INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

CJARS Criminal Justice Administrative Records System

# Modernizing Person-Level Entity Resolution with Biometrically Linked Records

Matthew Gross
Twitter

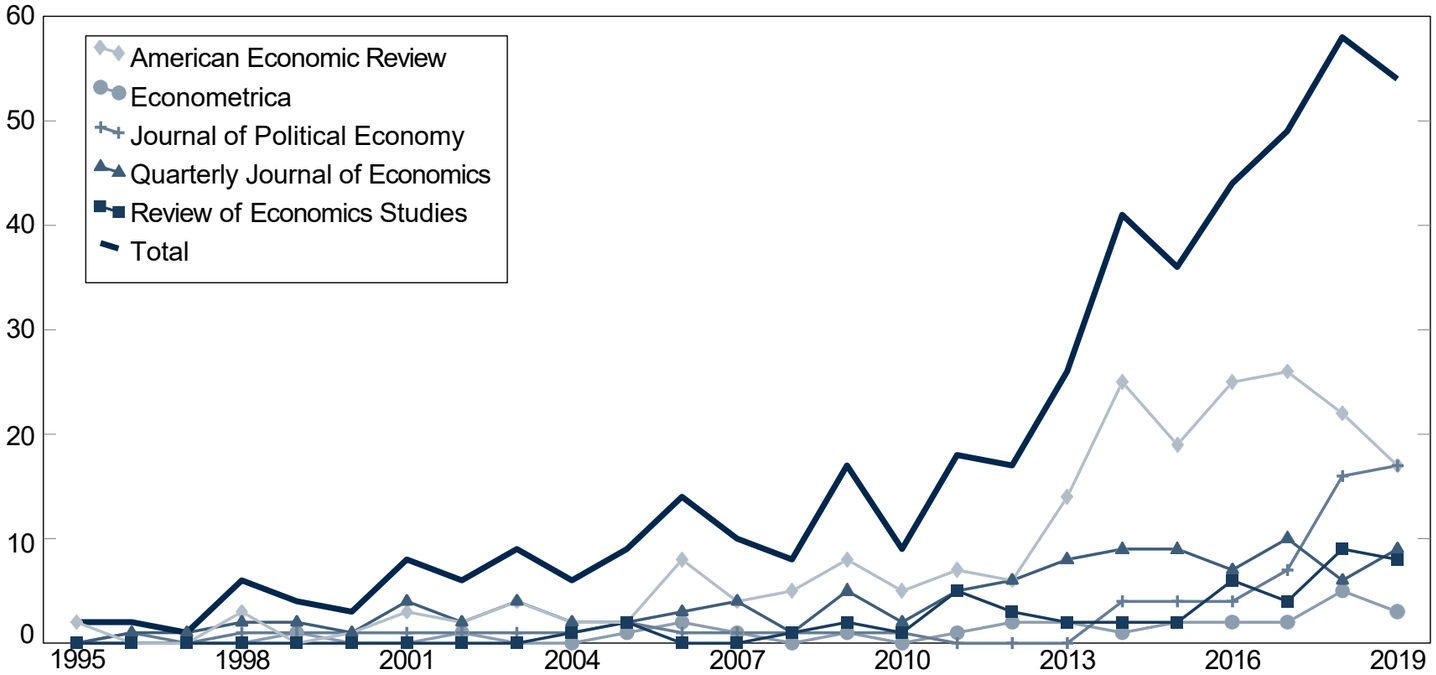
Michael Mueller-Smith
University of Michigan

March 21, 2022

# Motivation

- Research increasingly relies on “big data” and administrative records
  - Data linkage is an empirical necessity
- Frequently do not have access to unique linking identifiers
  - Rise of fuzzy or probabilistic matching techniques
  - Often limited discussion of matching strategies in research
- Implications of imperfect linkage in causal inference can be problematic
  - Introduces (potentially non-trivial) measurement error into analysis

# Rise of Administrative Data in Economics Research




- Increase in papers published in "top 5" economic journals that mention the term "administrative data"
- Total number of papers increases by a factor of 5 between 2010 and 2019

# Overview of Project

In this paper, we:

