Modernizing Person-Level Entity Resolution with Biometrically Linked Records

Matthew Gross* University of Michigan

Michael Mueller-Smith† University of Michigan

Draft date: December 3, 2020

Abstract

The increased utilization of administrative data in economics research has led economists to learn and deploy new empirical methods to prepare non-research data for analysis purposes. This paper focuses on record linkage, a common procedure to merge records from separate databases using common identifiers. We propose a supervised learning, record linkage algorithm that is trained using a large, novel dataset that includes biometric identifiers. We show that this large training data substantially improves model performance compared to the smaller training samples frequently reported in the literature. Next, we show that the model's performance degrades only slightly when attempting to link datasets with different underlying characteristics than the training sample. Lastly, we run simulations to explore how precision and recall, two commonly measured concepts in the matching literature, are directly related to internal and external validity of estimates. We show that the choice of matching algorithm can affect a researcher's ability to estimate unbiased and statistically precise parameters, reinforcing the importance of choosing the correct record linkage algorithm in applied work.

Keywords: Record Linkage, Probabilistic Matching, Administrative Data, Criminal Justice
JEL Codes: C18, C81, C88

*Department of Economics, University of Michigan, mbgross@umich.edu
†Department of Economics, University of Michigan, mgms@umich.edu

Acknowledgements: We thank Martha Bailey, John Bound, Charlie Brown, Connor Cole, Keith Finlay, Sara Heller, Kirabo Jackson, Emily Nix, Joseph Price, Mel Stephens, Jesse Rothstein, and Chelsea Temple for their comments and suggestions on this project. Jay Choi, Madeleine Danes, Francis Fiore, Jordan Papp, Benjamin Pyle, Lyllian Simerly, David Smith, Brittany Street, and Peixin Yang provided excellent research assistance. This research was performed with the generous financial support of the National Science Foundation (#1925563), the Bill and Melinda Gates Foundation, the Laura and John Arnold Foundation, the Michigan Institute for Teaching and Research in Economics, and the University of Michigan Population Studies Center.

1 Introduction

With the revolution in information technology, social science and policy researchers now have access to more data and computing power than ever. Increasing data availability, especially when linked, gives us the ability to answer new important questions. Recent work on economic mobility (Chetty *et al.*, 2016; Chetty and Hendren, 2018a,b), crime prevention (Heller *et al.*, 2016), health (Finkelstein *et al.*, 2012), environmental policy (Keiser and Shapiro, 2018) and the long term impacts of the great recession (Yagan, 2019) represent a small sample of topics being advanced through the utilization of linked data.

The number of papers citing "administrative data" among "top five" economics journals has rapidly increased in recent decades, especially since 2010 (see Figure 1).[1] These outlets together published 7 articles mentioning administrative data per year between 1995 and 2010; by 2017-2019, the corresponding figure grew to 54. Yet, the fastest growing type of cutting edge data – administrative records – are created without the intention of research applications and are instead a byproduct of the regular activities of public agencies, private businesses, or non-profits. So while administrative data clearly is now a major component of modern economic research, social scientists are regularly confronted with needing to develop and deploy an array of empirical methods to prepare non-research data for analysis purposes.

One of the most common tasks is record linkage, which merges rows of observations from two or more data sources using common identifiers available in the different data sources.[2] In the absence of accurate, unique identifiers, researchers must rely on similarity comparisons of plausibly identifying variables common to all data.[3] This raises questions of how best to quantify similarity, which variables to weigh more or less, and what index threshold should be established to merit a statistical match. Traditionally, most researchers rely on either deterministic rules (e.g. perfect match on first name, last name, and date of birth) for the sake of simplicity or probabilistic linear models trained on a subset of hand-coded records that undergo a clerical review to establish plausible true match status (see Table 1).[4]

[1] These are The American Economic Review, Econometrica, The Journal of Political Economy, The Quarterly Journal of Economics, and the Review of Economic Studies.

[2] Another form of record linkage, as in the focus of this paper, is identifying who is the same individual across rows in the same dataset without a reliable unique identifier (e.g. deduplication). This distinction is somewhat arbitrary as any deduplication problem can be restated as a matching problem.

[3] Also referred to as probabilistic matching, entity resolution, or fuzzy matching.

[4] Depending on the setting and application, hand-coded training samples can range from as little as 50 to as many as 80,000 observations. For example, recent work by Abowd *et al.* (2019), Wisselgren *et al.* (2014), and Feigenbaum (2016) hand-code 1,000, 8,000, and 80,000 observations respectively. In general however, the hand-coded samples used to estimate supervised learning models are somewhere between 500 and 10,000 observations.

This paper introduces a different approach leveraging a unique source of previously unexploited data. We use biometrically linked (fingerprint-matched) records from the U.S. criminal justice system to construct unbiased measures of true match status. While the administrative data is drawn from a highly selected portion of the general population (i.e. those accused of criminal activity or in prison for criminal conduct), it provides trillions of training pairs to fine tune a high-dimensional, non-linear, machine learning based linkage model that would otherwise be cost-prohibitive or impractical to estimate. The data is comprised of decades of personally identifiable information (PII) from two separate sources: (1) misdemeanor and felony defendants in criminal cases from a large district court and (2) incarcerated individuals from a state Department of Corrections. Both data sets include biometric ID numbers as well as the inconsistent, flawed PII information as originally entered into the data system.[5]

We compare the performance of a range of matching strategies from simple deterministic rules to more sophisticated prediction algorithms like random forests and neural networks. Our preferred specification is a demographic enhanced random forest specification that allows the determinants of PII match quality to flexibly vary by race/ethnicity and sex, tailoring the prediction according to the differential naturally occurring and error-induced variation in PII by demographic group. We also evaluate the relative gains of integrating a large, biometrically verified training sample compared to a feasible set of hand-coded training data, conditional on matching algorithm. We find that human coders tend to be overly conservative in assigning true match status through the process of clerical review, especially for Hispanic individuals and women.[6] Allowing our machine learning algorithm to train on 1 million observations strengthens performance on both recall and precision, demonstrating significant gains over typical sample sizes for model estimation.

Because our training data is highly selected, non-representative of the general population, and drawn only from the state of Texas, it is appropriate to question its general relevance beyond criminal justice applications and in the U.S. overall. We conduct three exercises to evaluate the degree of performance degradation as we extrapolate to other contexts with increasingly dissimilar populations: (1) a deduplication of multi-state prison data from a single date in time to assess national scaleability,[7] (2) a one-to-one linkage of registered Washington voters in 2008 and 2012 to assess performance in a more representative population, and

[5]While there are a number of criminal justice data repositories that leverage fingerprint based IDs, often the incorrect PII is overwritten to standardize entries therefore eliminating its use as training data.

[6]This is largely the result of an over-reliance on name similarity over date of birth similarity when determining hand-coded match status.

[7]The goal in this exercise is to evaluate whether the algorithm incorrectly identifies a single person as being in two places at the same time.

(3) entity resolution applied to corrupted synthetic data created from all deaths in the U.S. between 2000 and 2009 from the Social Security Administration's Death Master File (DMF) to assess model degradation among large populations with higher likelihood of naturally occurring PII similarity.[8] Across all three exercises, we surprisingly observe strong performance close to matching or exceeding the effectiveness of the model in our main application.

While details on matching often get shortchanged in academic publications, the common matching performance metrics of recall and precision directly relate to concepts of internal and external validity in causal inference, which empirical researchers should care about.[9] To illustrate these points, we conduct a series of simulation exercises that increasingly corrupt the record linkage process and track the resulting impact on parameter bias and precision. We consider two common scenarios: (1) designs where a matched record is an indication that an outcome has occurred (e.g., recidivism, employment, or public program take-up) for an individual, and (2) situations where analysis is conditioned on being in the matched sample (e.g. wage effects among those who file taxes, or health care utilization among those with Medicaid coverage). In the first scenario, we show that errors in recall and precision systematically attenuate the estimated coefficient of interest and impair statistical precision, making it less likely that the null hypothesis of a null effect will be rejected. In the second scenario, errors in precision lead to a similar attenuation effect and lack of statistical precision; however, errors in recall lead to overinflated estimates of the effect of interest. This last fact is closely related to the concept of external validity, where the observations that are successfully matched and included in the analysis sample are not representative of the general population.

The remainder of the paper is organized as follows. The next section of the paper reviews relevant literature. The third section discusses the algorithm methodology, while section four reports results from our out of sample tests. The fifth section reports results from performing the algorithm on synthetic data, and the sixth section concludes.

2 Statement of Linkage Problem and Related Literature

There is a large and diverse literature devoted to record linkage and probabilistic matching. Whether aiming to identify common entities within a given dataset (i.e. deduplication) or

[8] We generate several synthetic data sets by corrupting names and dates of birth in the spirit of Tran *et al.* (2013) to determine performance in the event of different transcription and data entry errors. See Appendix C for more details.

[9] Recent literature has explored the concept of data matching strategies and it's implications for empirical research in specific contexts. See, for example, Bailey *et al.* (2017) and Abramitzky *et al.* (2019) for a discussion of historical data linkage and Tahamont *et al.* (2019) for a discussion on linking an experimental intervention to administrative data.

combining two or more datasets without a unique linking variable (i.e. record linkage), a range of statistical techniques and methodologies have been developed.[10]

Record Linkage and Economic Research

In most economic applications, researchers leverage matching techniques as tools to support the analysis of two or more linked datasets. As researchers have noted, however, the linkage process itself becomes an added source of error that can have serious implications on estimated coefficients and standard errors (Scheuren and Winkler, 1993). For example, Scheuren and Winkler (1997) report a simulation where a naive estimator based on a faulty match is attenuated by as much as 60%. The authors propose an iterative methodology that corrects for errors in the match stage. Although based on an "ad hoc" modeling intervention, their method allows them to account for matching errors when estimating the regression of interest. They show that their proposed method allows them to recover nearly all of the attenuated coefficient. Lahiri and Larsen (2005) propose an unbiased estimator using match probabilities estimated by the linkage procedure as regression weights.[11] In addition, they also propose a bootstrapping method to achieve closer coverage of the unbiased confidence interval. In the simulation exercises, the estimator proposed by Lahiri and Larsen outperforms the one proposed in Scheuren and Winkler (1997); however, the assumption of independence between match probabilities and outcome variables is somewhat restrictive and likely to be violated in many cases. Lastly, Abowd *et al.* (2019) use a multiple imputation method to build 10 imputed datasets to account for errors in the linkage process. As an application of their methodology they show that the wage-firm size gradient as measured by surveys is overstated.

Bailey *et al.* (2017) review some common algorithms used to link historical datasets and show how different linking strategies can attenuate estimates of the intergenerational income elasticity by as much as 20%. In the context of historical record linkage, matching algorithms are often used to perform a one-to-one match between successive Census waves. The authors link the LIFE-M data and the 1940 Census to measure the intergenerational income elasticity of men with regard to their fathers.[12] Then, they attempt the link using methodologies previously published in the historical record linkage literature to see how each method yields different intergeneration elasticity of income estimates. They show that the choice of linkage

[10] Although the goals of deduplication and record linkage are different in practice, the underlying theory and methodology is equivalent as either problem can be restated as the other.

[11] This estimator is unbiased under the condition that the match probabilities are uncorrelated with the outcome variable of interest. This assumption is likely not to hold in many settings. We explore this further in our simulation section.

[12] Information about the LIFE-M data linking project can be found at https://sites.lsa.umich.edu/life-m/

method can lead to attenuation bias, with some estimates off by as much as 20% of the underlying true value.

Tahamont *et al.* (2019) show how in modern settings – e.g. linking administrative data with a randomized control trial to track binary outcomes – the linkage choices can impact statistical precision and attenuate the estimated treatment effect. The relevant research design occurs when a researcher attempts to link a treatment to an external (often administrative) dataset where the match status determines the outcome variable of interest. There are numerous examples of this type of design, such as measuring the effects of crime policy on recidivism and labor market outcomes (Mueller-Smith and T. Schnepel, 2020), the effects of job retraining programs on employment (Biewen *et al.*, 2014) or the effects of payday loans on financial outcomes Skiba and Tobacman (2019) among many others. Tahamont *et al.* show that overly conservative matching strategies, such as mandating a match only on perfect agreement of comparison variables, can attenuate estimated causal treatment effects and reduce statistical power. The authors also show that probabilistic algorithms, despite increasing the number of false positive matches, perform better than strict algorithms by increasing the number of true positive matches.

Defining Record Linkage

Fellegi and Sunter (1969) provide one of the earliest formalizations of the record linkage problem. Specifically, given two sets, $\mathbf{A}$ and $\mathbf{B}$, which contain elements a and b, one seeks to identify which elements of $\mathbf{A}$ and $\mathbf{B}$ are common to both sets. The full set of ordered pairs

$$\mathbf{A} \times \mathbf{B} = \{(a, b); a \in \mathbf{A}, b \in \mathbf{B}\}$$

is the union of two disjoint sets

$$\mathbf{M} = \{(a, b); a = b, a \in \mathbf{A}, b \in \mathbf{B}\}$$

and

$$\mathbf{N} = \{(a, b); a \neq b, a \in \mathbf{A}, b \in \mathbf{B}\}$$

which together account for all *matches* and *non-matches* among the ordered pairs.

The elements of $\mathbf{A}$ and $\mathbf{B}$ are assumed to contain a vector of common variables that provide identifying information (e.g. names, addresses, demographic traits, etc), but lack the certainty of a known unique identifier. A comparison function γ is defined to quantify the

similarity of the identifying information for a given pair

$$\gamma(a,b) = \left\{\gamma^1\left[\alpha(a),\beta(b)\right],\cdots,\gamma^K\left[\alpha(a),\beta(b)\right]\right\}$$

over K dimensions from the full set of ordered pairs in $\mathbf{A}\times\mathbf{B}$.

To complete the algorithm, one must define a decision rule mapping the comparison space, Γ, to one of three possible designations: a statistical match ($\mathbf{M^S}$), a statistical non-match ($\mathbf{N^S}$) , or statistical uncertainty ($\mathbf{U^S}$).

$$\mathbf{M^S} = \left\{(a,b); P(\mathbf{M}|\gamma) > P(\mathbf{N}|\gamma), P(\mathbf{M}|\gamma) > \rho^U, a\in\mathbf{A}, b\in\mathbf{B})\right\}$$

$$\mathbf{N^S} = \left\{(a,b); P(\mathbf{N}|\gamma) > P(\mathbf{M}|\gamma), P(\mathbf{N}|\gamma) > \rho^U, a\in\mathbf{A}, b\in\mathbf{B})\right\}$$

$$\mathbf{U^S} = \left\{(a,b); \rho^U > P(\mathbf{M}|\gamma) + P(\mathbf{N}|\gamma), a\in\mathbf{A}, b\in\mathbf{B})\right\}$$

where ρ^U represents a baseline probability threshold for asserting statistical match or non-match status. Fellegi and Sunter (1969) define these together as the *linkage rule*.

