# Optimal Probabilistic Record Linkage: Best Practice for Linking Employers in Survey and Administrative Data*

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

January 2019

**Abstract**

This paper illustrates an application 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 firms is highly asymmetric. To address these difficulties, this paper uses a supervised machine learning model to probabilistically link survey respondents in the Health and Retirement Study (HRS) with employers and establishments in the Census Business Register (BR) to create a new data source which we call the CenHRS. Multiple imputation is used to propagate uncertainty from the linkage step into subsequent analyses of the linked data. The linked data reveal new evidence that survey respondents' misreporting and selective nonresponse about employer characteristics are systematically correlated with wages.

**Keywords:** Probabilistic record linkage; survey data; administrative data; multiple imputation; measurement error; nonresponse

*Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed (DRB release number 6961 and 7357). 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.

[†]U.S. Census Bureau
[‡]Labor Dynamics Institute, Cornell University
[§]Institute for Social Research, University of Michigan
[¶]Department of Economics, University of Michigan
[‖]Department of Biostatistics, 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 (e.g., households or businesses). In this paper, we seek to link across frames by matching individual respondents in household survey data to administrative data on the universe of employers. How does one go about linking a survey response 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.

The setting that we consider is subject to a striking empirical fact: 0.3 percent of all firms employ 54 percent of all workers in the United States.[1] These firms are large, each with 500 or more employees, and typically operate in multiple locations. Because of this asymmetry, the dominant fraction of individuals in household survey samples are employed at a very small subset of employers. In the absence of unique identifiers that facilitate linkage, matching individuals to employers is challenging and inherently noisy because a large number of small employers within a given geographical unit constitute feasible matches for any given individual. To address this challenge, we design, test, and implement a general methodology for linkage of survey responses to employers absent unique identifiers. Our application matches survey respondents in the Health and Retirement Study (HRS) to their employers in the Census Business Register (BR) to create a new data source which we call the CenHRS.

Our linkage procedure has three distinct steps: We begin by estimating supervised machine learning models of employer and establishment (exact business location) match probability using a rich set of covariates drawn from both the HRS and the BR. The models are tuned to deliver high out-of-sample performance thereby generalizing their value beyond the sample used for estimation. Because of the flexibility built into the covariate space, these models capture important non-linearities inferred from human judgment in the training data. The second step of the procedure explicitly accounts for link uncertainty by using the prob-

[1] See, e.g., Statistics of U.S. Businesses (SUSB), U.S. Census Bureau. Reported statistics are based on data from 2015.

ability distribution over potential matches to multiply impute BR firms and establishment links for each HRS job. By relying on multiple imputation, the method that we propose allows analysts to incorporate link uncertainty when conducting inference. Third, we implement a data-driven procedure that provides optimal cutoffs to isolate the set of BR match candidates for which the linkage algorithm proves to be too noisy. Culling the set of candidates that fail to attain these cutoffs dramatically reduces the number of viable matches thereby reducing between-implicate variability. By eliminating low probability matches, which are overwhelmingly dominated by small employers, it also mitigates biases induced by linking a household-level survey to an establishment-level frame. This procedure stands in contrast with standard, but ad hoc, procedures such as selecting the one match with the highest probability or considering cutoffs based on a pre-selected probability such as 0.5. We will show that our statistical procedure leads to different and arguably better inferences.

The plan of this article is as follows. Section 2 discusses existing record-linkage approaches. Section 3 explains why we employ probabilistic linkage in our setting and provides detail on how we train and fit a match probability model. Section 4 discusses the importance of match uncertainty, explains how we apply multiple imputation to address this uncertainty, and documents a data driven procedure that we apply to mitigate it. Section 5 compares selected employer characteristics derived from HRS self-reports to BR imputations obtained using a variety of single and multiple imputation methods. Section 6 illustrates an application of the matched data to shed new light on the incidence of nonclassical measurement error and nonresponse bias in HRS respondent's reports of employer and establishment size. Section 7 concludes.

## 2 Earlier Procedures

In deterministic file matching applications record linkage is accomplished by isolating a set of variables that are common to a given record in both files. This procedure, known as blocking, constitutes both the first and the last step in deterministic linking. 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 after blocking are retained. In some instances, a sufficiently rich set of accurately measured blocking variables can allow a large fraction of the original file to be unequivocally matched (see, e.g., Warren et al. (2002), Hammill et al. (2009), Lawson et al. (2013), and Setoguchi et al. (2014)). In other cases, the matched file consists of a small 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) and Bailey et al. (2018).

The Fellegi and Sunter (1969) (FS) algorithm is a popular probabilistic linking method that picks the best match from the set of multiple potential matches. In this method, researchers estimate the probability that a particular characteristic agrees in the two files, given that the records should link (match) and given that they should not link (nonmatch).[2] 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. Candidate pairs 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. While manual review of the entire set of blocked 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 algorithm.

The Census Bureau has a long history of using FS-style probabilistic record linkage for business data. The best known examples are the Longitudinal Business Database (Jarmin and Miranda (2002)) and its predecessor the Longitudinal Research Database (McGuckin (1990)). These efforts began by using exact identifier record linkage based on the known relationship between the Census File Number, the identifier used in the pre-2002 BR historical data, and the Federal Employer Identification Number (EIN). FS methods were used when EINs produced many-to-many linkages, as in the case of multiple establishment employers, or no linkages, as in the case of births, deaths, and changes in the legal identifiers of a business.

Unlike the FS algorithm, the Bayesian approach to record linkage incorporates match uncertainty from the linkage step into the analysis step. The Bayesian framework proceeds as follows. The first step relies on comparing variables appearing in both sets of records and estimating a match probability model conditional on those variables. The second step resamples from the posterior distribution of first step parameters to iteratively estimate parameters of interest in the analysis step, thereby propagating uncertainty from the linkage into the analysis (see, e.g., Fortini et al. (2001), Tancredi and Liseo (2011), Liseo and Tancredi (2015) and Steorts et al. (2016)). Some Bayesian applications rely on training data while others use the

[2] A drawback of the probability estimation method in FS is that predictors are assumed to be independent conditional on true match status. In the HRS-BR employer matching problem that we address, this assumption would mean, for instance, that conditional on being a true match, the probability of agreement on 3-digit zip codes was independent of the probability of agreement on 4-digit zip codes. Such an assumption would be untenable.

expectation-maximization (EM) algorithm to develop record linkage without a training dataset.

The work that is more germane to this paper's method 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 employer businesses to the job histories in the 1990-1996 Surveys of Income and Program Participation (SIPP). This work also developed improved linkages within the 1990-1993 SIPP job histories, and integrated data from the BR into the SIPP (Stinson (2003)). Abowd and Stinson (2013) evaluates this linkage and uses it to compare self-reports and administrative reports of earnings.

Using methods that are much closer to the ones we develop in this paper, the LEHD program also linked employer establishments 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 also started with exact identifier methods using the SEIN. When these methods did not resolve the linkage, a Bayesian posterior predictive distribution was used to generate ten implicates linking establishments to the candidate worker employment history (Abowd et al. (2009)). These ten implicates were used to associate workplace characteristics to each worker history. Other incomplete data in the LEHD infrastructure was completed using similar Bayesian methods. The ten implicate threads were processed according to the Rubin (1987) combining formulas to produce the Quarterly Workforce Indicators (QWI). McKinney et al. (2017) provides a complete assessment of the total variability in the QWIs due to the multiple imputation and other edit procedures.

## 3 Building a prediction model

In our setting, exact identifier or deterministic matching is not a feasible strategy as the set of matched pairs is greater than 1 for almost every HRS job. We do not adopt manual review or FS because of scalability issues and also because they fail to fully incorporate link uncertainty. Since we are interested in generalizing our procedure to subsequent waves of data collection, the importance of high-quality out-of-sample match probability prediction is paramount in our application. To accomplish this, we proceed by flexibly estimating match probabilities using a supervised machine learning algorithm.

The estimator that we adopt allows for the inclusion of potentially more predictors than observations and relies on out-of-sample error minimization to guide model selection and parameter estimation. We characterize uncertainty in the linkage not through the posterior predictive distribution of the parameters

of the matching model but through the posterior predictive distribution of the fitted pair-specific match probabilities. To account for link uncertainty, we use multiple imputation methods when making inferences from the matched dataset.[3] This section describes the data and method employed to estimate the supervised learning algorithm and shows metrics of its out-of-sample accuracy.

### 3.1 Data

The HRS 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. The BR is the Census Bureau's list of essentially all employers in the country, and it is in turn linked to other Census Bureau survey and administrative data on employers. The HRS elicits information about employer identity from respondents to construct measures of pension wealth. These data are obtained at the 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. Although the names, addresses, and phone numbers captured in these reports were originally intended to aid the HRS in contacting employers about pension benefits, they also provide us with valuable data on employer identity and location. Matching names and addresses of employers reported in the HRS with the names and addresses of establishments (individual business locations) in the BR constitutes the basis of our linking algorithm.[4]

### 3.2 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 with a unique employer. 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

[3]The posterior predictive distribution of the Elastic Net estimates of the parameters has not been fully characterized in the literature. The estimation uncertainty is not propagated through to the match uncertainty because we know of no practical way to do so.