1. Utilize a large, novel training set to develop a highly non-linear model to match individuals based on name and date of birth
   - Compare a range of commonly used classifiers to determine which performs best
   - Compare the performance to models trained with smaller and hand-coded training sets
2. Evaluate the algorithm's out-of-sample performance using data from other contexts to determine external validity and suitability to both *record linkage* and *deduplication* applications
3. Simulation exercise to show how match performance statistics relate directly to estimation bias and statistical precision

**Note:** Baseline algorithm developed in this paper is used by CJARS to link individuals across criminal justice databases

# Defining Algorithm Performance Statistics

| | True Match | True Nonmatch |
|---|---|---|
| Algorithm Match | True Positive (TP) | False Positive (FP) |
| Algorithm Nonmatch | False Negative (FN) | True Negative (TN) |

**Precision** = % of algorithm matches that are "correct" = $\left(\frac{TP}{TP+FP}\right)$

**Recall** = % of the true matches correctly identified by the algorithm = $\left(\frac{TP}{TP+FN}\right)$

**F1 Score** = measure of overall performance = $2 \times \frac{Precision \times Recall}{Precision+Recall}$

# Main Takeaways

1. Develop a random forest model which outperforms other standard prediction algorithms
   - Models trained with large training sets up to 250,000-500,000 observations exhibit increased stability and higher performance
   - Model trained with biometric ID linked pairs outperforms hand coded training sets
     - greatly improves recall at modest-to-no cost to precision
   - Gains vary by demographic subgroups, suggesting that method of producing training data is particularly important when working with a heterogeneous sample

2. Performance of algorithm remains high in different contexts suggesting that the model may be applicable to non-criminal justice settings

3. Simulation demonstrates how the match performance measures of precision and recall are directly related to internal and external validity
   - Depending on the setting, errors in match recall and precision lead to biased estimators and incorrect confidence intervals

# Defining The Matching Problem

Given two sets **A** and **B** containing elements $a$ and $b$ :

- A record linkage algorithm seeks to identify which elements of **A** and **B** are common to both sets.

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

- **A**, **B** must include a vector of common variables
  - From this common vector, one can define a comparison function $\gamma$ to measure similarity between elements, and a decision rule to determine algorithmic matches

# The Motivating Matching Problem

This research arises out of the production needs of the Criminal Justice Administrative Records System (CJARS):

- Collect individual-level data from criminal justice organizations
- Identify the same individual across time and jurisdiction
- Lack unique identifiers, but have name and date of birth
- Build a matching model to predict matches across data sets

# Novel Training Data From Criminal Justice Sources

Our matching model is created using two sources of training data containing biometric (fingerprint) IDs and original (flawed) PII:

- Harris County District Clerk (Houston)
  - Criminal defendant booking data for cases between 1980 and 2017
  - 1,722,575 unique combinations of name and date of birth (1,317,063 unique IDs)
- Texas Department of Criminal Justice
  - State prison inmates between 1978 and 2014
  - 1,042,450 unique combinations of name and date of birth (905,528 unique IDs)
- Fingerprint ID is a source for knowing true match (TM) status for millions of records

# Redefining the Match Problem as Deduplication

- Approximately 2.8 million unique PII combinations between the two sources
  - No crosswalk linking the IDs across datasets
  - Two disjoint sets when identifying possible matches
- Predict the match status of pairs of observations within each data set
- Known as *data deduplication*
  - Note that any deduplication can be restated as a record linkage problem
  - Instead of matching set A to set B, we simply are matching set A to set A eliminating pairwise exact matches

# Blocking Strategy To Limit the Match Space

Cannot compare every possible training match since (approximately 2 trillion pairs; 670k TM).

