Finding Needles in Haystacks: Multiple-Imputation Record Linkage Using Machine Learning*

John M. Abowd† Joelle Abramowitz‡ Margaret C. Levenstein‡ Kristin McCue§
Dhiren Patki¶ Trivellore Raghunathan‡ ‖ Ann M. Rodgers‡
Matthew D. Shapiro**¶ †† Nada Wasi‡‡ Dawn Zinsser‡

July 2024

Abstract

This paper considers the problem of record linkage between a household-level survey and an establishment-level frame in the absence of unique identifiers. Linkage between frames in this setting is challenging because the distribution of employment across establishments is highly skewed. To address these difficulties, this paper develops a probabilistic record linkage methodology that combines machine learning (ML) with multiple imputation (MI). This ML-MI methodology is applied to link survey respondents in the Health and Retirement Study to their workplaces in the Census Business Register. The linked data reveal new evidence that non-sampling errors in household survey data are correlated with respondents' workplace characteristics.

Keywords: Administrative data; machine learning; multiple imputation; probabilistic record linkage; survey data.

---

*Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the Federal Reserve Bank of Boston, the principals of the Board of Governors, the Federal Reserve System, or the U.S. Census Bureau. We are grateful to the Associate Editor and two anonymous referees for helpful comments. We thank Jamie Fogel, Dyanne Vaught, and Sara Zobl for research assistance. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Project 6000576: CBDRB-FY21-CED006-0019, CBDRB-FY24-0153, CBDRB-FY24-CED006-0015). This research is supported by the Alfred P. Sloan Foundation through the CenHRS project at the University of Michigan with additional support from the Michigan Node of the NSF-Census Research Network (NCRN) under NSF SES 1131500. The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. Corresponding author: Matthew D. Shapiro (shapiro@umich.edu), Department of Economics, 611 Tappan Ave., Ann Arbor, MI 48109.

†Labor Dynamics Institute, Cornell University

‡Institute for Social Research, University of Michigan

§U.S. Census Bureau (retired)

¶Federal Reserve Bank of Boston

‖Department of Biostatistics, University of Michigan

**Department of Economics, University of Michigan

††NBER

‡‡Puey Ungphakorn Institute for Economic Research, Bank of Thailand

1 Introduction

Increasingly, researchers are interested in linking survey and administrative data for measurement and analysis. In most record linkage applications, the units being linked originate from the same frame. For instance, individuals in a given dataset are linked to the same individuals in a different dataset, or businesses in one dataset are linked to the same businesses in another dataset. In this paper, we consider the problem of linking across frames. We match individual respondents in household survey data to administrative data on the universe of employers. How does one use a household report of a business to link to the correct employer? It would be possible to build in these linkages from the start, especially where a sampling frame is created from administrative data. In that case, linkage is part of the design. This paper addresses the problem of linking individuals and employers where the linkage is not pre-designed into a survey. This situation typically arises in surveys of households, which are built from sampling frames of household addresses, often without the purpose of linkage as part of the design. Even in an idealized world where the survey and administrative frames were developed in tandem, additional linkages to other administrative data, that are not part of the design, may be desirable.

As in Abowd et al. (2009), Gutman et al. (2013), and Gutman et al. (2016) we treat record linkage as a missing data problem where true match status is unknown and must be imputed. To do this, we need to accomplish two related tasks. First, we need a way to predict whether any given pair of records drawn from the two datasets constitutes a true match. Second, we need a way to characterize uncertainty in the prediction of true matches and propagate that uncertainty into inferences drawn from the linked dataset by subsequent analyses.

The task of predicting true match status is difficult because the size distribution of firms is very skewed. Consider the striking empirical fact that 0.3 percent of all firms employ 54 percent of all workers in the United States.[1] There are however more than six million firms, so it follows that most firms are very small. With so many small firms, matching individuals to employers is inherently noisy because a large number of small employers among a set of potential candidates are feasible matches for any given survey respondent. This is our needle in the haystack problem. For example, imagine a set of candidate matches included a large insurance company, an independent credit union at the same location using the name of the insurance company, and a cafeteria operated by a third-party vendor also at the same location and using the name of the insurance company. The names and addresses of the candidate matches are all similar. When presented with this information, a human reviewer may use auxiliary information such as the respondent's industry, occupation, or reported firm size to guess the correct match. We automate and speed-up what a human reviewer would do by using a supervised machine learning (ML) approach to predict the matching firm. ML is particularly valuable for record linkage because it makes flexible use of a very large number of predictors, including auxiliary information, to mimic the heuristics used by a human reviewer. Furthermore, relying on a rich set of predictors lends support to the assumption

[1] 2021 Statistics of U.S. Businesses (SUSB), U.S. Census Bureau.

that imputation errors are ignorable thereby improving inferences in subsequent analyses of the imputed data. Finally, our cross-validated ML estimator is tuned to deliver high out-of-sample accuracy.

Multiple imputation (MI) allows analysts to propagate match uncertainty when conducting inference. For each record in household survey data, our procedure samples multiple candidates from administrative employer-level data by using ML-based match probability estimates as weights. In cases where the match probability estimates are highly concentrated, household survey records are linked to just one employer. Conversely, for cases where the match probability estimates are highly diffuse, household survey records are linked to many different employers. In the completed (matched) dataset, variability between implicates for a given household survey record captures uncertainty associated with the linkage for that household. Subsequent analyses of the completed dataset can then combine the multiple implicates for valid statistical inference as in Rubin (1987).

In this paper we apply our novel combined ML-MI approach to record linkage to match the Health and Retirement Study (HRS), a longitudinal household-level survey of older Americans, and the Census Business Register (BR), an administrative dataset that covers the universe of employers in the United States. This new linked household-employer dataset will provide researchers new ways to investigate wide-ranging questions about the role of employer- and workplace-specific factors in influencing wages, consumption and savings decisions, health outcomes, and retirement choices of older workers. We re-examine the well-known positive gradient between hourly wages and workplace size to provide an example of the type of analysis that the matched data can facilitate. We find that both non-classical measurement error and selective non-response in the HRS survey reports of workplace size generate upward bias in this gradient.

The article proceeds as follows. Section 2 describes record linkage methodologies in deterministic and probabilistic contexts. Section 3 provides details on the files that we link and explains the three major steps of our record linkage procedure. Section 4 assesses the fit of our match prediction model and evaluates the degree of uncertainty in our linkage. Section 5 illustrates an application of the matched data to shed new light on the incidence and consequences of nonclassical measurement error and selective nonresponse in household survey respondent reports of workplace size. Section 6 concludes.

2 Essentials of Record Linkage

This paper builds on an important literature that developed widely-used techniques for record linkage. The simplest approaches are non-probabilistic. In these deterministic file matching applications, researchers accomplish record linkage by isolating a set of variables that are common to a given record in both files. This procedure constitutes both the first and the last step in the linkage. It is the first step because it enumerates the set of possible matches. It is the last step because only those records that have exactly one match conditional on variable agreement are retained. In some instances, a sufficiently rich set of accurately measured variables can allow a large fraction of the original file to be unequivocally matched (see, e.g., Lawson et al. (2013) and Setoguchi et al. (2014)). In other cases, the matched file consists of a smaller and potentially non-random subset of the original file that limits the usefulness of the matched dataset for analysis. This concern is highlighted in the context of linking historical data, for example, in Bailey et al. (2017).

The Fellegi and Sunter (1969) (FS) method is an early and widely-used probabilistic linking approach that picks the best match from the set of multiple potential matches. In this method, researchers estimate the probability that a particular characteristic (such as gender or first and last name) agrees in the two files, given that the records should link (match) and given that they should not link (non-match). To estimate match probabilities the FS method relies on the strong, and sometimes untenable, assumption that the agreement status of each characteristic is independent conditional on true match status. Next, the data are used to determine log odds cutoffs above which potential matches are coded as true matches and below which they are treated as non-matches. Potential matches that fall between the cutoffs are evaluated manually, a procedure which has been criticized, for example, in Belin and Rubin (1995) because the error properties of manual review are unknown, may be subject to inconsistent standards across reviewers, and may fail to yield a substantial number of unequivocal matches.[2]

ML methods for record linkage constitute a growing alternative to the FS approach. These methods estimate highly flexible non-parametric functions and classify record pairs into matches and non-matches. For example, Cochinwala et al. (2001) and Elfeky et al. (2002) use decision trees for classification while Christen (2008a) and Christen (2008b) rely on support vector machines. ML approaches have been implemented with training data (supervised) and without it (unsupervised), with the former typically yielding more accurate linkage (see, e.g., Christen (2008b)). The key advantage of the ML-based record linkage approach is its high degree of accuracy. These implementations of ML create a deterministic classifier. Hence, like FS, existing ML-based record linkage applications select the best match among a set of candidate matches. That is, conditional on the matching algorithm, matches are treated as deterministic.

The Bayesian approach to record linkage characterizes uncertainty associated with parameters in the linkage process (Fortini et al. (2001) and Larsen (2004)). In this method, researchers specify prior distributions of parameters that govern the mixture of matches and non-matches that generate the comparison vector of agreement status for variables observed in both files. Draws from the posterior predictive distribution of the parameters are then used to produce estimates of pair-specific match probability. One-to-one matching is enforced using the mode of the posterior predictive distribution or by minimizing a loss function. Tancredi and Liseo (2011) refine this procedure by relying on observed discrete matching variables rather than a comparison vector of agreement status for those variables. Steorts et al. (2016) provide a method of linking multiple files, each with potentially duplicated records, within the Bayesian framework. Gutman et al. (2013) and Gutman et al. (2016) further develop the Bayesian approach by applying it to situations where variables used in the

[2]While comprehensive manual review of records has been adopted in some applications (e.g., Ferrie (1996)), it is prohibitively expensive in many settings and remains subject to the same criticisms as the manual review step of the FS method.

linkage model are available in both files as well as variables available in only one file. Moreover, they jointly model the linkage step as well as relationships between variables in the linked dataset (the analysis step). Then, by repeatedly sampling from the posterior distribution of the linkage step parameters they generate multiple implicates of linked datasets that are used in the analysis step and combined using the formulas in Rubin (1987). This procedure has the advantage of propagating uncertainty in the linkage step parameters into the analysis step.

Work that is highly germane to the household-employer record linkage problem we consider in this paper began as a part of the Longitudinal Employer Household Dynamics (LEHD) program in two projects that were initiated in the early years of that effort. The first of these projects linked employers to job histories in the 1990-1996 Surveys of Income and Program Participation (SIPP).[3] Abowd and Stinson (2013) evaluate this linkage and use it to compare self-reports and administrative reports of earnings. The LEHD program also links establishments (i.e. specific workplaces for a given employer) in the Quarterly Census of Employment and Wages, called the Employer Characteristics File in LEHD, to individual workers via the state unemployment insurance account number, called the SEIN in LEHD. This linkage starts with deterministic methods using the SEIN. When these methods do not find a one-to-one match, a Bayesian posterior predictive distribution is used to generate ten implicates linking establishments to the candidate worker's employment history (Abowd et al. (2009)).[4] These ten implicates are used to associate workplace characteristics to each worker history.[5] The ten implicate threads are processed according to the Rubin (1987) combining formulas to produce the Quarterly Workforce Indicators (QWI). McKinney et al. (2021) provide a complete assessment of the total variability in the QWIs due to the MI and other edit procedures.

The methodology we develop relies on the accuracy of the ML approach to record linkage while using MI to characterize uncertainty in the linkage and to propagate that uncertainty into subsequent analyses. To our knowledge, this combination of methods has not been previously employed in record linkage applications.[6] Our ML approach allows us to leverage a very large number of predictors to estimate match probabilities including both discrete and continuous observed variables from either file as well as agreement status variables constructed using both files. Furthermore, the flexibility inherent in this method accommodates rich complementarity between predictors and allows us to dispense with the assumption that predictor variables are independent conditional on true match status as has been posited in many prior applications. In addition, tuning our prediction models to achieve high out-of-sample accuracy facilitates scalability and precision linkage in a setting where there can be tens of thousands of candidate matches per survey respondent. Record linkage with such

[3] This work also developed improved linkages within the 1990-1993 SIPP job histories, and integrated data from the Census Business Register into the SIPP (Stinson (2003)).

[4] See Goldstein et al. (2012) for a similar approach applied to medical records.

[5] Other incomplete data in the LEHD infrastructure, such as incomplete data for education, are completed using similar Bayesian methods.

[6] ML methods have been used to improve MI in applications that do not involve record linkage. See, e.g., Reiter (2005) for the creation of partially synthetic public use microdata and Burgette and Reiter (2010) for missing variable imputation.

a large number of permutations is computationally challenging using existing Bayesian methods. Finally, unlike prior ML-based record linkage methods that use binary classification to select the single best match, the ML model that we use provides a match probability estimate for each record pair. By using a Bayesian bootstrap procedure (Rubin (1981)) to repeatedly sample candidate matches from the estimated match probability distribution when constructing MI linkages, our procedure allows us to approximate parameter uncertainty in the ML model while also characterizing uncertainty regarding latent match status.