[4]The original design of the CenHRS was predicated on having Federal employer identification numbers (EINs). EINs would provide tight, but not perfect, linkage to an employee's firm. The reliance on business name and address matching was necessitated by challenges in receiving permissions to use EINs for linkage. The Business Register includes EINs, and most HRS respondents have given permission to the Social Security Administration (SSA) to provide EINs for their employers to HRS for purposes of enhancing the HRS data infrastructure. We expect that we will evaluate the approach implemented here with a comparison to the matches achieved using EINs when access is obtained. This paper aims to link to both firm and establishment (i.e. the specific location at which they work), so even were EINs available, the method developed here would be necessary to link to establishments. We also want to have the capacity to link households that provide employer names and addresses, but may not consent to linkage with SSA administrative data.

a set of $N_{\text{BR}} \times N_{\text{HRS}}$ pairs. Because this set is of the order $10^6 \times 10^4$ the computational cost of such a prior is prohibitive.[5] To reduce the dimensionality of the search problem, we establish a blocking strategy.

Blocking reduces the size of the set of potential records in the BR that match a given record in the HRS. We block on 3-digit zip code, 10-digit phone number, telephone area code, and city-state: Any BR record that fails to share at least one of the blocking values with an HRS record is assumed to have 0 probability of being a true match. Employing this strategy substantially reduces the number of potential matches associated with each HRS job.

### 3.3 Training

The prediction task implied by probabilistic linkage in a supervised setting is to estimate a statistical model that takes data on the characteristics of a given candidate pair as an input and outputs the probability that the candidate pair is a true match. Defining a vector of pair characteristics by $\boldsymbol{X}_{ij}$ and true match status by $m_{ij}$ (i.e. $m_{ij} = 1$ if the pair is truly a match and 0 otherwise). The model provides us with an estimate of $P(m_{ij} = 1|\boldsymbol{X}_{ij})$.

To feasibly estimate this model, we need data on true match status for a sample of pairs, i.e. given a set of pairs with known characteristics $\boldsymbol{X}_{ij}$, we need to establish what $m_{ij}$ is. We construct this training dataset by drawing a stratified sample of approximately 1000 pairs (based on approximately 500 HRS jobs) from the blocked set and subjecting each pair to review by two different human experts. Details on the construction of the training dataset are documented in Appendix A. Exposing each pair to two different sets of eyes ensures that observationally equivalent cases (i.e. with the same $\boldsymbol{X}_{ij}$) can receive different evaluations about true match status (i.e. different $m_{ij}$). This data structure incorporates uncertainty about true match status into the error term of the prediction model, which is important for inference.

We consider employer and establishment matches separately. An employer match means that the employer's identity (e.g., Dunder Mifflin Paper Company) in the HRS corresponds to the employer's identity in the BR. In contrast, an establishment match implies that, in addition to an employer match, the workplace identified 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, dif-

[5] See, e.g., Statistics of U.S. Businesses (SUSB), U.S. Census Bureau. The number of establishments in the United States as of 2015 was 7.6 million, whereas the number of firms was 5.9 million.

ferent 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, in the training step, experts examine the information shown in Table 1 and separately determine whether a pair is truly an establishment match and/or an employer match.

### 3.4 Predictors

The purpose of the model is to automate the manual review process and to mimic the human judgment that underpins the training sample as closely as possible. To build this model we assemble a set of predictors that include not only variables directly observed by reviewers but also variables that capture latent institutional knowledge that reviewers may have relied on in their evaluation of candidate matches.

Table 2 shows the predictors used to estimate the employer and establishment match probability models. The first two variables are cubic splines of Jaro Winkler (JW) comparator scores for employer name and establishment address which flexibly capture reviewers' assessments of the similarity in the HRS and BR names and addresses.[6] The next two variables capture the importance of specific employers in the local (i.e. within blocking variable) and national labor market on match probability. These variables account for institutional factors such as specific knowledge about dominant employers that reviewers may have relied upon in ascertaining match status. To flexibly capture all complementarities across name and address similarity and the role of specific employers, we fully interact all four cubic splines together, expanding the set of predictors substantially.

The lower panel of Table 2 shows the set of binary predictors. These variables capture agreement between the HRS-BR candidate match on a number of dimensions. Some predictors, such as 10-digit phone agreement can be highly influential in predicting match probability, but it is rare for candidate pairs to share such granular characteristics. On the other hand, sharing SIC industry codes or 4-digit zip codes is more likely but less predictive of a match. The final two variables — employer provision of health insurance and retirement plans — incorporate information obtained purely from HRS respondents. We include these predictors because they are typically associated with large employers and serve as proxies for employer size when such information is missing in the HRS. Prior to model selection, there are a total of 1413 continous and binary predictors.

[6]The Jaro-Winkler score, which ranges from 0 to 1, measures the number of perturbations necessary to change one string into another string (Jaro (1989)). See Winkler (2006) for a recent overview.

### 3.5 Model selection

Given that our training data set consists of approximately 2000 observations, estimating a model with a high dimension of predictors is likely to yield unstable parameter estimates. To solve this dimensionality problem and, more importantly, to avoid over-fitting our model, we use machine learning tools to aid in prediction. Because we intend to apply our prediction model to the entire set of available HRS jobs and to future waves of the longitudinal survey, a key concern is to ensure that it generalizes well outside of the training data set. 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.[7]

In our setting, the complexity of the prediction model is indexed by the dimension of the covariate vector. Reducing model complexity by shrinking the number of covariates 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 developed by Zou and Hastie (2005). The EN estimator is the solution to the minimization problem posed in (1); $i$ indexes observations in the training set, while $j$ indexes regressors in the model:

$$\begin{aligned} &\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \\ &\text{st: } \sum_{j=1}^{p} \beta_j^2 \leq t_1, \ \sum_{j=1}^{p} |\beta_j| \leq t_2 \end{aligned} \tag{1}$$

In (1), the typical least squares minimization criteria 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 maximum 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 particularly important when some predictors are correlated. $t_2$, as in the LASSO, sets a maximum 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 shrinking the dimensionality of the model. The optimal prediction

[7] We use 10-fold cross validation to obtain test error estimates.

model is chosen by finding the pair of tuning parameters that jointly minimize out of sample error.[8]

The EN estimator can also be conveniently summarized in Lagrangian form as shown in equation (2). 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:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \left(\alpha|\beta_j| + (1-\alpha)\beta_j^2\right) \tag{2}$$

We obtain our establishment and employer prediction models by implementing the EN estimator with the logistic link function in MATLAB using the `lassoglm` function. This particular implementation of the EN estimator takes a given value of $\alpha$ and finds the value of $\lambda$ that delivers the lowest out-of-sample deviance. To obtain the best prediction model, we perform a grid search by iterating $\alpha$ from 0.05 to 0.95 in 0.05-unit increments. For each value of $\alpha$, we obtain the model associated with the lowest test deviance estimate. The optimal model is the one with the lowest test deviance across all the values of $\alpha$.

### 3.6 Model fit

Figure 1 shows the partial effects of the employer and the establishment matching models. The top row of the figure shows the partial derivative of the employer prediction model for four selected covariates holding all other variables at their means. The bottom row shows analogous effects for the establishment prediction model. For both models, similarity between HRS and BR names and addresses deliver the largest effect on the likelihood that a pair is a true match. This effect only manifests at very high levels of similarity and does so in a highly non-linear fashion. The partial effects in the employer model are uniformly higher than those in the establishment model, reflecting that employer matches are easier to confirm than establishment matches.

The partial effects of the two models underscore the value of employing the EN estimator and relying on cubic splines to model the covariate space. The models we estimate capture sharp inflection points in the curvature of the match likelihood, reflecting non-linearities in reviewer decisions that would be infeasible to replicate using a simpler parametric approach.

[8] Note that the shrinkage imposed by both the LASSO and Ridge penalty terms is not invariant to the scale of the regressors. Standard practice is to studentize the regressors before the model is estimated. The constant and re-scaling factors are estimated after model selection is complete.

### 3.6.1 Out-of-sample fit evaluated using cross-validated ROC curves

Although we do not use our models as binary classifiers, we illustrate their predictive performance by showing receiver operating characteristics (ROC) curves in Figure 2. For probability thresholds ranging from 0 to 1, the ROC curve plots the true positive rate (sensitivity) on the vertical axis against the false positive rate (1-specificity) 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. As such, the c-statistic captures the predictive ability of the model in a single number which aids in evaluating in the quality of the model relative to the two extremes of 0.5 and 1.0.

The left panel of Figure 2 compares employer match prediction performance using ROC metrics. The blue curve is based on the EN estimator while the red curve is based on a traditional logistic regression model estimated using JW scores of name and address. Applying tuning parameters from the optimal model, we estimate model coefficients using 9/10ths of the training data and compute sensitivity and specificity estimates by projecting the model on the remaining 1/10th of the sample. Iterating through each hold-out tenth yields the out-of-sample ROC estimate. The c-statistics from the two models are 0.98 and 0.94 respectively. The right panel shows the same contrast for the establishment prediction model; the c-statistics for these two models are 0.94 and 0.88 respectively. Employer matches are easier to ascertain than establishment matches. Employer matches depend mainly on address while establishment matches depends importantly on address in addition to name, so establishments are unconditionally less likely to be found relative to employers.

Table 3 compares the relative precision of each model at pre-selected sensitivity levels of 0.85, 0.90, 0.95 and 0.99. For the employer match model, EN attains false positive rates (1-specificity) which are 2-to-5 times lower than the corresponding values attained by traditional logistic regression. At the same sensitivity levels for the establishment match model, EN attains false positive rates which are 1.5-to-3 times lower than the corresponding values attained by traditional logistic regression. Taken together, the results shown in Figure 2 and Table 3 indicate that the EN prediction models deliver very high predictive ability out-of-sample, outperforming the simpler logit models in both cases.

