25]. These estimates can then be used to adjust for linear regression models [16] as well as for generalized linear models [20]. Recently, Goldstein et al. [26] treated the Y_{Bj} as missing values in file A. Assuming missing at random (MAR) [27], Goldstein et al. [26] combined scores based on $P(\gamma_{ij}|(i, j) \in M, \theta_M)$ and $P(\gamma_{ij}|(i, j) \in U, \theta_U)$ with prior information on the association between $\mathbf{Y}_{Ai}$, $\mathbf{Y}_{Bj}$, $\mathbf{X}_{Ai}$ and $\mathbf{X}_{Bj}$ to multiply impute the $\mathbf{Y}_{Bi}$ for each record in A. We take a different approach that relies on a mixture distribution to multiply impute for each eUCI linked record whether it is a true or a false link.

In contrast to probabilistic linking, the eUCI linking method does not provide any scores that inform the designation of a linked pair of records as a true link or a false link. Moreover, available $\mathbf{X}_{Ai}$ and $\mathbf{X}_{Bj}$ are practically similar, which does not enable us to rely on different similarity measures to distinguish between true and false links. However, the relationship between $\mathbf{Y}_{Ai}$ and $\mathbf{Y}_{Bj}$, possibly conditional on $\mathbf{X}_{Ai}$ and $\mathbf{X}_{Bj}$, may enable to distinguish between true and false links. While the form of $P(\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj}|g_{ij} = 1)$ in Equation (5) is specific to the application, it will often be convenient to express it as a mixture of products of conditional and marginal distributions, for example

$$\eta P_{M \cdot X}((\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj})|\mathbf{X}_{Ai}, \mathbf{X}_{Bj}, (i, j) \in M, g_{ij}=1, \theta_{M \cdot X}) P_{MX}(\mathbf{X}_{Ai}, \mathbf{X}_{Bj}|(i, j) \in M, g_{ij}=1, \theta_{MX}) + (1-\eta) P_{U \cdot X}((\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj})|\mathbf{X}_{Ai}, \mathbf{X}_{Bj}, (i, j) \in U, g_{ij}=1, \theta_{U \cdot X}) P_{UX}(\mathbf{X}_{Ai}, \mathbf{X}_{Bj}|(i, j) \in U, g_{ij}=1, \theta_{UX}) \quad (6)$$

where $\theta_M = \{\theta_{M \cdot X}, \theta_{MX}\}$ and $\theta_U = \{\theta_{U \cdot X}, \theta_{UX}\}$ are the parameters governing these distributions, and η is the probability of a true link among all linked record identified by the algorithm.

Following the work of many authors [e.g. 28, 29, 30] we formulate the mixture model in terms of unobserved indicators of component membership Z_l , $l = 1, \dots, n$, where $Z_l = 1$ if pair l , composed of records (i, j) , is a true link and $Z_l = 0$ if pair l is a false link. The mixture model can then be expressed as the following model,

$$\begin{aligned} (\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj})|g_{ij}=1, \theta_M, Z_l=1 &\stackrel{iid}{\sim} P_{M \cdot X}(\cdot|\theta_{M \cdot X}) P_{MX}(\cdot|\theta_{MX}) \\ (\mathbf{Y}_{Ai}, \mathbf{Y}_{Bj}, \mathbf{X}_{Ai}, \mathbf{X}_{Bj})|g_{ij}=1, \theta_U, Z_l=0 &\stackrel{iid}{\sim} P_{U \cdot X}(\cdot|\theta_{U \cdot X}) P_{UX}(\cdot|\theta_{UX}) \\ Z_l &\stackrel{iid}{\sim} \text{Bernoulli}(\eta). \end{aligned}$$

The “complete-data” likelihood, which assumes that the “missing data” $Z_1, \dots, Z_n$ are observed, can be written as

$$L(\theta_M, \theta_U, \eta | Y_A, Y_B, X_A, X_B, Z_1, \dots, Z_n, \forall (i, j) \text{ s.t. } g_{ij} = 1) = \prod_{l=1}^n [\eta P_{M \cdot X}(\cdot | \theta_{M \cdot X}) P_{MX}(\cdot | \theta_{MX})]^{Z_l} [(1-\eta) P_{U \cdot X}(\cdot | \theta_{U \cdot X}) P_{UX}(\cdot | \theta_{UX})]^{1-Z_l} \quad (7)$$

We use full Bayesian analysis to estimate θ_M , θ_U , η and impute the missing indicators $Z_1, \dots, Z_n$.

3.3. Bayesian Computation

Bayesian approaches to record linkage have been suggested by Larsen [31, 32], Fortini et al. [33, 34], and McGlinchey [35]. We adopt a Bayesian approach to inference because it enables us to create linked datasets that distinguish between true and false links. These datasets are obtained by sampling from the posterior distribution of $\mathbf{Z} = \{Z_l\}$, thus reflecting posterior uncertainty about $(\theta_M, \theta_U, \eta)$. The imputed datasets can be analyzed independently by researchers to provide summaries of the linkage times and their association with other characteristics. The results obtained from the different datasets can then be merged using common combination rules. In this formulation, we treat the unknown classification of linked records as missing data, and use a Data Augmentation (DA) [36] scheme that iterates between sampling the unknown classification and sampling the parameters $(\theta_M, \theta_U, \eta)$. While the algorithms for sampling $(\theta_M, \theta_U, \eta)$ given $\mathbf{Z}$ and the observed data are model-specific, the posterior distribution of Z_l corresponding to a linked record pair (i, j) given the parameters and the observed data is Bernoulli with probability:

$$P(Z_l=1 | Y_A, Y_B, X_A, X_B, \eta) = \frac{\eta P_{M \cdot X}(Y_{Ai}, Y_{Bj} | X_{Ai}, X_{Bj}) P_{MX}(X_{Ai}, X_{Bj})}{(\eta P_{M \cdot X}(Y_{Ai}, Y_{Bj} | X_{Ai}, X_{Bj}) P_{MX}(X_{Ai}, X_{Bj}) + (1-\eta) P_{U \cdot X}(Y_{Ai}, Y_{Bj} | X_{Ai}, X_{Bj}) P_{UX}(X_{Ai}, X_{Bj}))} \quad (8)$$

where for simplicity we omit conditioning on θ_M and θ_U .

One simplifying assumption is that the distribution of covariates among the true links and the false links are similar. Formally,

$$P_{MX}(X_{Ai}, X_{Bj} | (i, j) \in M, g_{ij}=1, \theta_{MX}) = P_{UX}(X_{Ai}, X_{Bj} | (i, j) \in U, g_{ij}=1, \theta_{UX}) = P_X(X_{Ai}, X_{Bj} | (i, j) \in M, g_{ij}=1, \theta_X). \quad (9)$$

Using this assumption, P_X cancels out in Equation (8) and hence, modeling the marginal distribution of X_{Ai} and X_{Bj} is inessential unless it informs the estimation of the parameters $(\theta_{M \cdot X}, \theta_{U \cdot X}, \eta)$.

To complete the Bayesian modeling we specified a prior distribution for the unknown parameters. Informative prior distribution for the unknown parameters is especially useful when using finite mixture models, because these models are not identifiable, in the sense that the distribution is unchanged if the group labels are permuted [37]. Information from previous record linkage operations has been used informally to select models [15] as well as to restrict parameters [38, 39]. Belin and Rubin [14] used “training data” to inform values of certain “global” parameters in the model. Although prior distributions are model-specific

and will be discussed in the following sections, we will utilize the validation study to inform these distributions when analyzing the complete dataset.

A byproduct of the Bayesian modeling is that at each DA algorithm iteration, $k, k \in \{1, \dots, K\}$, a vector $\mathbf{Z}^{(k)}$ is sampled. These vectors can be used to generate K datasets in which the designation of true links and false links are known. Analysis for an estimand of interest ν is performed on each dataset separately and the results are combined using Rubin's Rules for MI. Formally, let $\nu^{(\hat{k})}$ be the estimate for ν , and let $V^{(\hat{k})}$ be its sampling variance at dataset k using only links for which $Z_l^{(k)}=1$. The combined estimate for ν across the K datasets is

obtained by $\hat{\nu} = \frac{1}{K} \sum_{k=1}^K \hat{\nu}^{(k)}$ and its standard error by $\sqrt{\frac{1}{K-1} \sum_{k=1}^K (\hat{\nu}^{(k)} - \hat{\nu})^2 + \frac{1}{K} \sum_{k=1}^K \hat{V}^{(k)}}$. For small K , an interval estimate is calculated using the t approximation for the posterior variance of ν given by Barnard and Rubin [40], and for large K an interval estimate is calculated using percentiles of $\nu^{(\hat{k})}$.

4. Application to Linkage to Care Time

The proposed procedure was applied to files obtained from the RI Department of Corrections and The Miriam Hospital HIV care program between July 1, 2009 and December 31, 2013. The total number of unique incarcerated individuals who were released is 11,854 and the total number of individuals who are HIV+ and who visited the Miriam Immunology clinic was 1,634. Using eUCI linking that is based on individual's gender, birth date, first name, last name and alias, 134 pairs were identified. For each of those individuals we recorded $\mathbf{X}_A = \mathbf{X}_B = \mathbf{X} = (\text{Gender}, \text{Race}, \text{Ethnicity}, \text{Age})$, $\mathbf{Y}_A = (\text{First Clinic Visit})$ and $\mathbf{Y}_B = (\text{Release Date})$. Other variables that appear in either files A or B are related to the type of release, as well as clinical characteristics, income and housing status. The linkage time, $T_l = Y_{Bl} - Y_{Al}$, $l \in \{1, \dots, n\}$, has policy and clinical implications. Thus, instead of defining models for $P_m((Y_{Al}, Y_{Bl})|\mathbf{X}_l, \boldsymbol{\theta}_m)$, $m \in \{M, U\}$, in equation (6), we will define models for $P_m(T_l|\mathbf{X}_l, \boldsymbol{\theta}_m)$. It is important to note that some of the linkage times in the validation data and among the 134 linked individuals are censored. These linkage times belong to individuals that have visited the clinic in the past, and now either receive their care elsewhere or have not been linked to care post release from incarceration when the Miriam Hospital file was formulated. Throughout, we will refer to the dataset composed of the 134 linked individuals as the study's data.