3 The Machine Learning, Multiple Imputation (ML-MI) Procedure

3.1 Overview

In this section, we describe our ML-MI record linkage procedure for matching household-level survey data to establishment-level administrative data. Our approach acknowledges that many matches are uncertain and is explicit about uncertainty at all steps. It produces a dataset of multiply-imputed links that, if used appropriately, will allow analysts to produce statistics and inferences that account for the uncertainty of matches.

We integrate machine learning and multiple imputation using the following steps, which are shown in Figure 1. First, we enumerate the set of candidate establishments that constitute feasible matches for each survey report about a particular job using a technique known as blocking. This yields a set of blocked pairs. Second, we create training data for supervised ML based on a sample drawn from the set of blocked pairs. Third, using the training data, we estimate an ML model nested within a weighted Bayesian bootstrap (WBB). Fourth, we use the model to obtain match probabilities for of the blocked pairs, which are candidate matches. Finally, we draw an implicate using the match probabilities as weights when sampling among the candidates. This step is repeated to create M implicates.

Our procedure accounts for match uncertainty in two ways. First, there is parameter uncertainty in the ML model because it is based on finite data. Second, there is match uncertainty conditional on the parameters because the probability distribution over potential matches is not degenerate. Both types of uncertainty propagate through our procedure since the model is re-estimated and an implicate is drawn using match probabilities as weights within each bootstrap iteration.

At the conclusion of this section, we describe how the multiply-imputed matches should be used in analysis in order to propagate match uncertainty.

3.2 Dataset structure

Before delving further into the details of our methodology, we briefly describe the datasets that we use in our application. The household survey that we use is the Health and Retirement Study (HRS) that surveys more than 22,000 Americans over the age of 50 every two years. It is a large-scale longitudinal project that studies the labor force participation and health transitions that individuals undergo toward the end of their work lives and in the years that follow. About 70 percent of HRS respondents give permission to the Social Security Administration (SSA) to provide earnings records, which include U.S. Federal Employer Identification Numbers (EINs), to the HRS for purposes of enhancing the HRS data infrastructure.[7] In addition to EINs provided by the earnings records, the HRS elicits information about employer name, establishment address, and telephone number for the respondent's "main job" or the job about which a bulk of work-related survey questions are asked. Respondent reports of employer identity and address are obtained at the survey baseline (i.e., when new respondents are enrolled in the study, generally every six years when a new cohort is added to the study) and in each subsequent wave if the respondent reports having changed jobs.[8] We use EINs along with employer names and establishment addresses to match HRS respondents' employers and workplaces to the Census Business Register (BR) which is the Census Bureau's list of essentially all establishments in the United States. Note that the establishment is the workplace, and a given firm may operate many establishments. The BR contains information on EIN, employer name and establishment address, company affiliation, size, payroll, industry classification and other employer-level and establishment-level characteristics and can be linked to other Census Bureau survey and administrative data.[9] We refer to the dataset created by matching the HRS to the BR and associated Census Bureau data as the CenHRS.

Our procedure has three cases for linking HRS jobs to the BR: deterministic match based on EIN and probabilistic match with or without EIN. Table 1 shows these cases and their characteristics.

- In the first case, respondents consent to SSA linkages and can be deterministically matched to an establishment in the BR. This case happens when the respondent has just one job (and therefore just one EIN) in a given year and that EIN corresponds to exactly one establishment in the BR.

- In the second case, respondents consent to SSA linkages, but cannot be deterministically matched to an establishment in the BR. This case happens either because the respondent has multiple jobs in a given year (and consequently has multiple EINs), or because the respondent's EIN does not uniquely identify an establishment in the BR, or both.

- In the third case, respondents do not consent to SSA linkages and therefore we do not have their EINs.

For respondents in the second and third categories, which represent about 60 percent of our sample, we implement probabilistic record linkage using ML and MI. In the next section, we discuss how the availability of the EIN affects this procedure.

[7]For consenting respondents, in addition to earnings records, SSA provides retirement and disability benefit claims data.

[8]The HRS collects employer names, establishment addresses, and phone numbers to contact respondents' employers about retirement benefit provisions.

[9]EINs are tax identification numbers; they do no uniquely identify establishments except in two special cases. The first case is for employers that operate just one establishment (single-unit employers). The second case is for employers that operate multiple establishments (multi-unit employers), have multiple EINs, and where a specific EIN points only to one establishment. An example of the latter case would be if Dunder Mifflin Paper Company was a two-establishment firm with one establishment located in Scranton, PA, and another in New York, NY, and if each of those establishments had its own EIN.

3.3 Procedure

3.3.1 Blocking

Let jobs in the HRS be indexed by $i = 1, ..., N_{\text{HRS}}$. A job in the HRS is defined as a spell of employment at a unique establishment. Let establishments in the BR be indexed by $j = 1, ..., N_{\text{BR}}$. If we start with the prior that every record in the BR is a potential match for each job in the HRS, we would need to search over a set of $N_{\text{BR}} \times N_{\text{HRS}}$ pairs. This set is of the order $10^6 \times 10^4$.

To reduce the dimensionality of the search problem, we follow a blocking strategy. Blocking groups record pairs that share specific characteristics wherein pairs that agree on at least one characteristic are regarded has having a positive probability of being matches, while pairs that fail to agree on any characteristics are deemed as non-matches (see, e.g., Christen (2012)). That is, the blocking strategy assigns zero probability to candidates outside of the block. If the block has only one candidate, the linkage is deterministic. For HRS respondents who consent to SSA linkage, we block on EIN. For HRS respondents who do not consent to SSA linkage, we block on 10-digit phone number, 3-digit zip code, telephone area code, and city-state.[10]

3.3.2 Training data

The blocking variables we use strongly influence the level of ex-ante match uncertainty. As seen in Table 1, blocking on EINs generates an average of 436 (sd = 1,280) candidate matches (i.e. unique establishments in the BR) per HRS respondent. In contrast, when EINs are not available, blocking on location-specific variables generates 30,790 candidate matches (sd = 29,630) per HRS respondent. This large increase in the number of candidate matches arises because of the skewed distribution of employment across establishments. In particular, when EINs are unavailable, a large number of establishments populate the location-specific blocks – this is the so-called "needle-in-haystack" problem. Because of the stark difference in ex-ante match uncertainty associated with the two types of blocking schema, the relationship between predictors and match status varies substantially based on whether the HRS-BR pair is blocked using EIN or not. We account for these differences by creating two different training samples and train separate models: one based on pairs blocked using EINs and the other based on pairs blocked without EINs. Each training sample consists of $N^T \approx 1000$ randomly-selected HRS-BR unlabeled pairs. We oversample pairs with a higher likelihood of being true matches using the data and methodology described in Appendix A.

We now specify the procedure for creating the training data by human review to label HRS-BR pairs. Define $\boldsymbol{x}_i^H$ as a vector of individual demographic characteristics, employment-related variables, self-reported employer characteristics, and survey paradata for HRS respondent i. Define $\boldsymbol{x}_j^B$ as a vector of characteristics for establishment j drawn from administrative data in the BR. Let $k(ij) = 1, \ldots, N^T$ index HRS-BR pairs in each of the unlabeled samples. These data are examined by reviewers who observe certain pair characteristics, $\boldsymbol{x}_k$, which are a subset of $(\boldsymbol{x}_i^H, \boldsymbol{x}_j^B)$; Table

[10] We do not model blocking uncertainty. For respondents who consented to SSA linkage, EINs come directly from the SSA data (no uncertainty). For respondents who did not consent to SSA linkage, the blocking variables were respondent provided. Accounting for uncertainty in these data is outside the scope of our current models.

2 lists the elements of $\boldsymbol{x}_k$. Each HRS-BR pair is evaluated by two reviewers. Define $y_{k,r} = 1$ if reviewer r scores pair k as a match and $y_{k,r} = 0$ otherwise. To the extent that they disagree, the two reviewer assessments—i.e. the $y_{k,r}$—reflect uncertainty about latent match status.[11]

For each HRS-BR pair in the unlabeled samples, reviewers consider employer and establishment match status separately. An employer match means that the employer identity (e.g., Dunder Mifflin Paper Company) in the HRS corresponds to the employer identity in the BR. In contrast, an establishment match implies that, in addition to an employer match, the workplace reported by the HRS respondent exactly corresponds to the physical location in the BR (e.g., Dunder Mifflin Paper Company, 1460 Main Street, Scranton, PA). This distinction is important because workplace characteristics can differ substantially even at different locations of a single employer. For example, different establishments of a given employer may experience differential expansion or contraction, produce different types of goods or services, or employ workers of different skill types or ages. Consequently, we construct four different training datasets: employer match for EIN blocked pairs, establishment match for EIN blocked pairs, employer match for non-EIN-blocked pairs, and establishment match for non-EIN-blocked pairs. In the employer match datasets, $y_{k,r}$ refers to employer match status; in the establishment match datasets, $y_{k,r}$ refers to establishment match status.

Once reviewers complete their assessments, each of the four training datasets can be represented by the following matrix:

$$\boldsymbol{T} = \begin{bmatrix} y_{1,1} & \boldsymbol{x}_1 \\ y_{1,2} & \boldsymbol{x}_1 \\ \vdots & \vdots \\ y_{N^T,1} & \boldsymbol{x}_{N^T} \\ y_{N^T,2} & \boldsymbol{x}_{N^T} \end{bmatrix}. \tag{1}$$

Because there are two reviewer ($r \in \{1, 2\}$) outcomes associated with each HRS-BR pair in the training sample indexed by $k(ij) = 1, \ldots, N^T$, there are $l(ij) = 1, \ldots, 2N^T$ rows in the training data set $\boldsymbol{T}$.

3.3.3 Estimating the ML model using weighted Bayesian bootstrap

We estimate the ML models using a set of variables that supplements the information observed by reviewers ($\boldsymbol{x}_l$). The supplemented set of predictors is given by the vector $\tilde{\boldsymbol{x}}_{l(ij)} = f(\boldsymbol{x}_l, \boldsymbol{x}_i^H, \boldsymbol{x}_j^B)$ where the function $f(\cdot)$ supplements and transforms observed data. Table 3 shows the elements of $\tilde{\boldsymbol{x}}_{l(ij)}$.

The first set of predictors are pair-specific. Cubic splines of Jaro-Winkler (JW) scores for employer name and establishment address, and a linear JW score for city jointly capture reviewers' assessments of the similarity in the HRS and BR names and addresses.[12] We include cubic splines of

[11] A total of seven reviewers, including most of the co-authors of this paper, conducted these reviews inside the Federal Statistical Research Data Center (FSRDC) computing environment.

[12] JW scores, which range from 0 to 1, combine edit distance and q-gram-based comparison techniques to measure string similarity. The JW score for establishment address is based on the number and street and does not include information on city, state, or zip code similarity.

the share of employment within the blocking variable accounted for by a candidate BR establishment (or employment share). This variable accounts for the fact that individuals in household surveys are more likely to be employed at larger establishments for any given blocking schema. For EIN-blocked cases, we also include cubic splines of the share of annual earnings accounted for by a candidate EIN (or earnings share). When HRS respondents have multiple jobs, this variable aids in disambiguation by accounting for the fact that the "main job"—which respondents provide answers about in survey questions—is more likely to be associated with a larger share of total earnings. Finally, we include variables measuring pair-specific agreement status for a number of characteristics: 7- and 10-digit phone number, 3-, 4-, and 5- digit zip code, city-state, one-digit industry codes, and employer and establishment size class. Some pair-level predictors, such as 10-digit phone agreement, can be highly influential in predicting match probability, but it is very rare for candidate matches to share such granular characteristics. On the other hand, sharing industry codes or 4-digit zip codes is more likely but less predictive of a match.

The second set of variables comes purely from the BR and includes the log size of the employer and whether the BR candidate match is a single-unit (SU) or multi-unit (MU) business.[13] Size and MU status variables are intended to capture the higher unconditional probability that individuals in a household survey will be employed at larger, MU employers.

The third set of variables comes purely from the HRS and includes the respondent's age, gender, race and ethnicity, education, nativity, marital status, survey interview mode (in person or telephone), survey interview language (English or Spanish), log hourly real wage (using the consumer price index for deflation to 1983 dollars), years of tenure, weeks worked per year, hours worked per week, whether the respondent's employer provides health insurance and/or a retirement savings plan, and the respondent's two-digit occupation and one-digit industry. We include these variables to control for job-specific determinants of match status as well as the quality of identifying information about the employer and establishment reported by the HRS respondent.[14]

To account for the fact that household survey respondents are more likely to be employed at larger establishments and at employers which provide a larger share of annual earnings, we fully interact the cubic splines of name and address JW scores, employment share, and earnings share.[15] Including this rich set of higher order interactions substantially increases the number of variables we use to predict match likelihood. Combining HRS-BR pair-specific variables, variables only from the BR, and variables only from the HRS, and all the interaction terms, we have a total of 9,200 predictors in the vector $\tilde{\boldsymbol{x}}_{l(ij)}$.

Having defined the set of predictors, we use the logistic function to model match probabilities

[13] SU businesses operate only one establishment, whereas MU-businesses operate multiple establishments.

[14] We include information on health insurance and/or retirement plan provision because these variables are correlated with employer size, which is often missing in the HRS.