### 3.6.2 Out-of-sample fit evaluated using 2010 HRS hold-out sample

We exploit the availability of EINs for respondents from the 2010 wave of the HRS as a secondary check of the fit of our employer prediction model that is entirely independent of the training data set.[9] These EINs were obtained from Form 5500 (F5500) pension plan data through the HRS pension project which seeks to match respondents to their employer's pension plans. While some pension EINs represent the respondent's current employer, others could represent EINs for union-sponsored pensions that do not have corresponding entries in the BR. In these instances, the validation exercise we propose here understates the accuracy of our matching algorithm.

To evaluate our model using the EINs, we first fit the employer match model on blocked candidates for HRS respondents from 2010. Then, foreshadowing the multiple imputation procedure that we introduce in the next section, we draw 10 implicates with replacement from the posterior predicitive distribution of candidate matches. Next, we match the F5500 EIN to the BR to determine the employer associated with the F5500 EIN.[10] Finally, for each HRS job, we ascertain whether the employer identified by the F5500 link appears among the set of 10 implicates. Table 4 shows that the average of this concordance over approximately 1900 jobs with linked F5500 EINs in the 2010 HRS is 0.42. When we impose data driven cutoffs to filter away low quality matches (discussed in section 4.1), the concordance rate rises to 0.63.

## 4 Multiple imputation of links

Having illustrated the methodology and precision of our prediction model, we now turn to the issue of how to use match probabilities to construct links between the HRS and BR. In our discussion, we emphasize the idea that variables obtained via the linkage procedure are best thought of as imputed values rather than true values. This distinction underscores the role of uncertainty in the linkage procedure.

Conventional linking methods, including FS and manual review, singly impute linkages using the most likely match. The central concern with single imputation is that it fails to properly account for uncertainty in the linkage. This failure can bias confidence interval estimates and lead to invalid inference.

In contrast to single imputation methods, we rely on multiple imputation (MI) to link records in the HRS

[9] Pension sponsors are legally required to report information about their plans on Internal Revenue Service Form 5500. EINs obtained through the HRS pension project were assigned by clerical review that matched names and addresses reported by HRS respondents to names and address of pension plan sponsors listed on Form 5500.

[10] Employer identity is ascertained on the basis of a variable known as the Census firm identifier. All establishments associated with a particular employer have the same Census firm identifier even if they have different EINs.

and the BR. MI potentially provides a way to account for linkage uncertainty, but it requires a fully specified posterior predictive model, which some methods (e.g., FS) do not provide. The MI solution involves repeatedly estimating a parameter of interest using different draws from the posterior predictive distribution of the imputed variable. This procedure propagates randomness in the imputed value into the estimand of interest, thereby yielding valid inferences (e.g., Rubin (1977), Rubin and Schenker (1986), and Rubin (1987)). In complementary work in a regression context, potential matches are aggregated by using match probability estimates as weights as in Lahiri and Larsen (2005).

To implement our MI procedure, we obtain the posterior predictive distribution of potential matches by normalizing the estimated match probabilities to sum to one for each HRS job. We do this separately for employer matches and establishment matches. Next, instead of selecting the single best employer or establishment, we draw a sample of $M = 10$ matches with replacement using the normalized match probabilities as sampling probabilities. This procedure yields $M$ multiply imputed establishment and employer links for each HRS job. Together, these links constitute $M$ completed data sets. For any statistic generated using imputed data, we can use all $M$ completed data sets along with the formulae in Rubin and Schenker (1986) to compute the variance owing to sampling noise (within-implicate variability) and the variance due to linkage uncertainty (between-implicate variability).

For some scalar parameter $\theta$, 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 $\theta$ is

$$\hat{\theta} = M^{-1} \sum_{m=1}^{M} \hat{\theta}_m \tag{3}$$

The within-implicate variance is

$$\hat{\sigma^2_W} = M^{-1} \sum_{m=1}^{M} \hat{\sigma^2_m} \tag{4}$$

The between-implicate variance is

$$\hat{\sigma^2_B} = (M-1)^{-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}\right)^2 \tag{5}$$

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

$$\hat{\sigma^2} = \hat{\sigma^2_W} + (1 + M^{-1})\hat{\sigma^2_B} \tag{6}$$

The ratio of between-to-total variance is also known as the missingness ratio and summarizes the extent to which a given parameter estimate is influenced by between-implicate uncertainty. Greater model accuracy in matching an HRS job to a BR candidate translates into a lower missingness ratio.

### 4.1 A data driven procedure to reduce match uncertainty

While the blocking algorithm we describe above dramatically shrinks the number of potential BR candidates, the number of candidate matches is still very large for many HRS jobs. This empirical regularity is a consequence of seeking to match household-level survey data to an establishment-level frame. Most individuals are employed at a relatively small number of large employers whereas the vast majority of employers are, in fact, very small. As a consequence, matching blocks are populated with many small employers each of which receive a trivial match probability, a problem that is made more acute when EINs cannot be employed for blocking. To illustrate the impact of block size on match uncertainty, consider an example where an HRS job is blocked with 1000 BR candidate matches with one large employer candidate being the correct match and 999 small employer candidates being non-matches. Suppose the large employer candidate obtains a match probability of 0.5 while the remaining 999 small employer candidates receive match probability $= \frac{0.5}{999}$. In this instance, random small employers will populate half of the implicates even though they are two orders of magnitude less likely to be the right match relative to the large employer. To address this concern, we propose a data driven procedure that mitigates the impact of block-size induced noise in the linkage process.

The procedure we adopt is an intuitive combination of binary classification and multiple imputation. For each HRS job, the procedure entirely eliminates a set of BR candidates whose estimated match probability falls below a minimum threshold. As such, we refine the posterior predictive distribution of potential matches by concentrating the remaining mass on a smaller set of candidates. Sampling from this, more concentrated, distribution lowers between-implicate variability.

The details of the procedure are as follows. First, we estimate ROC curves using the training data as

shown in section 3.6. We then pick the probability threshold $p^*$ that minimizes the following criterion

$$D(p) = \Big((1 - \text{sensitivity})^2 + (1 - \text{specificity})^2\Big)^{1/2}. \tag{7}$$

The cutoff probability $p^*$ minimizes the distance between the ROC curve and the upper left corner of the graph. Put differently, $p^*$ is the feasible cutoff closest to the infeasible point where the sensitivity and specificity are both 1 (see, e.g., Coffin and Sukhatme (1997) and Youden (1950)). While the criterion we use places equal weight on sensitivity and specificity, this choice is arbitrary and can be modified depending on the objective of the analyst. We estimate these cutoffs separately for each quartile of the block size distribution in the training data. This stratification allows candidate pairs for HRS jobs associated with larger blocks to have lower thresholds and vice versa. Table 5 shows the average number of pairs in each quartile of the block-size distribution along with estimates of the associated cutoff probabilities.

Having obtained block-size dependent cutoffs, we discard any candidate pairs whose estimated match probability is below the cutoff. Finally, we re-normalize the match probabilities to sum to one over the set of surviving candidates and draw multiple implicates from this set. For 33 percent of HRS jobs, exactly 0 BR candidates survive the employer match threshold. For the establishment match model, 8 percent of HRS records have 0 BR candidates. These HRS jobs represent cases where there is not enough information to produce a plausible match candidate from within the blocked set. Nevertheless, to the extent that exclusion from the imposition of cutoffs is non-random, it can generate selection bias. We investigate this issue in Appendix B. The following subsections illustrate the extent of match uncertainty before and after imposing thresholds.

## 4.2 Concentration as a measure of uncertainty

Table 6 illustrates the degree of concentration among implicates obtained in the 2010 wave of the HRS. Concentration among the implicates is defined as the proportion of unique matches among the 10 multiply imputed matches for each HRS job. The left panel shows employer-match concentration whereas the right panel shows establishment-match concentration; for both models, we show concentration rates without cut-offs and with cutoffs. The first row of the table shows the fraction of HRS jobs for which a single BR record populated all 10 implicates, which is the maximum level of concentration. Subsequent rows show the share of implicates associated with successively higher numbers of unique BR entities. Cases with 5 or more

unique matches are binned together.

The table shows that the imposition of minimum match probability cutoffs increases the concentration of BR entities across the implicates; that is, less disparate BR entities are realized as potential links after imposing the cutoff. For both models, the fraction of HRS jobs mapped to a single BR candidate rises by a factor of approximately 50 after cutoffs are used to exclude low quality matches. Concentration increases in the range of 10-to-30 fold are seen for jobs mapped to two or three unique BR entitites. Because employer matches are easier to ascertain, the level of concentration is higher overall as compared with establishment matches.

### 4.3 Concordance in employer identity as a measure of uncertainty

Our models predict employer and establishment match status independently. We use the model outputs to compare the concordance between employer identity from the employer match model with employer identity from the establishment match model.[11] A high degree of overlap in this dimension indicates that both models select the same set of employers. Greater agreement between the two models is therefore not only an internal consistency check, but also a measure of match certainty.

Table 7 shows the average fraction (out of 10) of employer identities that are common to both the employer and establishment predictions for each HRS job. These rates are reported for each quartile of the block-size distribution, and for the sample as a whole. Prior to imposing probability cutoffs, smaller block sizes yield greater concordance between the models: the average rate in the bottom quartile is 14 percent whereas the average rate in the top quartile is almost halved to 6.6 percent. This decline occurs because uncertainty grows in the number of potential matches. Once cutoffs are imposed, these concordance rates increase by 4-to-8 fold. Furthermore, the monotonic decline in match uncertainty vanishes in the cutoff based sample. Taken together with the concentration improvements highlighted earlier, these statistics show that the application of a simple filtering technique can dramatically reduce between-implicate variability among multiply imputed matches.