Figure 1a displays the distribution of log linkage times for TP, FP, and FN groups in the validation study as well as the distribution of log linkage times in the study's data for non-censored individuals. The FP group has shorter average linkage times than either the TP or the FN groups. While, the TP and the FN groups seem to have relatively similar average linkage times, with the smaller mode in the FN group arising because of the small number of uncensored records in this group. The study's dataset shows that the linkage time in the population is composed of two modes. One mode is close to the modes of the TP and FN groups, and the other is close to the mode of the FP group. Figure 1b displays the proportion of patients that have not linked to care by the number of days post release. In the validation data, all of the censored individuals are either at the TP or FN groups. In addition, the curve

of the study's data is close to the TP curve when the number of days post release is small, and is between the FN and the TP curves when the number of days post release is large. These observations provide support to the conjecture that linkage times can be used to distinguish between TP and FP links.

The log-normal distribution is commonly used to approximate time to event data [41]. Figure 1a demonstrates that for non-censored observations, $\log(T_l)$ has an approximate normal distribution within the FP and TP groups. Using the validation dataset we examined a few log-Normal models that condition on different combinations of the covariates. Let I_l be an indicator function that is equal to 1 if T_l is censored and zero otherwise. Assuming non-informative censoring [42], the likelihood function of all linked pairs within the TP group or the FP group is:

$$L(\boldsymbol{\theta}_m | \mathbf{T}, \mathbf{I}, \mathbf{Z}, \mathbf{X}_A, \mathbf{X}_B) = \prod_{l=1}^n f(T_l | \boldsymbol{\theta}_m, Z_l=m)^{(1-I_l)} \times \left(\int_{T_l}^{\infty} f(x | \boldsymbol{\theta}_m, Z_l=m) dx \right)^{I_l} \quad (10)$$

$$f(T_l | \boldsymbol{\theta}_m, Z_l=m) \propto \frac{1}{T_l \sigma_m} \exp \left(- \left(\frac{(\log(T_l) - \mu_{lm})^2}{2\sigma_m^2} \right) \right)$$

where $\mu_{lm} = \beta_{m0} + \mathbf{X}_l' \boldsymbol{\beta}_m$, $\mathbf{X}_l' \subseteq \mathbf{X}_p$, and $\boldsymbol{\beta}_m = \{\beta_{m1}, \dots, \beta_{mP'}\}$, $P' = |\mathbf{X}_l'|$, $m \in \{M, U\}$.

To complete the Bayesian model we assume that $\beta_{mp} \stackrel{i.i.d}{\sim} N(0, 100)$, $\forall m \in \{M, U\}$, $p \in \{0, \dots, P'\}$, $\sigma_m^2 \sim \text{Inverse-Gamma}(1, 1)$, and that $\eta \sim \text{Beta}(3, 1)$. The prior distribution for η was chosen because it is generally known that a large proportion of identified links are true links. All of the other prior distributions are proper and relatively diffused, resulting in a proper posterior distribution.

The validation data did not provide any evidence that Assumption (9) is violated. Thus, as noted in Section 3.3, modeling P_X is inessential for classification of linked record as TP or FP. Plugging (10) into mixture model (7) and combining with the corresponding prior distributions result in a posterior distribution that does not have a close form. Censored linkage times could be seen as missing data [37] that can be imputed given the parameters and the classification of links. Sampling from the full posterior distribution requires the use of Markov chain Monte Carlo (MCMC) algorithm. Our MCMC algorithm iterated between three major steps. First, sampling the parameters, $\boldsymbol{\theta}_m$, $m \in \{M, U\}$, given the classification of links and the imputed censored linkage times. Second, sampling the missing classifications, given the parameters and imputed censored linkage times, from Bernoulli distributions with probabilities (8). Lastly, given the parameters and the classification of links we sampled the censored linkage times from a truncated log-Normal distribution. We applied this algorithm for each model using three MCMC chains starting from different positions with 10,000 samples from each chain. The Gelman and Rubin [43] potential scale reduction statistics projected little potential improvement in the estimates by increasing the number of iterations ($\hat{R} < 1.05$ for all scalar parameters). MCMC sampling was performed using JAGS 3.4.0 [44] and the samples were analyzed using the R 3.0 software [45].

Based on the validation dataset, for each of the conditional models, we estimated its predictive accuracy and the deviance information criterion (DIC) [46]. Table 2 displays the results for the different mixture of regression models. The model that includes the intercept

and an indicator for *Race* is the best in distinguishing between true and false links, and it also has the smallest DIC. However, the model that includes *Race*, *Gender* and *Age* has similar prediction accuracy and slightly larger DIC. For all models, changing the prior distribution of η to *Beta*(1, 1), resulted in similar trends to *Beta*(3, 1), with slightly worse classification of links as TP or FP (data not shown). Based on this analysis, the models

$$\log(T_l)|Z_l=m \sim N(\beta_{m0} + \beta_{m1} \times \text{Race}, \sigma_m^2) \forall l \in \{1, \dots, L\} \text{ and } \quad (11)$$

$$\log(T_l)|Z_l=m \sim N(\beta_{m0} + \beta_{m1} \times \text{Race} + \beta_{m2} \times \text{Gender} + \beta_{m3} \times \text{Age}, \sigma_m^2) \forall l \in \{1, \dots, L\}, \quad (12)$$

where N is the normal distribution, seemed the most appropriate. When using MI it is recommended to include all of the available variables [e.g. 23, 47, 48, 49]. Using more covariates in the model reduces the bias of unaccounted for correlations towards zero, as well as increases the validity of the imputation model [48]. Rubin [49] also noted that MI inferences are confidence-valid, even if mildly important predictors are left out of the imputation model. Thus, we will examine the performance and the results of models (11) and (12) on the study's dataset.