Putting aside the issue of pairs with uncertain designations, the linkage result will be a statistical designation of match status, which may contain type I and type II errors.

	$(a,b)\in\mathbf{M^S}$	$(a,b)\in\mathbf{N^S}$
$(a,b)\in\mathbf{M}$	True Positive	False Negative
$(a,b)\in\mathbf{N}$	False Positive	True Negative

Algorithmic Approaches to Record Linkage

Operationalizing record linkage requires defining comparison functions and threshold values for determining predicted match status. Two related approaches are most frequently used in modern record linkage: (1) deterministic and (2) probabilistic.

In a deterministic algorithm, two records are classified as a match or non-match based on the exact match of one or more variables common to both records. In some deterministic models, paired observations must match on all common variables to be classified as a match. In other settings with a rich set of matching variables, multiple linkage rules are defined to allow for more flexibility in the match process (Setoguchi *et al.*, 2014, for example). Lastly, some deterministic models utilize an "iterative method" of rules to identify matches (Ferrie, 1996; Abramitzky *et al.*, 2012, 2014; Dahis *et al.*, 2019, for example). In general, the researcher determines the rules used to classify matches based on the setting and the variables available.

For example, data that includes Social Security numbers will leverage this variable at the expense of agreement on address or middle name; however, researchers attempting to link data that includes only name and date of birth may specify that the last name must be the same to consider two records a match.

Probabilistic algorithms, on the other hand, attempt to predict the match probability of any two observations based on the relative agreement of their matching variables. This requires the additional step of defining comparator functions that measure the degree of non-exact similarity between two potential comparison values (e.g. "Mike" as opposed to "Michael"). But, this approach has benefits over the purely deterministic models in that it more flexibly sets a decision rule that optimizes the tradeoff between making more matches and limiting false matches (Mèray *et al.*, 2007; Tromp *et al.*, 2011; Moore *et al.*, 2016), especially in settings where there is no direct identifier such as Social Security Number (Dusetzina *et al.*, 2014).

Fellegi and Sunter propose a weighting system that places different value on each variable used to determine a statistical match or link. These weights are based on the underlying probability that a variable will match given that a,b are a true match and the probability that a variable will match given that a,b are a true non-match. Once the weights are estimated, it is possible to calculate a composite score for any pair of observations from **A** and **B**, and use a threshold system where observations above a certain cumulative score are classified as a statistical match.

Building on Fellegi and Sunter (1969), Jaro (1989) and Larsen and Rubin (2001) use an Expectation-Maximization (EM) routine to estimate the underlying match weights in the classic Fellegi-Sunter (F-S) framework. Sadinle and Fienberg (2013) extends the model by proposing a F-S model that matches observations between three different sets instead of two. The EM routine is especially useful when the researcher does not have access to training data, as the match weights are determined through a process of picking weights to maximize an objective function, followed by clerical review. The process is repeated until the researcher is satisfied with the identified matches.

More recently, researchers have estimated match weights using insights from machine or supervised learning. These algorithms typically require training data to estimate a model for out of sample prediction (Feigenbaum, 2016; Abowd *et al.*, 2019) with the resulting match predictions depending both on the quality of the model as well as the accuracy of the training data.

Usually training data is created by manually determining match status for a sample of

paired observations through a process referred to as clerical review.[13] This process can be time consuming and expensive, which limits the available sample size for training models. With training data in hand, however, one can extrapolate predicted match status for the remainder of paired observations using one of many possible statistical models. Feigenbaum (2016) attempts to match individuals from the 1915 Iowa State Census to the 1940 Federal Census. He runs a probit regression of true match status on a host of match variables using available training data. The probit model estimates the predicted probability of a match given a vector of match variables. Once this model is recorded, he uses it to estimate matches from the full sample of the data. Other non-regression based classifier algorithms that can be used to make predictions include neural networks, Naive Bayes Classifiers (NBC), Support Vector Machines (SVM) and Random Forests.[14] We discuss these alternative algorithms in the latter part of the paper.

Lastly, a newer class of probabilistic models have recently been proposed utilizing Bayesian techniques (Steorts *et al.*, 2016; Fortini *et al.*, 2001, for example). From a practical perspective, the complexity of these algorithms require more computational power and lack scaleability for administrative data application which often contain hundreds of thousands if not millions of observations; however, one of the benefits of Bayesian models is that they more naturally allow the researcher to directly characterize and account for matching error in the analysis stage (Steorts, 2015).

3 Data and Background

We utilize a novel source of variance, finger-print based identifiers, found in two data sets to generate training data for our matching algorithm described in Section 4. The first comes from the Harris County Justice Information Management System (JIMS) in Texas and includes personally identifying information (PII) for all criminal defendants for cases charged between 1980 and 2017. Harris County creates a system person number (SPN) to track individuals across interactions within JIMS. This SPN is a biometric ID that is tied to one's fingerprints, meaning that it should uniquely identify individuals [15] and remain relatively constant over

[13] A notable exception is the paper by Price *et al.* (2019), which leverages a public family-tree website to generate a large training sample of "true links." This method is an improvement over typical clerical-review generated training sets since the people identifying matches have more information and a higher incentive to create correct links than a standard hand-coder.

[14] Feigenbaum (2016) also estimates versions of Random Forest and SVM models.

[15] Fingerprint uniqueness is generally accepted; however, there is some concern that the automated methods used to match fingerprints use substantially less information than a full print and therefore increases the chances of false positives (Pankanti *et al.*, 2002). Comparisons of fingerprint matching technology suggest that the state of the art systems have false positive and negative rates of approximately 0.1% (Maltoni *et al.*,

time.[16] An individual with multiple charges and appearances in court will show up many times in the Harris County data. The SPN number links the same individual across charges; however, the PII recorded for each individual charge has not been synchronized. This creates a data system where the same SPN can have different combinations of PII. These differences could be caused by typographic errors, legal name changes, or the use of an alias. Our data contains 1,317,063 unique SPN, and 1,722,575 unique combinations of name and date of birth, indicating approximately 1.31 combinations of PII within each SPN.[17]

The second data source we use to generate our matching algorithm is from the Texas Department of Criminal Justice (TDCJ). This data includes PII for inmates in the Texas state prison system between 1978 and 2014. In addition to PII, the TDCJ data also contains a biometric identifier, the Texas State Identification Number (SID), which is also built off of fingerprints. Similar to the Harris County data, the recorded PII varies within a given ID. In total, there are 905,528 unique IDs and 1,042,450 unique PII combinations, implying slightly less PII variation within a given ID relative to Harris County data.[18]

While there are overlapping populations between the TDCJ and JIMS data systems and their biometric IDs are built off of the same underlying variation (fingerprints), the systems have not been unified and so there does not exist a unique SPN to SID crosswalk. As such, throughout our analysis, we treat these data as appended but disjoint sets, generating training pair matches and statistical matches only within a given dataset rather than across the TDCJ and JIMS records.

Individuals involved in the criminal justice system are a highly selected group in the general population, which raises important questions about the general relevance of our empirical models to other settings. Table 2 describes the demographic traits of these data sources as compared to the general population in the United States. Not surprisingly, the Harris County court and Texas prison data have a disproportionate number of men and people of color compared to the population at large. As a result, the types of within-biometric ID variation in PII may differ systematically with a broader population. For instance, women more regularly change their last names due to marriage. Because women are not well represented in the

2017; Watson *et al.*, 2014).

[16]Recent work by Yoon and Jain (2015) and Galbally *et al.* (2019) raise some concerns about the permanence of fingerprints as the subject ages and the duration of time between imprints grows. The lack of criminal activity by the elderly should reduce the set of individuals that offend over long periods of time, making this concern relatively minor in our setting. Large numbers of individuals with multiple assigned IDs would likely indicate that our precision estimates are a lower bound, but we see limited evidence that this is the case.

[17]When conditioning on individuals who have more than one appearance in the court system, this ratio increases to 1.47 combinations of PII within each SPN.

[18]When conditioning on individuals who have more than one appearance in the prison data, this ratio increases from 1.15 to 1.17 combinations of PII within each SID.

criminal justice system, our prediction algorithm may not be optimized to recognize these errors as much as we might hope for a general population application.

Given this discrepancy, in Section 5.3 we evaluate whether performance degrades when applying our prediction algorithm to several settings beyond the scope of our training data. This analysis sheds light on the general suitability of our model for non-criminal justice applications.

4 Matching Algorithm

We define our match problem in terms of data deduplication: identifying which records in a given dataset belong to the same individual. We start with set $\mathbf{D}$ containing N total observations, each with unique combinations of full name and date of birth.[19] The potential match space $\mathbf{\Delta}$ of all records $d_i \in \mathbf{D}$ contains $\frac{N \times (N-1)}{2}$ unique ordered pairings:

$$\mathbf{\Delta} = \{(d_i, d_j); i < j, d_i \in \mathbf{D}, d_j \in \mathbf{D}\}$$

We seek to identify the subset $\mathbf{\Omega}$ containing pairings of observations that belong to the same latent identity

$$\mathbf{\Omega} = \{(d_i, d_j); d_i = d_j, (d_i, d_j) \in \mathbf{\Delta}\}$$

where there are $\mu \leq N$ total entities in dataset $\mathbf{D}$.

Assessing the match potential for every pair of observations is impractical due to the size of most administrative data applications (including our own). In order to save computational resources and focus our search on pairs with likely matches, we utilize a blocking method to reduce the number of comparisons. Specifically, we propose a simple blocking strategy $\mathbb{B}$, comprised of $\mathbb{B}_1 \cup ... \cup \mathbb{B}_L$ individual blocks. Each block $\mathbb{B}_{l \leq L}$ creates a partition of $\mathbf{\Delta}$. An example of a block partition could be the subset of records that exactly match on date of birth. Another might be those that share the exact same first and last names. The more specific a blocking criteria is the fewer comparisons that are made and the greater chance that an underlying set of matched records is missed by the algorithm. The goal in building in multiple (potentially overlapping) blocks is to restrict the comparison space for computational feasibility while also providing flexibility to identify matches that may not satisfy the criteria for any given block. Any pair of records that are not in the subset created by $\mathbb{B}$ are automatically classified as a non-match.

[19] As is common in the matching literature, we eliminate duplicative observations with the exact same combination of PII.

In practice, we utilize the union of 10 blocks described in Table 3. For a given pair of observations to be compared, it must appear in the same block group for at least 1 of the 10 block definitions. The first four blocks are created by limiting comparisons to those with the same date of birth and either the phonex or soundex code (described in more detail below) for the first or last name. The next six blocks rely on pairings that share the first and last name soundex or phonex code with a single component of the date of birth also being common (i.e. day of birth, month of birth, or year of birth). Given the reliance on the soundex and phonex codes to generate candidate pairs, our algorithm will perform poorly in the event of simultaneous typos in the first syllable of both the first and last names. Together, these steps reduce the comparison space of $\mathbf{\Delta}$ from over 4 trillion observation pairs to just 17,577,515 observation pairs, with 95.2% of actual matches included in the blocked subset of paired observations.[20]

We also must introduce a comparison function $\gamma(d_i, d_j)$ to quantify record similarity and create predictions regarding match status. For each pair of observations, we generate *46 variables* that apply different comparators to various components of the PII. We include dummy variables for whether there is an exact match for first name, last name, middle name or the soundex or phonex code matches for any of the three name components. We include a dummy variable for whether the standardized first or middle name is an exact match. The standardized name is created using a U.S. Census Bureau crosswalk of nicknames. For example, the standardized name for someone named Matt or Mike would be Matthew and Michael respectively. This allows us to account for common nicknames when creating matching weights. Vick and Huynh (2011) provide evidence that using standardized names can improve the performance of matching algorithms.

In addition to binary match variables, we calculate a number of distance metrics to measure the similarity of names and dates of birth. For each name component we include the Jaro-Winkler, Levenshtein Edit Distance normalized by the string length and raw edit distance. Lastly, we calculate a measure of uniqueness for each first, middle and last name in our data. We then take an average of the uniqueness measure within the comparison pair and interact it with the relevant Jaro-Winkler comparison score and the number of raw edits

[20] A review of the non-blocked true-match pairs indicate stark differences in PII that would likely be impossible to resolve with any probabilistic matching technique. We believe two potential phenomenon might contribute to this pattern. First, errors can be made with fingerprint entry creating a false biometric link between two distinct individuals. Second, justice involved individuals may intentionally falsify their PII through the use of an alias. Both of these issues in the data are likely non-trivial given that the source data extends back to the 1980's, before advances in information technology infrastructure reduced the risk of these problems. As a consequence, the external (non-criminal justice) relevance of our model may be best characterized when focusing on just the hold-out blocked sample, excluding the non-blocked observations, which presents an even more optimistic view of the model's performance.

to create two different measures. The idea behind these variables is to give extra weight to rare names that match. For example, two observations with the last name "Smith" (the most common last name in the 2010 Census) are less likely to be the same person than two observations with the last name "Cooke" (the 1,000th most common last name in the 2010 Census). This should give extra weight to individuals with rare names that are similar. To measure date similarity, we include raw string edit distances and absolute numerical distance between the month, day and year of the dates of birth as well as the date of birth overall. See Table A.1 for a list of each variable included in the model.

A variety of linear and non-linear prediction algorithms as well as rules of thumb could be applied to the data at this point to determine which comparators receive more or less weight in generating a prediction of true match probability. We are agnostic with regard to empirical methods and explore a range of candidate algorithms in Section 5, ultimately settling on a *random forest* classifier as our preferred specification.

The random forest algorithm, as proposed by Breiman (2001), allows for classification by building many decision trees using random draws of the training data such that each decision tree is constructed using a different bootstrapped sample.[21] In addition, the variables used to split the tree are randomly selected in each tree. The bootstrapped samples combined with the randomly selected splitting variables allow for the construction of a large number of prediction models with minimal correlation between them. Classification is based on the mode prediction over the full sample of trees. [22]

A single decision tree effectively captures non-linearities and interactions among terms; however, predictions based on individual trees often have high variance. Building many trees based on bootstrapped samples allows us to build a non-linear model while also alleviating concerns of overfitting (Hastie *et al.*, 2016). Based on these properties, a random forest classifier is particularly well-suited to our application of building a non-linear model for entity resolution.

While random forest models are commonly used by computer scientists, they are utilized relatively infrequently in applied economics research. In most cases, they are used as tools to predict macroeconomic trends (Alessi and Detken, 2018, for example), though other applications include predicting future criminal recidivism (Grogger *et al.*, 2020) or the determinants of preferences for income distribution (Keely and Tan, 2008). There is also a recent theoretical literature showing the benefits of using random forests over more traditional linear regression or matching models in the estimation of heterogeneous treatment effects

[21] The technique of utilizing multiple draws of a random sample is also known as bagging.
[22] See appendix B for more details about the random forest classification methodology.