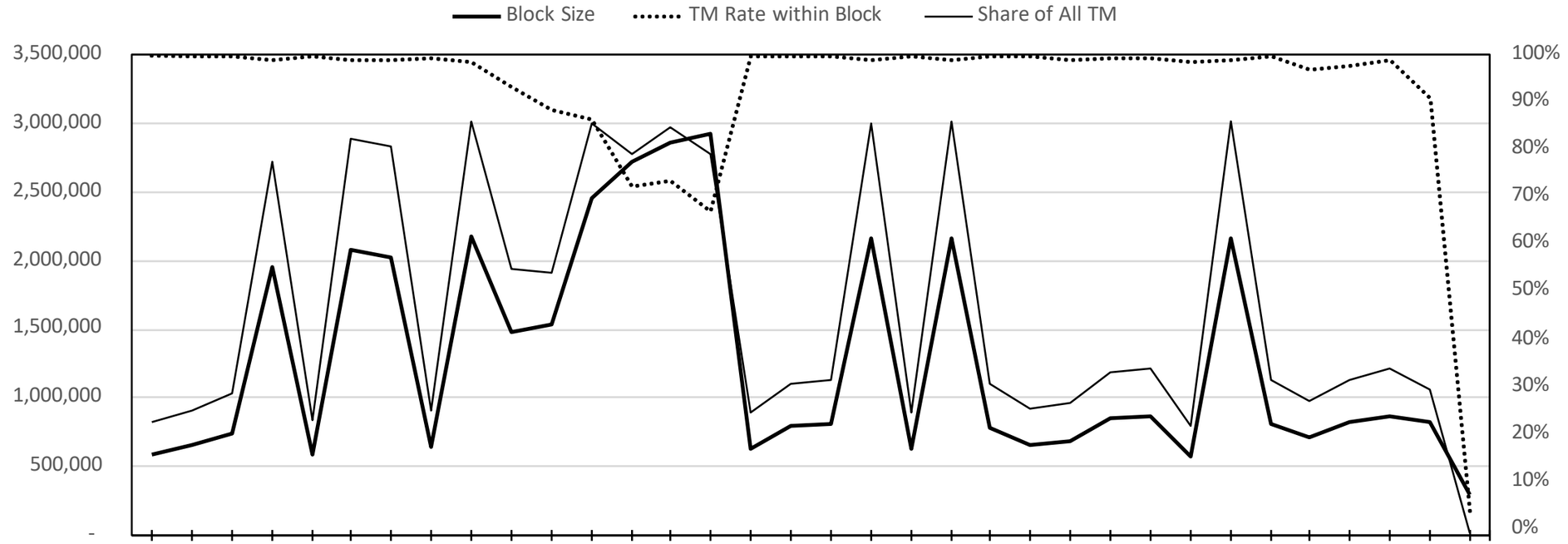
Limit the potential match space using a *blocking strategy*:

- Pairs are only evaluated for TM status if they meet a deterministic criteria:
  - Exact match on date of birth + last name soundex
- Union of 10 overlapping blocks
- 2 trillion → 17.5 million pairs
  - > 95% of the 670 thousand true matches
- Currently working to leverage *learning disjunctive normal form (DNF) blocking*
  - Very cool/efficient! Limited set of results on this at the moment

# Overlapping Blocks Identify Most True Matches

| 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 |

# Learning disjunctive normal form (DNF) blocking




**Blocking characteristics**

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Full Date of Birth | x | x | x | x | x | x | x | x | x | x | x | x | | x | | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| Day of Birth | | | | | | | | | | | | | | | x | | | | | | | | | | | | | | | | | |
| Month of Birth | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Year of Birth | | | | | | | | | | | | | x | | | | | | | | | | | | | | | | | | | |
| First Name | x | | x | x | | | x | | | | x | | x | x | x | | | | | | | | | | | | | | | | | |
| First Name Soundex | | x | | | | x | | | | | | | | | | | x | x | | x | | | | x | | | | | | | x | |
| First Name Phonex | | | | | | | | x | | | | | | | | | | | x | | | x | | | | | x | x | | | x | |
| Last Name | | | | x | x | x | | | | x | | x | x | | x | | | | | | | | | | | | | | | | | |
| Last Name Soundex | | | | | | | x | | | | | | | | | x | x | | | | x | | | | x | | | x | | | | x | |
| Last Name Phonex | | | | | | | | x | | | | | | | | | | | x | | | x | | | | | | | | | | |
| Name Swap Match | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | x |
| Address - Building Number | | | x | | | | | x | | | | | | | | | x | x | | | | x | | x | | | | | x | | x | |
| Address - Street Name | x | x | | | x | | | x | | | | | | | x | | | | x | | | x | | | | x | | | | x | | |
| Address - City | | | | | | | | | | x | x | | | | | | | | | | | | | | | | | | | x | x | | |
| Address - Zip Code | | | | | | | | | | | | | | | | | | | | | | | | x | x | x | x | | | | x | | |