[15] When EINs are unavailable, earnings share cannot be defined. For these cases, we interact JW scores for name and address with cubic splines of log employer size from the BR.

as:

$$P(y_{l(ij)} = 1 | \tilde{\boldsymbol{x}}'_{l(ij)}) = \frac{\exp \tilde{\boldsymbol{x}}'_{l(ij)} \boldsymbol{\beta}}{1 + \exp \tilde{\boldsymbol{x}}'_{l(ij)} \boldsymbol{\beta}}. \tag{2}$$

To approximate posterior uncertainty in $\boldsymbol{\beta}$, we use weighted Bayesian bootstrap (WBB) (see, e.g., Rubin (1981), Newton and Raftery (1994), and Newton et al. (2021)). For $m = 1, \ldots, M$ replications, we construct the WBB weights for each of the four training data sets in two steps. First, we draw $k = 1, \ldots, N^T$ i.i.d. random variates, $\nu_k^{(m)}$, from an exponential distribution with mean 1. Next, we assign each pair k a random weight given by

$$w_k^{(m)} = \nu_k^{(m)} / \bar{\nu}^{(m)}, \tag{3}$$

where $\bar{\nu}^{(m)}$ is the sample mean of the $\nu_k^{(m)}$. Following Newton and Raftery (1994), the weights obtained in Equation (3) are equivalent to drawing weights from a uniform Dirichlet distribution. Note that the WBB weights are constructed at the pair-level, indexed by $k = 1, \ldots, N^T$, and not the pair-reviewer level, indexed by $l = 1, \ldots, 2N^T$. Consequently, pair-specific duplicates in the training dataset receive exactly the same Bayesian bootstrap weight. Implementing the WBB procedure this way accounts for pair-level clustering within the training dataset.

Define the dimension of $\tilde{\boldsymbol{x}}_{l(ij)}$ as q. Since $q \gg 2N^T$, i.e., the number of predictors exceeds the number of observations, we rely on the Elastic Net (EN) shrinkage estimator for model selection and estimation of $\boldsymbol{\beta}$ (Zou and Hastie (2005)). The EN-based parameter estimate for the m-th WBB replication is obtained by maximizing the constrained likelihood function:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(m)} = \underset{\beta \in \mathbb{R}^q}{\operatorname{argmax}} \sum_{l=1}^{2N^T} w_l^{(m)} \left(y_l \log \left(\frac{\exp(\tilde{\boldsymbol{x}}'_l \boldsymbol{\beta})}{1 + \exp(\tilde{\boldsymbol{x}}'_l \boldsymbol{\beta})} \right) + (1 - y_l) \log \left(\frac{1}{1 + \exp(\tilde{\boldsymbol{x}}'_l \boldsymbol{\beta})} \right) \right) \\ + \lambda \sum_{p=1}^{q} \left(\alpha |\beta_p| + (1 - \alpha) \beta_p^2 \right). \end{aligned} \tag{4}$$

In Equation (4), l indexes observations in the training data, while p indexes predictors. $w_l^{(m)}$ is the random weight attached to observation l in WBB replication m. The two tuning parameters, α and λ, which either zero out or shrink the elements of $\hat{\boldsymbol{\beta}}^{(m)}$, are estimated using 10-fold cross validation to optimize out-of-sample predictive performance; see Appendix B for additional details. We plug $\hat{\boldsymbol{\beta}}^{(m)}$ from the respective models (employer and establishment, EIN and non-EIN) into Equation (2) to obtain an estimate of the probability that a human reviewer would regard an unlabeled $i - j$ pair as a match, which we denote by $\hat{p}_{ij}^{(m)}$.[16] We iterate this WBB step $M = 10$ times.

Our match imputation relies on the assumption of strongly non-informative linkage or linkage at

[16] Parameter estimates reflect an average of the two reviewer evaluations. We found that the fraction of cases for which reviewers agreed varied from a low of 93 percent to a high of 97 percent, depending on the training dataset. Because reviewer disagreement in the training data was so unlikely, we do not model between-reviewer uncertainty or propagate it in our MI procedure.

random (see Gutman et al., 2016 and Han and Lahiri, 2019 respectively). That is, unobserved determinants of match status are ignorable conditional on the predictors in the model. This assumption can be stated as:

$$P(y_{ij} = 1|\tilde{\boldsymbol{x}}_{ij}) = P(y_{ij} = 1|\tilde{\boldsymbol{x}}_{ij}, \boldsymbol{z}_{ij}), \tag{5}$$

where $\boldsymbol{z}_{ij}$ represents a vector of all other variables that influence match status for a given pair. The high dimension of $\tilde{\boldsymbol{x}}_{ij}$ with rich information at the pair-level, from the BR, and from the HRS, makes assumption (5) more tenable and facilitates valid inferences for a wide class of subsequent analyses with the linked data.[17] In the next section, we discuss how we use these probabilities to multiply impute matches.

3.3.4 Multiple imputation of matches

Existing probabilistic record linkage procedures which either select the highest probability match or enforce one-to-one matching in other ways effectively treat the match, conditional on the procedure, as deterministic. In common with Bayesian record linkage approaches, our procedure captures both linkage uncertainty as well as the uncertainty in the parameters of the matching model. In the final step, we propagate uncertainty when selecting candidate matches by using multiple imputation.

Selecting matches for EIN-based cases

When EINs are available for blocking, we normalize the $\hat{p}_{ij}^{(m)}$ to sum to one for each HRS respondent. For each outcome (employer or establishment match), this provides M different estimates of the match probability for each candidate. Then, for each HRS respondent, we draw one implicate from each of the M normalized match probability distributions. We do this separately for employer matches and establishment matches.

Selecting matches for non-EIN-based cases

When EINs are unavailable and blocking is based on location-specific variables, the skewed distribution of employment across establishments implies that there are many more candidate matches to consider, and the task of match selection is substantially harder. Consider an example where an HRS respondent is paired with 10,000 BR candidate matches where one large-employer candidate is the correct match, and 9,999 small-employer candidates are non-matches. Suppose the large-employer candidate obtains a normalized match probability of 0.5 while each of the 9,999 small-employer candidates receive normalized match probabilities of $\frac{0.5}{9999}$. In this example, random small-employer candidates are as likely to be sampled as the large-employer candidate even though they are three orders of magnitude less likely to be correct.[18]

[17] See, e.g, issues of congeniality in imputation models as noted by Meng (1994), Rubin (1996), and Murray (2018).

[18] The logistic function is bounded below by zero, which causes the model always to return a non-zero match probability. The confounding effect of low probability matches on linkage precision is driven by the large numbers of

To mitigate the confounding effect of large numbers of candidates, we apply a minimum match probability threshold to eliminate low-quality matches from consideration. For each of the 10 WBB replications, we estimate this threshold by relying on the set of record pairs where EINs yield deterministic matches between HRS respondents and BR establishments. Although we know the true match for these respondents, we proceed as if the EIN were unavailable. That is, we block on location-specific variables, obtain match probabilities for each pair using the estimated linkage models, and use those match probabilities to select implicates for each HRS respondent.[19] This approach relies on the assumption that the distribution of candidate matches derived from deterministically linked data is similar to the distribution of candidate matches obtained for respondents without a unique link, which may not be guaranteed in practice. In our setting, we find that the mean and standard deviation of the number of potential matches in the two datasets are, in fact, quite close: for respondents without EINs blocked on location-specific characteristics there are, on average, 30,790 candidate matches (sd = 29,630). For the validation dataset of deterministically linked respondents for whom we construct candidate pairs using location-specific characteristics there are, on average, 30,050 candidate matches (sd=27,560).

The top left panel of Figure 2 illustrates how we determine the match probability threshold at the employer level. The horizontal axis of the figure plots the sample analog of the fraction linked, averaged across WBB replications. For the m-th WBB replication this variable measures the share of HRS respondents with minimum match probabilities above a given probability threshold, p:

$$\mathcal{L}(p)^{(m)} = \mathbb{E}[\mathbb{1}\{\min_j(\hat{p}_{ij}^{(m)}) > p\}], \tag{6}$$

where $\hat{p}_{ij}^{(m)}$ represents the estimated match probability for pair ij. The vertical axis of the figure plots the sample analog of the precision rate, also averaged across WBB replications. For the m-th WBB replication, the precision rate is the share of HRS respondents with minimum match probabilities above a given probability threshold that are correctly matched:

$$\mathcal{P}(p)^{(m)} = \mathbb{E}[\mathbb{1}\{y_{ij} = 1 | \min_j(\hat{p}_{ij}^{(m)}) > p\}], \tag{7}$$

where $y_{ij} \in \{0, 1\}$ is the match status for pair ij that is observed in the validation dataset.

In the top left panel of Figure 2, the solid circle labeled "Naive" shows match quality without imposing any threshold. Although the entire HRS sample is deemed to be linked, the precision rate is extremely low. Moving away from the naive case, we iterate over progressively higher probability thresholds. For each new choice of threshold, we re-normalize the estimated match probabilities over the set of candidates that survive the threshold, and sample matches with probability proportional to the re-normalized match probabilities. Progressively raising the match probability threshold

candidates each of which has a trivially small match probability.

[19]The resulting dataset enumerates *all* HRS-BR pairs conditional on the blocking variables for a given HRS respondent. This location-blocked validation dataset allows us to evaluate the challenge of selecting the right match from a very large number of potential matches. It is a conceptually different dataset than the training dataset, which is a stratified random sample of pairs.

generates movement up and to the left along the solid blue line. For higher thresholds, the share of HRS respondents with maximum match probabilities above the threshold decreases, and, consequently, the fraction of the sample that can be linked falls. The solid blue line therefore traces out the *precision frontier* which is the trade-off between the precision rate and the share of the sample that can be linked. We define the optimal point on the precision frontier as the threshold which yields a precision rate and a sample link rate that are each closest to their maximum values of 1, or the top right corner of the graph.[20] Formally, the optimal probability threshold for the m-th WBB replication is

$$\hat{p}^{*(m)} = \underset{p\in[0,1]}{\operatorname{argmin}} \left(\left(1 - \mathcal{P}(p)^{(m)}\right)^2 + \left(1 - \mathcal{L}(p)^{(m)}\right)^2 \right)^{1/2}. \tag{8}$$

The optimal trade-off is shown with a hollow circle in Figure 2. With respect to the quality of inferences drawn from non-EIN based linkages, the optimal threshold in Equation (8) places equal weight on controlling selection bias induced by incomplete linkage as it does on controlling incorrect linkage.

The top right panel of Figure 2 shows that the extremely large number of candidates per HRS respondent drives the low level of precision attained with unrestricted match selection. The solid black circle shows that there are between 10^4 and 10^5 candidates from which to select implicates when we block on location-specific variables. The hollow circle shows that the application of our optimal match probability threshold lowers the average number of candidate matches by three orders of magnitude and yields a very large increase in the precision rate. The lower row of Figure 2 shows analogous statistics for the establishment matching model.

For the sub-sample of HRS respondents without EINs, we leave as unmatched respondents for which all BR match candidates' estimated match probability is below the optimal threshold. Any matching procedure, of course, should admit the possibility that there is no reasonable match. Our procedure handles this possibility systematically based on the estimated matching model, its uncertainty, and a well-specified objective function estimated using validation data. Hence, determining that a case is a non-match is entirely integrated into the procedure. It does not rely on ancillary determination, for example, that a case is treated as a non-match if the match probability is below an externally specified threshold.

3.4 Using the multiply-imputed dataset

Our procedure yields $M = 10$ multiply imputed employer and establishment links for each HRS respondent thereby constituting M completed datasets.[21] For any statistic generated using imputed data, we can combine estimates obtained from each of the M completed data sets using the

[20] This criterion is related to the literature on cutoffs for ROC curves, which has used proximity to the maximum attainable values of the true positive rate and the false positive rate as way to choose the optimal cutoff. See, e.g., Coffin and Sukhatme (1997).

[21] When EINs are sufficient to yield a one-to-one BR match for an HRS respondent, record linkage is deterministic and trivial. We include these cases in each of the M completed datasets.

formulas in Rubin (1987) to compute the variance owing to sampling uncertainty (within-implicate variability), and the variance due to linkage uncertainty (between-implicate variability).[22] For some scalar parameter of interest θ, let $\hat{\theta}_m$ represent estimates derived from the $m = 1, \ldots, M$ completed data sets. Let $\hat{\sigma^2_m}$ represent the variances associated with each of the M parameter estimates. The multiply imputed estimate of θ is

$$\hat{\theta} = M^{-1} \sum_{m=1}^{M} \hat{\theta}_m. \tag{9}$$

The within-implicate variance is

$$\hat{\sigma^2_W} = M^{-1} \sum_{m=1}^{M} \hat{\sigma^2_m}. \tag{10}$$

The between-implicate variance is

$$\hat{\sigma^2_B} = (M-1)^{-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}\right)^2. \tag{11}$$

The total variance associated with $\hat{\theta}$ is

$$\hat{\sigma^2} = \hat{\sigma^2_W} + (1 + M^{-1})\hat{\sigma^2_B}. \tag{12}$$

4 Assessing model fit and linkage accuracy

In this section we implement our record linkage methodology by matching employed respondents in the 2010 wave of HRS to the BR. We begin by showing selected partial effects of the EN-based employer and establishment matching models and compare the predictive accuracy of EN-based models with simpler logit models. We then show statistics that quantify the degree of linkage uncertainty under different types of blocking schemes. For non-EIN blocked matches, we provide evidence that our threshold-based procedure reduces bias in imputed employer and establishment characteristics. Finally, we show characteristics of matched and unmatched respondents in the CenHRS.