(Taddy *et al.*, 2016; Wager and Athey, 2018).

To assess model performance, we take a 1,000,000 pair random sample from the blocked pair set, which is slightly more than 5% of blocked pairs. The same million observations are used to train each of the candidate prediction models, while the remaining 16,577,515 blocked pairs are held back for out-of-sampling testing purpose. This is especially important in the context of highly non-linear machine learning models, which can have a tendency to overfit training data.

The sequence of steps in the data construction, model training, and out-of-sample testing is presented in Figure 2.

5 Evaluating classification performance

5.1 Baseline results

Table 4 presents six performance metrics for evaluating the relative strength of ten different prediction algorithms. These range from a basic deterministic model, which requires exact matching on 5 out of 6 variables (first name, middle name, last name, day of birth, month of birth, and year of birth), to more sophisticated machine learning algorithms like neural networks and random forests. A description of each prediction algorithm is described in detail in Appendix B.

We evaluate performance along six criteria, five of which focus on the quality of statistical matches while the sixth measures computational intensity. Statistical match quality criteria are measured using various combinations of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) in the out-of-sample blocked pairs as well as the universe of non-blocked pairs.[23] Defining these outcomes for the trillions of candidate matched pairs is accomplished through comparing predicted statistical match status against the fingerprint-based measure of true match status. Through excluding the 1 million training observations, we avoid conflating model performance with concerns about data overfitting. We also impose linkage transitivity in our algorithm to ensure that if record A matches to record B, and record B matches to record C, then we count records A and C as matches regardless of whether the model determines them to be a match. This will affect measures of the model's performance by increasing the number of true and false positives.

[23] All non-blocked pairs are defaulted to be a statistical non-match meaning they can only be classified as TN or FN. This saves substantial computing resources.

Accuracy ($\frac{TP+TN}{TP+FP+TN+FN}$) and the False Positive Rate ($\frac{FP}{FP+TN}$) although widely reported in the record linkage literature are relatively meaningless in this context due to the large number of true negatives. We utilize slightly modified definitions that replace TN with $10 \times (TP + FN)$ implying that we cap the number of true negatives at a ratio of 10:1 relative to the number of true matches in the data. Because of the need for this modification, we focus instead primarily on Precision ($\frac{TP}{TP+FP}$), which captures the true match rate among statistical matches, and Recall ($\frac{TP}{TP+FN}$), which captures the statistical match rate among true matches.

No single algorithm dominates all performance criteria. Most algorithms deliver precision rates in the 0.92 to 0.94 range, suggesting most classifiers generate reliable statistical matches. A much wider performance range is observed for recall (0.72 to 0.88) meaning that "better" and "worse" algorithms distinguish themselves by being able to better identify marginal matches where the similarity of PII between two records may not be clearly obvious.

Non-linear machine learning algorithms (random forests, neural networks) outperform other classifiers with regard to recall. The flexibility provided in these models in accounting for non-linearities drives this result. Our preferred specification enhances the standard random forest model with 8 additional comparison variables accounting for the shared demographic traits (sex, race/ethnicity) between the pairs, which adds another dimension of comparison but also adds flexibility in the treatment of existing comparison variables (e.g. pairs of female records may rely less on last name matching in establishing a statistical match given the higher natural rate of last name changes in the female population). We call this model, which is our preferred specification, the demographic-enhanced random forest (DE-RF) model.

Figure 3 shows a variety of diagnostic graphs from the training data for the DE-RF model. Figure 3a plots a histogram of the predicted match probabilities as well as the underlying true match rate across the distribution. High performing binary classification models differentiate likely matches from non-matches (visible from the clear bimodal distribution in this probability density in this figure) as well as efficiently sort ambiguous pairings into those more and less likely to be true matches (visible from the monotonic increase in true match rate throughout the distribution as well as the fairly sharp increase in true match rate starting around roughly 0.4).

The receiver operating characteristic (ROC) curve plots the true positive rate (also known as recall) against the false positive rate for varying thresholds in the predicted index for establishing a statistical match (see Figure 3b). Improving the true positive rate comes at the expense of the false positive rate and vice versa, which is also reflected in the tradeoff between recall and precision shown in Figure 3c. At very high thresholds, the few statistical matches

made are almost always true matches, which raises precision; however, such high thresholds means many true matches are missed lowering recall. The only method of simultaneously improving both recall and precision is through model improvements that better predict matches and non-matches in the first place.

The F-Statistic balances these tradeoffs through combining the concepts of recall and precision into a single statistic that takes the harmonic mean of both components:

$$\text{F-Statistic} \quad = \quad 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

We utilize the F-Statistic for two purposes. First, in the training data, we establish our statistical match threshold using the predicted probability that maximizes the in-sample F-Statistic (0.42 in the case of our demographic-enhanced random forest model as seen in Figure 3d). Second, we identify the preferred algorithm in Table 4 based on the routine that delivers the best out-of-sample F-Statistic.

It happens that our preferred specification also performs relatively well on our second performance criteria: the duration of combined estimation and prediction routines. In our application, the demographic-enhanced random forest model processing time took 0.28 hours to complete. While this is longer than the deterministic model, which entailed no model estimation or prediction, it is significantly lower than many of the other candidate algorithms, sometimes by a matter of days.

Figure 4 examines the relative contribution of the individual PII components on statistical match probability in the DE-RF model. We consider first name, middle name, last name, and the complete date of birth. To assess the impact of each variable, we residualize the predicted match probability by the non-focal string edit distances and plot resulting residual against the string edit distance for the focal variable using a local polynomial plot. This approach abstracts from the variety of comparators used in the model to assess similarity for the same pair of variables as well as the inherent non-linear nature of the model given the random forest specification, but should generally capture the first order contribution of each component of PII.

Perfect or close to perfect alignment on date of birth has the strongest overall contribution to statistical match probability. Once two or more edits are necessary to align a given pair of dates of birth, there is no relative contribution to match probability. There are a variety of reasons that contribute to this pattern. First, exact dates of birth are highly unique within the population. Second, there is no naturally occurring reason why a date of birth should change over time (as opposed to a name change or nickname), making it a more reliable

predictor of match status. Finally, although speculative, numeric information may be less prone to data entry error again making this a better predictor.

On the name components, similarity on last name followed by first name followed by middle name have strongest predictive power for match status. Last names are more unique in the United States compared to first names, which have a tendency to cluster around commonly occurring names. Both last and first names though exhibit a degree linearity with more edits further decreasing the likelihood of a statistical match. Middle name, however, has minimal to no contribution to match status after 3 or more edits are necessary to align the pair of variables, likely reflecting the fact that middles names are irregularly collected, often receiving less oversight than other components of PII.

5.2 Decomposing model performance and assessing demographic heterogeneity

We compare the DE-RF model's performance to three alternatives in Table 5, which include standard practices in the economics literature: (1) a deterministic model, (2) a random forest model trained on 5,000 hand-coded paired observations, and (3) a random forest model trained on the same 5,000 paired observation sample but using the biometrical identifier.[24] The point in making these comparisons is to respectively highlight factors that together contribute to the overall success of the DE-RF model: (1) model flexibility, (2) elimination of potential human bias in the training sample, and (3) depth of training data. We also examine how performance changes overall as well as across various demographic groups (race/ethnicity, sex, and birth decade).

The deterministic and hand-coded models share similar features. Both strategies yield results with high precision rates and low recall rates, meaning the quality of statistical matches is quite high but many potential matches are missed. In practice, this suggests that both exact matching as well as probabilistic strategies built off of human-driven clerical review may be overly conservative.[25]

The random forest model trained on just 5,000 candidate pairs (the "slim biometric model") represents a sizable shift towards recall at a slight cost to precision. The fingerprint-based measure of true match status pushes the model to identify more marginal candidate pairs as statistical matches, increasing recall. Overall, the deterministic, hand-coded, and the slim

[24]See Appendix A for details on our hand-coding procedure.

[25]A review of the discrepancy between the hand-coded and biometric match statuses indicate that the reviewers systematically favored name similarity over date of birth similarity, which consequently lead to both false positive and false negatives. This lines up with the results from Figure 4 that date of birth information is more uniquely identifying that other components of PII.

biometric model deliver F-statistics that are roughly similar, falling into the 0.8 to 0.9 range across the various demographic subgroups.

The DE-RF model performs quite well relative to these three comparison models. There is a 12 to 13 percentage point improvement on recall relative to the deterministic and hand-coded strategies. When compared to the "slim biometric model," the DE-RF model achieves an even greater improvement in recall at a similarly small cost of 1 to 2 percentage points of precision. The improvement on recall without substantial penalty to precision indicates the DE-RF is better able to predict match status and sort candidate pairs accordingly.

Figure 5 further investigates performance gains as training sample size is increased. This figure shows the convergence of out-of-sample model performance as the size of the training sample is increased incrementally from 5,000 training observations up to 1 million training observations.[26] Models trained on fewer than 250,000 observations show a surprising degree of inconsistency, especially regarding precision when using 50,000 training observations or fewer. One challenge these models face is insufficient coverage of marginal match and non-match training pairs in order to identify the optimal statistical match threshold. Even so, performance gains accrue at each larger sample size, pushing the production frontier higher in terms of both recall and precision, demonstrating the benefit of combining highly non-linear machine learning models with large sample sizes.

Returning to Table 5, the DE-RF model is the clear choice in the sample overall and for all demographic subgroups. Interestingly, the largest improvements are observed for demographic groups with the lowest baseline statistics (e.g. female, Hispanic, 1960's births) from the deterministic model. As a result, match rate statistics across various demographic subgroups exhibit lower variance than traditional strategies yield. We will return to this theme in Section 6 where we discuss the implications of match quality for causal inference.

5.3 Assessing performance degradation in external applications

One contribution of this paper is practical in nature. We have designed and estimated a model that could be applied to other settings where quality training data may not be available. For example, the model could be used to match education records (Zimmerman, 2019), credit bureau records (Miller *et al.*, 2020), home financing records (Cloyne *et al.*, 2019) or health

[26]For this specific exercise, 16.6 million observations of the total 17.6 million blocked matched pairs were selected at random to be eligible for use in the training sample; the remaining 1 million observations were held back as out-of-sample testing data. 100 independent models were estimated for each given level of training data, with training observations selected at random (with replacement) from the 16.6 million pool of eligible pairs in order to gauge the speed of model convergence.

records (Duggan *et al.*, 2018). To the extent that a given target application resembles the Texas criminal justice system, the algorithm should perform well. Whether the model works in dissimilar populations remains an open question.

We develop three exercises that tests the limits of the model. In the first, we take the universe of prisoners incarcerated on July 1, 2017 from 9 states[27] (excluding Texas where the training data comes from), run the deduplication, and measure the number of false positives created by the model. Because we know each record is from a distinct individual on that day, matching the data to itself can only produce false positives. The goal of the exercise is to assess the performance in non-Texas criminal justice settings.

The second exercise attempts a one-to-one match among voter registration records in the state of Washington from 2008 to 2012. This is a special case of deduplication, and is particularly relevant for social scientists linking individuals across multiple survey waves. Voter registration IDs create a measure of true match status while the PII retains its original non-synchronized values, meaning there is variation of PII within voter IDs. The goal of this application is to assess model performance in a non-criminal justice setting.

The final exercise selects all deaths in the United States from 2000 to 2009 as captured in the Social Security Administration's Master Death File. We apply a corruption algorithm that introduces phonetic, typographic, and nickname errors into the data and try to reconcile the corrupted files with their original source observations using the matching algorithm.[28] Our focus in this exercise is testing model degradation under increasingly large sample sizes. With a fixed set of names and dates of birth, large populations present a particular challenge as there is increasing risk that any given entity has an exact or close match in PII with another entity. As the PII space becomes more crowded, it becomes increasingly difficult to differentiate true matches from true non-matches.

Table 6 shows the results of these three exercises. Out of 330,756 inmates incarcerated on July 1, 2017 in non-Texan prisons that we can track, we create 463,969 blocked pairs which generate to 2001 predicted statistical matches, or a 0.00% false positive rate (0.4% if conditioning on being the the blocked pair sample).[29] Fewer than 1 percent of the statistically generated identifiers are in two places at the same time.

The Washington voter registration experiment pushes our algorithm further along a number of dimensions. The population is more demographically representative of the general

[27]The nine states are Arkansas, Connecticut, Florida, Illinois, Michigan, Mississippi, North Carolina, Nebraska and Ohio.

[28]See Appendix C for a more detailed description of the corruption algorithm.

[29]The effective false positive rate in the data overall is 0.00%, but this is a relatively meaningless statistic.

population and larger overall than the criminal justice records in our main results (7,551,570 registration records from 2008 and 2012 combined). This latter issue can be quite challenging as with a larger population, there can be higher density in the space of PII, making it more difficult to differentiate marginal true positives from marginal false positives. In spite of these challenges, we observe precision at 0.92, recall at 0.88, and a combined f-statistics at 0.90. A degree of performance loss is to be expected as there are more women in this general population dataset, who are harder to link based on higher rates of naturally occurring legal name changes compared to men.

The final exercise scales up the issue of PII density to the national scale using records from the national Master Death File. Based on 20,298,659 unique deaths between 2000 and 2009, we generate roughly 4 million corrupted records, bringing the total sample for the exercise up to 24,300,530 records. If the algorithm is working properly, the statistical matches will be able to link the corrupted records back to their unique source information without also being linked to other, unrelated individuals. The table reports promising performance statistics: 0.97 precision, 0.93 recall, and a combined 0.95 f-statistic. This suggests that scaling up the potential applications well beyond the original training data is feasible, in spite of the lack of uniqueness in names and dates of birth in the general population.

6 Data Simulation

In this final section, we conduct two groups of simulation exercises to examine how recall and precision errors can impact estimated treatment effects, and how these biases relate directly to the concepts of external and internal validity in causal inference. The first scenario considers a research setting where a matched record is an indication that an outcome has occurred (e.g., recidivism, employment, or public program take-up) for an individual.[30] In the second setting, the analysis sample itself is conditioned on being matched because a given outcome is only observed in the linked data. Examples include studying the impact of an intervention on wage effects among those who file taxes, health care utilization among those with Medicaid coverage, or consumer behavior among those holding a specific brand of credit card.

For the first scenario, we use the following data generating process:

$$y_i = \mathbb{1}\left(\beta d_i + \epsilon_i > F^{-1}(\mu)\right)$$

[30] Tahamont *et al.* (2019) provides an example of how conservative deterministic matching techniques can bias estimated treatment effects in a randomized control trial.

where, outcome y_i is a function of individual i's treatment status (d_i) and a random shock term ($\epsilon_i \sim N(0,1)$). The outcome is normalized by taking the inverse standard normal CDF of the parameter μ, which sets the average rate of the outcome (i.e. the match rate) in the non-treated control group.