The less than perfect concordance reflects intrinsic uncertainty in record linkage. 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.

[11] As noted earlier, employer identity is ascertained on the basis of Census firm identifiers. All establishments associated with a particular employer have the same Census firm identifier even if they have different EINs.

# 5 Comparing self-reported and imputed employer characteristics

In Table 8 we show moments of the employer and establishment size distribution using the linkage strategy outlined in Section 4. We consider five different methods of measuring employer and establishment size which we illustrate by reporting averages within selected percentiles of employer and establishment size distribution. The top panel of the table shows employer size statistics while the bottom panel shows establishment size statistics. The percentiles of the size distribution are re-computed separately for each row. The table hence illustrates what a researcher would infer about the size characteristics of HRS respondents using either the HRS self-report or the population of linked employers or establishments.

For each panel, the first row is the HRS self-reports and the next four rows use different imputation methods for linkage to the BR:

1. SI-random draws a single implicate at random from the set of 10 multiply imputed candidate matches reflecting the most naive imputation method.

2. SI-best selects the implicate with the highest predicted match probability; this procedure is similar to FS methods where subjective judgment is used to reconcile the presence of multiple potential matches. The added benefit of SI-best over purely subjective match selection is that match probability estimates provide a well-defined metric to select between alternative candidates.

3. MI-conventional draws 10 implicates with replacement using estimated match probability as the sampling probabilities.

4. MI-optimal first imposes data driven optimal cutoff probabilities on the estimated posterior predictive distribution, eliminates cases below the cutoff, and then re-samples with replacement from the re-normalized posterior distribution.

For the MI-based statistics we show, in addition to means and standard errors, the fraction of variance that owes to between-implicate uncertainty (the missingness ratio). The missingness ratio captures variability that is relevant for inference that is ignored using SI-based methods. The standard errors and missingness ratio for MI-based estimates are computed using the formulae in section 4.

Comparing moments of the employer and establishment statistics using each of these procedures facilitates comparison with other linkage applications and highlights the value of our preferred method. Employer and establishment size as reported by HRS respondents is consistently larger than SI-random and

MI-conventional imputation from the BR.[12] In contrast, the SI-best and MI-optimal based estimates of employer and establishment size are larger than SI-random and MI-conventional estimates. Reflecting the approximately log-normal distribution of firm size, most matching blocks contain many small firms. In contrast, the dominant fraction of workers are employed at large firms. This dichotomy has the potential to generate non-trivial bias especially because our task is to match household survey data to an establishment level frame. Both the SI-random and MI-conventional procedures over-represent small firms because they draw from a set of candidates where the count of small firms far exceeds large firms. Because self-reports of establishment size are likely to be more reliable than employer size (individuals know how many workers are at their workplace more readily than how many workers a firm employs across workplaces), we use the lower panel of the table to establish the bias reduction gains of SI-best and MI-optimal imputation strategies. Across much of the establishment size distribution, these two measures more accurately correspond to self-reports while SI-random and MI-conventional are consistently downward biased. Improvement in imputation accuracy obtains because SI-best and MI-optimal select larger employers with greater probability and, therefore, produce estimates that are closer to HRS self-reports.[13]

While both SI-best and MI-optimal mitigate bias, only MI-optimal incorporates linkage uncertainty into the standard error estimate. For the middle tenth of the employer and establishment size distribution, about 3.5 percent of the variance in average size owes to linkage uncertainty respectively. This variability is ignored in the SI-best procedure thereby downward biasing the associated standard errors. Secondary to the improvement of MI-optimal over SI-best, we see that the cutoffs based procedure generally reduces between variability relative to conventional MI, reinforcing earlier measures of concentration and concordance among the 10 implicates.

## 6 Application: The wage-firm size gradient

Using both household and firm 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

[12] Employer size in the HRS is first elicited as a continuous variable. If respondents do not report a number, they are given the option of reporting one of six bins: [1,4], [5,14], [15,24], [25,99], [100,499], and 500+. In Table 8, we convert binned reports of employer and establishment size to continuous values by using the midpoint of the interval. For respondents who report "500+", we impute a continuous value by randomly drawing an employer size from the set of continuous valued reports that are above 500.

[13] As we show in the next section, HRS self-reports of employer and establishment size are downward biased due to nonclassical measurement error and nonresponse bias.

we illustrate an application of CenHRS by re-examining the relationship between wages and employer size. In particular, our approach reveals how systematic biases generated by measurement error and nonresponse would remain hidden without establishing linkages to administrative data.

We begin by showing non-parametric evidence of the positive wage-size gradient in our sample of working HRS respondents in the 2010 wave. The left panel of Figure 3 shows average log hourly wages for each decile of the log employer-size distribution, the right panel shows average log hourly wages for each decile of the log establishment-size distribution. In both panels the gradient based on self-reported (HRS) size is steeper than the gradient based on multiply imputed size obtained from linkages to administrative data (MI-BR).[14] If employer and establishment size were subject to classical measurement error — as is often the case in self-reports of earnings — one would expect the survey based gradient to be attenuated relative to the administrative data based gradient. However, the converse is true.

Figure 4 explains the amplification bias by revealing nonclassical measurement error and nonresponse bias in HRS self-reports of employer characteristics. The top left panel shows average log employer size for each decile of the log wage distribution and illustrates a stark pattern: workers in lower deciles of the wage distribution underreport the size of their employer. This error diminishes as wages increase but does not vanish even at the top of the wage distribution. As such, self-reporting error about employer size is positively correlated with wages thereby generating amplification bias in the survey-based wage-size relationship. The lower left panel of Figure 4 shows the same qualitative relationship between self-reported error in establishment size and wages. Relative to employer-size discrepancies, the magnitude of these errors are substantially smaller and the confidence interval estimates for the two curves overlap across the entire wage distribution. Biases such as these could occur if low-wage workers are less informed about the employment structure of a firm than are say, high-wage, managerial workers who have more institutional knowledge of the firm's operations. They could also emerge if low-wage workers at multi-establishment firms tend to report establishment size as a proxy for employer size more frequently than do high-wage workers.

To illustrate the role of nonresponse bias, the two right-hand panels of Figure 4 show similar contrasts but restrict the sample of administrative data imputations to coincide with the sample where HRS respondents provide self-reports. This restriction eliminates nonresponse bias as a reason for the difference between self-reports and administrative data imputation by focusing purely on reporting error. The similarity between the

[14] In this and subsequent statistics, the MI-BR values are obtained after imposing cutoffs.

plots in the left half of the figure and those in the right half indicate that reporting error is, in fact, the main driver of amplification bias.

To examine these differences more carefully, Table 9 shows measurement error and nonresponse bias within each decile of the wage distribution.[15] Measurement error is consistently negative and declining in wages for employer and establishment size, whereas nonresponse errors are typically positive and are largest in the 3rd, 4th, and 5th deciles of the log wage distribution. With a few exceptions in the tails of the wage distribution, these data reveal that nonresponders in the HRS predominantly work at larger firms and establishments than do responders. Furthermore, because selective nonresponse is concentrated in lower deciles of the log wage distribution, relying purely on self-reports would make it appear that lower-wage workers are employed at smaller firms and establishments than is actually the case. This bias further amplifies the survey-based wage-size gradient.

The patterns discussed here provide new evidence on how survey responses about employer characteristics are selectively misreported or not reported at all. With linkages to administrative information on employers in CenHRS, we are able to characterize measurement and nonresponse errors that are unobservable in other household survey datasets.

# 7 Conclusion

This paper describes the construction of a new dataset, 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 asymmetries in the distribution of employment across firms that makes matching inherently noisy. To address these issues, we resort to probabilistic linkage and utilize a supervised machine learning model to estimate the probability that specific employers and establishments in the BR are matches for individuals in the HRS. Our prediction model relies on a rich set of covariates and a high degree of flexibility to replicate important non-linearities inherent in the training data. Using probabilities estimated from the model, we employ MI to characterize uncertainty in the linkage. To further refine the posterior distribution of candidate matches we estimate probability cutoffs that provide the best sensitivity and specificity combination out-of-sample. Eliminating candidate matches that fail to meet these cutoffs dramatically reduces between-implicate variability while also reducing biases inherent in the

[15] Appendix C formalizes how measurement error and nonresponse bias terms are computed using moments from HRS and MI-BR data.

between-frame linkage that we construct. We use these newly linked data to provide new evidence that reporting errors as well as selective nonresponse to survey questions on employer characteristics vary systematically with wages.

Beyond issues related to record linkage, 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. To validate and extend the linkages that we have developed in this paper we will exploit the availability of EINs in subsequent efforts, substantially improving the quality of these data for future research.

## References


ABOWD, J. M., B. E. STEPHENS, L. VILHUBER, F. ANDERSON, K. L. MCKINNEY, M. ROEMER, AND S. WOODCOCK (2009): "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators," in *Producer Dynamics: New Evidence from Micro Data*, ed. by T. Dunne, J. B. Jensen, and M. J. Roberts, University of Chicago Press, 149–230.

ABOWD, J. M. AND M. H. STINSON (2013): "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data," *Review of Economics and Statistics*, 95, 1451–1467.

BAILEY, M., C. COLE, M. HENDERSON, AND C. MASSEY (2017): "How Well Do Automated Methods Perform in Historical Samples? Evidence From New Ground Truth," NBER Working Paper No. 24019.

BAILEY, M., C. COLE, AND C. MASSEY (2018): "Simple Strategies for Improving Inference With Linked Data: A Case Study of the 1850-1930 IPUMS Linked Representative Historical Samples," Unpublished.