4.1 ML matching model estimates

4.1.1 Partial effects of matching models

Figure 3 shows partial effects of JW scores for name and address on employer (top row) and establishment (bottom row) match probability. These partial effects plot the numerical derivative of

[22] In complementary work in a regression context, potential matches can be aggregated using match probability estimates as weights as in Lahiri and Larsen (2005).

the estimated model with respect to each JW score, holding all other predictors at their sample means. 95 percent confidence intervals represent posterior uncertainty in the parameters that index the matching model and are based on the 10 WBB replications of the training data. The first two graphs in each row show partial effects in EIN-blocked training data, while the second two graphs show partial effects in non-EIN-blocked data. Conditional on having EINs available for blocking, JW scores for name are informative about employer match status, while JW scores for address are not. The reverse is true for establishment match status, where isolating the right workplace from a set of potential workplaces loads more heavily on address information. In the absence of EINs, we see that JW scores for both name and address are important, although they matter only at very high levels of similarity.

The partial effects shown in Figure 3 underscore the value of using the EN estimator and relying on cubic splines with dense interactions to model match status. These higher order terms capture sharp inflection points in the match likelihood, thereby mimicking non-linearities in reviewer decisions that would be infeasible to replicate using a simpler parametric approach.

4.1.2 Predictive performance evaluated using cross-validated ROC curves

We illustrate the predictive performance of our models by showing receiver operating characteristics (ROC) curves in Figure 4. For probability thresholds ranging from 0 to 1, the ROC curve plots the true positive rate on the vertical axis against the false positive rate on the horizontal axis. A model that was only as good as chance in classifying matches would have an ROC curve that ran along the 45-degree line, while a perfect classifier would have an ROC curve that hugged the left and top edges of the graph. The area under the curve (the c-statistic) would be 0.5 for the good-as-chance classifier, while it is 1.0 for a perfect classifier.

The top row of Figure 4 compares employer and establishment match prediction performance using ROC metrics in training data that are blocked on EINs. The lower row shows analogous ROC metrics for training data that are blocked using location-specific variables (i.e. in the absence of EINs). In each graph, the blue curve shows the performance of the EN estimator with the full suite of predictors, while the red curve shows the performance of traditional logistic regression that uses only JW scores for name and address. ROC curves for both the EN and the logit models are constructed using 10-fold cross validation to show out-of-sample performance. We further compare the supervised linkage models with the well-known unsupervised Fellegi-Sunter (FS) record linkage model, which is estimated using the Expectation-Maximization (EM) algorithm. The ROC curve for the FS model is plotted in green. We use binary name and address agreement as linking variables for the FS model, which, following Winkler (1990), are set to equal one when the respective JW score is above 0.94. Comparing the three sets of ROC curves, we see that the supervised models outperform the FS model by a healthy margin. Comparing across the two supervised models, we see that EN consistently outperforms logit, although the gain is not as pronounced for establishment matching when EINs are available.

4.2 Evaluating the linkage

4.2.1 Quantifying match uncertainty for probabilistically linked respondents

Table 4 provides a simple way to summarize the degree of linkage uncertainty in our probabilistically matched sub-samples. The upper panel shows statistics for employer linkage whereas the lower panel shows statistics for establishment linkage. In each panel, we divide respondents into four different groups. The first group, shown in the first column, refers to respondents who are probabilistically matched using EINs. The next set of columns shows respondents who are probabilistically matched without EINs using three different thresholds: the naive case of no threshold, the optimally-chosen threshold, and an extreme threshold that delivers a precision rate of 80 percent.[23]

The first row of the table shows the fraction of HRS respondents for which a single employer populated all 10 implicates; that is, there is no uncertainty about the linkage. Subsequent rows show the share of implicates associated with successively higher numbers of unique matches. Cases with 4 or more unique matches are binned together. With EIN-based linkage almost 90 percent of respondents have no linkage uncertainty. In the absence of EINs, unrestricted sampling generates a very high level of uncertainty where nearly 80 percent of respondents are matched to 10 different employers. Applying thresholds, however, leads to a sharp reduction in linkage uncertainty, moving from what is effectively random matching to near-EIN levels of match quality when extreme thresholds are applied.

The statistics in the lower panel of the figure paint a qualitatively similar picture. Unsurprisingly, the overall degree of establishment-level linkage uncertainty is higher because of the added difficulty of finding the correct location in addition to the correct employer. With EINs, we see approximately 45 percent of respondents have no linkage uncertainty at the establishment level, only about half of what we attain at the employer level. Nevertheless, the gains in linkage accuracy are very substantial once we use thresholds as shown in the non-EIN columns.

The statistics in Table 4 summarize the degree of linkage uncertainty in the CenHRS and quantify the extent to which it can be mitigated using principled, data-driven, techniques. Researchers might not be happy with this uncertainty, but making it explicit is clearly superior to choosing a deterministic procedure and proceeding as if it were exact.

4.2.2 Using thresholds reduces bias in imputed variables

In the previous section we showed how we used probability thresholds to improve linkage precision when EINs were unavailable. With an objective function to select optimal thresholds that placed equal weight on precision and the linkage rate, we obtained a precision level of about 0.6. Although this is several times larger than what we would find without thresholds, it is still far from 1 which leaves open the possibility that employer- and establishment-level variables may be biased.

In top panel of Figure 5 we show the relationship between MI-log employer size from the BR and

[23] Recall from Figure 2 that 80 percent precision is the approximate upper bound of what the location-based blocking variables and matching models can attain.

true log employer size in our validation sample. We divide the true log employer size distribution into 20 equally sized bins and plot the bin-level mean of the true value on the horizontal axis against the bin-level mean of the MI value on the vertical axis. An unbiased imputation procedure would accurately fit each ventile of unobserved variable and therefore lie along the 45-degree line. We see that non-EIN-based linkage without thresholds generates biased imputations across the entire employer size distribution and is therefore far from ideal. In contrast, non-EIN-based linkage with an optimally chosen threshold accurately imputes the missing variable of interest.

In the lower panel of Figure 5, we conduct the same exercise but with log establishment size as the target variable. As with the employer-level imputations, non-EIN-based linkage without thresholds is biased across the entire establishment size distribution. In contrast, MI with an optimally chosen threshold is accurate in the lower half of the establishment size distribution but understates true establishment size at extremely large workplaces. Although the confidence intervals widen substantially in the right tail, MI-log establishment size is understated which is driven in part by the difficulty of finding extremely large establishments. In some instances, like public school districts or other types of public-sector employers, BR establishment data can represent aggregations of workplaces across different locations, so the meaning and use of the establishment match is less clear. Matches at the employer level are not subject to this problem.

4.2.3 Individual characteristics for matched and unmatched respondents

Table 5 shows selected characteristics of employed HRS respondents in the 2010 wave. The first column shows statistics for the full sample while the next four columns show statistics for linked and non-linked sub-samples at the employer- and establishment-level respectively. Because the overall linkage rate ranges from 85 to 90 percent, successfully linked sub-samples are broadly representative of the full sample of respondents. There are, however, systematic differences between the characteristics of linked and unlinked respondents that are informative about reasons for non-linkage. Moving down the rows of the table, one sees that unlinked respondents are less likely to be White, and more likely to be Hispanic and foreign born. They have, on average, about one less year of education, 11-22 percent lower annual earnings, and between 2 and 3 years less in tenure with their employer relative to linked respondents. Markedly, while about 20 percent of linked respondents are employed in the public sector, the same statistic for unlinked respondents is 3 percent. Finally, looking at the paradata, one sees that although there are no differences in the mode of interview, linked respondents are more likely to answer the survey instrument in English. Combined with the higher Hispanic and foreign born share, this statistic indicates that immigrants are likely to be over-represented in the non-linked sub-sample.

The data presented in Table 5 point at two potential drivers for selective linkage in the CenHRS. First, it is possible that non-linked respondents simply have less identifying information about their employers and therefore provide lower quality data to the HRS. Second, it is possible that these respondents intentionally withhold identifying information. The second possibility is consistent with the fact that non-linked respondents are non-consenters to SSA linkages by construction and may therefore prefer to maintain a higher level of anonymity relative to consenters.

4.3 The Census-Enhanced Health and Retirement Study (CenHRS) crosswalk

Using the methods described in this paper, we have created a resource that is available to researchers. The CenHRS crosswalk links HRS respondents to their employers in the BR. The crosswalk is multiply-imputed using the techniques described in this paper. In addition to employer information available in the BR, the crosswalk permits linkage of the HRS to other datasets linked to the BR, including summary measures of workforce characteristics from Longitudinal Employer-Household Dynamics (LEHD) data, Form 5500 data, and Census business surveys.

The crosswalk is available to the research community through Federal Statistical Research Data Centers (FSRDCs). To apply for access, researchers must contact their local FSRDC administrator to develop a proposal. Interested researchers will ultimately need to submit a proposal to the Census Bureau via the ResearchDataGov portal and to the HRS via their Restricted Data Application portal and to obtain Special Sworn Status (SSS).

5 Application: The wage-size gradient

Using both household and employer-level survey data as well as administrative employer-employee linked data, a number of studies have established that larger employers pay observationally equivalent workers higher wages (see, e.g., Brown and Medoff (1989), Oi and Idson (1999), and Bloom et al. (2018)). In this section, we discuss an application of the CenHRS by re-examining the relationship between wages and establishment size. In particular, our approach reveals that non-sampling errors in survey data are correlated with workplace characteristics and would remain hidden without constructing linkages to administrative data.

5.1 Wage-size gradient in household-survey data

Consider the following statistical model for the relationship between worker wages and establishment size in the cross section

$$w_{ij} = \gamma_0 + \gamma_1 s^*_{ij} + v_{ij}, \tag{13}$$

where w_{ij} is the log hourly nominal wage of worker i employed at establishment j, s^*_{ij} is an error-free measure of the log of worker $i's$ establishment's size, and v_{ij} is an error term that captures other factors influencing worker wages.[24] The HRS provides household survey-based measures of hourly wages, w_{ij}, as well as establishment size. s_{ij} is often missing and is potentially error-ridden when

[24] Although we ignore control variables when writing Equation (13), our empirical implementation includes control variables, which we describe in detail below. Control variables may be subject to measurement errors of their own, which makes the bias in the variable of interest hard to characterize. We ignore the added effect of measurement errors in the control variables, to the extent they exist, in the discussion that follows.

it is reported. Survey-based measures of log establishment size can be written as

$$s_{ij} = s^*_{ij} + u_{ij}, \tag{14}$$

where u_{ij} is the measurement error. Assume throughout that $Cov(u_{ij,}v_{ij}) = 0$. The probability limit of $\hat{\gamma}_1$ estimated using the survey-based measure of establishment size by ordinary least squares (OLS) is

$$\begin{aligned} \text{plim}\, \hat{\gamma}_{1,\text{S}} &= \frac{Cov(s_{ij}, \gamma_0 + \gamma_1 s^*_{ij} + v_{ij})}{V(s_{ij})} \\ &= \gamma_1 \frac{V(s^*_{ij}) + Cov(s^*_{ij}, u_{ij})}{V(s^*_{ij}) + V(u_{ij})}. \end{aligned} \tag{15}$$

Under the classical measurement error model, discrepancies in survey reports are not systematically related with the underlying variable of interest implying that $Cov(s^*_{ij}, u_{ij}) = 0$. Given this framework, the presence of added noise in the explanatory variable attenuates $\hat{\gamma}_{1,\text{S}}$. Alternatively, if discrepancies in survey reports are systematically related to the underlying variable of interest—i.e. if the measurement error is non-classical—then $\hat{\gamma}_{1,\text{S}}$ may be either amplified or attenuated depending on the sign of $Cov(s^*_{ij}, u_{ij})$ and its magnitude relative to $V(u_{ij})$. In the next subsection, we present evidence that survey response error is non-classical.

5.2 Using MI variables from administrative data to assess bias in the wage-size gradient

Define $\hat{s}^{*(m)}_{ij}$ as the m-th implicate of log establishment size obtained using our MI-based procedure. We can write the true value of log establishment size under our imputation procedure as

$$s^*_{ij} = \hat{s}^{*(m)}_{ij} + \eta^{(m)}_{ij}. \tag{16}$$

Given the rich suite of variables used in the imputation, it is reasonable to assume that the condition posited in Equation (5) is met. Consequently, the following moment conditions hold:

$$Cov(\hat{s}^{*(m)}_{ij}, \eta^{(m)}_{ij}) = 0 \tag{17}$$

$$Cov(\hat{s}^{*(m)}_{ij}, v_{ij}) = 0. \tag{18}$$

That is, the imputed variable is uncorrelated with imputation error, $\eta^{(m)}_{ij}$, as well as the error in the regression model, v_{ij}.