The econometrician is interested in estimating the following linear probability model:

$$y_i = \Delta d_i + \nu_i$$

but, only observes $\tilde{y}_i$ which is contaminated by both problems of recall and precision. To operationalize these ideas, we introduce two match quality shock terms: ρ_i, $\pi_i \in U(0,1)$.

$$\tilde{y}_i = \begin{cases} 0 & \text{if } y_i = 1 \quad \& \quad \rho_i \geq \bar{\rho} \\ 1 & \text{if } y_i = 0 \quad \& \quad \pi_i \geq \bar{\pi} \\ y_i & \text{otherwise} \end{cases}$$

where matched outcome $y_i = 1$ is replaced with 0 creating a false negative if the recall shock (ρ_i) exceeds the recall threshold of $\bar{\rho}$. Similarly, the match outcome $y_i = 0$ is replaced with 1 creating a false positive if the precision shock (π_i) exceeds the precision threshold of $\bar{\pi}$. This setup allows us to examine the potential interactions of better and worse match quality on these two important dimensions simultaneously.

We conduct 1,000 empirical simulations of this model, where d_i is assigned at random (i.e. $d_i \perp \epsilon_i, \quad \nu_i$) to 50 percent of 5,000 observations. For each individual simulation, we estimate a number of distinct parameterizations, cycling over a control outcome mean (μ) of 0.25, 0.50, and 0.75, a β of 0.05, 0.10, and 0.25, a recall threshold ($\bar{\rho}$) ranging from 0.50 to 1.00, and a precision threshold ($\bar{\pi}$) ranging from 0.50 to 1.00.

Figures 6 and 7 report the average estimated $\hat{\Delta}$ and corresponding p-value testing the null hypothesis that $\Delta = 0$ over the 1,000 independent simulations. Worse precision and recall rates bias estimates of $\hat{\Delta}$ towards zero systematically,[31] and impair statistical precision increasing the likelihood that there is a failure to reject the null hypothesis. With larger control means and low precision rates, there is increased likelihood of actually flipping the sign of $\hat{\Delta}$ and rejecting the null hypothesis. Note the saddle-like shape in the bottom row of Figure 7, where depending on precision and recall parameters, the same model will lead to rejecting the null in favor of both positively and negatively signed $\hat{\Delta}$'s.

For the second scenario, we use the following data generating process, which introduces

[31] An unbiased measure of $\hat{\Delta}$ is included in the top right hand corner of each plot where precision and recall rates are both 100% and there is effectively no data corruption in place.

a covariate ($x_i \sim N(0,1)$) into the model resulting in heterogenous treatment effects of the intervention:

$$y_i = \mu + \beta(d_i - d_i \times x_i) + \gamma x_i + \epsilon_i$$

Again, the econometrician is interested in estimating the linear model ($y_i = \Delta d_i + \nu_i$), but only observes $\tilde{y}_i$ which is contaminated by both problems of recall and precision. In this setting, we operationalize the match quality problems in the following way:

$$\tilde{y}_i = \begin{cases} missing & \text{if } \rho_i \geq \bar{\rho} \\ y_{\tilde{i}} & \text{if } \pi_i \geq \bar{\pi} \\ y_i & \text{otherwise} \end{cases}$$

Because the outcome now is dependent on the match in the first place, low recall rates will result in a larger share of the outcome data being missing and reducing the sample size consequently. The term $y_{\tilde{i}}$ represents a completely different draw of the y_i outcome from the population distribution (both in terms of d_i, x_i, and ϵ_i) in order to align with thought experiment that a record has matched to the outcome database, but simply randomly matched to the wrong row.

We also allow the correlation of x_i with ρ_i and π_i to be positive, creating a scenario where those least likely to benefit from a given intervention are most likely to face issues in match quality. As we saw in Section 5.2, match quality does vary by key demographic traits that in many settings drive heterogeneous response to interventions, making this setup uncontrived.

Like the first scenario, we conduct 1,000 empirical simulations of this model, where d_i is assigned at random (i.e. $d_i \perp \epsilon_i$, ν_i) to 50 percent of 5,000 observations. For each individual simulation, we estimate a number of distinct parameterizations, cycling over a control outcome mean (μ) of 0.25, 0.50, and 0.75,[32] a β of 0.05, 0.10, and 0.25, a recall threshold ($\bar{\rho}$) ranging from 0.50 to 1.00, and a precision threshold ($\bar{\pi}$) ranging from 0.50 to 1.00.

Figures 8 and 9 report the average estimated $\hat{\Delta}$ and corresponding p-value testing the null hypothesis that $\Delta = 0$ over the 1,000 independent simulations. Due to the heterogeneous treatment effects and the correlation of demographic traits with match quality, lower recall rates exclude those least likely to benefit from the intervention resulting in estimates that exaggerate the average treatment effect of d_i. As the estimate of $\hat{\Delta}$ is pushed higher, it is more likely to reject the null hypothesis, which could facilitate a more opaque form of data mining in social science. The exclusion of these records though from the empirical analysis

[32]Because the change in μ is essentially just a level shift in the regression intercept, we should not expect this to create meaningfully different patterns across the simulations.

exactly invokes the challenge of external validity, creating an internally valid estimate that just does not apply to the population overall.

Worse precision operates similarly to the first experiment, where lower precision rates bias the estimated $\hat{\Delta}$ closer to zero and reduce statistical precision.

Across both sets of thought experiments, a wide range of match quality parameterizations are considered. In practice, it may be unrealistic to think that moving from a 50% recall rate and precision rate to the full elimination of match quality errors is a feasible improvement. In our setting (Table 5), we observe several groups that experience recall improvements on the order of 20 percentage points going from deterministic matching (which is still common in the literature) to our proposed DE-RF model without meaningful sacrifice to precision. As seen in the figures, this can have meaningful implications for both bias in the estimation of treat effects as well as precision in evaluating null hypotheses.

7 Conclusion

This paper addresses the increasingly common challenge of integrating individual-level records from disparate administrative datasets for the purposes of cutting edge social science research. We leverage a novel source of variation, millions of fingerprint-based biometric identifiers, to train a flexible machine learning-based entity resolution model that outperforms a variety of standard practices in the literature. Evidence suggests continuing returns to utilizing a large training sample well beyond current recommendations in the literature.

We show how the model's performance extrapolates to non-criminal justice contexts, including settings with significantly more records which could in principle reduce performance due to crowding in the PII space. While there are many theoretical reasons why we should observe performance degradation, the model manages to yield match rates at or exceeding our baseline results, suggesting broader potential returns to the model through a range of fields of economic research that rely on linked administrative records.

Model simulations connect the statistical matching performance criteria of precision and recall to the concepts of external and internal validity in causal inference. This is especially important given the documented inconsistent performance of standard matching techniques across demographic groups, where individuals with limited naturally occurring name variation or name confusion (e.g. white men) are easiest to match. Without affording a more flexible matching strategy, results may be biased towards these demographic groups depending on the exact model specification.

Future work is needed to further test the limits of the model's effectiveness, including its ability to successfully differentiate non-deceased individuals in the full national population in the United States, for which there is no public roster currently available. That said, this research represents an important first step in bringing discipline to an increasingly common aspect of empirical social science research in the U.S.

References

Abowd, J. M., Abramowitz, J., Levenstein, M. C., McCue, K., Patki, D., Raghunathan, T., Rodgers, A. M., Shapiro, M. D. and Wasi, N. (2019). *Optimal Probabilistic Record Linkage: Best Practice for Linking Employers in Survey and Administrative Data*. Working Papers 19-08, Center for Economic Studies, U.S. Census Bureau.

Abramitzky, R., Boustan, L. P. and Eriksson, K. (2012). Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *The American Economic Review*, **102** (5), 1832–1856.

—, — and — (2014). A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy*, **122** (3), 467–506.

—, —, —, Feigenbaum, J. J. and PÃ©rez, S. (2019). *Automated Linking of Historical Data*. Working Paper 25825, National Bureau of Economic Research.

Ahrens, A., Hansen, C. B. and Schaffer, M. E. (2018). LASSOPACK: Stata module for lasso, square-root lasso, elastic net, ridge, adaptive lasso estimation and cross-validation. Statistical Software Components, Boston College Department of Economics.

Alessi, L. and Detken, C. (2018). Identifying excessive credit growth and leverage. *Journal of Financial Stability*, **35**, 215 – 225, network models, stress testing and other tools for financial stability monitoring and macroprudential policy design and implementation.

Bailey, M., Cole, C., Henderson, M. and Massey, C. (2017). *How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data*. Working Paper 24019, National Bureau of Economic Research.

Biewen, M., Fitzenberger, B., Osikominu, A. and Paul, M. (2014). The effectiveness of public-sponsored training revisited: The importance of data and methodological choices. *Journal of Labor Economics*, **32** (4), 837–897.

Breiman, L. (2001). Random forests. *Machine Learning*, **45** (1), 5–32.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, **95** (3), 759 – 771.

Chetty, R. and Hendren, N. (2018a). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*. *The Quarterly Journal of Economics*, **133** (3), 1107–1162.

— and — (2018b). The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates*. *The Quarterly Journal of Economics*, **133** (3), 1163–1228.

—, — and Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *The American Economic Review*, **106** (4), 855–902.

Christen, P. (2012). Data matching.

— and Churches, T. (2002). *Febrl - Freely extensible biomedical record linkage*. Tech. rep.

Cloyne, J., Huber, K., Ilzetzki, E. and Kleven, H. (2019). The effect of house prices on household borrowing: A new approach. *American Economic Review*, **109** (6), 2104–36.

Dahis, R., Nix, E. and Qian, N. (2019). *Choosing Racial Identity in the United States, 1880-1940*. Working Paper 26465, National Bureau of Economic Research.

Doherr, T. (2018). Brain: Stata module to provide neural network.

Duggan, M., Gruber, J. and Vabson, B. (2018). The consequences of health care privatization: Evidence from medicare advantage exits. *American Economic Journal: Economic Policy*, **10** (1), 153–186, date revised - 2017-12-01; Availability - URL:http://www.aeaweb.org.proxy.lib.umich.edu/aej-policy/] Publisher's URL; Last updated - 2018-03-01.

Dusetzina, S. B., Tyree, S., Meyer, A.-M., Meyer, A., Green, L. and Carpenter, W. R. (2014). *Linking Data for Health Services Research: A Framework and Instructional Guide*. Tech. Rep. 14-EHC033-EF.

Feigenbaum, J. J. (2016). *A Machine Learning Approach to Census Record Linking*. Tech. rep.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64** (328), 1183–1210.

Ferrante, A. and Boyd, J. (2012). A transparent and transportable methodology for evaluating data linkage software. *Journal of Biomedical Informatics*, **45** (1), 165 – 172.

Ferrie, J. P. (1996). A new sample of males linked from the public use microdata sample of the 1850 u.s. federal census of population to the 1860 u.s. federal census manuscript schedules. *Historical Methods*, **29** (4), 141, last updated - 2013-02-23.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K. and Group, O. H. S. (2012). The oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics*, **127** (3), 1057–1106.

Fortini, M., Liseo, B., Nuccitelli, A. and Scanu, M. (2001). On bayesian record linkage. *Research in official statistics*, **4**, 185–198.

Galbally, J., Haraksim, R. and Beslay, L. (2019). A study of age and ageing in fingerprint biometrics. *IEEE Transactions on Information Forensics and Security*, **14** (5), 1351–1365.

Grogger, J., Ivandic, R. and Kirchmaier, T. (2020). *Comparing Conventional and Machine-Learning Approaches to Risk Assessment in Domestic Abuse Cases*. Tech. Rep. Discussion Paper No 1676, Centre for Economic Performance.

Guenther, N. and Schonlau, M. (2016). Support vector machines. *The Stata Journal*, **16** (4), 917–937.

Hastie, T., Tibshirani, R. and Friedman, J. (2016). The elements of statistical learning.

Heller, S. B., Shah, A. K., Guryan, J., Ludwig, J., Mullainathan, S. and Pollack, H. A. (2016). Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago*. *The Quarterly Journal of Economics*, **132** (1), 1–54.

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** (406), 414–420.

Keely, L. C. and Tan, C. M. (2008). Understanding preferences for income redistribution. *Journal of Public Economics*, **92** (5), 944 – 961.

Keiser, D. A. and Shapiro, J. S. (2018). Consequences of the Clean Water Act and the Demand for Water Quality*. *The Quarterly Journal of Economics*, **134** (1), 349–396.

Lahiri, P. and Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, **100** (469), 222–230.

Larsen, M. D. and Rubin, D. B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, **96** (453), 32–41.

Maltoni, D., Cappelli, R. and Meuwly, D. (2017). Automated fingerprint identification systems: From fingerprintsto fingermarks. In M. Tistarelli and C. Champod (eds.), *Handbook of Biometrics for Forensic Science*, pp. 37 – 61.

Mèray, N., Reitsma, J. B., Ravelli, A. C. and Bonsel, G. J. (2007). Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *Journal of Clinical Epidemiology*, **60** (9), 883.e1 – 883.e11.

Miller, S., Wherry, L. R. and Foster, D. G. (2020). *The Economic Consequences of Being Denied an Abortion*. Working Paper 26662, National Bureau of Economic Research.

Moore, C. L., Gidding, H. F., Law, M. G. and Amin, J. (2016). Poor record linkage sensitivity biased outcomes in a linked cohort analysis. *Journal of Clinical Epidemiology*, **75**, 70 – 77.

Mueller-Smith, M. and T. Schnepel, K. (2020). Diversion in the Criminal Justice System. *The Review of Economic Studies*, rdaa030.

Pankanti, S., Prabhakar, S. and Jain, A. K. (2002). On the individuality of fingerprints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24** (8), 1010–1025.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pérez, A., Larrañaga, P. and Inza, I. (2009). Bayesian classifiers based on kernel density estimation: Flexible classifiers. *International Journal of Approximate Reasoning*, **50** (2), 341 – 362, special Section on The Imprecise Dirichlet Model and Special Section on Bayesian Robustness (Issues in Imprecise Probability).

Price, J., Buckles, K., Van Leeuwen, J. and Riley, I. (2019). *Combining Family History and Machine Learning to Link Historical Records*. Working Paper 26227, National Bureau of Economic Research.

Sadinle, M. and Fienberg, S. E. (2013). A generalized fellegi-sunter framework for multiple record linkage with application to homicide record systems. *Journal of the American Statistical Association*, **108** (502), 385–397.

Sayers, A., Ben-Shlomo, Y., Blom, A. W. and Steele, F. (2015). Probabilistic record linkage. *International Journal of Epidemiology*, **45** (3), 954–964.

Scheuren, F. and Winkler, W. (1997). Regression analysis of data files that are computer matched - part ii. *Survey Methodology*, **23**.