Following the recommendation of Belin and Rubin [14] and Larsen [50], when analyzing the study's dataset, we will replace the diffused prior distributions for $\beta_m = (\beta_{m0}, \beta_{m1})$, σ_m^2 and η , that were used earlier, by prior distributions that are based on the validation dataset. Formally, we assume that

$$\begin{aligned} \sigma_M^2 &\sim \text{Inv-Gama}(1, 1) \\ \beta_M | \sigma_M^2 &\sim N_P(\hat{\beta}_M, V_{M\hat{\beta}} \times 10) \\ \sigma_U^2 &\sim \text{Inv-Gamma}(1, 1) \\ \beta_U | \sigma_U^2 &\sim N_P(\hat{\beta}_U, V_{U\hat{\beta}} \times 10) \\ \eta &\sim \text{Beta}(\alpha_0, \beta_0) \end{aligned}$$

where $(\hat{\beta}_M, V_{M\hat{\beta}})$ and $(\hat{\beta}_U, V_{U\hat{\beta}})$ are the maximum likelihood estimates and their corresponding sampling covariance matrices from regression models (11) and (12) based on the validation data in the TP and FP groups, respectively. N_P is a P-dimensional Normal distribution, and (α_0, β_0) were chosen such that the mean of η will be equal to the proportion of true links among all links in the validation data, and the standard deviation will be 0.04. These prior distributions incorporate estimates from the validation study in a diffused fashion. No significant changes in the analysis were observed when the prior covariance matrices of β_m were multiplied by either 1 or 20 instead of 10, as well as increasing the standard deviation of η to be 0.1 (data not shown).

4.1. Simulation

To examine the performance of the proposed methods in comparison to an analysis that includes all of the linked units, we conducted a set of simulations. Let, X_l and Y_l be observed scalar covariate and scalar outcome for linked pair $l \in \{1, \dots, n\}$, respectively. In addition,

let W_l be an unobserved indicator variable representing whether the l th linked pair is a true link or a false link. The values of (X_l, W_l, Y_l) for the n linked subjects were generated from the following distributions:

$$\begin{aligned} X_l &\sim N(0, 1) \\ W_l &\sim \text{Bern}(\pi) \\ Y_l &\sim N(1 + X_l + 0.5W_l + \delta X_l W_l, \omega), \end{aligned}$$

where N represents the Normal distribution, Bern represents the Bernoulli distribution, δ defines the change in slopes between true links and false links, and ω the conditional variance of Y_l . For simplicity, we did not assume any censored observations. Table 3 describes the four factors and their two levels that were examined, resulting in a full factorial design with 2^4 levels.

The estimands of interest are the intercept, β_0 , and slope, β_1 of the linear regression of Y on X among the true-positives. Each factors' configuration was replicated 100 times and for each replication we recorded, $(\hat{\beta}_0, \hat{\beta}_1)$, the estimates of (β_0, β_1) , their sampling variance $(\hat{\sigma}_{\beta_0}^2, \hat{\sigma}_{\beta_1}^2)$, and determined whether 95% interval estimates covered or did not cover (β_0, β_1) for true links ($W_l = 1$). Using these values, we calculated for each configuration and each estimand, the mean coverage rate, the bias, the mean estimated sampling variance, and the mean squared error (MSE).

Three methods were examined in the simulations:

- a. Linear regression that included all linked units.
- b. The proposed method where we assume that

$$Y_l | X_l, \theta \sim \eta N(\mu_M, \sigma_M^2) + (1 - \eta) N(\mu_U, \sigma_U^2).$$

- c. The proposed method where we assume that

$$Y_l | X_l, \theta \sim \eta N(\mu_{0M} + \mu_{1M} X_l, \sigma_M^2) + (1 - \eta) N(\mu_{0U} + \mu_{1U} X_l, \sigma_U^2).$$

Method (c) represents the model that is used to generate the data and Method (b) represents a mis-specified model. The two mixture models used prior distributions similar to the ones used for the validation data analysis.

Table 4 summarizes the results for the different models across factors' configurations. In all of the configurations both methods (b) and (c) have coverages that are close to nominal for both β_0 and β_1 . Method (a) generally has coverages that are significantly lower than nominal, thus making this method statistically invalid for interval estimation [51]. Among the statistically valid methods, Method (c) has smaller biases, sampling standard error and MSEs, in comparison to Method (b). These results are expected, because Method (c) is similar to the model that generated the data. Despite the imperfect model, Method (b) provides valid interval estimates, but with sampling standard errors that are larger than the true model. Similar results have been observed in other studies where mis-specified models are used for imputations [52, 30, 49].

To address the issue of possible misspecification of the mixture model, we also examined the DIC for methods (b) and (c) as well as their success rate in identifying true and false links correctly. Across all of the configurations, Method (c) had an average DIC of 1,313 while Method (b) had an average DIC of 2,293. The range of differences between the average DIC for Method (b) to Method (c) across configurations is (233, 2,490), indicating significant preference to Method (c). In addition, the proportion of correctly classified subjects was 76% when using Method (c) to 60% when using Method (b). This analysis suggests that the observed study and validation data can identify models with better fit, thus reducing the negative effects of mis-specified models in the imputation process. This provides additional support to examine the performance of both models (11) and (12).