BELIN, T. R. AND D. B. RUBIN (1995): "A Method for Calibrating False-Match Rates in Record Linkage," *Journal of the American Statistical Association*, 90, 694–707.

BLOOM, N., F. GUVENEN, B. S. SMITH, J. SONG, AND T. VON WACHTER (2018): "Inequality and the Disappearing Large Firm Wage Premium," *American Economic Association Papers and Proceedings*, 108, 317–322.

BROWN, C. AND J. MEDOFF (1989): "The Employer Size-Wage Effect," *Journal of Political Economy*, 97, 1027–1059.

COFFIN, M. AND S. SUKHATME (1997): "Receiver Operating Characteristic Studies and Measurement Errors," *Biometrics*, 53, 823–837.

FELLEGI, I. P. AND A. B. SUNTER (1969): "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

FERRIE, J. 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, 141–156.

FORTINI, M., B. LISEO, A. NUCCITELLI, AND M. SCANU (2001): "On Bayesian Record Linkage," *Research in Official Statistics*, 4, 185–198.

HAMMILL, B. G., A. F. HERNANDEZ, E. D. PETERSON, G. C. FONAROW, K. A. SCHULMAN, AND L. H. CURTIS (2009): "Linking Inpatient Clinical Registry Data to Medicare Claims Data Using Indirect Identifiers," *American Heart Journal*, 157, 995–1000.

JARMIN, R. S. AND J. MIRANDA (2002): "The Longitudinal Business Database," Center for Economic Studies, U.S. Census Bureau, Working Paper 02-17.

JARO, M. A. (1989): "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

LAHIRI, P. AND M. D. LARSEN (2005): "Regression Analysis With Linked Data," *Journal of the American Statistical Association*, 100, 222–230.

LAWSON, E. H., C. Y. KO, R. LOUIE, L. HAN, M. RAPP, AND D. S. ZIGMOND (2013): "Linkage of a Clinical Surgical Registry with Medicare Inpatient Claims Data Using Indirect Identifiers," *Surgery*, 153, 423–430.

LISEO, B. AND A. TANCREDI (2015): "Regression Analysis With Linked Data: Problems and Possible Solutions," *Statistica*, LXXV, 19–34.

MCGUCKIN, R. H. (1990): "Longitudinal Economic Data At the Census Bureau," Center for Economic Studies, U.S. Census Bureau, Working Paper 90-1.

MCKINNEY, K., A. GREEN, L. VILHUBER, AND J. ABOWD (2017): "Total Error and Variability Measures with Integrated Disclosure Limitation for Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in On The Map," Center for Economic Studies, U.S. Census Bureau, Working Paper 17-71.

OI, W. Y. AND T. L. IDSON (1999): "Firm Size and Wages," in *Handbook of Labor Economics 3B*, ed. by O. C. Ashenfelter and D. Card, North Holland, 2165–2214.

RUBIN, D. B. (1977): "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.

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

RUBIN, D. B. AND N. SCHENKER (1986): "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.

SETOGUCHI, S., Y. ZHU, J. J. JALBERT, L. A. WILLIAMS, AND C.-Y. CHEN (2014): "Validity of Deterministic Record Linkage Using Multiple Indirect Personal Identifiers," *Circulation: Cardiovascular Quality and Outcomes*, 7, 475–480.

STEORTS, R. C., R. HALL, AND S. E. FEINBERG (2016): "A Bayesian Approach to Graphical Record Linkage and De-duplication," *Journal of the American Statistical Association*, 111, 1660–1672.

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

TANCREDI, A. AND B. LISEO (2011): "A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems," *Annals of Applied Statistics*, 5, 1553–1585.

WARREN, J. L., C. N. KLABUNDE, D. SCHRAG, P. B. BACH, AND G. F. RILEY (2002): "Overview of the SEER-Medicare Data: Content, Research Applications, and Generalizability to the United States Elderly Population," *Medical Care*, 40, IV3–IV18.

WINKLER, W. E. (2006): "Overview of Record Linkage and Current Research Directions," Statistical Research Division, U.S. Census Bureau, Working Paper 2006-2.

YOUDEN, W. J. (1950): "Index for Rating Diagnostic Tests," *Cancer*, 3, 32–35.

ZOU, H. AND T. HASTIE (2005): "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society Series B*, 67, 301–320.

**Table 1:** Reviewer's information set

| Category | Review variables (HRS and BR) |
|---|---|
| 1 | Employer name, establishment address, and phone number |
| 2 | Employer single or multi-unit status |
| 3 | Employer and establishment size |
| 4 | Employer industry code and code description |

**Table 2:** Key predictors in the matching models

| Predictor | Description |
|---|---|
| Cubic spline JW score name | Similarity in HRS and BR name |
| Cubic spline JW score address | Similarity in HRS and BR address |
| Cubic spline block share | Importance of establishment in local (within blocking variable) labor market |
| Cubic spline employer size (BR) | Importance of employer in national labor market (employer size from BR) |
| Full interaction of cubic splines | All complementarities between continuous variables |
| Employer size agreement | 1 if employer size in HRS and BR agree, 0 if missing or disagree |
| Establishment size agreement | 1 if establishment size in HRS and BR agree, 0 if missing or disagree |
| Multi-unit status in BR | 1 if employer is multi-unit, 0 if single-unit |
| 7-digit phone number agreement | 1 if 7-digit phone number in HRS and BR agree, 0 if missing or disagree |
| 10-digit phone number agreement | 1 if 10-digit phone number in HRS and BR agree, 0 if missing or disagree |
| 3-digit zip code agreement | 1 if 3-digit zip code in HRS and BR agree, 0 if missing or disagree |
| 4-digit zip code agreement | 1 if 4-digit zip code in HRS and BR agree, 0 if missing or disagree |
| 5-digit zip code agreement | 1 if 5-digit zip code in HRS and BR agree, 0 if missing or disagree |
| SIC industry code agreement | 1 if SIC code in HRS and BR agree, 0 if missing or disagree |
| Employer provides health insurance | 1 if HRS respondent indicates employer provides health insurance, 0 if missing or no provision |
| Employer provides retirement plan | 1 if HRS respondent indicates employer provides retirement plan, 0 if missing or no provision |

**Notes:** The cubic splines for JW scores for name and address have 10 cut points each. The cubic splines for block share and employer size have 3 cut points each. There are a total of 1413 predictors prior to model selection.

**Table 3:** Prediction accuracy: Elastic Net versus Logistic Regression

| | 1-Specificity | | | |
|---|---|---|---|---|
| | Employer match | | Establishment match | |
| Sensitivity | Elastic Net | Logit | Elastic Net | Logit |
| 0.85 | 0.026 | 0.057 | 0.080 | 0.266 |
| 0.90 | 0.034 | 0.136 | 0.145 | 0.361 |
| 0.95 | 0.068 | 0.354 | 0.309 | 0.697 |
| 0.99 | 0.311 | 0.634 | 0.605 | 0.887 |

**Notes:** Sensitivity and specificity estimates are computed using 10-fold cross-validation.

**Table 4:** Validation with pension EINs

| | No cutoffs | Cutoffs |
|---|---|---|
| Employer ID agreement | 0.421 | 0.625 |
| $N$ | 1900 | 1250 |

**Notes:** Pension EINs are obtained through an independent linkage exercise where the HRS used employer names to search for IRS Form 5500 pension filings.

**Table 5:** Receiver Operating Characteristics Curve based cutoff estimates

| | | Cutoffs | |
|---|---|---|---|
| Quartiles of block size | Avg. block size | Employer match | Establishment match |
| 1 | 13620 | 0.154 | 0.066 |
| 2 | 26390 | 0.343 | 0.154 |
| 3 | 44340 | 0.600 | 0.171 |
| 4 | 99270 | 0.534 | 0.123 |
| Full sample | 46220 | 0.236 | 0.157 |

**Notes:** ROC estimates are computed using 10-fold cross-validation. Cutoff estimates provide probability thresholds which minimizes the distance to the top left corner of the ROC graph (i.e. yield maximum sensitivity and specificity). The training data set has $\approx$ 2000 observations.

**Table 6:** Concentration of multiple implicates

| | Employer match | | Establishment match | |
|---|---|---|---|---|
| Unique matches | No cutoffs | Cutoffs | No cutoffs | Cutoffs |
| 1 | 0.009 | 0.478 | 0.001 | 0.057 |
| 2 | 0.015 | 0.283 | 0.002 | 0.058 |
| 3 | 0.020 | 0.158 | 0.002 | 0.051 |
| 4 | 0.031 | 0.055 | 0.002 | 0.060 |
| 5-10 | 0.926 | 0.027 | 0.993 | 0.774 |
| $N$ | 5700 | 3700 | 5700 | 5200 |

**Notes:** This table is based on the set of working HRS respondents in the 2010 wave who provided names and addresses of their employers. Totals may not sum to 1 because each cell is independently rounded. HRS jobs with 5 or more matches are binned together to prevent disclosure of information in small cells.

**Table 7:** Concordance between employer and establishment models

| Quartiles of block size | No cutoffs | Cutoffs |
|---|---|---|
| 1 | 0.140 | 0.547 |
| 2 | 0.103 | 0.694 |
| 3 | 0.066 | 0.718 |
| 4 | 0.066 | 0.522 |
| Full sample | 0.110 | 0.586 |

**Notes:** This table is based on the set of working HRS respondents in the 2010 wave who provided names and addresses of their employers. Block size is defined here as the number of candidate BR pairs within a block defined on an HRS job (i.e. 3-digit zip, 10-digit phone number, telephone area code, or city-state).