The probability limit of $\hat{\gamma}_1$ estimated using the m-th implicate of establishment size by OLS is

$$\begin{aligned}\text{plim}\,\hat{\gamma}_{1,\texttt{MI}}^{(m)} &= \frac{Cov(\hat{s}_{ij}^{*(m)}, \gamma_0 + \gamma_1 s_{ij}^* + v_{ij})}{V(\hat{s}_{ij}^{*(m)})} \\ &= \gamma_1 \frac{Cov(\hat{s}_{ij}^{*(m)}, s_{ij}^*)}{V(\hat{s}_{ij}^{*(m)})}, \end{aligned} \tag{19}$$

where the second expression follows from Equation (18). Finally, from Equations (16) and (17) it follows that $Cov(\hat{s}_{ij}^{*(m)}, s_{ij}^*) = V(\hat{s}_{ij}^{*(m)})$ and hence,

$$\text{plim}\,\hat{\gamma}_{1,\texttt{MI}}^{(m)} = \gamma_1. \tag{20}$$

It follows that $\hat{\gamma}_{1,\texttt{MI}}$, which averages across the $m = 1, \ldots, M$ parameter estimates obtained from each of the completed datasets, is also consistent.

Having shown the conditions under which the MI-based estimator is consistent, we can write the inconsistency arising from the use of survey-based reports of establishment size as

$$\text{plim}\,\hat{\gamma}_{1,\texttt{S}} - \text{plim}\,\hat{\gamma}_{1,\texttt{MI}} = \gamma_1 \left(\frac{V(s_{ij}^*) + Cov(s_{ij}^*, u_{ij})}{V(s_{ij}^*) + V(u_{ij})} - 1 \right). \tag{21}$$

As noted previously, non-classical measurement error, which engenders a correlation between s_{ij}^* and u_{ij}, can generate either attenuation or amplification bias in $\hat{\gamma}_{1,\texttt{S}}$.

Table 6 shows coefficient estimates obtained using the different measures of establishment size across sub-groups of respondents. Panel A considers all respondents with wage data who can be matched to establishments in the BR, either deterministically or probabilistically. Of these respondents, column (1) is restricted to those who report establishment size, while column (2) includes all respondents. Panel B considers only those respondents who can be deterministically linked to establishments in the BR. Of these respondents, column (3) is restricted to those who report establishment size, while column (4) includes all respondents. The first row of the table shows estimates based on survey measures of establishment size, which can only be obtained when respondents report them. The second row shows estimates based on administrative data on establishment size from the BR. To control for variation in individual characteristics that affect hourly wages, all the regression models we estimate condition on age, gender, race, Hispanic ethnicity, partnered/coupled status, years of education, tenure, hours worked per week, weeks worked per year, one-digit occupation fixed effects, and one-digit industry fixed effects.[25] We focus on the variable of interest imputed from the BR and do not report coefficients for these control variables in the table.

Looking first at column (1), we see that $\hat{\gamma}_{1,\texttt{S}}$ is significantly larger than $\hat{\gamma}_{1,\texttt{MI}}$. The presence of amplification bias in the survey-based coefficient is consistent with HRS respondents' reports

[25] The sample sizes reported in the table represent cases for which both hourly wage data as well as all control variables are non-missing.

of workplace size being subject to non-classical measurement error. Notably, this bias is present not only in the full sample (column 1), but also in the perfectly matched sub-sample (column 3), indicating that our finding of non-classical measurement error is not merely an artifact of the probabilistic record linkage process.

Obtaining establishment size from administrative data linkages provides two measurement advantages relative to survey data. The first, as described above, is to address measurement error concerns. The second is that we are able to observe establishment size even when survey respondents may not have reported it. This is evident from the increase in sample size that results from moving from column (1) to column (2). The MI-based coefficients in these two columns are comparable, indicating that selective non-response does not contribute to bias. In contrast, comparing the MI-based estimates in column (3) to those column (4), we see that selective non-response amplifies the estimated coefficient by about 0.01. This represents about half of the overall amplification bias of 0.021 (0.044-0.023), with the remainder being due to non-classical measurement error. Finally, we note that the estimates of $\hat{\gamma}_{1,\mathtt{MI}}$ in columns (2) and (4) are very close. The proximity of these estimates serves as a useful validity check of the record linkage algorithm because it shows that the wage-size gradient in the perfectly linked sub-sample is similar to the gradient in the full sample.[26]

Figure 6 illustrates the nature of non-sampling errors in the full sample (left column) and the perfectly matched sub-sample (right column). In each of the four panels we divide MI-log establishment size from the BR into 20 equally-sized bins and plot the bin-level mean on the horizontal axis. In the top two panels, we plot the bin-level mean of log establishment size reported by HRS respondents on the vertical axis. The bin scatter plot shows that measurement error in survey reports is *negatively* correlated with the true values: That is, workers at the smallest establishments overstate size, and workers at larger establishments understate size. In the lower tail, differences between survey and administrative data could reflect seasonal volatility in small establishments.[27] In the upper tail, where errors are more pronounced and size is less volatile over calendar months, the pattern is consistent with the idea that individual employees may be unaware of the full scale of operations and may therefore underestimate workplace size when answering the survey. In the bottom two panels, we plot the bin-level mean of the rate of non-response about establishment size. In these panels, we see that non-response is higher on average in the perfectly linked sub-sample relative to the broader sample of all linked respondents.

The estimates we report here provide new evidence on how household survey responses about workplace characteristics are selectively misreported, or not reported at all. With linkages to administrative information on workplaces in the CenHRS, we are able to characterize measurement and non-response errors that are not observed in other household survey datasets.

[26] Patki and Shapiro (2023) propose a two-stage least squares (TSLS) estimator for dealing with failures of the strongly non-informative linkage assumption. In the data used here, the MI- and TSLS-based estimates are essentially identical, which provides additional evidence that the assumptions underpinning the linkage algorithm appear to hold. See Appendix C.

[27] BR size information is based on payroll tax information reported in March, while three-quarters of HRS interviews are conducted in the summer months between May and September.

6 Conclusion

This paper describes the construction of a new dataset, the CenHRS, that is obtained by linking a household-level survey to an establishment-level frame in the absence of unique identifiers. The between-frame linkage task that we undertake is complicated by skewness in the distribution of employment across firms that makes matching much more difficult. To address these issues, we use probabilistic linkage based on supervised machine learning models to estimate the probability that specific employers and establishments in the BR are matches for individuals in the HRS. Our models rely on a rich set of predictors and a high degree of flexibility to replicate important non-linearities inherent in human-reviewed training data. Using probabilities estimated from the models, we employ multiple imputation to characterize uncertainty in the linkage. To further refine the set of candidate matches, we estimate probability thresholds that provide the best trade-off between precision and the sample linkage rate. Eliminating candidate matches that fail to meet these thresholds dramatically reduces both linkage uncertainty as well as bias in the imputed variables. We use these newly linked data to provide new evidence that reporting errors and non-response propensity vary systematically with workplace characteristics.

Beyond issues related to record linkage, the CenHRS opens new avenues for research by extending pre-existing measures of activities, experiences, and outcomes for individuals from their family and home context to the work context. These new measures will provide data necessary for a more comprehensive understanding of the determinants of health and well-being over the lifespan.

References

Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Anderson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock (2009) "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators," in Dunne, Timothy, J. Bradford Jensen, and Mark J. Roberts eds. *Producer Dynamics: New Evidence from Micro Data*, 149–230: University of Chicago Press.

Abowd, John M. and Martha H. Stinson (2013) "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data," *Review of Economics and Statistics*, 95 (5), 1451–1467.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey (2017) "How Well Do Automated Methods Perform in Historical Samples? Evidence From New Ground Truth," NBER Working Paper No. 24019.

Belin, Thomas R. and Donald B. Rubin (1995) "A Method for Calibrating False-Match Rates in Record Linkage," *Journal of the American Statistical Association*, 90 (430), 694–707.

Bloom, Nicholas, Fatih Guvenen, Benjamin S. Smith, Jae Song, and Till von Wachter (2018) "Inequality and the Disappearing Large Firm Wage Premium," *American Economic Association Papers and Proceedings*, 108, 317–322.

Brown, Charles and James Medoff (1989) "The Employer Size-Wage Effect," *Journal of Political Economy*, 97, 1027–1059.

Burgette, Lane. F. and Jerome P. Reiter (2010) "Multiple Imputation for Missing Data via Sequential Regression Trees," *American Journal of Epidemiology*, 172 (9), 1070–1076.

Christen, Peter (2008a) "Automatic Record Linkage Using Seeded Nearest Neighbour and Support Vector Machine Classification," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 151–159: Association for Computing Machinery.

——— (2008b) "Automatic Training Example Selection for Scalable Unsupervised Record Linkage," in *Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 511–518: Springer-Verlag.

——— (2012) *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*: Springer-Verlag.

Cochinwala, Munir, Verghese Kurien, Gail Lalk, and Dennis Shasha (2001) "Efficient Data Reconciliation," *Information Sciences*, 137 (1-4), 1–15.

Coffin, Marie and Shashikala Sukhatme (1997) "Receiver Operating Characteristic Studies and Measurement Errors," *Biometrics*, 53, 823–837.

Elfeky, M. G., V. S. Verykios, and A. K. Elmagarmid (2002) "TAILOR: A Record Linkage Toolbox," in *Proceedings 18th International Conference on Data Engineering*, 17–28: Institute of Electrical and Electronics Engineers.

Fellegi, Ivan P. and Alan B. Sunter (1969) "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64 (328), 1183–1210.

Ferrie, Joseph P. (1996) "A New Sample of Males Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript Schedules," *Historical Methods*, 29 (4), 141–156.

Fortini, Marco, Brunero Liseo, Alessandra Nuccitelli, and Mauro Scanu (2001) "On Bayesian Record Linkage," *Research in Official Statistics*, 4 (1), 185–198.

Goldstein, Harvey, Katie Harron, and Angie Wade (2012) "The Analysis of Record-Linked Data Using Multiple Imputation with Data Value Priors," *Statistics in Medicine*, 31 (28), 3481–3493.

Gutman, Roee, Christopher C. Afendulis, and Alan M. Zaslavsky (2013) "A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs," *Journal of the American Statistical Association*, 108 (501), 34–47.

Gutman, Roee, Cara J. Sammartino, Traci C. Green, and Brian T. Montague (2016) "Error Adjustments for File Linking Methods Using Encrypted Unique Client Identifier (eUCI) with Application to Recently Released HIV+ Prisoners," *Statistics in Medicine*, 35 (1), 115–129.

Han, Ying and Partha Lahiri (2019) "Statistical Analysis with Linked Data," *International Statistical Review*, 87, S139–S157.

Lahiri, Partha and Michael D. Larsen (2005) "Regression Analysis With Linked Data," *Journal of the American Statistical Association*, 100 (469), 222–230.

Larsen, Michael D. (2004) "Record Linkage Using Finite Mixture Models," in Gelman, Andrew and Xiao-Li Meng eds. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, 309–318: John Wiley.

Lawson, Elise H., Clifford Y. Ko, Rachel Louie, Lein Han, Micheal Rapp, and David S. Zigmond (2013) "Linkage of a Clinical Surgical Registry with Medicare Inpatient Claims Data Using Indirect Identifiers," *Surgery*, 153 (3), 423–430.

McKinney, Kevin, Andrew Green, Lars Vilhuber, and John M. Abowd (2021) "Total Error and Variability Measures for the Quarterly Workforce Indicators and LEHD Origin-Destination Employment Statistics in OnTheMap," *Journal of Survey Statistics and Methodology*, 9 (5), 1146–1182.

Meng, Xiao-Li (1994) "Multiple-Imputation Inferences with Uncongenial Sources of Input," *Statistical Science*, 9 (4), 566–573.

Murray, Jared S. (2018) "Multiple Imputation: A Review of Practical and Theoretical Findings," *Statistical Science*, 33 (2), 142–159.

Newton, Michael A., Nicholas G. Polson, and Jianeng Xu (2021) "Weighted Bayesian Bootstrap for Scalable Posterior Distributions," *Canadian Journal of Statistics*, 49 (2), 421–437.

Newton, Michael A. and Adrian E. Raftery (1994) "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society, Series B*, 56 (1), 3–48.

Oi, Walter Y. and Todd L. Idson (1999) "Firm Size and Wages," in Ashenfelter, Orley C. and David Card eds. *Handbook of Labor Economics 3B*, 2165–2214: North Holland.

Patki, Dhiren and Matthew D. Shapiro (2023) "Implicates as Instrumental Variables: An Approach for Estimation and Inference with Probabilistically Matched Data," *Journal of Survey Statistics and Methodology*, 11 (3), 597–618.

Reiter, Jerome P. (2005) "Using CART to Generate Partially Synthetic Public Use Microdata," *Journal of Official Statistics*, 21 (3), 441–462.

Rubin, Donald B. (1981) "The Bayesian Bootstrap," *Annals of Statistics*, 9 (1), 130–134.