To examine the scalability of the proposed approach we have estimated the average running time of our algorithm across configurations using a single processor. The results for Model (c) were similar to the ones observed for Model (b). For sample sizes of 250, 500, and 2000 linked records, we observed an average running time of 15, 35, 205 seconds with standard errors of 0.3, 0.5, 104. As the sample size increases there is an approximately linear increase in the CPU time required. However, even for relatively large sample sizes the average CPU time required is well under 5 minutes. The usage of more efficient software can decrease this time even further.

4.2. Results

Linking the two files using eUCI linking resulted in 134 subjects, of which 11 did not visit the clinic after being released (e.g. censored). For inmates that had several release episodes, we selected the earliest release episode for which they actually linked to care, or in cases when the inmate did not link to care we selected the one with the longest time since release. Using models (11) and (12) we created K datasets in which each link is classified as a true link or a false link. In each of these K imputed datasets we estimated the proportion of linked inmates that have AIDS, have stable housing conditions, and are below federal poverty level. These characteristics are only observed in the Ryan White dataset. In addition, we also estimated the median linkage time for the entire set of true links, as well as for subgroups of truly linked released inmates. The median linkage times and their sampling standard errors were estimated using a parametric log-Normal model and assuming non-informative censoring. We also used the Weibull model to estimate the median linkage time, but the fit of the log-Normal model was better in terms of the Q-Q plot and resemblance of the Kaplan-Meier curve.

Table 5 depicts the results after adjustment for false positive links by Model (11), Model (12) and the unadjusted dataset. The point estimate for the proportion of released inmates with AIDS, the proportion that have stable housing, and the proportion that are below the federal poverty level is similar for both the adjusted datasets and the unadjusted dataset. However, the interval estimates are smaller for the unadjusted dataset. This is expected, because the unadjusted dataset does not take into account the possible errors in the linkage.

The overall estimated median linkage is higher by 10 days after adjustment for possible linking errors using Model (11) in comparison to the unadjusted dataset. There is also a significant increase in the width of the interval estimate. Similar trends are observed for the

different sub-populations. Model (12) has similar overall median linkage time to the unadjusted dataset. However, the median linkage times for females, and for younger released inmates (less than 40 years old) is higher than Model (11) and the unadjusted dataset, and shorter for older released inmates and males. For all models, the differences in linkage times between males and females and older and younger released inmates are notable, but not statistically significant at the 5% level.

5. Conclusion

Analysis of linked datasets is likely to become increasingly important as researchers and policy analysts seek to integrate two data sources while adapting to privacy regulations that limit access to unique identifiers. In our application, two data sources were linked using eUCI inclusive of aliases algorithm because of limited access to original identifiers. This linked dataset is used to identify the performance of different states in linking HIV+ released inmates to care. When using eUCI inclusive of aliases for record linking, a record in one file may be linked to two different records in the other file. In our experience, after examining several linked datasets from different states, the linked eUCIs were distinct and we did not encounter such issues. If such a problem arises, we advise checking both records manually to identify possible discrepancies in one of them.

Using a validation dataset, we showed that the eUCI linking algorithm provides comparable results to other commonly used file linking methods. However, all of the methods did not link some records that should have been linked, while linking records that should not have been linked. From the validation data, we were not able to identify variables that appear in both files that discriminate between falsely linked and truly linked records. However, we observed that the linkage time, which is a function of two variables, each coming from a different file, enables us to distinguish between true links and false links. Using this relationship, we developed an algorithm that provides multiple linked datasets in which each pair of linked records is identified as a true link or a false link. Each of these datasets can be analyzed by any model and combined using standard MI rules. Further analysis performed with these imputed datasets includes estimation of the correlations of clinical characteristics and linkage time for released inmates, as well as application of this method to other states. Adjusting for false positive linked individuals using the MI method separates the statistical analysis from the linkage error estimation. This separation makes fewer demands on the time and expertise of researchers and provides flexibility in the selection of statistical models.

The proposed procedure does not adjust for records that were supposed to be linked but were not. Identifying these records is difficult because it requires manual clerical processes that are labor intensive especially when only a small proportion of the records in the two data sources include similar individuals. The validation data did not identify any significant differences between true links that were identified and true links that were not identified in terms of linkage time as well as other covariates. This supports our assumption of no difference in distributions between false negatives and true positives. However, because of the relatively small number of subjects in the validation data, it would be useful to examine this issue in further research. Without knowledge on the true number of HIV+ inmates that

are released each year, our findings are only limited to those that either received care before incarceration or after being released. Thus, we are missing a significant group of subjects that received treatment in prison and never linked to our outside clinic. This group of inmates may increase the average linkage time, and may be regarded as censored by the number of years in our dataset. The validation study has shown that this group is about 22% of the total released inmates. The validation data can be used to adjust the estimate of the average linkage time under the assumptions of non-informative censoring [42] and that the population of RI released inmates does not change over time. In states where the total number of HIV+ inmates released can be obtained from prison records, the estimate can be adjusted without using the assumption that the population of released inmates is constant over time and only relying on the non-informative censoring assumption.