# Building and Testing a Model

- Starting with the blocked pairs, we take a 1 million observation random sample to use as a training data set
  - Represents a substantial increase over standard training sample sizes
- Large training data allows us to estimate complicated non-linear model with many features
  - We define 46 matching variables, based on name components and date of birth, to determine similarity of records
  - Include name standardization variables to account for common nicknames
- Test performance of the model using reserved data not included in the training sample
- Estimate different models using the same training sample and compare performance using the same reserved sample
  - Allows for a clean comparison across models

# Comparing Different Classification Models

We are agnostic about the choice of model and compare across a wide range of options

- Simple deterministic for exact matches
- Machine learning - SVM, Random Forrest, Neural Networks
- Naive bayes classifiers
- Regression based - LASSO

In all models, we train on the same 1,000,000 observation sample of pairs and test out of sample performance on the remaining pairs

# Performance of Different Models

| Model | Precision (% of matches are correct) | Recall (% of correct matches made) | F-Statistic |
|---|---|---|---|
| Deterministic | 0.93 | 0.76 | 0.84 |
| Naive Bayes Classifier (Discrete) | 0.90 | *0.72* | *0.80* |
| Naive Bayes Classifier (Kernel) | *0.88* | 0.81 | 0.84 |
| Support Vector Machine | **0.94** | 0.83 | 0.88 |
| Lasso Shrinkage Model | 0.90 | 0.82 | 0.86 |
| Random Forest | 0.93 | 0.88 | 0.90 |
| Random Forest (Demog. Enhanced) | 0.93 | **0.89** | **0.91** |
| Neural Net Perceptron | 0.93 | 0.85 | 0.89 |
| Neural Net | 0.92 | 0.88 | 0.90 |

- Demographic enhanced random forest model achieves the highest out of sample performance
- Substantial improvement to recall with minimal cost to precision
- Fast to estimate!

# Adjusting the Size of Training Data

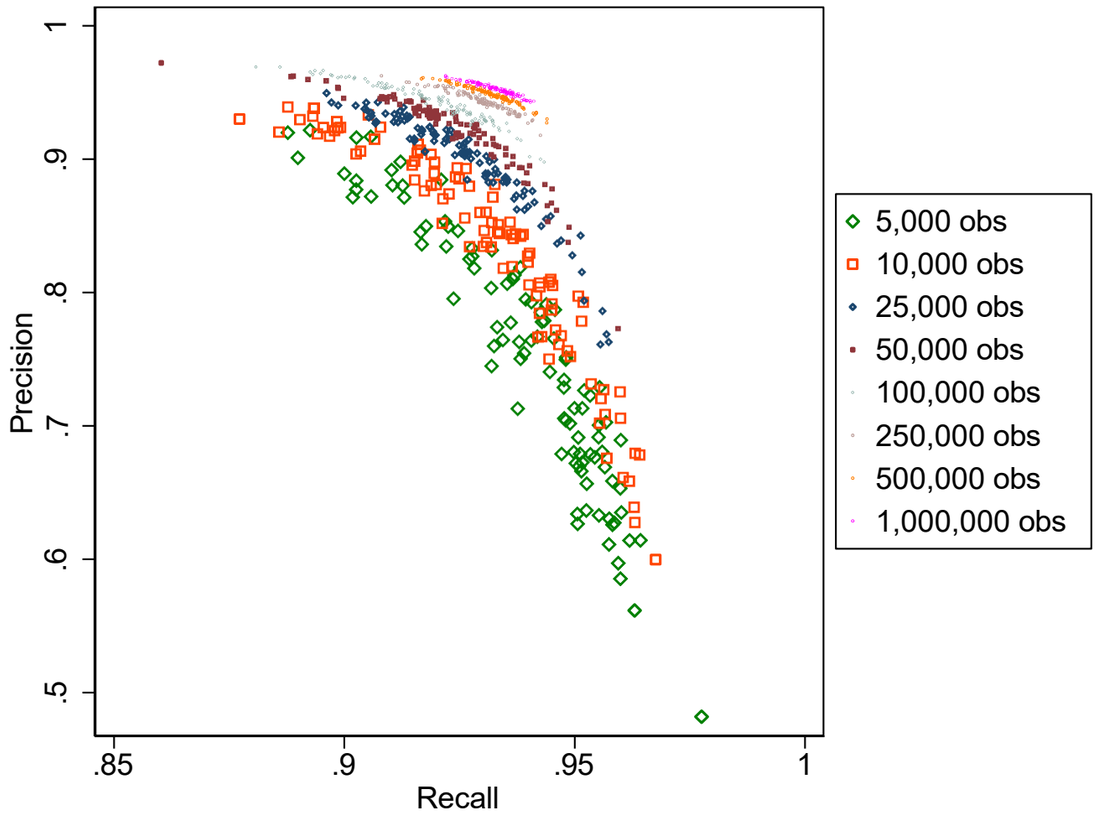
Significant improvement over prior matching algorithms... why?