— and Winkler, W. E. (1993). Regression analysis of data files that are computer matched - part i. *Survey Methodology*, **19** (1), 39–58.

Setoguchi, S., Zhu, Y., Jalbert, J., Williams, L. A. and Chen, C.-Y. (2014). Validity of deterministic record linkage using multiple indirect personal identifiers: Linking a large registry to claims data. *Circulation: Cardiovascular Quality & outcomes*, **7** (3), 475–480.

Skiba, P. M. and Tobacman, J. (2019). Do payday loans cause bankruptcy? *The Journal of Law and Economics*, **62** (3), 485–519.

Steorts, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Anal.*, **10** (4), 849–875.

—, Hall, R. and Fienberg, S. E. (2016). A bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, **111** (516), 1660–1672.

Taddy, M., Gardner, M., Chen, L. and Draper, D. (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, **34** (4), 661–672.

Tahamont, S., Jelveh, Z., Chalfin, A., Yan, S. and Hansen, B. (2019). *Administrative Data Linking and Statistical Power Problems in Randomized Experiments*. Working Paper 25657, National Bureau of Economic Research.

Tran, K.-N., Vatsalan, D. and Christen, P. (2013). Geco: An online personal data generator and corruptor. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, New York, NY, USA: ACM, pp. 2473–2476.

Tromp, M., Ravelli, A. C., Bonsel, G. J., Hasman, A. and Reitsma, J. B. (2011). Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. *Journal of Clinical Epidemiology*, **64** (5), 565 – 572.

Vick, R. and Huynh, L. (2011). The effects of standardizing names for record linkage: Evidence from the united states and norway. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, **44** (1), 15–24.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, **113** (523), 1228–1242.

Watson, C., Fiumara, G., Tabassi, E., Cheng, S. L., Flanagan, P. and Salamon, W. (2014). *Fingerprint Vendor Technology Evaluation*. Tech. Rep. NISTIT 8034, National Institute of Standards and Technology.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage.

Wisselgren, M. J., Edvinsson, S., Berggren, M. and Larsson, M. (2014). Testing methods of record linkage on swedish censuses. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, **47** (3), 138–151.

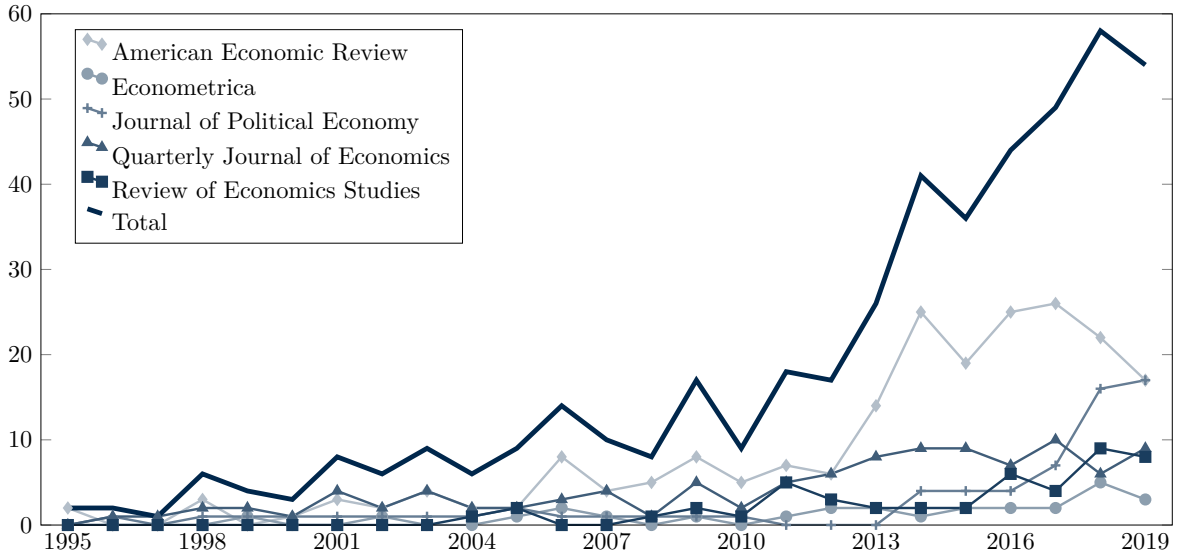
Yagan, D. (2019). Employment hysteresis from the great recession. *Journal of Political Economy*, **127** (5), 2505–2558.

Yoon, S. and Jain, A. K. (2015). Longitudinal study of fingerprint recognition. *Proceedings of the National Academy of Sciences of the United States of America*, **112** (28), 8555–8560.

Zimmerman, S. D. (2019). Elite colleges and upward mobility to top jobs and top incomes. *American Economic Review*, **109** (1), 1–47.

Figures

Figure 1: Total publication mentioning "administrative data" in the top 5 economic journals, 1995-2019.



Note: The figure was compiled by searching the top 5 economic journals for papers that contain the exact phrase "administrative data." We used search functions provided by Oxford Journals, JSTOR, Wiley Online Library and University of Chicago Press to cover the relevant journals and years.

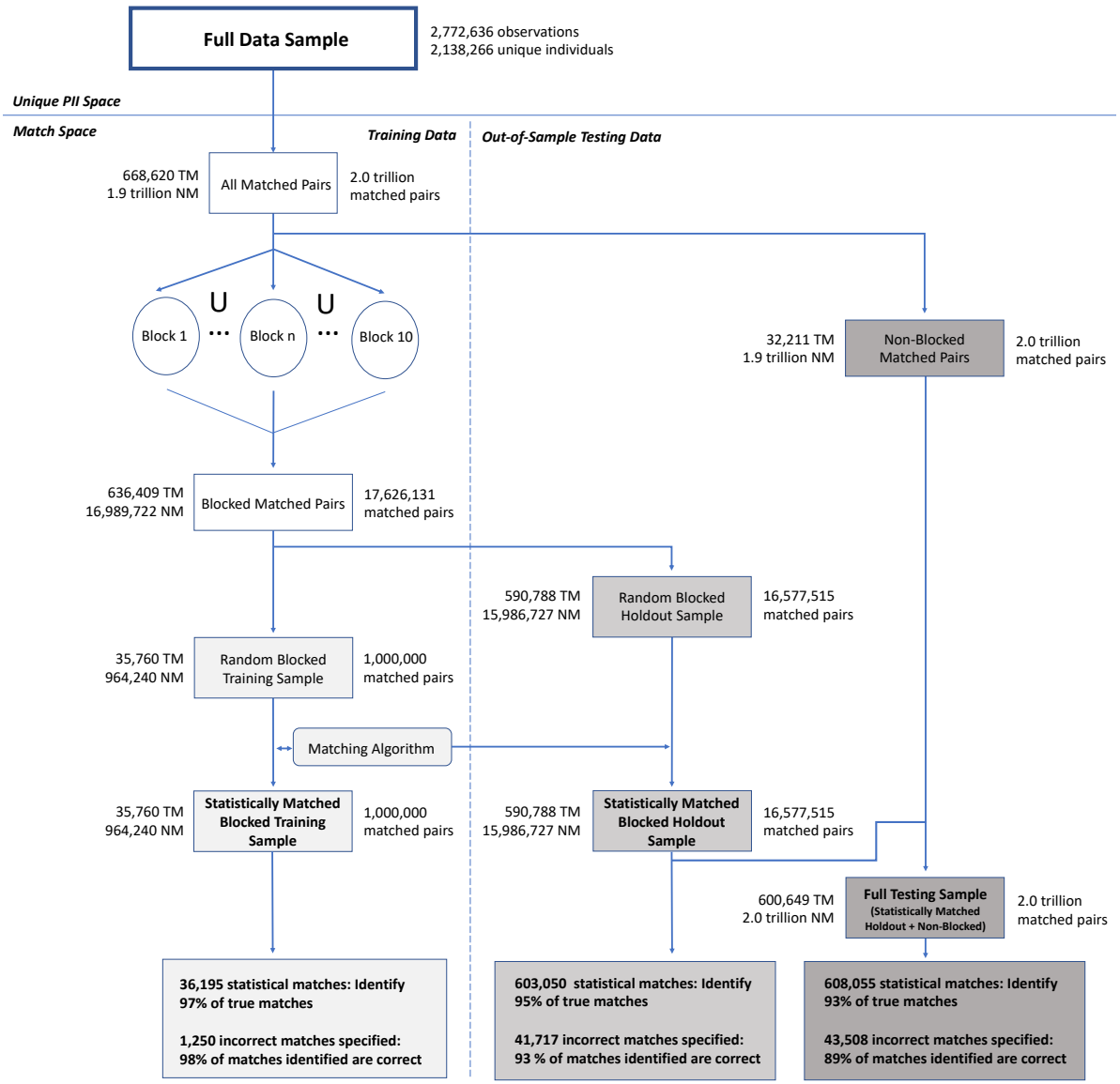
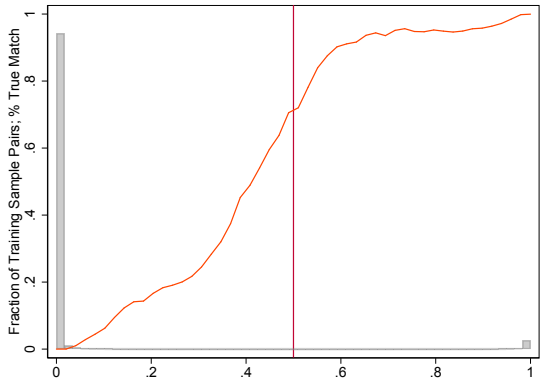
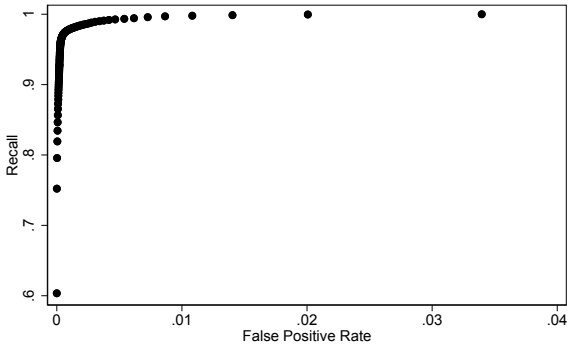


Figure 2: Model Training and Testing Overview

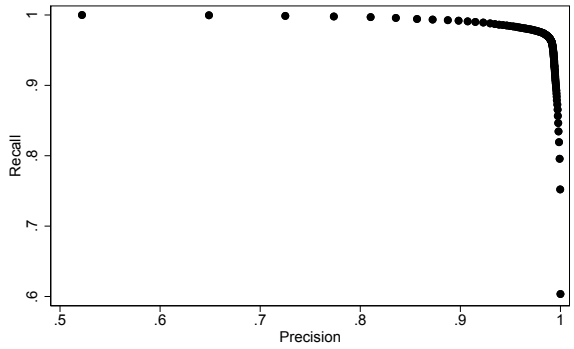
Notes: TM - True Match; NM - True Non-Match. Performance statistics based on the demographic enhanced random forest model described in Section 4. Starting with the original court and inmate data, this flow chart shows how the model is trained and tested to generate out of sample predictions. The blocking strategy cuts down the potential match space from 2 trillion to 17.6 million matches at the cost of removing approximately 5% of the total true matches. Once the blocking has identified candidate matches, the pairs are split into a training sample and a testing sample. The demographic enhanced random forest algorithm is used to train a predictive model. The recall and precision of the training set is shown on the bottom left box. The results from the testing blocked pairs is shown in the middle gray box, while the full out-of-sample results (including pairs that are not matched together) are shown in the box on the bottom right.



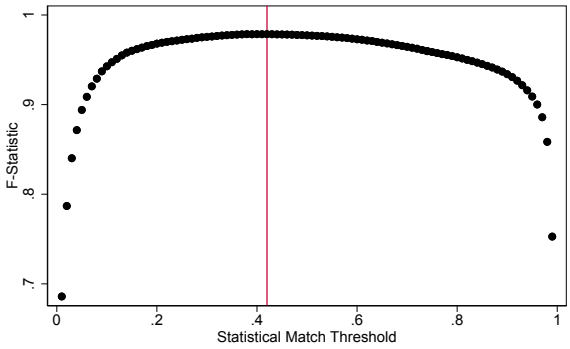
(a) Histogram of match probability index and underlying true match rate



(b) Receiver operating characteristic (ROC) curve



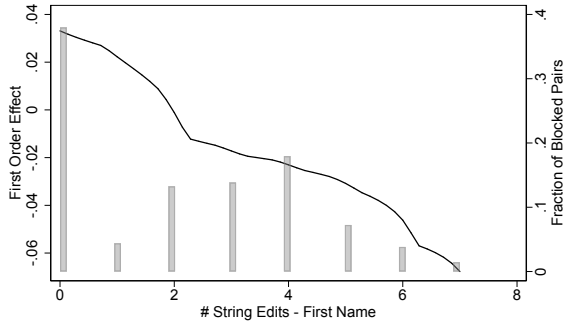
(c) Recall/precision tradeoff curve



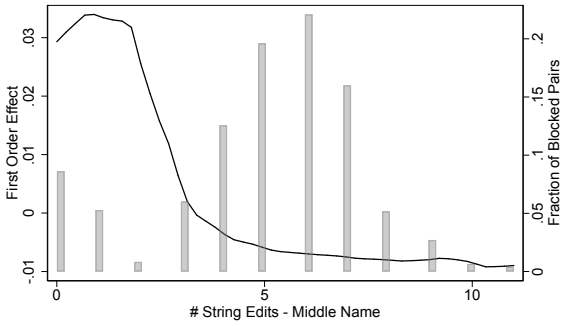
(d) Implied F-Statistic at varying threshold values

Figure 3: Diagnostic performance for varying statistical match thresholds

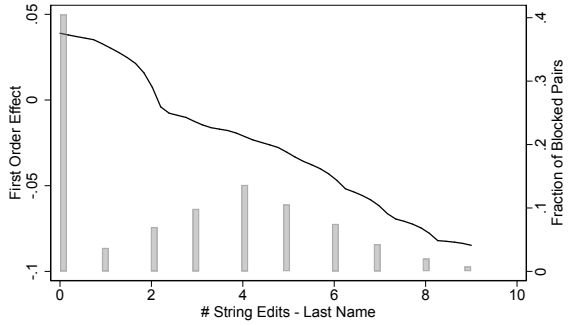
Panel a shows that the probability of correctly classifying a match increases with the underlying true match rate, though the increase levels off around a true match rate of 0.6. Panels b and c show the ROC curve and precision vs. recall curves, respectively. These plots illustrate the tradeoffs between conservative and aggressive matching thresholds. Panel d illustrates the maximization process used to select the optimal match threshold. The red line indicates the statistical match threshold that maximizes the F-statistic in the training sample.