Here, we assumed that the marginal distribution of covariates that appear in both files is similar between true links and false links (Assumption (9)). The validation dataset did not provide any evidence to the contrary. In cases where the marginal distribution of $\mathbf{X}$ is different between TP and FP links, it is essential for the classification of links and should be modeled. Flexible models such as the general location model of Olkin and Tate [53] and its extensions [54, 55] that has been proposed for missing data imputation [e.g. 56, 57] can be used. These types of models were also recently extended for latent class mixture models as well [58].

The interval estimates using Rubin's Rules for MI are best suited for estimands that have posterior distribution, or, for likelihood-oriented statisticians, sampling distribution of the complete-data estimator that is approximately Gaussian [49, 59, 30]. When the assumption of normality of the complete-data posterior/sampling distribution is not justified, Rubin's Rules may result in inappropriate inferences. To address this issue we have calculated interval estimates on the transformed log survival scale [57], and exponentiated the endpoints of the interval. When a transformation is not attainable, and the normality assumption is not justifiable, Zhou and Reiter [60] proposed a three stage procedure to approximate the posterior distribution by (i) simulate many draws from the posterior distribution in each imputed dataset; (ii) mix all of the draws; and (iii) determine the quantiles of the posterior distribution from the mixed sample. The latter procedure requires a large number of imputations, but was shown to approximate the posterior distribution well.

Deterministic matching using eUCI may not always be the optimal choice for linking two files. It is plausible that using deterministic matching with complete information such as full name, address, etc., or relying on probabilistic linking algorithms, will generate results that are more accurate. All of these algorithms will generally suffer from erroneously matched records. The procedure presented here enables investigators to adjust for possible errors obtained from any record linking algorithm without the need to assume non-informative linking. The method relies on the correlation between variables that appear in either file to identify erroneously linked records, and it provides a principled approach to include prior knowledge on the linking error. The method is only required to be applied once before any analysis is performed, which reduces the computational complexity of the analysis as well as making it tractable and efficient.

References

1. Springer SA, Pesanti E, JJH, Macura T, Doros G, Altice FL. Effectiveness of antiretroviral therapy among hiv-infected prisoners: reincarceration and the lack of sustained benefit after release to the community. *Clinical Infectious Diseases*. 2004; 38(12):1754–1760. [PubMed: 15227623]
2. Baillargeon J, Giordano TP, Rich JD, Wu ZH, Wells K, Pollock BH, Paar DP. Accessing antiretroviral therapy following release from prison. *JAMA*. 2009; 301(8):848–857. [PubMed: 19244192]
3. Rich JD, Holmes L, Salas C, Macalino G, Davis D, Ryczek J, Flanigan T. Successful linkage of medical care and community services for hiv-positive offenders being released from prison. *Journal of Urban Health*. 2001; 78(2):279–289. [PubMed: 11419581]
4. Wohl DA, Scheyett A, Golin CE, White B, Matuszewski J, Bowling M, Smith P, Duffin F, Rosen D, Kaplan A, Earp J. Intensive case management before and after prison release is no more effective than comprehensive pre-release discharge planning in linking hiv-infected prisoners to care: a randomized trial. *AIDS and Behavior*. 2011; 15(2):356–364. [PubMed: 21042930]
5. Montague BT, Rosen DL, Solomon L, Nunn A, Green T, Costa M, Baillargeon J, Wohl DA, Paar DP, Rich JD. Tracking linkage to hiv care for former prisoners. *Virulence*. 2012; 3(3)
6. Data and Reporting Team (DART). The eUCI and You. Oct. 2011 <https://careacttarget.org/library/euci-and-you>
7. Information Technology Laboratory. Technical report. Federal Information Processing Standards; 2012. Secure hash standard (shs). <http://csrc.nist.gov/publications/fips/fips180-4/fips-180-4.pdf>
8. Data and Reporting Team (DART). Technical report. HRSA HIV/AIDS Bureau; Dec. 2014 Encrypted unique client identifier (euci): Application and user guide. <https://careacttarget.org/library/encrypted-unique-client-identifier-euci-application-and-user-guide>
9. Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. *Statistics in Medicine*. 2002; 21(10):1485–1496. [PubMed: 12185898]
10. Fellegi IP, Sunter AB. A theory for record linkage. *Journal of the American Statistical Association*. 1969; 64:1183–1210.
11. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records computers can be used to extract “follow-up” statistics of families from files of routine records. *Science*. 1959; 130(3381):954–959. [PubMed: 14426783]
12. Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: A comparison of link plus, the link king, and a basic deterministic algorithm. *Health Informatics Journal*. 2008; 14(1):5–15. [PubMed: 18258671]
13. Newman LM, Samuel MC, Stenger MR, Gerber TM, Macomber K, Stover JA, Wise W. Practical considerations for matching std and hiv surveillance data with data from other sources. *Public Health Reports*. 2009; 124 (Suppl 2):7–17.
14. Belin TR, Rubin DB. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*. 1995; 90(430):694–707.
15. Larsen MD, Rubin DB. Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*. 2001; 96(453):32–41.
16. Lahiri P, Larsen MD. Regression analysis with linked data. *Journal of the American Statistical Association*. 2005; 469:222–230.
17. Neter J, Maynes ES, Ramanathan R. The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*. 1965; 60(312):1005–1027.
18. Scheuren F, Winkler WE. Regression analysis of data files that are computer matched. *Survey Methodology*. 1993; 19:39–58.
19. Scheuren F, Winkler WE. Regression analysis of data files that are computer matched ii. *Survey Methodology*. 1997; 23:157–165.
20. Chambers, R.; Chipperfield, J.; Davis, W.; Kovacevic, M. Technical Report Working Paper 18-09. Centre for Statistical and Survey Methodology, University of Wollongong; 2009. Inference based on estimating equations and probability-linked data.