- Machine learning technique?
- Training sample size?
- Hand-coding or biometric trained model?

To evaluate the role of sample size:

- Estimate series of bootstrapped models off using a range of 5,000-1,000,000 training observations
  - 100 times per candidate training sample size to measure performance stability
- Shows that recall does not stabilize until $\sim$250,000 training observations

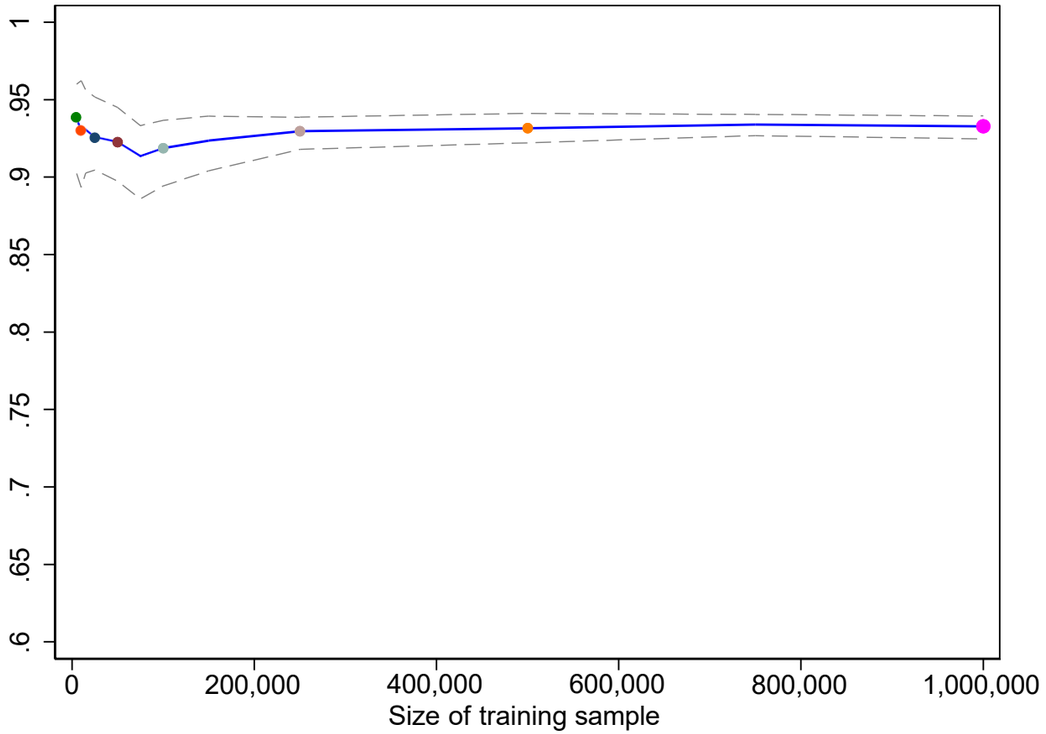
# Comparing Performance By Training Set Size