——— (1987) *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

——— (1996) "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91 (434), 473–489.

Setoguchi, Seko, Ying Zhu, Jessica J. Jalbert, Lauren A. Williams, and Chih-Ying Chen (2014) "Validity of Deterministic Record Linkage Using Multiple Indirect Personal Identifiers," *Circulation: Cardiovascular Quality and Outcomes*, 7 (3), 475–480.

Steorts, Rebecca C., Rob Hall, and Stephen E. Feinberg (2016) "A Bayesian Approach to Graphical Record Linkage and De-duplication," *Journal of the American Statistical Association*, 111 (516), 1660–1672.

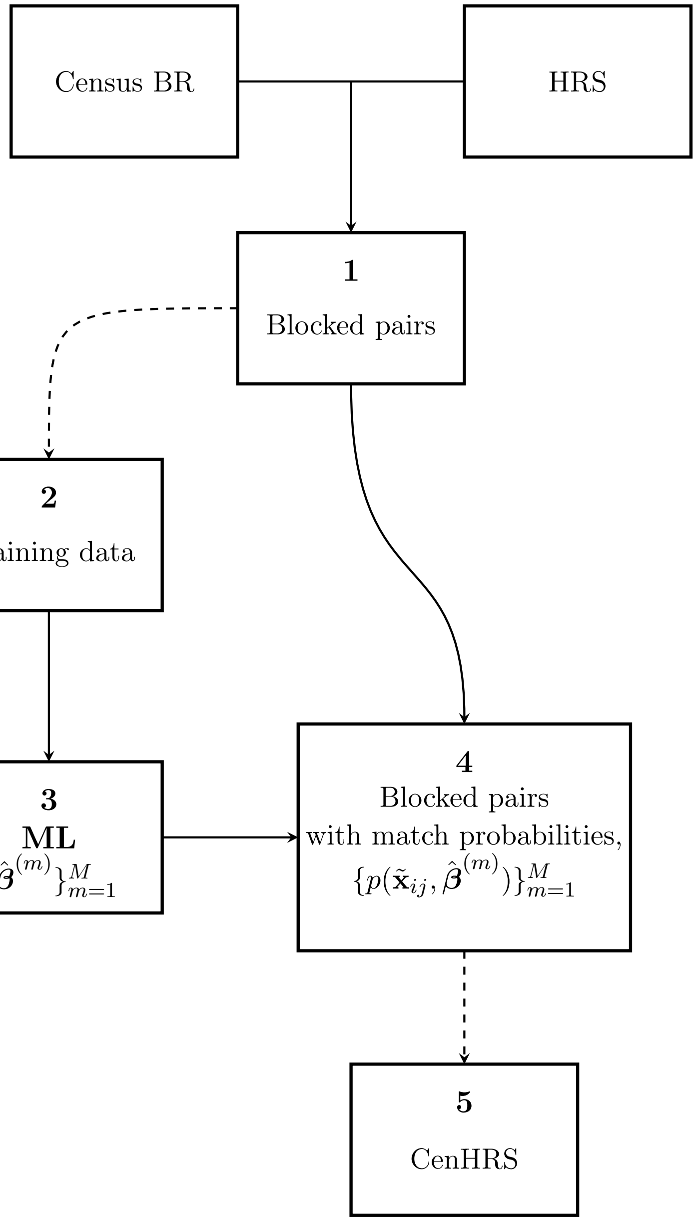
Stinson, Martha (2003) "Technical Description of SIPP Job Identification Number Editing, 1990-1993 SIPP Panels," SIPP Technical Paper, U.S. Census Bureau.

Tancredi, Andrea and Brunero Liseo (2011) "A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems," *Annals of Applied Statistics*, 5 (2B), 1553–1585.

Winkler, William (1990) "String Comparator Metrics and Enhanced Decision Rules in the Fellegiâ-Sunter Model of Record Linkage," in *ASA Proceedings of the Survey Research Methods Section*, 354–359: American Statistical Association.

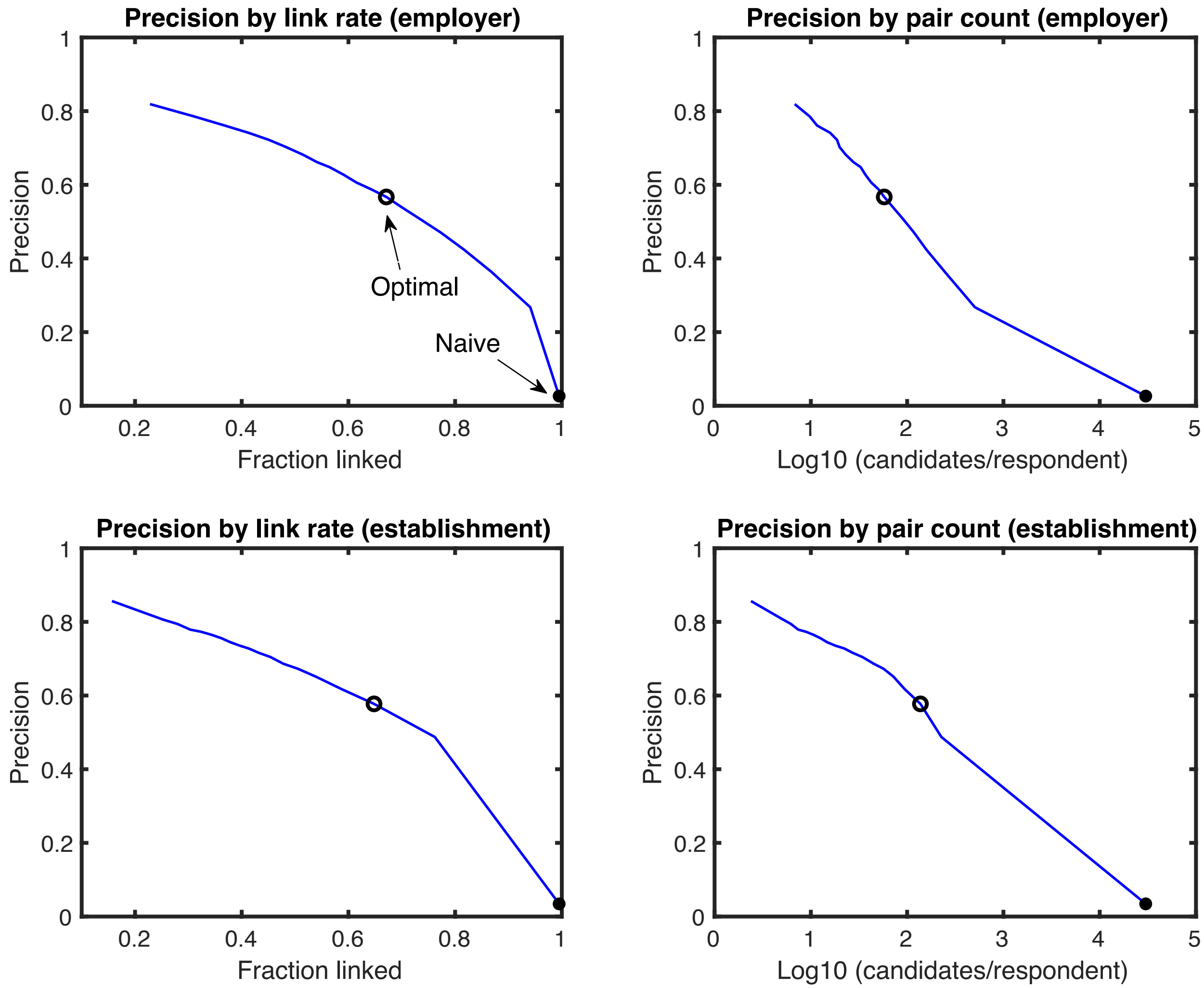
Zou, Hui and Trevor Hastie (2005) "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

Figure 1: CenHRS Record Linkage Process



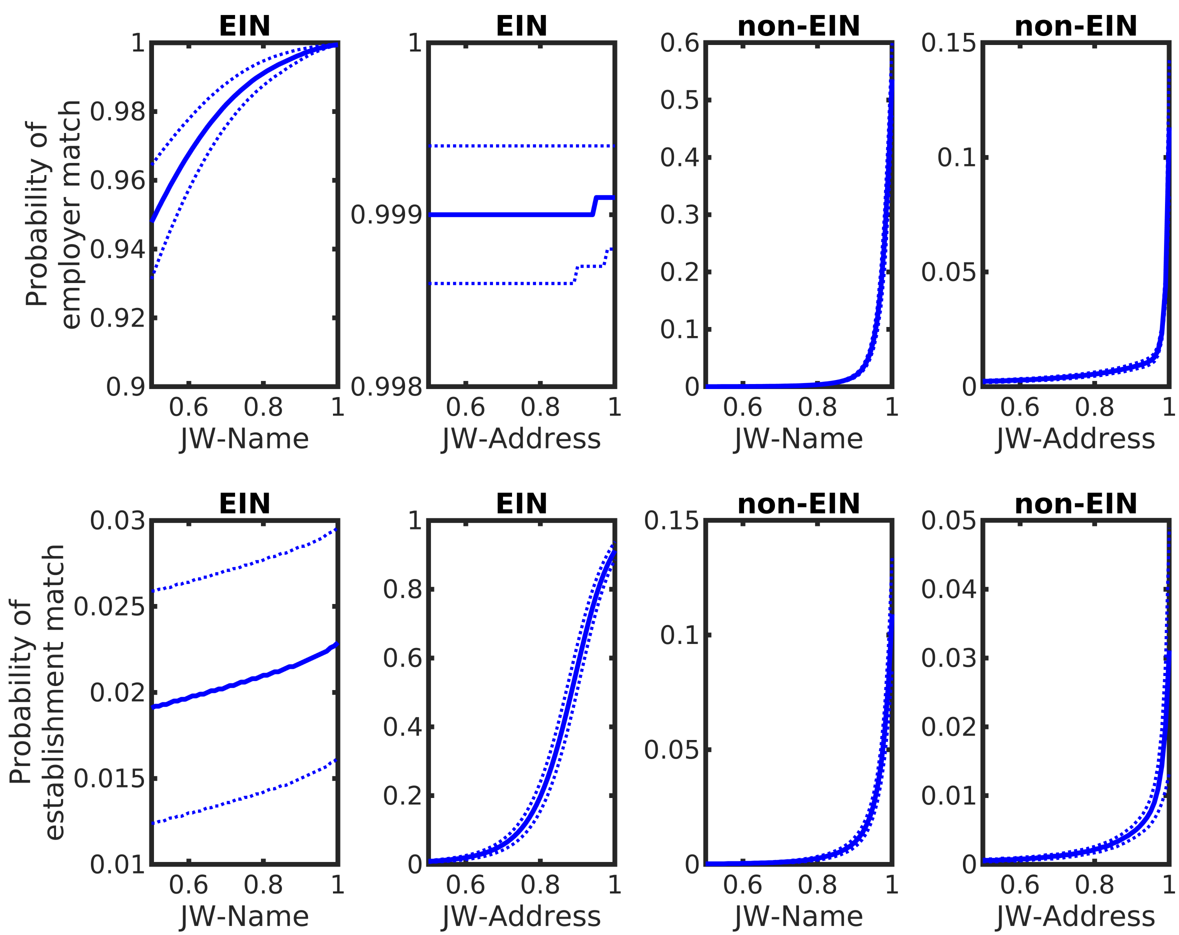
Notes: This figure shows the record linkage process for the CenHRS. Dashed lines represent sampling. Solid lines represent operations performed on all observations in the data set. Key steps are labeled with numbers. 1: blocking pairs of records from the HRS and BR files, 2: drawing unlabeled pairs and creating a training data set through human review, 3: estimating the linkage model using machine learning, 4: obtaining match probability estimates, 5: drawing multiple implicates with probability proportional to the estimated match probability.