21. Hof MHP, Zwinderman AH. Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Statistics in Medicine*. 2012; 31(30):4231–4242. [PubMed: 22807060]
22. Gutman R, Afendulis CC, Zaslavsky AM. A bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association*. 2013; 108(501):34–47. [PubMed: 23645944]
23. Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc; New York, NY: 1987.
24. Winkler, WE. US Bureau of the Census, Statistical Research Division Report. 2006. Overview of record linkage and current research directions.
25. Winkler, WE. Proceedings of the Section on Survey Research Methods. American Statistical Association; 2002. Methods for record linkage and bayesian networks; p. 3743-3748.
26. Goldstein H, Harron K, Wade A. The analysis of record-linked data using multiple imputation with data value priors. *Statistics in Medicine*. 2012; 31(28):3481–3493. [PubMed: 22807145]
27. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63(3):581–592.
28. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1977:1–38.
29. Titterington, DM.; Smith, AFM.; Makov, UE. *Statistical Analysis of Finite Mixture Distributions*. Wiley; Chichester, NY: 1985.
30. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2. Wiley-Interscience; Hoboken: 2002.
31. Larsen MD. Multiple imputation analysis of records linked using mixture models. *Proceedings of the Survey Methods Section*. 1999:65–71.
32. Larsen, MD. Record linkage using finite mixture models. In: Gelman, A.; Meng, XL., editors. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*. Wiley; Chichester, UK: 2004. p. 309-318.
33. Fortini M, Liseo B, Nuccitelli A, Scanu M. On bayesian record linkage. *Research in Official Statistics*. 2001; 4 (1):185–198.
34. Fortini M, Nuccitelli A, Liseo B, Scanu M. Modelling issues in record linkage: a bayesian perspective. *Proceedings of the American Statistical Association, Survey Research Methods Section*. 2002:1008–1013.
35. McGlinchy MH. A bayesian record linkage methodology for multiple imputation of missing links. *ASA Proceedings of the Joint Statistical Meetings*. 2004:4001–4008.
36. Tanner MA, Wong WH. The calculation of posterior densities by data augmentation. *Communications of the Association for Computing Machinery*. 1987; 5(398):563–567.
37. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian data analysis*. CRC press; Boca Raton, Florida: 2004.
38. Winkler WE. Near automatic weight computation in the fellegi-sunter model of record linkage. *Proceedings of the Bureau of the Census Annual Research Conference*. 1989; 5:145–155.
39. Winkler WE. Advanced methods for record linkage. *American Statistical Association Proceedings of Survey Research Methods Section*. 1994:467–472.
40. Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. *Biometrika*. 1999; 86(4):948–955.
41. Ibrahim, JG.; Chen, M-H.; Sinha, D. *Bayesian survival analysis*. Wiley Online Library; 2005.
42. Kalbfleisch, JD.; RLP. *The statistical analysis of failure time data*. John Wiley & Sons; Hoboken, NJ: 2011.
43. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science*. 1992; 7(4):457–472.
44. Plummer, M. *Jags: A program for analysis of bayesian graphical models using gibbs sampling*. 2003.
45. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2014. <http://www.R-project.org/>

46. Spiegelhalter DJ, Best NG, Carlin BP, Linde AVD. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2002; 64(4):583–639.
47. Rubin DB, Schafer JL, Schenker N. Imputation strategies for missing values in post-enumeration surveys. *Survey Methodology*. 1988; 14(1):2.
48. Xiao-Li M. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*. 1994:538–558.
49. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association*. 1996; 91(434):473–489.
50. Larsen MD. An experiment with hierarchical bayesian record linkage. 2012 arXiv preprint arXiv:1212.5203.
51. Lehmann, EL.; Romano, JP. *Testing Statistical Hypotheses*. 3. Springer; 2005.
52. Raghunathan TE, Rubin DB. Roles for bayesian techniques in survey sampling. *Proceedings of the Silver Jubilee meeting of the Statistical Society of Canada*. 1998:51–55.
53. Olkin I, Tate RF. Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*. 1961:448–465.
54. Krzanowski WJ. Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*. 1980; 36:493–499. [PubMed: 7213911]
55. Krzanowski WJ. Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis-testing approach. *Biometrics*. 1982; 38:991–1002.
56. Little RJA, Schluchter MD. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*. 1985; 72(3):497–512.
57. Schafer, JL. *Analysis of Incomplete Multivariate Data*. CRC press; 1997.
58. Mitra R, Reiter JP. Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine*. 2011; 30(6):627–641. [PubMed: 21337358]
59. Schafer JL. Multiple imputation: a primer. *Statistical methods in medical research*. 1999; 8(1):3–15. [PubMed: 10347857]
60. Zhou X, Reiter JP. A note on bayesian inference after multiple imputation. *The American Statistician*. 2010; 64(2):159–163.