**Table 8:** Employer and establishment size statistics

| | A: Employer size percentile | | | | |
|---|---|---|---|---|---|
| Source | [0,25) | [24,45) | [45,55) | [55,75) | [75,100] |
| HRS | 17.6 | 177.8 | 710.5 | 2945 | 164800 |
| | (.51) | (3.56) | (10.92) | (50.89) | (13040) |
| SI-random | 0.8 | 7.5 | 73.0 | 4272 | 138800 |
| | (.02) | (.14) | (2.11) | (129.90) | (6192) |
| SI-best | 2.4 | 124.6 | 917.1 | 6668 | 190100 |
| | (.08) | (3.47) | (15.63) | (129.10) | (8270) |
| MI-conventional | 0.8 | 7.1 | 61.4 | 3537 | 126700 |
| | (.02) | (.13) | (1.79) | (107.80) | (5643) |
| | [0.001] | [0.007] | [0.029] | [0.008] | [0.001] |
| MI-optimal | 63.4 | 804.6 | 2673 | 8474 | 247300 |
| | (2.27) | (15.97) | (31.74) | (156.20) | (12210) |
| | [0.002] | [0.008] | [0.033] | [0.003] | [0.000] |
| | **B: Establishment size percentile** | | | | |
| Source | [0,25) | [24,45) | [45,55) | [55,75) | [75,100] |
| HRS | 5.6 | 25.4 | 54.5 | 127.3 | 1848 |
| | (.11) | (.29) | (.32) | (1.70) | (229) |
| SI-random | 0.5 | 2.4 | 4.5 | 12.11 | 3478 |
| | (.01) | (.02) | (.02) | (.12) | (1379) |
| SI-best | 0.8 | 5.0 | 13.5 | 58.5 | 14910 |
| | (.02) | (.06) | (.14) | (1.02) | (2613) |
| MI-conventional | 0.5 | 2.4 | 4.1 | 8.6 | 2158 |
| | (.01) | (.02) | (.02) | (.09) | (983) |
| | [0.001] | [0.067] | [0.472] | [0.023] | [0.001] |
| MI-optimal | 1.0 | 7.8 | 26.45 | 120.4 | 42020 |
| | (.02) | (.12) | (.29) | (1.97) | (4880) |
| | [0.022] | [0.030] | [0.034] | [0.008] | [0.001] |

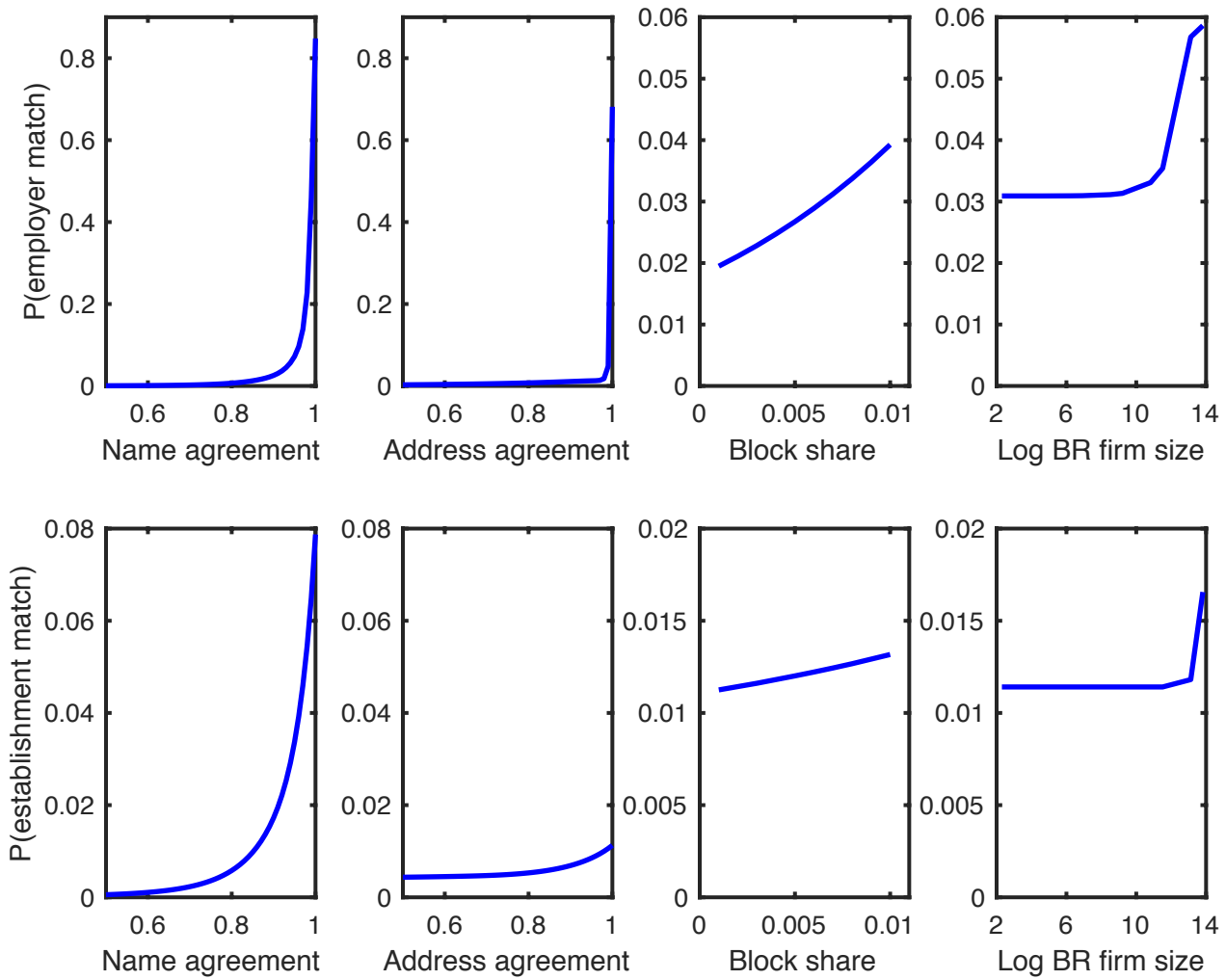
**Notes:** This table is based on the set of working HRS respondents in the 2010 wave who provided names and addresses of their employers. SI-random selects one match at random from the set of 10 multiply imputed matches. SI-best selects the BR candidate associated with the highest predicted match probability. MI-conventional uses standard multiple imputation with 10 implicates. MI-optimal imposes optimal cutoff probabilities before employing conventional multiple imputation. Standard errors are shown in parentheses. MI standard errors incorporate within and between variablility. Missingness ratios (ratio of between-implicate variance to total variance) are shown in square brackets.

**Table 9:** Measurement error and nonresponse bias across the wage distribution

| | Log employer size | | Log establishment size | |
|---|---|---|---|---|
| Log wage decile | Measurement error | Nonresponse bias | Measurement error | Nonresponse bias |
| 1 | -2.308 | -0.094 | -0.258 | 0.038 |
| 2 | -1.798 | 0.091 | -0.349 | 0.055 |
| 3 | -1.264 | 0.261 | -0.123 | 0.068 |
| 4 | -1.084 | 0.339 | -0.370 | 0.132 |
| 5 | -1.334 | 0.217 | -0.376 | 0.124 |
| 6 | -1.192 | 0.164 | -0.354 | -0.082 |
| 7 | -0.701 | 0.116 | -0.249 | -0.047 |
| 8 | -0.374 | 0.044 | -0.144 | 0.023 |
| 9 | -1.038 | -0.084 | -0.166 | 0.005 |
| 10 | -1.028 | -0.085 | 0.060 | -0.057 |