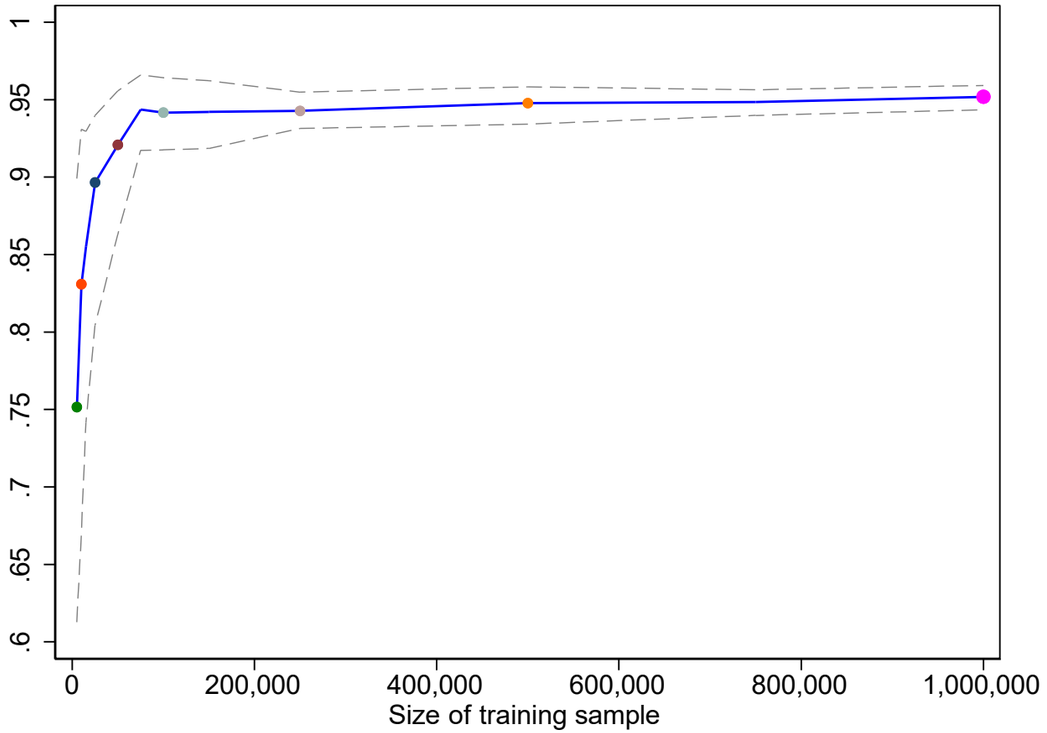

- As the training set grows, the trade off between precision and recall decreases
- Variation between independent models from the same sample size shrinks