Figure 2: Precision rates in non-EIN-based record linkage



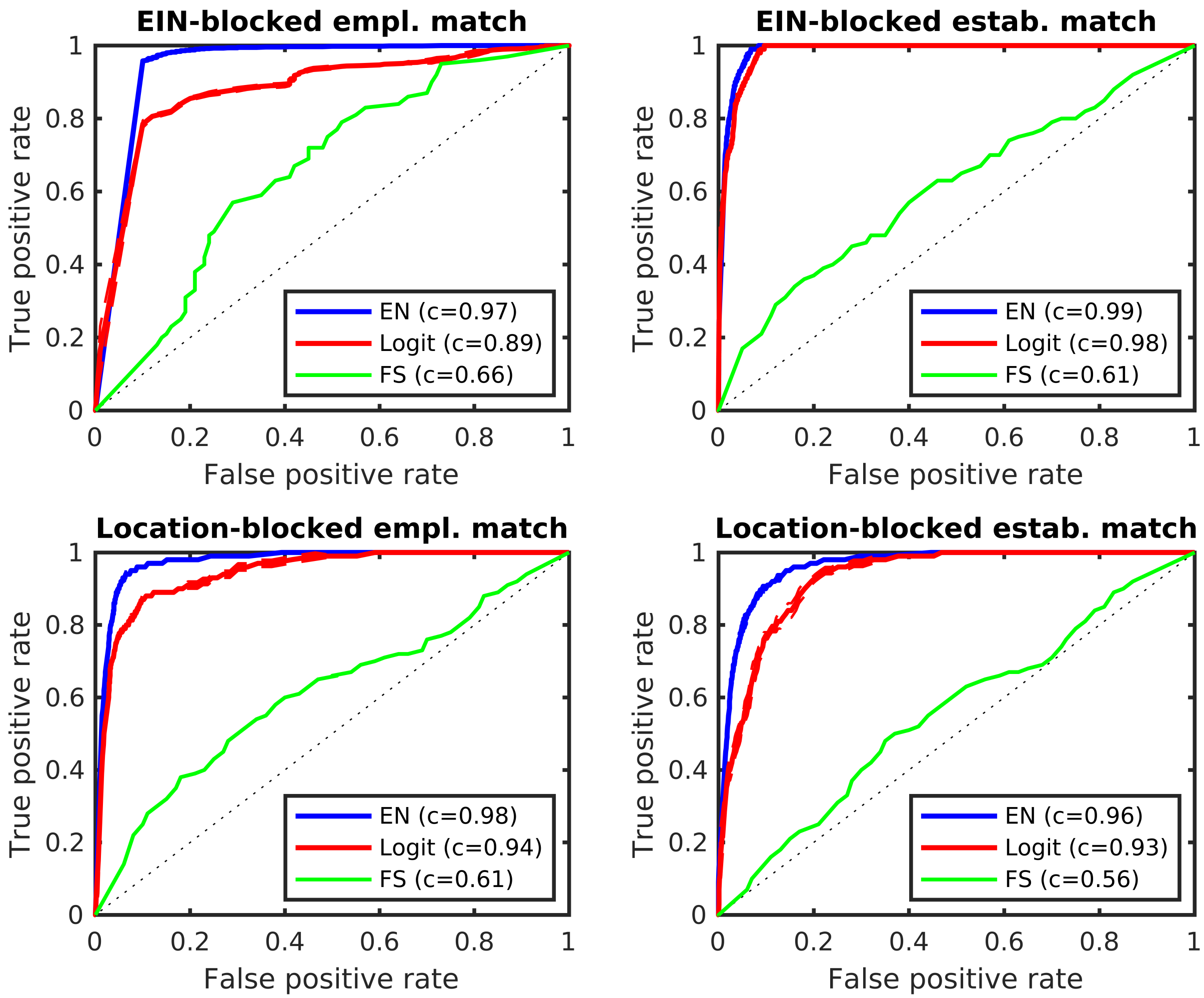
Notes: This figure shows the precision rate (i.e. the fraction of HRS respondents correctly matched) attained under different probability thresholds in a validation sample. The precision rate is the rate achieved by the EN estimator for different probability thresholds. Statistics are averages across 10 WBB replications.

Figure 3: Selected partial effects of the matching models



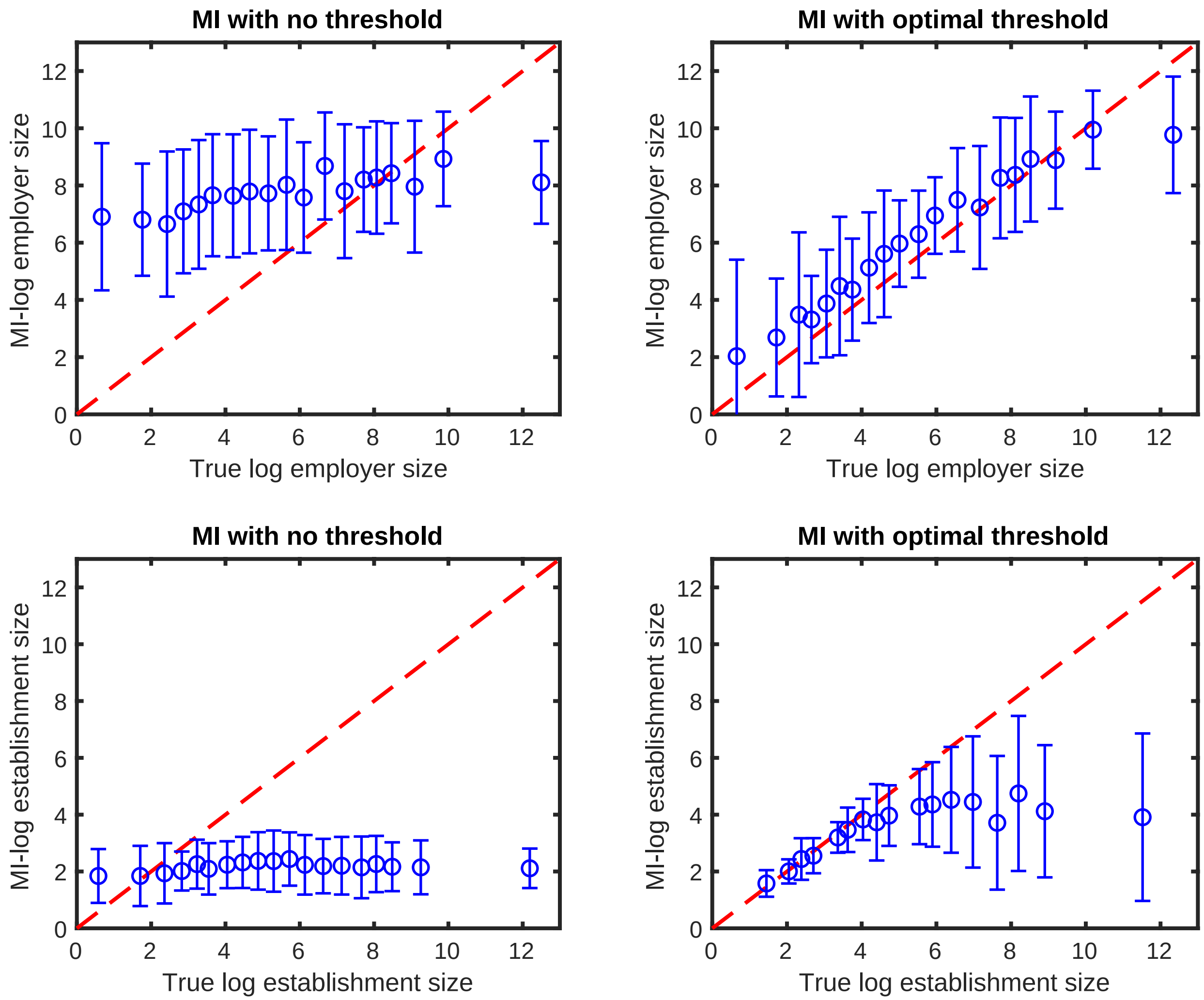
Notes: This figure shows partial effects of the matching models for a given predictor holding all other predictors at their mean. The vertical axes have different scales in each graph. The top row shows the effect of Jaro-Winkler (JW) scores for name and address similarity between the HRS and BR on the probability of employer match status, separately for EIN and non-EIN blocked training data. The bottom row shows the effect of the same predictors on establishment match status, separately for EIN and non-EIN blocked training data. 95 percent confidence intervals reflect posterior uncertainty in the parameters of the matching models and are estimated using Bayesian bootstrap replications of the training data.

Figure 4: Receiver operating characteristic curves of different matching models



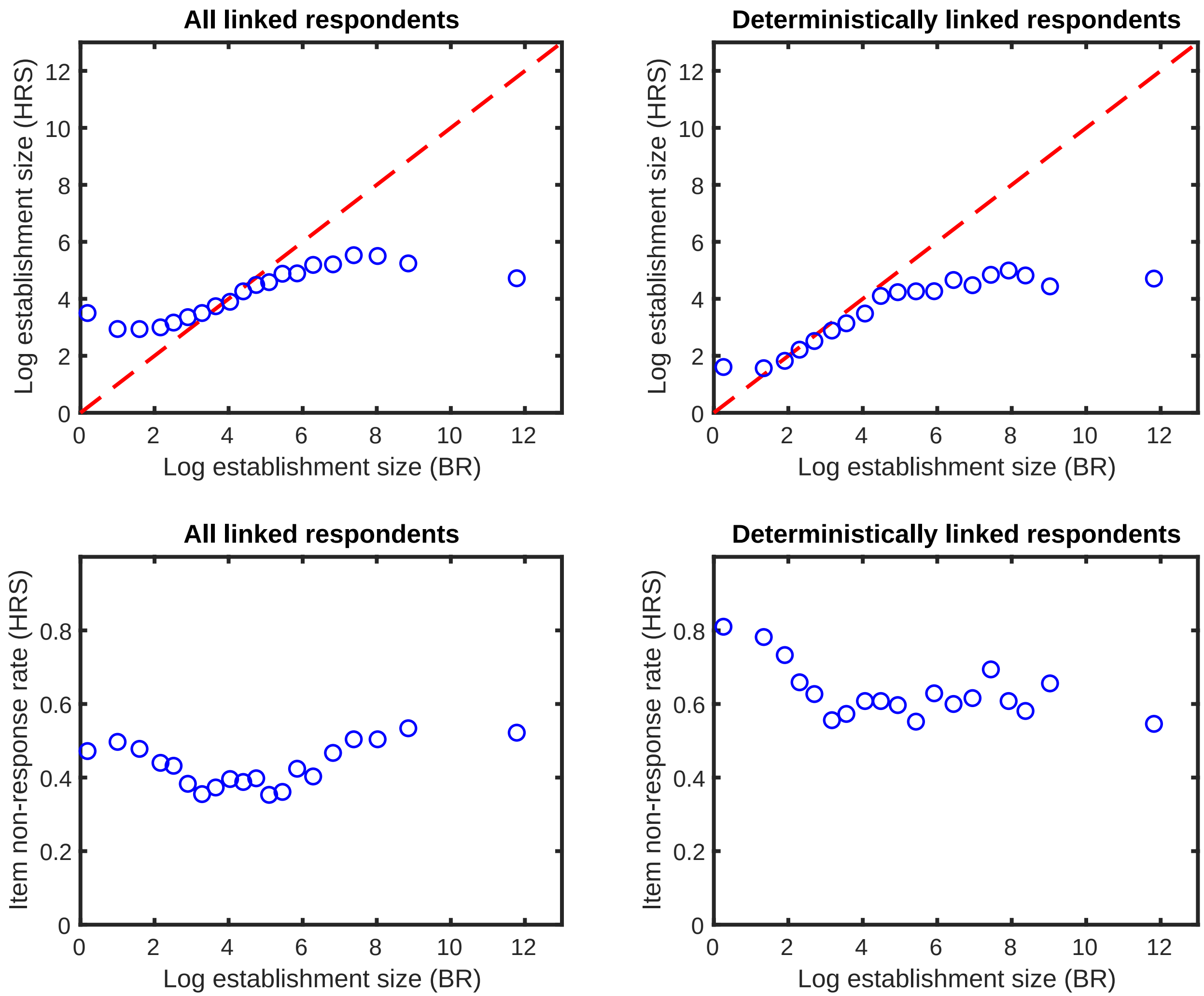
Notes: EN represents the Elastic Net estimator which uses the full suite of predictors, Logit is conventional logistic regression estimated using only Jaro-Winkler scores for name and address, and FS represents the Fellegi-Sunter algorithm using binary variables for name and address agreement. EN and Logit estimates show out-of-sample predictive performance, which is estimated using 10-fold cross-validation. c-statistics show the area under each curve which ranges from a minimum of 0.5 (random classifer) to a maximum of 1 (perfect classifier). 95 percent confidence intervals for EN and Logit reflect posterior uncertainty in the parameters of the matching models and are estimated using Bayesian bootstrap replications of the training data.

Figure 5: Reducing imputation bias by applying optimally chosen thresholds



Notes: This figure shows the relationship between true log size and multiply-imputed log size at the employer and establishment levels for naively and optimally chosen implicates in a validation sample. 95 percent confidence intervals for MI measures jointly account for within- and between-implicate variability using Rubin's combining formulas. Some cells are suppressed for statistical disclosure limitation.

Figure 6: Non-sampling errors in HRS reports of establishment size



Notes: This figure shows the relationship between log establishment size from the Business Register and self-reported log establishment size from the HRS in the top row, and item non-response rates about establishment size in the HRS in the bottom row.

Table 1: Types of record linkage in the 2010 wave of the CenHRS

	HRS respondents	Share of respondents	BR candidates per HRS respondent	
			Mean	Std. Dev.
Deterministic match, EIN available	2500	0.415	1	-
Probabilistic match, EIN available	1800	0.295	436	1280
Probabilistic match, EIN not available	1800	0.291	30790	29630

Notes: This table shows the record linkage strategy used for different sub-samples of working HRS respondents in the 2010 wave. Shares may not sum to 1 because each cell is independently rounded.

Table 2: Reviewer's information set

Review variable	Source	
	HRS	BR
Employer's name, establishment address, phone number	✓	✓
Whether employer is single unit or multi unit		✓
Employer and establishment size	✓	✓
Employer's industry description	✓	✓
Respondent's occupation description	✓	
Provision of health insurance and/or retirement plan	✓	
Number of EINs in respondent's earnings record	✓	

Notes: This table shows the types of variables that reviewers observe when determining match status for a HRS-BR pair in the training datasets. As shown in the source column, some variables are available in both the HRS and the BR, while others are available in only one of the two sources.