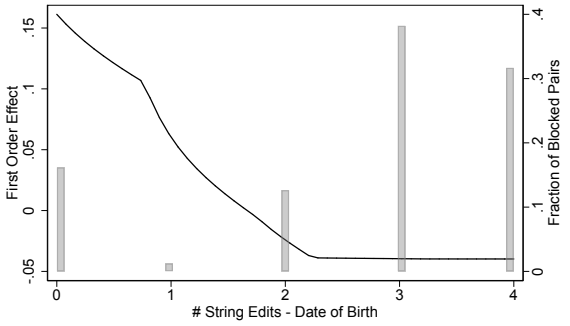
(a) First Name



(b) Middle Name



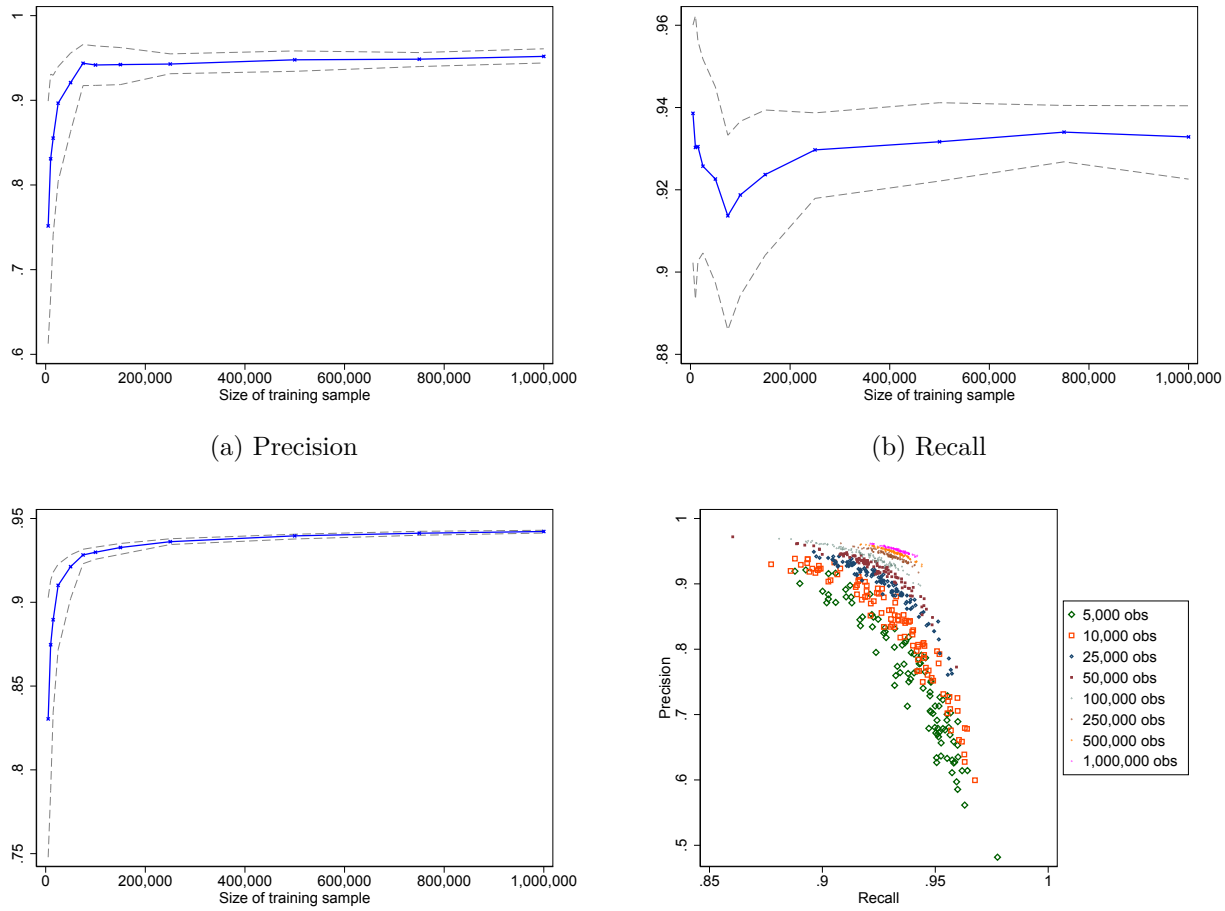
(c) Last Name



(d) Date of Birth

Figure 4: First order impact on predicted match probability

Panel a shows the first derivative first name raw edits on the predicted match probability, indicating that there is a non-linear and decreasing relationship between the first name edit distance and predicted match probability. Blocked pairs with the same first name are a predicted match between 3 and 4% of the time. Panel b shows the same first derivative but for the number of middle name edits, indicating that the relationship is cubic. Panel c shows that the predicted match status decreases with the number of last name edits, and panel d shows that predicted match status decreases with date of birth (string) edits.



(a) Precision

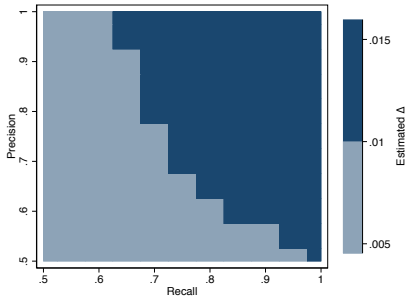
(b) Recall

(c) F-Statistic

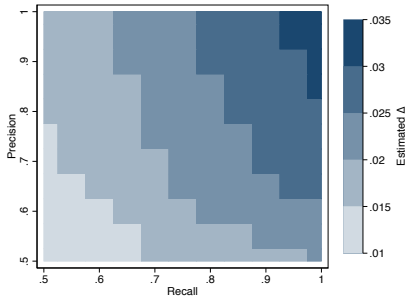
(d) Performance Distribution

Figure 5: Convergence of model performance as training sample increases

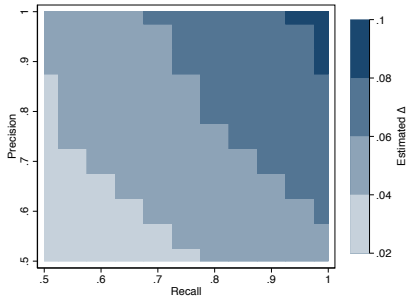
This figure shows the convergence of out-of-sample model performance as the size of the training sample is increased from 5,000 training observations up to 1 million training observations. 16.6 million observations of the total 17.6 million blocked matched pairs were selected at random to be eligible for use in the training sample; the remaining 1 million observations were held back as out-of-sample testing data for this exercise. 100 independent models were estimated for each given level of training data, with training observations selected at random (with replacement) from the 16.6 million pool of eligible pairs. Panels a, b, and c show the change in average as well as 5th/95th percentile model performance as the sample size grows. Panel d shows the full set of precision and recall results per bootstrapped sample for a subset of training sample levels evaluated.



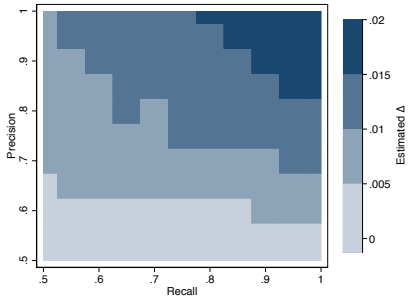
(a) Control Mean = 0.25; $\beta = 0.05$



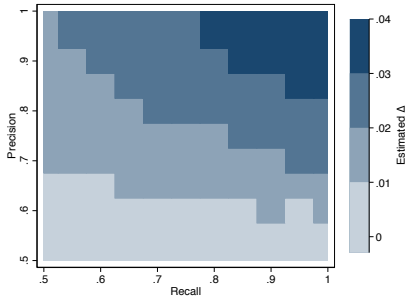
(b) Control Mean = 0.25; $\beta = 0.10$



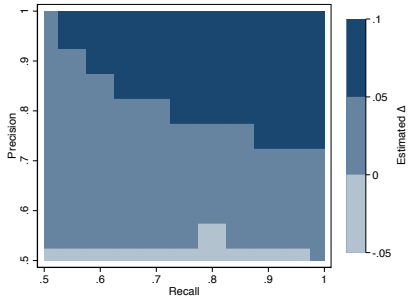
(c) Control Mean = 0.25; $\beta = 0.25$



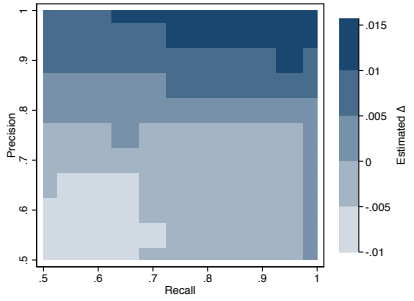
(d) Control Mean = 0.50; $\beta = 0.05$



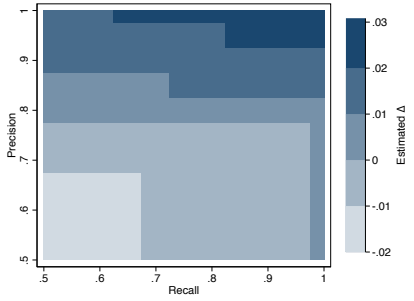
(e) Control Mean = 0.50; $\beta = 0.10$



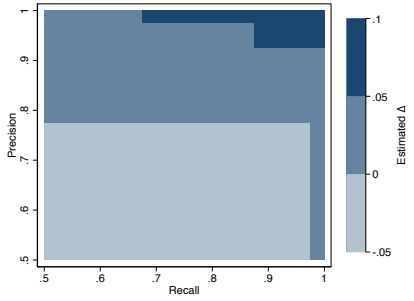
(f) Control Mean = 0.50; $\beta = 0.25$



(g) Control Mean = 0.75; $\beta = 0.05$



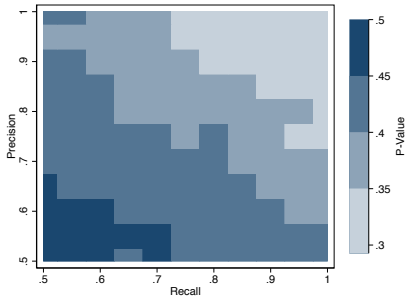
(h) Control Mean = 0.75; $\beta = 0.10$



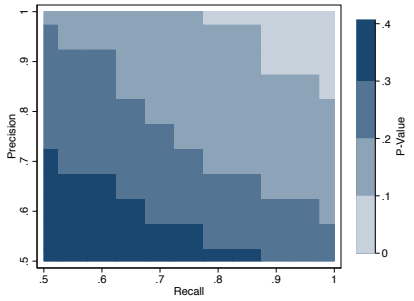
(i) Control Mean = 0.75; $\beta = 0.25$

Figure 6: Average estimated $\hat{\Delta}$ over 1,000 simulation runs with varying model parameterizations (Scenario 1)

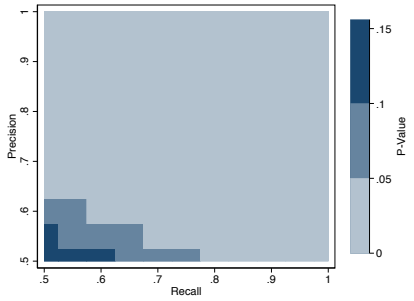
This figure reports the average estimated Δ over 1,000 independent simulations described in 6. The figure shows that worse precision and recall rates bias estimates of $\hat{\Delta}$ towards zero systematically.



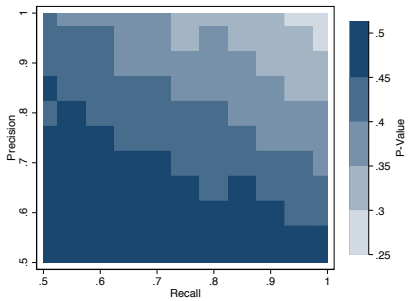
(a) Control Mean = 0.25; $\beta = 0.05$



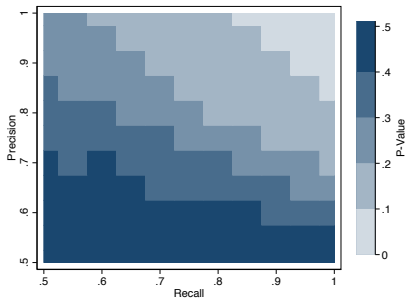
(b) Control Mean = 0.25; $\beta = 0.10$



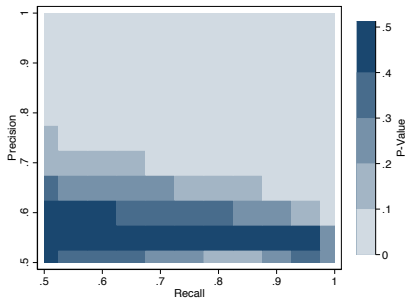
(c) Control Mean = 0.25; $\beta = 0.25$



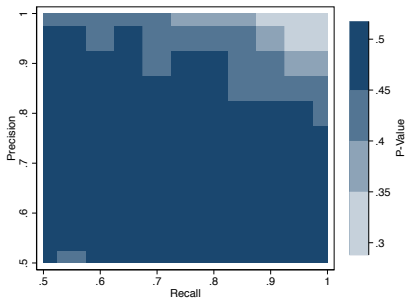
(d) Control Mean = 0.50; $\beta = 0.05$



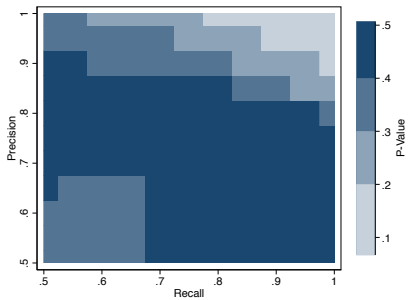
(e) Control Mean = 0.50; $\beta = 0.10$



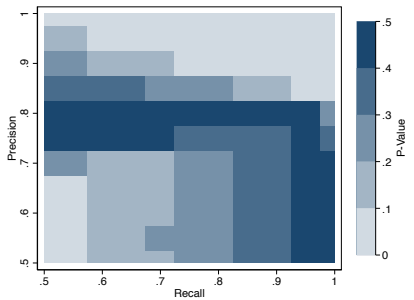
(f) Control Mean = 0.50; $\beta = 0.25$



(g) Control Mean = 0.75; $\beta = 0.05$



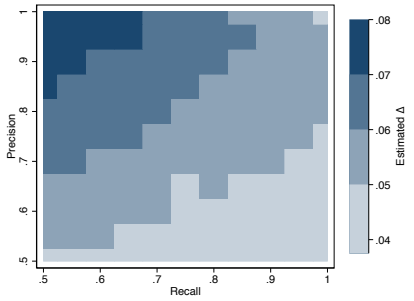
(h) Control Mean = 0.75; $\beta = 0.10$



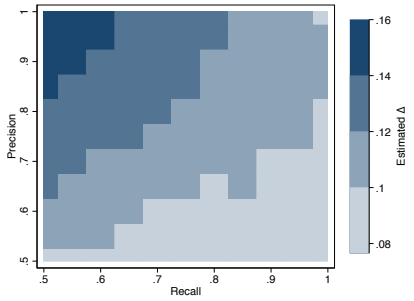
(i) Control Mean = 0.75; $\beta = 0.25$

Figure 7: Average estimated p-value over 1,000 simulation runs with varying model parameterizations (Scenario 1)

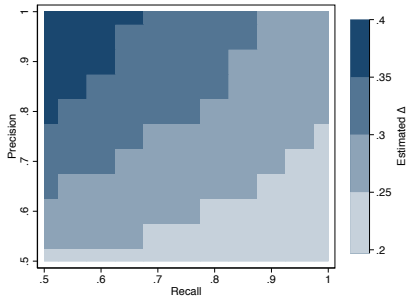
This figure reports the p-value testing the null hypothesis that $\Delta = 0$ over the 1,000 independent simulations described in 6. Worse precision and recall rates impair statistical precision, increasing the likelihood that there is a failure to reject the null hypothesis.