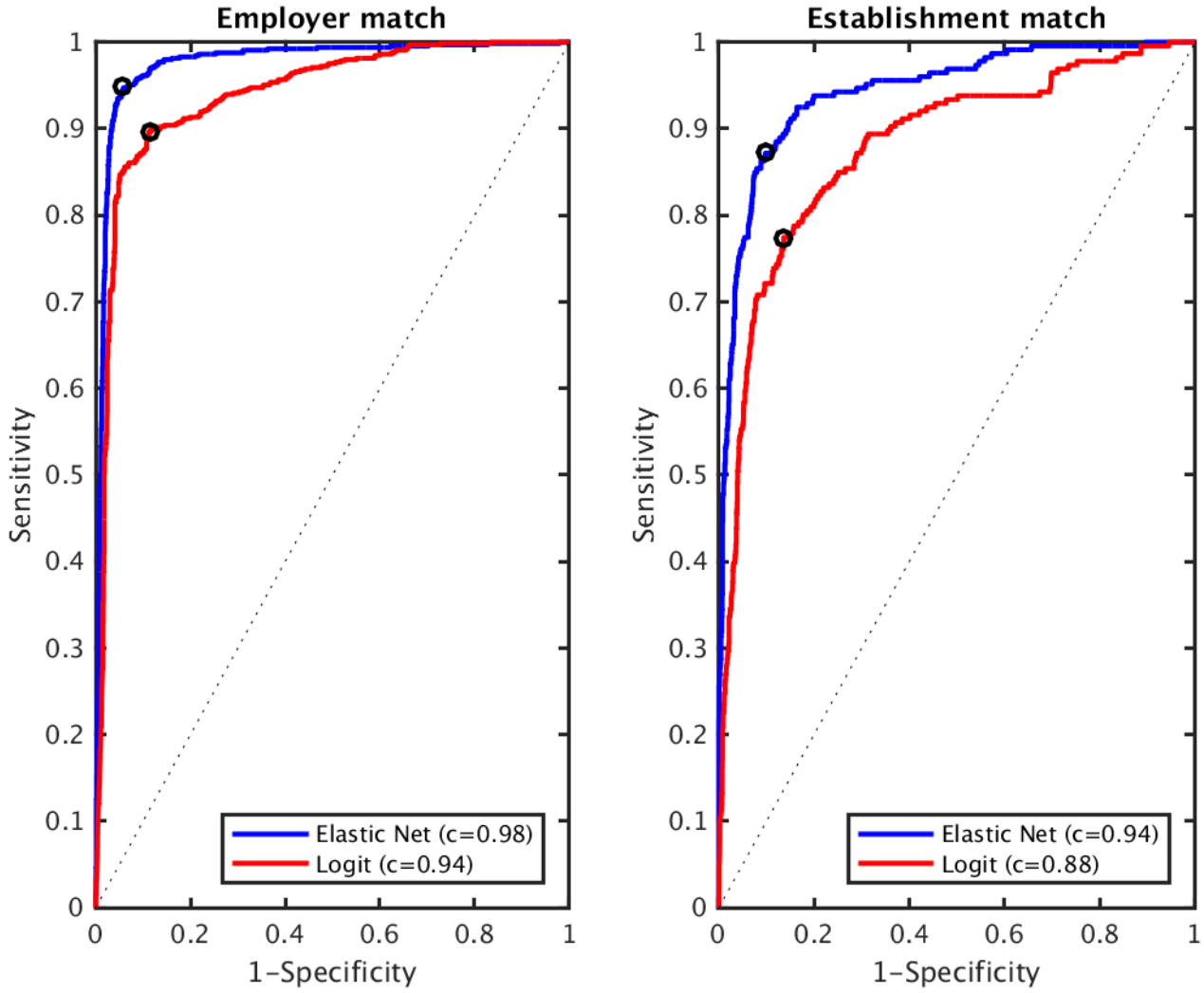
**Notes:** This table is based on the set of working HRS respondents in the 2010 wave who provided names and addresses of their employers. Measurement error is the difference between MI-optimal based imputations conditional on the sample where HRS self-reports are nonmissing and the averages based on HRS self-reports. Nonresponse bias is the difference between MI-cutoff based imputations for the whole sample and MI-optimal based imputations conditional on nonmissing HRS self-reports. See Appendix C for details.

**Figure 1:** Partial effects of the matching models



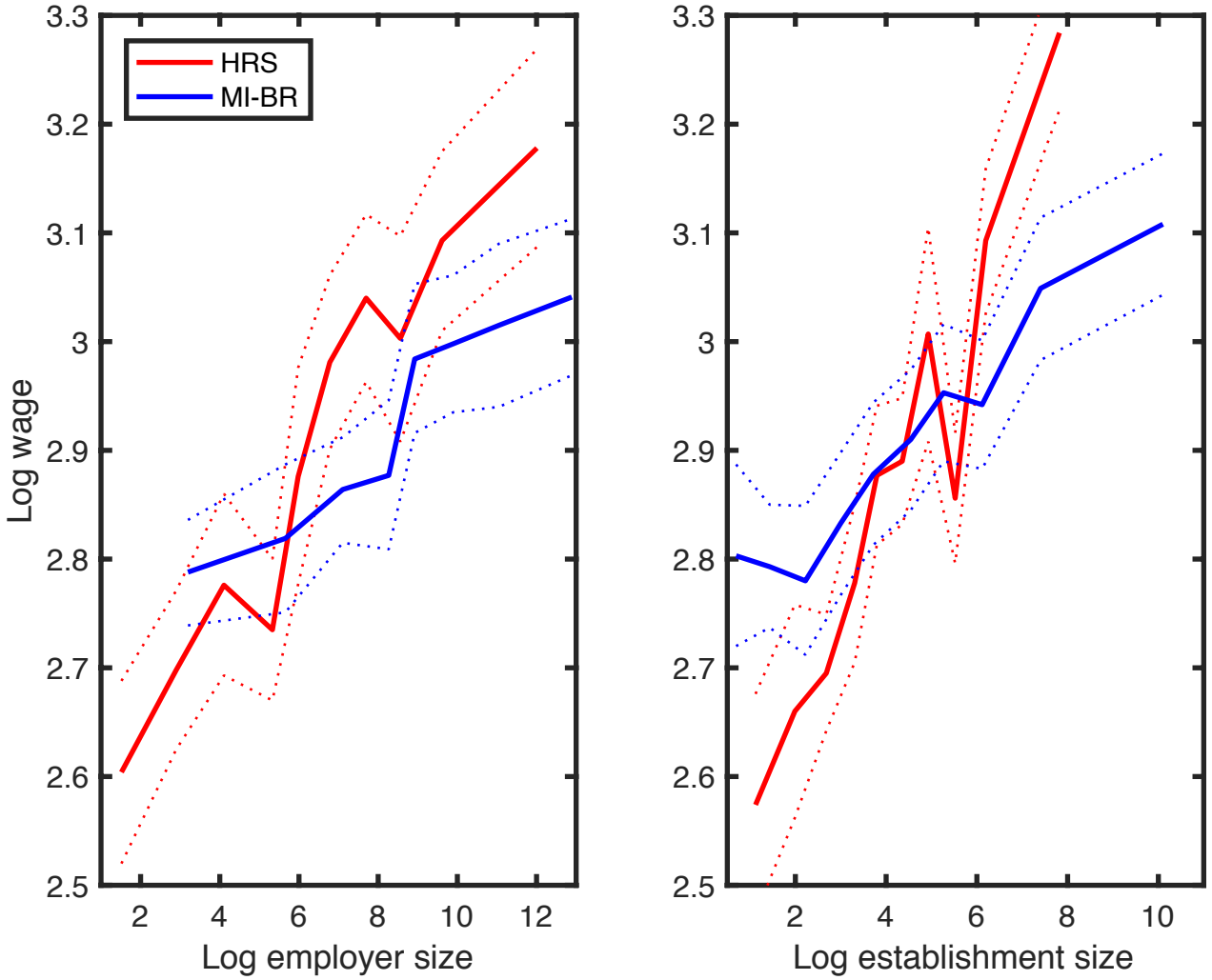

**Notes:** Each graph shows the partial derivative of the matching function for a given predictor holding all other predictors at their mean. Name and address agreement are based on Jaro-Winkler scores for similarity between HRS and BR names and addresses. These variables range from 0 to 1. Block share is the fraction of employment within the block (i.e. 3-digit zip, 10-digit phone number, telephone area code, or city-state that are common between a given HRS-BR pair) accounted for by a given establishment in the BR.

**Figure 2:** ROC curves of the matching models



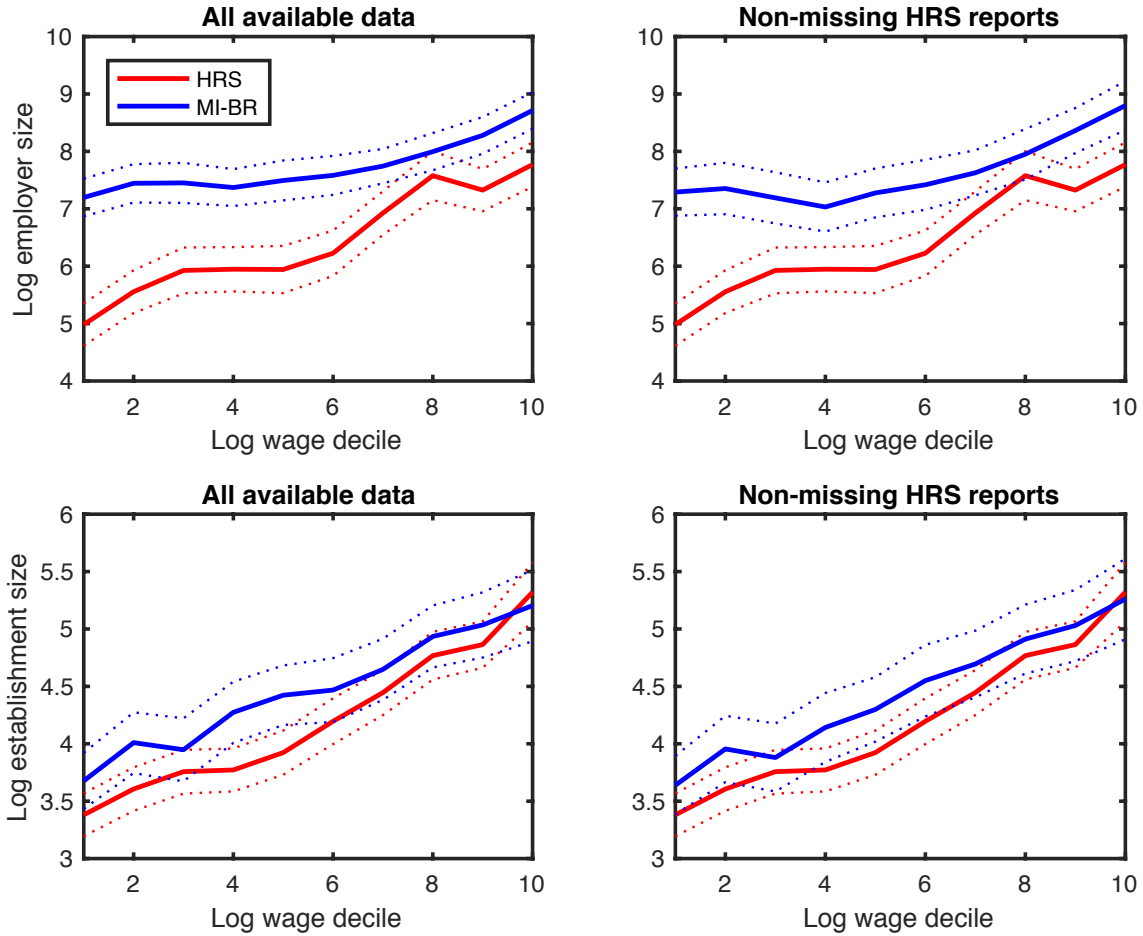

**Notes:** ROC estimates are computed using 10-fold cross-validation. Hollow black circles show the model's optimal cutoff probability which minimizes the distance to the top left corner of the graph (i.e. yield maximum sensitivity and specificity).