# Comparing Performance By Training Set Size

Recall

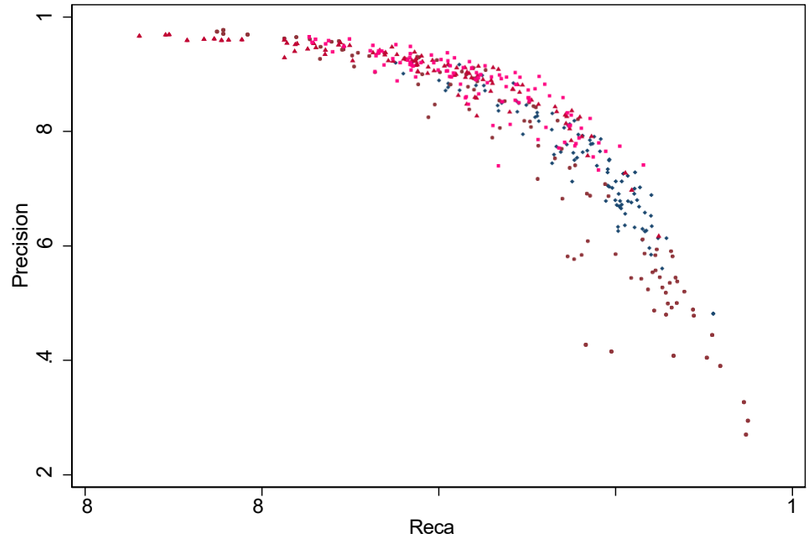



Precision

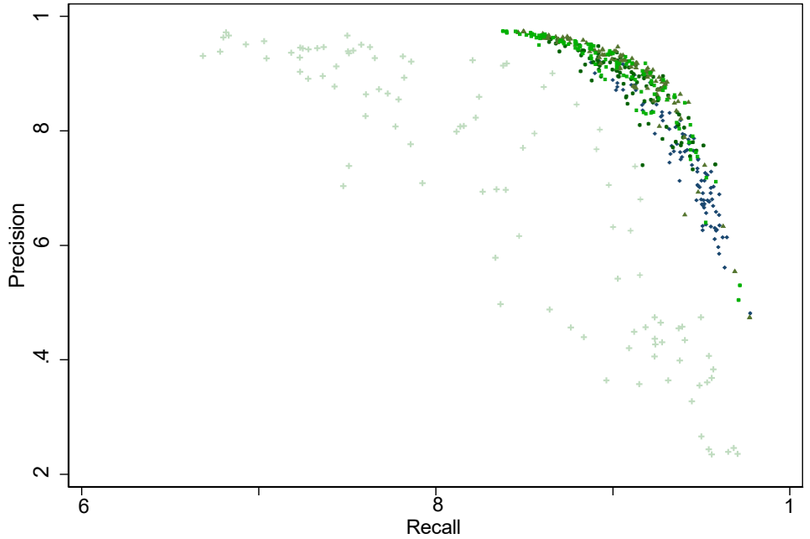



- Solid blue line is the average performance at each training set size
- Dashed lines represent the 5th/95th percentile performance

# Comparing Performance By Training Set Composition




(a) Excluding low, high, and marginal matches in training sample




(b) Adjusting marginal share in training sample

| | Symbol | Low Predicted Match Likelihood | Marginal Predicted Match Likelihood | High Predicted Match Likelihood | Average Precision | Std. Dev. Precision | Average Recall | Std. Dev. Recall |
|---|---|---|---|---|---|---|---|---|
| Baseline (5,000 obs.) | ◆ | 96% | 1% | 3% | 0.75 | 0.10 | 0.94 | 0.02 |
| *Panel A: Excluding low, high, and marginal matches in training sample* | | | | | | | | |
| Scenario A1 | ● | 0% | 50% | 50% | 0.73 | 0.20 | 0.93 | 0.04 |
| Scenario A2 | ▲ | 50% | 50% | 0% | 0.90 | 0.06 | 0.90 | 0.03 |
| Scenario A3 | ■ | 50% | 0% | 50% | 0.88 | 0.06 | 0.91 | 0.06 |
| *Panel B: Adjusting marginal share in training sample* | | | | | | | | |
| Scenario B1 | ● | 50% | 0% | 50% | 0.88 | 0.06 | 0.91 | 0.06 |
| Scenario B2 | ▲ | 33% | 33% | 33% | 0.90 | 0.08 | 0.90 | 0.03 |
| Scenario B3 | ■ | 25% | 50% | 25% | 0.89 | 0.09 | 0.90 | 0.03 |
| Scenario B4 | — | 0% | 100% | 0% | 0.69 | 0.24 | 0.83 | 0.09 |
| *Full Blocked Pair Comparison Sample Statistics* | | | | | | | | |
| Percent True Matches | | 1% | 51% | 89% | | | | |
| Total True Matches | | 93,943 | 103,557 | 429,048 | | | | |
| Total Blocked Pairs | | 16,889,570 | 204,344 | 483,601 | | | | |

# Adjusting the Method of Generating Training Data

Another difference between our approach and prior work is the method of defining TM in training data:

- Typically training sample is hand-coded
  - Raises issues of human bias or lack of familiarity with target linking population
  - Costly to produce large samples
- Estimate the preferred random forest model on hand-coded data to see how performance changes
  - 5,000 pair sample from blocked pairs and assigning 3 RAs to label match status of each pair
  - Method is similar to what is reported in the literature
- Especially interested in possible heterogeneous performance across demographic groups