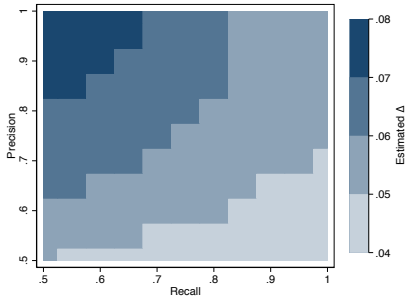
(a) Control Mean = 0.25; $\beta = 0.05$



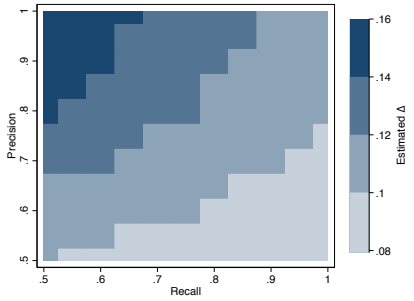
(b) Control Mean = 0.25; $\beta = 0.10$



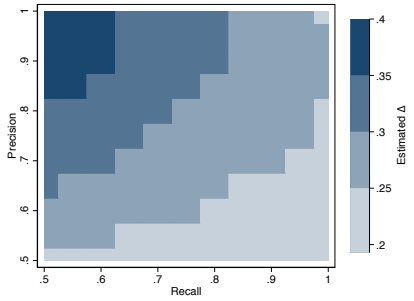
(c) Control Mean = 0.25; $\beta = 0.25$



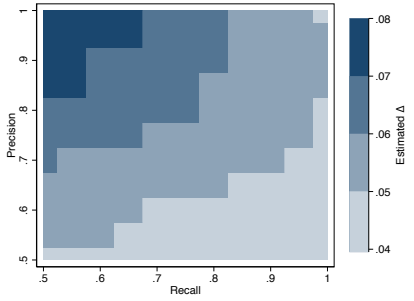
(d) Control Mean = 0.50; $\beta = 0.05$



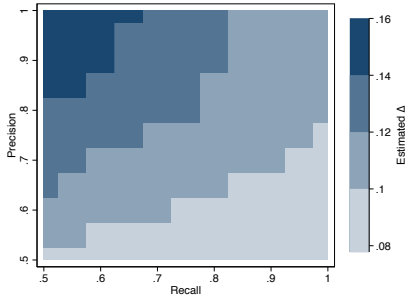
(e) Control Mean = 0.50; $\beta = 0.10$



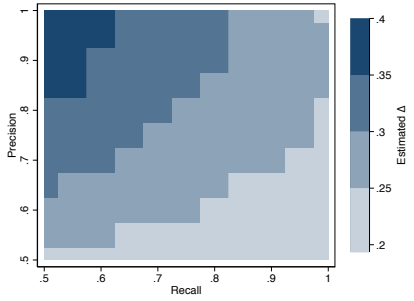
(f) Control Mean = 0.50; $\beta = 0.25$



(g) Control Mean = 0.75; $\beta = 0.05$



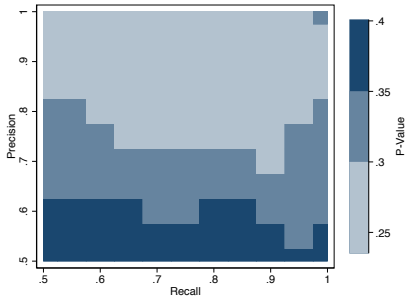
(h) Control Mean = 0.75; $\beta = 0.10$



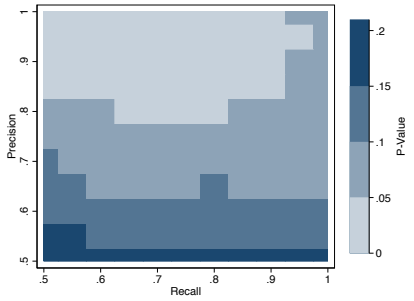
(i) Control Mean = 0.75; $\beta = 0.25$

Figure 8: Average estimated $\hat{\Delta}$ over 1,000 simulation runs with varying model parameterizations (Scenario 2)

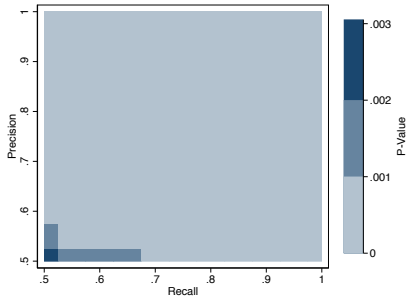
This figure reports the average estimated Δ over 1,000 independent simulations described in 6 for the scenario with heterogeneous treatment effects. Due to the heterogeneity of the treatment effect, decreases in recall lead to systematic overestimates of $\hat{\Delta}$. Similar to scenario 1, decreases in precision bias $\hat{\Delta}$ towards zero systematically.



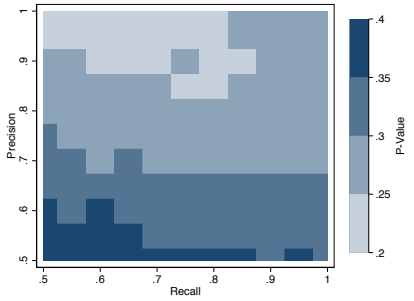
(a) Control Mean = 0.25; $\beta = 0.05$



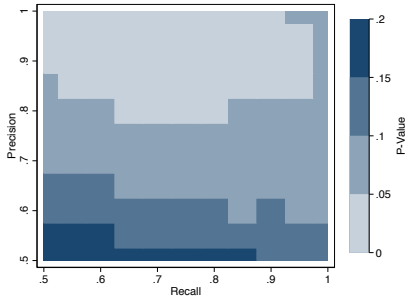
(b) Control Mean = 0.25; $\beta = 0.10$



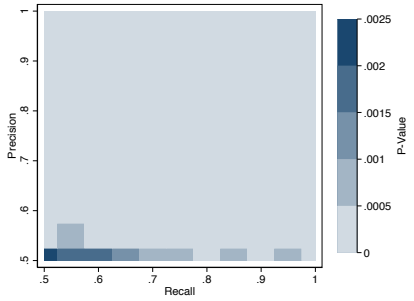
(c) Control Mean = 0.25; $\beta = 0.25$



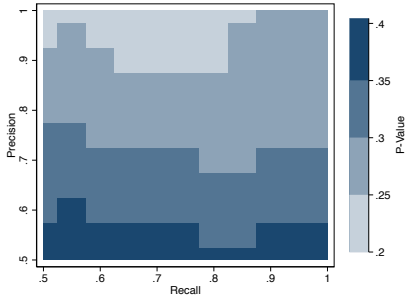
(d) Control Mean = 0.50; $\beta = 0.05$



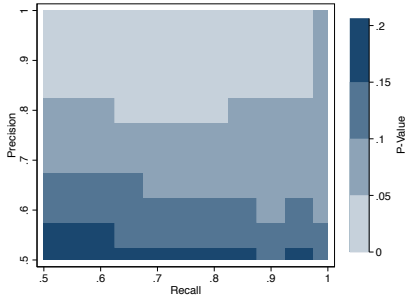
(e) Control Mean = 0.50; $\beta = 0.10$



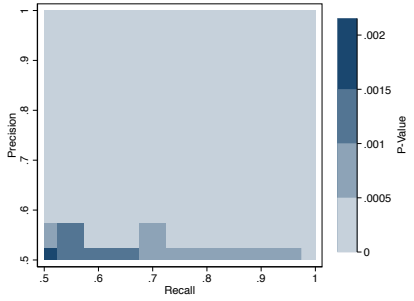
(f) Control Mean = 0.50; $\beta = 0.25$



(g) Control Mean = 0.75; $\beta = 0.05$



(h) Control Mean = 0.75; $\beta = 0.10$



(i) Control Mean = 0.75; $\beta = 0.25$

Figure 9: Average estimated p-value over 1,000 simulation runs with varying model parameterizations (Scenario 2)

This figure reports the p-value testing the null hypothesis that $\Delta = 0$ over the 1,000 independent simulations described in 6 for the scenario with heterogeneous treatment effects. Worse precision and recall rates impact statistical precision though the effects are not systematic.

Tables

Table 1: Matching Strategies Used in 2019 Administrative Data Papers From the "Top 5" Journals

	No Matching Required	Matching Required Not Discussed	Deterministic Matching	Fuzzy Matching	Total Papers
Papers	23	19	10	2	54

The table was compiled by searching the top 5 economic journals for papers that contain the exact phrase "administrative data." We used search functions provided by Oxford Journals, JSTOR, Wiley Online Library and University of Chicago Press to cover the relevant journals and years. Papers published in 2019 in the "top 5" economics journals are classified according to the matching procedure used to link the data. Approximately 40% of the papers do not require matches, while 35% of papers do not explicitly discuss the matching method used to create data. Most deterministic strategies use unique identifiers to execute merges.

Table 2: Summary Statistics of Training Data, External Testing Data, and the General U.S. Population

	TDCJ Inmates 1978-2014	Harris County JIMS 1980-2017	WA Voters 2008 and 2012	Multi-State Prison Snapshot, July 1, 2017	DMF Death Year, 2000	United States Population, 2010
Share Male	0.88	0.80	0.50	0.93	0.48	0.49
Share White	0.36	0.27	0.72	0.44	0.82	0.64
Share Black	0.37	0.33	0.35	0.51	0.08	0.12
Share Hispanic	0.26	0.32	0.11	0.04	0.06	0.16
Average Age	42.5	40.9	37.8	44.2	78.3	37.2
Share in Texas	100.0	100.0	0	0	0.05	0.08
Observations	3,152,630	4,119,621	11,808,233	330,756	8,922,820	308,745,538
Unique IDs	905,530	1,317,315	5,379,888	N/A	2,230,705	
Unique PII Combinations	1,095,054	1,723,008	6,164,621	329,088	4,202,455	N/A

Summary statistics of demographics for all relevant samples. USA data is measured using the 2010 Decennial Census. Washington data is measured using the 2008 and 2012 ACS 1 year sample. The average age is as of April 1, 2010, except in the WA ACS sample where it is the average across the 2008 and 2012 waves. The samples used to train the matching model have a higher proportion of men and people of color than the comparison populations.

Table 3: Description of Individual Blocks

Block	Fraction of True Matches	True Matches Not Included
Date of birth + last soundex	77.9	147,654
Date of birth + first soundex	81.5	123,808
Month of birth + first soundex + last soundex	72.7	182,324
Day of birth + First soundex + last soundex	72.1	186,694
Year of birth + first soundex + last soundex	72.1	186,798
Date of birth + last phonex	77.9	147,761
Date of birth + first phonex	82.1	119,861
Month of birth + first phonex + last phonex	73.2	179,241
Day of birth + First phonex + last phonex	72.5	183,624
Year of birth + first phonex + last phonex	72.5	183,720
Union of Blocks	95.2	32,211

Each row represents a separate block that is used to partition the data. The full match space is generated by taking the union of pairs created across all 10 blocks.

Table 4: Comparison of Out-of-Sample Model Performance

Model	Accuracy	Precision	Recall	F-Statistic	False Positive Rate	Estimation + Prediction Duration (Hours)
Deterministic	0.97	0.93	0.76	0.84	**0.006**	**0.00**
Naive Bayes Classifier (Discrete)	*0.97*	0.90	*0.72*	*0.80*	0.008	0.06
Naive Bayes Classifier (Kernel)	0.97	*0.88*	0.81	0.84	*0.011*	1.42
Support Vector Machine	0.98	**0.94**	0.83	0.88	0.006	*199.55*
Lasso Shrinkage Model	0.98	0.90	0.82	0.86	0.009	21.01
Random Forest	0.98	0.93	0.88	0.90	0.006	0.26
Random Forest (Demog. Enhanced)	**0.98**	0.93	**0.89**	**0.91**	0.007	0.28
Neural Net Perceptron	0.98	0.93	0.85	0.89	0.006	2.11
Neural Net	0.98	0.92	0.88	0.90	0.008	10.13

This table compares performance across a number of classifiers. Because there are roughly 2 trillion true negatives, which swamp comparison of accuracy and false positive rates across models, we limit the ratio of false negatives to true matches at a ratio of 10:1. Otherwise, the accuracy rate for all models would be 1.00 and the false positive rate would be 0.00. In either case, we focus on the precision, recall and F-statistic to differentiate model performance. Numbers in bold indicate the best performance across all models for a given statistic. Numbers in italics represent the worst performance across all models for a given statistic. The demographic enhanced random forest achieves the highest F-statistic and recall rate, and has a precision rate that is slightly lower than the SVM classifier.