Table 3: Predictors in the matching models

Predictor	Description
	Pair level
Cubic spline Jaro-Winkler score name	Similarity between HRS and BR name
Cubic spline Jaro-Winkler score street address	Similarity between HRS and BR address
Jaro-Winkler score city	Similarity between HRS and BR city
Cubic spline establishment's share of employment within blocking variable	Concentration of employment across candidate establishments
Cubic spline EIN share of respondent's total annual earnings	Concentration of earnings across candidate employers
Agreement on 7 digit and 10 digit phone number	Agreement status binary variable
Agreement on 3-, 4-, 5-digit zip code, city-state	Agreement status binary variable
Agreement on one-digit industry code	Agreement status binary variable
Agreement on employer size class	Agreement status binary variable
	BR only
Log employer size	
Whether single-unit or mult-unit employer	
	HRS only
Age, gender, race, ethnicity, nativity, years of schooling, marital status	
Survey interview mode and language	
Log hourly real wage, tenure, weeks worked/year, hours worked/week	
Provision of health insurance, provision of retirement plan	
Two-digit occupation, one-digit industry	

Notes: This table shows the types of predictors used in the Elastic Net matching models. Pair-level predictors are based on information that is specific to the HRS-BR pair. BR-only predictors are derived purely from the BR, and HRS-only predictors are derived purely from the HRS. Cubic splines for name and address Jaro-Winkler (JW) scores have 10 cut points each. The cubic splines for block share and earnings share have 3 cut points each. Earnings shares across jobs cannot be computed for respondents who do not consent to SSA linkage. For respondents who do not consent to linkage, we use cubic spline of log BR size with 3 cut points. All cubic spline variables are fully interacted with each other. After including interaction terms and indicator variables to account for missing values of HRS variables, there are a total of 9,200 predictors for the EIN and non-EIN based models.

Table 4: Concentration of multiple implicates

Employer-level linkage				
		Non-EIN-based		
N (unique implicates)	EIN-based	No threshold	Optimal threshold	Extreme threshold
1	0.89	0.1	0.44	0.83
2	0.09	0.05	0.23	0.12
3	0.01	0.05	0.14	0.03
4-10	0.01	0.79	0.19	0.02
N (respondents)	1800	1800	1000	350
Establishment-level linkage				
		Non-EIN-based		
N (unique implicates)	EIN-based	No threshold	Optimal threshold	Extreme threshold
1	0.44	0.01	0.28	0.69
2	0.17	0.01	0.18	0.13
3	0.06	0.01	0.13	0.07
4-10	0.32	0.97	0.4	0.11
N (respondents)	1800	1800	900	250

Notes: This table shows the concentration of implicates across HRS respondents. For non-EIN-based linkage, the table shows concentration of implicates across respondents for three different thresholds: no threshold (or the naive case), the optimally-chosen threshold, and an extreme threshold which is associated with a precision rate of 80 percent in the validation data.

Table 5: HRS respondent characteristics by linkage status

		Employer		Establishment	
	Full sample	Linked	Non-linked	Linked	Non-linked
Age	57.61	57.70	56.92	57.79	56.54
Male	0.45	0.45	0.48	0.44	0.50
White	0.67	0.69	0.57	0.69	0.59
Black	0.22	0.21	0.24	0.21	0.23
Other race	0.11	0.10	0.19	0.10	0.17
Hispanic	0.15	0.14	0.26	0.14	0.25
Partnered/coupled	0.73	0.73	0.73	0.72	0.73
Years of schooling	13.24	13.40	12.06	13.43	12.11
Native born	0.85	0.87	0.69	0.86	0.73
Annual earnings ($)	41,800	42,950	33,330	42,480	37,660
Hours worked per week	38.14	38.38	36.47	38.28	37.36
Weeks worked per year	48.77	49.02	47.04	48.91	47.98
Tenure (years)	11.22	11.66	8.23	11.57	9.24
Public sector worker	0.17	0.18	0.03	0.19	0.03
Interviewed in English	0.93	0.94	0.81	0.94	0.84
Interviewed in person	0.74	0.74	0.76	0.74	0.76
N	6100	5400	750	5200	850

Notes: This table shows HRS respondent characteristics for the full sample of working respondents in the 2010 wave, and for the linked and non-linked sub-samples at the employer and establishment level. Annual earnings are in 2010 dollars. Case counts are independently rounded.

Table 6: Log establishment size effect in log wage

		A: All linked respondents		**B:** Deterministically linked sub-sample	
		HRS respondents reporting size	All HRS respondents	HRS respondents reporting size	All HRS respondents
Parameter	Establishment-size source	(1)	(2)	(3)	(4)
$\hat{\gamma}_{1,\mathtt{S}}$	HRS	0.041		0.044	
		(0.005)		(0.009)	
$\hat{\gamma}_{1,\mathtt{MI}}$	BR	0.020	0.019	0.033	0.023
		(0.004)	(0.003)	(0.006)	(0.005)
N		2600	4200	850	1800

Notes: This table shows the effect of log establishment size on log wages. Regression samples are restricted to observations where HRS respondents reported hourly wages or provided sufficient information to infer hourly wages from reports of total earnings and total hours. All regression models include controls for weekly hours, annual weeks, tenure, years of schooling, partnered/coupled status, nativity, gender, race, Hispanic ethnicity, age, one-digit occupation fixed effects, and one-digit industry fixed effects.

Appendix A Constructing the training data set

Our training samples are composed of HRS-BR pairs generated by blocking the 1998 and 2004 waves of the HRS with the BR. The first sample blocks on EIN and is used to fit probabilistic matching models for cases where EINs are available, while the second sample blocks on 10-digit phone number, 3-digit zip code, telephone area code, and city-state and is used to fit probabilistic matching models for cases where EINs are unavailable. We choose 1998 and 2004 to create the training samples for two reasons. First, these were years in which the HRS drew fresh cohorts of survey respondents. Second, the file structure of the BR changed in substantive ways in 2002. As such, using HRS cohorts before and after 2002 to estimate the matching models allows us to account for unobserved variation in the quality of data drawn from the BR.

Simple random sampling of pairs for human review would produce very few true matches and therefore limit the predictive performance of our models. Instead, we follow a stratified random sampling approach to draw candidate matches (see, e.g., Christen (2012)). We begin by computing Jaro-Winkler (JW) scores for name and address similarity for each pair. We then divide the JW scores for name and address into 4 bins each, with grid points spaced closer together at the right tail of the respective JW score distributions. This binning exercise defines 16 strata from which we draw equally-sized samples to obtain a total sample size of $N^T \approx 1000$ pairs. Because the bins are concentrated at the top of the JW name and address score distribution, this stratified sampling methodology substantially increases the share of true matches in the training data set relative to a simple random sample.

Appendix B Model selection

Our training data sets consist of approximately 2000 observations each, which is substantially smaller than the number of predictors available to estimate match probabilities. To solve this dimensionality problem and, more importantly, to avoid over-fitting our model, we use ML tools to aid in prediction. While a complex model with many variables and interactions has the potential of reducing in-sample (training) errors substantially, this improvement is misleading because it considers the wrong model-fit criterion. To ensure that the model generalizes well, we consider out-of-sample (test) error which we estimate using 10-fold cross validation.

In our setting, the complexity of the model is indexed by the number of predictors. Reducing model complexity by shrinking the number of predictors increases the bias component of the test error, but has the potential to reduce the variance component substantially. In order to obtain a model with the optimal degree of complexity, we employ the Elastic Net (EN) shrinkage estimator.

The EN estimator solves the constrained maximum likelihood problem posed in (22):

$$\max_{\beta\in\mathbb{R}^q}\sum_{l=1}^{2N^T} w_l^{(m)}\left(y_l\log\left(\frac{\exp(\tilde{\boldsymbol{x}}_l'\boldsymbol{\beta})}{1+\exp(\tilde{\boldsymbol{x}}_l'\boldsymbol{\beta})}\right)+(1-y_l)\log\left(\frac{1}{1+\exp(\tilde{\boldsymbol{x}}_l'\boldsymbol{\beta})}\right)\right)$$
$$\text{st: }\sum_{p=1}^{q}\beta_p^2\leq t_1,\ \sum_{p=1}^{q}|\beta_p|\leq t_2, \tag{22}$$

where l indexes observations in the training dataset, and p indexes predictors. In (22), the typical maximum likelihood problem is supplemented with two constraints, each of which constitutes a tuning parameter for the estimator. Together, these tuning parameters control the level of model complexity: t_1, as in Ridge Regression, sets a threshold on the sum of squared values of the coefficients. The Ridge penalty term has the effect of controlling the variance component of test error by preventing any one predictor from exhibiting too strong of an effect on the outcome. This penalty is important when some predictors are correlated. t_2, as in the LASSO, sets a threshold on the sum of the absolute values of the coefficients. When this second constraint binds, some of the coefficients are set exactly to zero, thereby reducing the complexity of the model.

To find the optimal model, we re-cast the EN estimator in Lagrangian form, as shown in equation (23). The two tuning parameters discussed above are replaced by a Lagrange multiplier, $\lambda\in\mathbb{R}_+$, and a parameter $\alpha\in[0,1]$ that controls the degree of mixing between the Ridge constraint and the LASSO constraint:

$$\max_{\beta\in\mathbb{R}^q}\sum_{l=1}^{2N^T}\overbrace{w_l^{(m)}\left(y_l\log\left(\frac{\exp(\tilde{\boldsymbol{x}}_l'\boldsymbol{\beta})}{1+\exp(\tilde{\boldsymbol{x}}_l'\boldsymbol{\beta})}\right)+(1-y_l)\log\left(\frac{1}{1+\exp(\tilde{\boldsymbol{x}}_l'\boldsymbol{\beta})}\right)\right)}^{\ell(y_l,\tilde{\boldsymbol{x}}_l;\boldsymbol{\beta})}$$
$$+\lambda\sum_{p=1}^{q}\left(\alpha|\beta_p|+(1-\alpha)\beta_p^2\right) \tag{23}$$

We obtain prediction models by implementing the EN estimator using the `glmnet` package in `R`. This particular implementation of the EN estimator takes a given value of α and finds the value of λ that delivers the lowest out-of-sample (test) deviance, which is defined as:

$$-2\sum_{f=1}^{10}\sum_{l\in f}\ell\left(y_{lf},\tilde{\boldsymbol{x}}_{lf};\hat{\boldsymbol{\beta}}_{f'}(\alpha,\hat{\lambda})\right). \tag{24}$$

In Equation (24), f indexes 10 equally-sized random partitions (folds) of the data. $\ell(\cdot)$ represents the log likelihood as defined in Equation (23). $\tilde{\boldsymbol{x}}_{lf}$ is the vector of predictors for observation l in fold f, $\hat{\boldsymbol{\beta}}_{f'}(\alpha,\hat{\lambda})$ is the parameter vector estimated using observations on all folds except for fold f, and $\hat{\lambda}$ is the test deviance-minimizing choice of λ. To obtain the best prediction model, we perform a grid search by iterating α from 0.05 to 0.95 in 0.05-unit increments and select the model with the lowest test deviance across all the values of α.

Appendix C Testing the strongly non-informative linkage assumption

Patki and Shapiro (2023) provide a two-stage least squares (TSLS) approach for estimating regression coefficients with probabilistically matched data when the strongly non-informative linkage assumption fails; i.e., when the imputed variable associated with implicate m is correlated with the imputation error for implicate m. As a weaker alternative to Equation (17), the TSLS-based approach requires the variable associated with implicate m be uncorrelated with the imputation error for any *other* implicate m':

$$Cov(\hat{s}_{ij}^{*(m)}, \eta_{ij}^{(m')}) = 0. \, \forall \, m \neq m' \tag{25}$$

We implement this approach to estimate the parameter of interest by TSLS, which we refer to as $\hat{\gamma}_{1,\texttt{TSLS}}$. Comparing columns (1) and (2) of C1, we see that $\hat{\gamma}_{1,\texttt{MI}}$ and $\hat{\gamma}_{1,\texttt{TSLS}}$ are very close, which suggests that the assumptions underpinning the linkage algorithm appear to hold in our application.

Table C1: Log establishment size effect in log wage

		HRS respondents reporting size	All HRS respondents
Parameter	Establishment-size source	(1)	(2)
$\hat{\gamma}_{1,\texttt{MI}}$	BR	0.020	0.019
		(0.004)	(0.003)
$\hat{\gamma}_{1,\texttt{TSLS}}$	BR	0.022	0.020
		(0.004)	(0.003)
N		2600	4200

Notes: This table shows the effect of log establishment size on log wages. Regression samples are restricted to observations where HRS respondents reported hourly wages or provided sufficient information to infer hourly wages from reports of total earnings and total hours. All regression models include controls for weekly hours, annual weeks, tenure, years of schooling, partnered/coupled status, nativity, gender, race, Hispanic ethnicity, age, one-digit occupation fixed effects, and one-digit industry fixed effects.