(a) Distribution of log linkage time for validation dataset by type of links and study's dataset link time

(b) Cumulative proportion of released individuals that are not linked to care as a function of time from release

Figure 1.
Distributions of linkage times and cumulative proportions of released individuals.

Table 1

Comparison of file linking methods in the validation dataset

Method & Characteristics used	# Records matched	# True Positives	True Positives of declared links	True Positives of true links	# False Positives	# False Negatives
Manual Review: Gold Standard	79	79	100%	100%	0	0
eUCI						
LN, FN, DOB, Gender	76	66	87%	84%	10	13
LN, FN, DOB, Alias, Gender	84	74	88%	93%	10	5
Deterministic						
LN	117	62	53%	79%	55	17
LN, Gender	114	62	54%	79%	52	17
LN, FN, Gender	67	58	87%	73%	9	21
LN, FN, Alias	133	68	51%	86%	65	11
LN, FN, Alias, Gender	129	68	53%	86%	61	11
LN, FN, Alias, DOB	73	64	88%	81%	9	15
Probabilistic						
LN, FN, Alias, DOB, Gender	82	71	87%	90%	11	8

LN – Last Name, FN – First Name, DOB – Date of Birth

Table 2

Performance of different mixture of linear regression models on the validation dataset

% Correct classification of links			
Model	Average	95% Interval Estimates	DIC
β_{m0}	0.61	(0.26, 0.84)	359
$\beta_{m0} + \beta_{m1}Age$	0.63	(0.26, 0.85)	351
$\beta_{m0} + \beta_{m1}Gender$	0.64	(0.26, 0.86)	362
$\beta_{m0} + \beta_{m1}Race$	0.66	(0.28, 0.86)	344
$\beta_{m0} + \beta_{m1}Age + \beta_{m2}Gender + \beta_{m3}Race$	0.66	(0.28, 0.86)	348

Table 3

Factors and corresponding levels used in the simulation analysis

Factor	Levels	Description
n	{250, 500, 2000}	Number of links identified
π	{0.8, 0.9}	Probability of true link
γ	{0.2, 0.6}	Change in correlation for the link group
η	{0.5, 1}	Conditional variance of Y_i

Table 4

Mean and range across configurations of 95% Coverage, bias, sampling standard error and MSE for the three methods described in Section 4.1

Method	Estimand	Summary statistic	95% Coverage	Bias	Sampling SE	MSE
No-adjustment (a)	Intercept	Mean	0.14	-0.15	0.04	0.03
		Range	(0.00,0.68)	(-0.21, -0.10)	(0.01,0.07)	(0.01,0.05)
	slope	Mean	0.59	-0.05	0.04	0.01
		Range	(0.0,0.94)	(-0.12, -0.02)	(0.01,0.07)	(0.00,0.02)
Mixture-Intercept (b)	Intercept	Mean	0.98	0.19	0.29	0.08
		Range	(0.95,1.00)	(0.01,0.45)	(0.11,0.47)	(0.01,0.28)
	slope	Mean	0.92	-0.15	0.18	0.03
		Range	(0.58,1.00)	(-0.27, -0.07)	(0.06,0.30)	(0.01,0.09)
Mixture-Intercept & Slope (c)	Intercept	Mean	0.97	-0.03	0.14	0.01
		Range	(0.93,1.00)	(-0.14,0.01)	(0.02,0.27)	(0.00,0.03)
	slope	Mean	0.98	0.00	0.09	0.00
		Range	(0.95,1.00)	(-0.06,0.01)	(0.01,0.19)	(0.00,0.01)

Table 5

Point and 95% interval estimate of the median number of days until linking to care after adjustment for linking errors and including all linked records for various sub-populations

Estimand	Adjustment for linking errors partial model (11)			Adjustment for linking errors full model (12)			All links		
	Value	95% Interval Estimate	Median	95% Interval Estimate	Median	95% Interval Estimate	Median	95% Interval Estimate	
Proportions									
Patients with AIDS	0.40	(0.30, 0.50)	0.40	(0.30, 0.49)	0.40	(0.32, 0.49)	0.40	(0.32, 0.49)	
Permanent Housing	0.48	(0.38, 0.58)	0.48	(0.39, 0.58)	0.49	(0.40, 0.57)	0.49	(0.40, 0.57)	
Below federal poverty level	0.89	(0.83, 0.96)	0.89	(0.84, 0.95)	0.90	(0.84, 0.95)	0.90	(0.84, 0.95)	
Median Link Time									
All	97	(53, 176)	85	(52, 138)	88	(64, 120)	88	(64, 120)	
Females	125	(35, 435)	135	(44, 411)	107	(45, 256)	107	(45, 256)	
Males	93	(51, 171)	78	(45, 137)	85	(61, 119)	85	(61, 119)	
Hispanic	106	(39, 288)	90	(38, 215)	98	(50, 190)	98	(50, 190)	
Non-hispanic	95	(51, 176)	84	(50, 143)	86	(60, 122)	86	(60, 122)	
<30	189	(51, 695)	193	(46, 803)	164	(65, 417)	164	(65, 417)	
30–40	121	(48, 302)	137	(52, 362)	113	(58, 217)	113	(58, 217)	
40–50	86	(42, 173)	69	(32, 149)	79	(50, 125)	79	(50, 125)	
>50	72	(27, 187)	50	(14, 180)	62	(33, 118)	62	(33, 118)	