Table 5: Demographic-Specific Performance Statistics

		Demographic Enhanced Random Forest		
	Deterministic	5,000 Hand-Coded Training Obs.	5,000 Biometric Training Obs.	1,000,000 Biometric Training Obs.
Panel A: Precision Rates				
Overall	0.93	**0.95**	*0.91*	0.93
Race/Ethnicity				
White	0.96	**0.97**	*0.94*	0.94
Black	0.97	**0.97**	0.96	*0.95*
Hispanic	*0.88*	**0.98**	0.95	0.93
Sex				
Male	*0.92*	**0.96**	0.95	0.94
Female	**0.97**	0.95	*0.83*	0.90
Decade of Birth				
1960s	0.93	**0.94**	*0.89*	0.92
1970s	0.94	**0.96**	*0.93*	0.93
1980s	0.97	**0.97**	*0.93*	0.94
1990s	**0.98**	0.98	*0.95*	0.96
Panel B: Recall Rates				
Overall	*0.76*	0.77	0.84	**0.89**
Race/Ethnicity				
White	0.81	*0.81*	0.89	**0.93**
Black	*0.80*	0.81	0.86	**0.91**
Hispanic	*0.73*	0.74	0.88	**0.93**
Sex				
Male	0.79	*0.77*	0.83	**0.90**
Female	*0.68*	0.76	0.86	**0.88**
Decade of Birth				
1960s	*0.72*	0.72	0.80	**0.86**
1970s	*0.80*	0.82	0.88	**0.92**
1980s	*0.86*	0.88	0.93	**0.96**
1990s	*0.88*	0.90	0.95	**0.98**
Panel C: F-Statistics				
Overall	*0.84*	0.85	0.87	**0.91**
Race/Ethnicity				
White	0.88	*0.88*	0.91	**0.94**
Black	*0.88*	0.89	0.90	**0.93**
Hispanic	*0.84*	0.84	0.92	**0.93**
Sex				
Male	*0.85*	0.87	0.89	**0.92**
Female	*0.80*	0.84	0.84	**0.89**
Decade of Birth				
1960s	*0.81*	0.81	0.84	**0.89**
1970s	*0.86*	0.88	0.90	**0.93**
1980s	*0.91*	0.92	0.93	**0.95**
1990s	*0.93*	0.94	0.95	**0.97**

This table compares performance across a number of classifiers and training data. Entries in bold represent the best performance compared to other models, while entries in italics represent the worst performance across models. The demographic enhanced model performs the best in terms of recall and the overall F-statistic for every demographic group. The model trained with hand-coded training data is more conservative in identifying matches since high precision comes at the expense of low recall. Similarly, the deterministic model is successful at limiting false matches (precision), though is unable to identify true matches as well as the random forest algorithms.

Table 6: Testing Model Performance in External Applications

Application	Accuracy	Precision	Recall	F-Stat.	False Pos. Rate
Multi-State Inmate Snapshot (July 1, 2017)	1.00	–	–	–	0.000
Washington State Voter Records (2008 & 2012)	0.98	0.92	0.88	0.90	0.008
Corrupted Death Master File (2000-2009)	0.98	0.97	0.93	0.95	0.003

Comparison of model performance across a range of external applications. Row 1 refers to the deduplication of all prisoners in incarceration in different states on July 1, 2017. The 2nd row refers to the one-to-one match of Washington state voter records using the 2008 and 2012 voter files. Row 3 refers to the deduplication of the corrupted DMF. For each exercise, we use the baseline random forest model generated from the 1,000,000 observation training sample. For the prisoner deduplication exercise, there are no "true matches" so precision and recall cannot be calculated. The low false positive rate in row 1 suggests that the model is not overly permissive when identifying matches. Rows 2 and 3 suggest that the model performance is dependent on the target data population, though the model performs well in both the Washington voter match and the corrupted DMF match. Because there are an excessive number of true negatives, which swamp the accuracy and false positive rates in each external application, we limit the ratio of false negatives to true matches at a ratio of 10:1 where possible. Since by construction there are no true matches in the July 1 prisoner application, this adjustment is not feasible. Otherwise, the accuracy rate for all models would be 1.00 and the false positive rate would be 0.00.

Appendix Tables

A Generating a hand-coded Sample

Many supervised learning algorithms are estimated using training data created through a process of hand-coding and clerical review. To quantify the benefit of our methodology, we construct a version of the training dataset that we would need to generate in the absence of a biometric ID linking observations.

We take a 5,000 observation random sample of the candidate pairs created from our blocking strategy, and have multiple research assistants code each observation to determine whether the two individuals represent the same person. Approximately 31% of the pairs are from the Harris County Court data and 69% are from the Texas Prison data. For each pair in the random sample, we include the name, date of birth and race of each individual to be used as match variables by the research assistant. For the court data, we also include information about the charge associated with each observation as well as the final disposition. For the prison movement data, we include whether the observation is a prison entry or exit and the date of the movement.

We instruct the RAs to code a match only when they are confident that a given pair represents the same person. Each observation in the 5,000 pair sample is coded independently by 3 analysts. For the final training sample, we take the mode designation for each pair, so if 2 analysts code it as a match, it is considered a match. If only one analyst codes it a match, we consider it a non-match.

Of the 5,000 candidate pairs, the RAs coded 161, or 3.2% as a match. The RAs correctly identified 92% of the "true-matches" while 4% of the hand-coded matches are incorrect, both according to the underlying biometric ID.

B Defining prediction algorithms

Deterministic

The deterministic model represents a conservative, ad-hoc strategy of record linkage based on exact matches. Using first name, last name, middle name and the three components of date of birth (month, day, year), we define a statistical match as any pair that has an exact match on 5 out of 6 non-missing components. As an example, two observations with the same birth date, first name and last name but a different middle name would be considered the same person by the deterministic algorithm.

Naives Bayes Classifier (Discrete and Kernel)

We use Sayers *et al.* (2015) as a template to implement a Naive Bayes Classification (NBC) model using string comparators. This model is functionally equivalent to the one proposed by Winkler (1990) and accounts for typographical errors in matching variables by utilizing a string distance function instead of a binary comparator. String distance comparators allow two strings to get a positive match weight, even if they are not identical. To run the NBC model, we must estimate a set of match weights that determine the odds that a pair is a match given a vector of matching variables. For each continuous comparison variable, we estimate the probability that the comparison variable is a match, conditional on the true match status. We consider partial agreement when a continuous distance metric measured on the [0,1] interval has a value of 0.85 or greater. We use the estimated weights to generate a score for each pair, and set a threshold by maximizing the F-statistic over the score space.

Next, we test a minor variant to the Naive Bayes Classifier, by estimating the conditional distribution of continuous comparison variables through kernel density estimation (Hastie *et al.*, 2016; Pérez *et al.*, 2009). This flexible NBC does not require us to discretize continuous match variables and instead allows to flexibly estimate their conditional probabilities. The match weights for discrete variables are unchanged in this algorithm.

To operationalize the continuous NBC, we estimate kernel density functions for the the distribution of each continuous variable, conditional on match status. This implies that for each variable, we estimate two kernel density distributions: one for the distribution conditional on a match, and the other for the distribution conditional on non-match. We use the Epanechnikov kernel function to estimate the each distribution.

Once we estimate the probability distribution functions for the continuous variables, we are able to construct weights at each point of the distribution by taking the natural log of the P(match)/P(non-match)for each value in the support of the contiuous variable. Once we have match weights for each variable, we aggregate the weights for all variables and determine the optimal threshold by maximizing the F-statistic over the score space.

Support Vector Machine

Support Vector Machine models (SVM) are another type of supervised classification algorithm. SVM models perform classification by using training data to construct a hyperplane that separates the training data into target classes. In ideal applications, the training data can be perfectly separated by a hyperplane; however, in many cases, a perfectly separating boundary is not possible. For example, one could imagine two pairs of observations with the same name and birthday. If one pair represents the same person, while the other pair represents two different people, it would be impossible to construct a hyperplane that would separate these two pairs.

We implement an SVM model using the Stata application written by Guenther and Schonlau (2016). We use the radial basis function kernel and conduct a grid search as described by Guenther and Schonlau to identify the optimal weight and scaling parameters

on a 1% sample of the training data set. For each parameter, we run the model at 8 evenly spaced points within the interval [0.001,10,000]. Since there are 2 parameters and 8 possible values for each, we run the model $8 \times 8 = 64$ times and pick the parameter values for the run with the highest resulting F-statistic. Once the optimal tuning parameters are established, we run the SVM on the full sample of 1,000,000 training pairs. This application of SVM takes a substantial amount of time to both train and estimate.

Lasso Shrinkage Model

Least absolute shrinkage and selection operator (Lasso) models are a popular method for variable selection and prediction. Lasso models are shrinkage estimators, meaning that some independent variables are essentially removed from the final model used for prediction. This helps to avoid overfitting in the presence of many explanatory variables (Hastie *et al.*, 2016). More formally, we estimate a linear probability model of the form:

$$TM_{i,j} = \beta \mathbf{X} + \epsilon_{i,j} \quad st \quad \sum_{\beta=1}^{K} |\beta_k| <= t$$

where $TM_{i,j}$ is the match status of observations i,j as measured by the biometric ID and $\mathbf{X}$ is a matrix of match variables. We use the Lasso command written by Ahrens *et al.* (2018) for Stata. The constraint, t, is selected using the extended Bayesian information criteria proposed by Chen and Chen (2008).

After estimating the Lasso model, we use the coefficients on the selected variables to predict the match probability of each pair in our training set. Note that since this is a linear probability model, the resulting score is not constrained to be in the [0,1] interval. We pick the match threshold that maximizes the F-statistic over the match space.

Random Forest (Standard, Demographic Enhanced, and hand-coded)

We implement a random forest machine learning algorithm proposed by Breiman (2001), and developed as a Python application in the Scikit-learn package by Pedregosa *et al.* (2011). The model is run using 4 parallel processors.

Our standard random forest model has 250 trees, where each tree is estimated on a bootstrapped sample of 1,000,000 observations with replacement. The maximum number of splitting variables is determined by the number of inputs/3 which is equal to 15. The splitting variables on each tree are chosen at random, so every tree will have a different group of input variables. Once the model is finished estimating on the training pairs, we are able to predict in sample and out of sample classification by taking the mode prediction over the 250 trees.

The demographic enhanced random forest model is the same as the standard model, except we add indicator variables to determine whether 1 or both observations is female, as well as whether 1 or both observations are white, black or hispanic. These extra demographic variables

raise the number of inputs so the maximum number of splitting variables to be selected is 18. We run this model on 250 trees where each tree is estimated using a bootstrapped sample of 1,000,000 pairs.

Lastly, we run a version of the random forest model with a 5,000 observation training sample that is hand-coded by research assistants. [33] In this model, we estimate 250 trees where each tree is split using a bootstrapped sample of 5,000 observations from the hand-coded pairs. Because we include demographic variables, the maximum number of variables that are eligible to be selected is 18.

Neural Net (Perceptron and Hidden Layers)

Neural networks are a class of prediction models designed to mimic the function of a human brain. Neural networks are capable of creating highly non-linear models through the use of hidden layers that receive signals from input (match) variables and then transmit a signal through a linking function. One can increase the complexity of a neural network by increasing the number of hidden layers and nodes within each hidden layer.

First, we implement a neural network with one hidden layer comprised of 24 nodes, and use Stata's BRAIN command (Doherr, 2018) to estimate the output layer. The initial signal value for each node in the hidden layer is randomly chosen in the interval [-0.25, 0.25]. For each iteration through the training sample, the observations are sorted randomly and then the signal weights for each node are updated subject to the training factor which is set at 0.25. After a full cycle through the training sample, the data are quasi-randomly resorted and then the same process of signal updating occurs. In total, we include 500 iterations through the training sample. After estimating the model, we are left with the predicted probability that each training pair is a match. To assign statistical matches based on the predicted score, we select the threshold that maximizes the F-statistic across the score space.

Next we implement a neural network with no hidden layers, sometimes referred to as a simple perceptron. The specifications for estimating the simple perceptron are the same as the hidden layer model, and we use the predicted probabilities after iterating 500 times through the training sample. The match threshold is assigned using the same F-stat maximization routine.

C Applying a corruption algorithm to the Social Security Administration's Master Death File

In many record linkage applications, it is prohibitively difficult to acquire data that can be used to test the out of sample performance of a matching algorithm. We follow a common strategy (Christen and Churches, 2002; Christen, 2012; Ferrante and Boyd, 2012; Bailey *et al.*, 2017, for example) by testing our algorithm on a synthetic, corrupted data set. As an input,

[33] See Appendix B for details on sample construction.

we use the Social Security Administration's Death Master File (DMF). The DMF records the social security number, birth date, name and date of death for all deaths reported to the SSA. We downloaded a publicly available copy of the file that goes through November 30, 2011, which contains approximately 85 million records. Using these variable inputs, we are able to construct a new data set that has been randomly edited to include a number of data errors common to large tables. Below is a description of the methodology used to create this synthetic data.

We limit our sample to individuals that died between the years 2000-2009, leaving us with a base file of approximately 20.3 million unique death records. The original data include very few middle names or middle initials. Because our main algorithm is estimated on data that includes middle names, we impute middle initials for those who are missing names based on the year and location of birth. [34]

Next, we identify three separate, common transcription errors -name standardization edits, phonetic edits and general edits- that we use to corrupt the DMF data file. The name standardization edits replaces a name with a common nickname or vice-versa. For example, a record with a first name of "Matt" could be adjusted to instead have the first name "Matthew". The phonetic edits identify character groups that are commonly used interchangeably due to their similar phonetic sound. For example the letters "ck" and "k' are often used to make similar sounds and therefore are a common source of misspelled names. The general edits are intended to mimic errors as a result of faulty data entry and optical charcter recognition (OCR). These include mostly typographic errors and account for mistakes common to users of a QWERTY keyboard. Common examples of OCR errors include interchanging "m" and "n" or "l" and "i". Note that the phonetic and general edits use data files from corruptor software written by Tran *et al.* (2013) to identify common errors in these two categories. These files have been supplemented by other common phonetic misspellings.

Beginning with our base file, we corrupt our data in the following order: (1) name standardization edits, (2) phonetic edits and (3) general edits. After removing observations that do not receive an edit, we are left with approximately 4 million observations (20%) that have at least one type of edit. Of the edited observations, 42% have a name standardization error, 34% have a phonetic error and 32% have a general error. Next, we append the base file to the corrupted observations, resulting in a dataset of 24.3 million records, where 20 million are original records, and 4 million represent corrupted records from at least one of the three possible edits.

[34]Based on the first three digits of the SSN, we are able to determine the individual's state of birth using the crosswalk published by the SSA at https://www.ssa.gov/employer/stateweb.htm. Note that the SSA stopped allocating SSN by geography in 2011.

Table A.1: Description of matching variables

Metric	**Variables**	**Number of features**
Jaro-Winkler distance (JW)	first, middle, last, first standardized, middle standardized	5
Levenshtein distance (LD)	first, middle, last, first standardized, middle standardized, birth month, birth day, birth year	8
Levenshtein distance normalized by string length (LDN)	first, middle, last, first standardized, middle standardized	5
Missing indicator	middle	1
Exact match indicator (EM)	first, middle, last, first standardized, middle standardized	5
Soundex match indicator	first, middle, last	3
Phonex match indicator	first, middle, last	3
Date distance	date,month, day, year	4
Uniqueness interactions	first (EM, JW, LD, LDN), middle (EM, JW, LD, LDN), last (EM, JW, LD, LDN)	12
Total variables		46

Each row represents a metric used in the matching algorithm. First, middle and last refer to individual name components, while first and middle standardized refer to the root name as determined by the census bureau crosswalk. Birth month, day and year refer to the individual components of birth date. For example, the matching algorithm includes the Jaro-Winkler distance for each name component listed under the variable column. In total, there are 46 variables used in the baseline model.