**Figure 3:** HRS versus CenHRS wage-size gradient




**Notes:** Dotted lines show 95 percent confidence intervals. Some HRS respondents report hourly wages directly. Others report compensation at daily, weekly, monthly, or annual levels. When compensation is reported at a different level than hourly, we convert it using the respondent's report of how many hours per week and weeks per year worked. Data in the 1-3 deciles and the 4-5 deciles of the BR imputed employer size distribution are binned together to prevent disclosure of information in small cells.

**Figure 4:** Measurement error and nonresponse bias across the wage distribution




**Notes:** Dotted lines show 95 percent confidence intervals. For multiply imputed statistics from the BR, the confidence interval accounts for within- and between-implicate variability. Some HRS respondents report hourly wages directly. Others report compensation at daily, weekly, monthly, or annual levels. When compensation is reported at a different level than hourly, we convert it using the respondent's report of how many hours per week and weeks per year worked. In the left-hand column, the MI-BR plot uses all available data, regardless of whether respondents provide a report of their employer's or establishment's size. In the right-hand column, the MI-BR plot is restricted to the sample of respondents that report employer and establishment size.

# Appendices

## A Constructing the training data set

Simple random sampling of pairs for manual review would produce very few true matches thereby limiting the predictive ability of our model. Instead, we oversample records with high levels of agreement on name and address to obtain a large set of possibly true links. The training sample is composed of HRS-BR candidate matches generated by blocking the 1998 and 2004 waves of the HRS with the BR on 3-digit zip code, 10-digit phone number, telephone area code, and city-state. We choose these specific years for two reasons. First, 1998 and 2004 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 train the model allows us to account for unobserved variation in the quality of data drawn from the BR.

Starting from a set of approximately 1000 HRS-BR candidate matches constituting about 500 unique HRS jobs, each record was evaluated by two different expert reviewers yielding approximately 2000 training observations. A total of eight reviewers conducted these reviews inside the Federal Statistical Research Data Center (FSRDC) computing environment.

## B Sample selection induced by cutoffs

As noted in section 4.1 the data driven cutoffs that we estimate and impose on the set of pairs prior to drawing multiple implicates sometimes results in 0 BR candidate matches for a given HRS job. For about 33 percent of employer matches and 8 percent of establishment matches where such a situtation occurs, we have little confidence about having selected the right employer or establishment using our matching models.[16] This appendix examines the extent to which HRS respondents for whom we find at least one BR match differ from HRS respondents for whom we cannot find any suitable matches.

Table B1 shows demographics, education, wages, annual hours of work, union membership, tenure, total labor market experience, and whether respondents work for public sector employers. The first column

[16] While employer matches are easier to confirm than establishment matches, the ROC-based optimal cutoffs trade off sensitivity and specificity. Furthermore, the employer and establishment match models are estimated independently. Thus, the share of HRS jobs with inadequate BR employer candidates (33 percent) is larger than the share of HRS jobs with inadequate BR establishment candidates (8 percent).

of the table shows characteristics for the full sample. The left panel shows characteristics for respondents matched to at least one employer (above cutoff) and respondents for whom no adequate BR candidate employer is available (below cutoff). The right panel shows the same information for establishment matches. Concentrating first on the left panel reveals some key differences: Nonwhite and foreign-born respondents are more likely to be unmatched. Respondents that are unmatched have about two years less in tenure with their current employer and have slightly lower lifetime labor market experience. Finally, unmatched workers are substantially less likely to be employed in the public sector. These differences point at employer attachment as an important source of signal strength in the data on employer names and addresses (obtained primarly for pension characteristics in the HRS). Furthermore, because public employers are more likely to maintain unified pension plans, it is easier to obtain sharper matching of public sector employees using our method.[17]

Patterns of selection in the right hand panel are similar to those highlighted above: Respondents who are nonwhite, foreign-born, have lower tenure, less labor market experience, and are more likely to be employed outside the public sector are less likely to be matched to any establishment. In addition, union membership is predictive of higher quality matches. On the whole, these statistics show that information elicited from survey responses may be garbled in ways that are correlated with individual characteristics. While our cutoff-based procedure filters away these sources of noise and hones in on higher quality matches for the majority of the sample, it does not refine information that is already garbled. Addressing this concern ultimately requires reliance on sharper identifiers such as EINs.

[17] Private employers often offer multiple pension plans with a variety of names. While reporting information about their pension plans, respondents may provide pension plan names that differ in small but meaningful ways from the employers name as it would appear in the BR. This source of variability could reduce the accuracy of our matching algorithm.

**Table B1:** Characteristics of matchable and non-matchable respondents

| | | Employer match | | Establishment match | |
|---|---|---|---|---|---|
| Variable | Full sample | Above cutoff | Below cutoff | Above cutoff | Below cutoff |
| N | 5700 | 3700 | 1900 | 5200 | 450 |
| Male | 0.446 | 0.432 | 0.472 | 0.439 | 0.531 |
| Age | 56.8 | 56.7 | 57.0 | 56.7 | 58.2 |
| Native born | 0.837 | 0.865 | 0.782 | 0.843 | 0.758 |
| White | 0.643 | 0.684 | 0.565 | 0.647 | 0.596 |
| Black | 0.236 | 0.210 | 0.287 | 0.235 | 0.249 |
| Other race | 0.121 | 0.107 | 0.148 | 0.118 | 0.155 |
| Schooling (years) | 13.3 | 13.5 | 13.0 | 13.4 | 12.5 |
| Wage ($/hr) | 29.5 | 28.9 | 30.6 | 29.0 | 34.8 |
| Hours | 1902 | 1901 | 1905 | 1902 | 1906 |
| Union | 0.143 | 0.139 | 0.151 | 0.147 | 0.096 |
| Tenure (years) | 10.2 | 10.7 | 9.2 | 10.3 | 7.5 |
| Experience (years) | 32.5 | 32.9 | 31.9 | 32.6 | 30.2 |
| Public employer | 0.247 | 0.280 | 0.184 | 0.258 | 0.108 |

**Notes:** This table is based on the set of working HRS respondents in the 2010 wave who provided names and addresses of their employers. Some HRS respondents report hourly wages directly. Others report compensation at daily, weekly, monthly, or annual levels. When compensation is reported at a different level than hourly, we convert it using the respondent's report of how many hours per week and weeks per year worked. Public sector employment is coded by the HRS based on the report of the employer name elicited from the respodent.

## C Measurement error and nonresponse bias

Let $s_{ij}^*$ represent the log of firm size for respondent $i$ employed at firm $j$. Define $R_i = 1$ if the HRS respondent reports a value for firm size and $R_i = 0$ if they do not. Assume that the HRS respondent reports firm size so that

$$s_{ij} = s_{ij}^* + v_i \tag{8}$$

where $s_{ij}$ is the log of self-reported firm size and $v_i$ is reporting error that is potentially correlated with other respondent level characteristics.

Denote the decile of the log hourly wage of respondent $i$ by $d_i$. The expectation of log firm size conditional on the decile of log wages is

$$E\left[s_{ij}^* | d_i\right] \tag{9}$$

The log of self-reported firm size conditional on the decile of log wages is

$$E\left[s_{ij} | d_i,\ R_i = 1\right] \tag{10}$$

The log of true firm size conditional on the decile of log wages among those who do respond is

$$E\left[s_{ij}^* | d_i,\ R_i = 1\right] \tag{11}$$

Measurement error is given by subtracting (11) from (10):

$$\underbrace{E\left[v_i | d_i,\ R_i = 1\right]}_{\text{Measurement error}} = E\left[s_{ij} | d_i,\ R_i = 1\right] - E\left[s_{ij}^* | d_i,\ R_i = 1\right] \tag{12}$$

Decompose term (9) by writing

$$E\left[s_{ij}^* | d_i\right] = p^d E\left[s_{ij}^* | d_i, R_i = 1\right] + (1 - p^d) E\left[s_{ij}^* | d_i, R_i = 0\right] \tag{13}$$

where $p^d = P\left[R_i = 1 | d_i\right]$ is the conditional response probability. Subtracting $E\left[s_{ij}^* | d_i, R_i = 1\right]$ from both

sides of equation (13) yields the following expression for bias due to nonresponse

$$\underbrace{E\left[s_{ij}^{*}|d_i\right] - E\left[s_{ij}^{*}|d_i, R_i = 1\right]}_{\text{Nonresponse bias}} = (1 - p^d)\left\{E\left[s_{ij}^{*}|d_i, R_i = 0\right] - E\left[s_{ij}^{*}|d_i, R_i = 1\right]\right\} \tag{14}$$

Positive values of the left side of equation (14) imply that nonresponders work at larger employers than do responders since $1 - p^d \in (0, 1)$. The converse is true for negative values of the left side.