# Precision Performance by Training Data

Precision = $(\frac{TP}{TP+FP})$ = percent of matches that are "correct"

| | | Demographic Enhanced Random Forest | | |
|---|---|---|---|---|
| | Deterministic | 5,000 Hand-Coded Training Obs. | 5,000 Biometric Training Obs. | 1,000,000 Biometric Training Obs. |
| *Precision by Race/Ethnicity* | | | | |
| Overall | 0.93 | **0.95** | *0.91* | 0.93 |
| 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 |

- Deterministic and model trained by hand-coded data have slightly higher precision across race/ethnicity

# Recall Performance by Training Data

Recall = $(\frac{TP}{TP+FN})$ = percent of the true matches made

| | | Demographic Enhanced Random Forest | | |
|---|---|---|---|---|
| | Deterministic | 5,000 Hand-Coded Training Obs. | 5,000 Biometric Training Obs. | 1,000,000 Biometric Training Obs. |
| *Recall by Race/Ethnicity* | | | | |
| Overall | *0.76* | 0.77 | 0.84 | **0.89** |
| 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** |

- Biometric ID → substantial gains in Recall

# Overall Performance by Training Data

$$\text{F statistic} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

| | | Demographic Enhanced Random Forest | | |
|---|---|---|---|---|
| | Deterministic | 5,000 Hand-Coded Training Obs. | 5,000 Biometric Training Obs. | 1,000,000 Biometric Training Obs. |
| *F-Statistics by Race/Ethnicity* | | | | |
| Overall | *0.84* | 0.85 | 0.87 | **0.91** |
| 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** |

- The large, biometric ID training data performs the best
- Large gains in recall make up for small declines in precision

# Takeaways

- Large, biometric ID training data yields highest performing model
- Smaller biometric ID training set outperforms hand-coded training set
  - Higher recall at only slight cost to precision
  - RAs are too conservative about matching and overweight name similarity vs. DOB similarity
- Heterogeneous performance by race (and other demographic groups)
  - Major implications for research on diverse populations with heterogeneous TE

# Out of Sample Exercises

We run 3 exercises to test the algorithm's out of sample performance - each exercise is progressively more different than the training sample

1. Non-Texas Prison population (333 thousand inmates)
   - Identify individuals in prison on July 1, 2017, and match across states to check for false positives
2. Washington voter records
   - Conduct a one-to-one match between 2008 and 2012 presidential voting records
   - Higher rate of females in this data will likely lead to performance declines due to higher probability of name changes
3. Social Security Master Death File (DMF) for 2000-2009 deaths
   - Start with 20.2 million deaths and generate 4 million corrupted records
   - PII is corrupted to include standard spelling, OCR and keyboard errors
   - Increasing density of PII space will make it harder to differentiate TM from TN

# Out of Sample Exercise Results

| 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 |

1. Out of 463,969 inmate blocked pairs, the algorithm makes 2001 matches (0.4% false positive rate within blocks)
2. Despite the higher proportion of females in the WA data, the algorithm recall is still 88%
3. The algorithm is able to match 93% of the corrupted DMF records back to the original file while avoiding almost all false positives

# The Importance of Match Performance in Economic Research

Poor recall and precision can impact the estimation of treatment effects

- Matching scenario where match is the dependent value of interest
  - Common example is criminal recidivism as an outcome variable - matching strategy to determine if individual appears in future arrest data (Tahamont et al, 2019)

- Scenario where the analysis sample is conditioned on being matched
  - For example, studying healthcare utilization among those with Medicaid

Determine how different precision and recall errors affect estimated treatment

# Scenario 1

For scenario 1, the outcome variable is generated by the following equation:

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

where $d_i$ indicates whether person i received a treatment, and $\mu$ determines the baseline outcome rate for the non-treated population

- To the econometrician, the outcome variable is 1 if a treated observation is matched to an external data set (such as arrest, hospital admission, etc)
- Errors in recall lead to fewer "correct" matches made
- Errors in precision lead to more "incorrect" matches made

The econometrician wants to estimate the simple linear probability model:

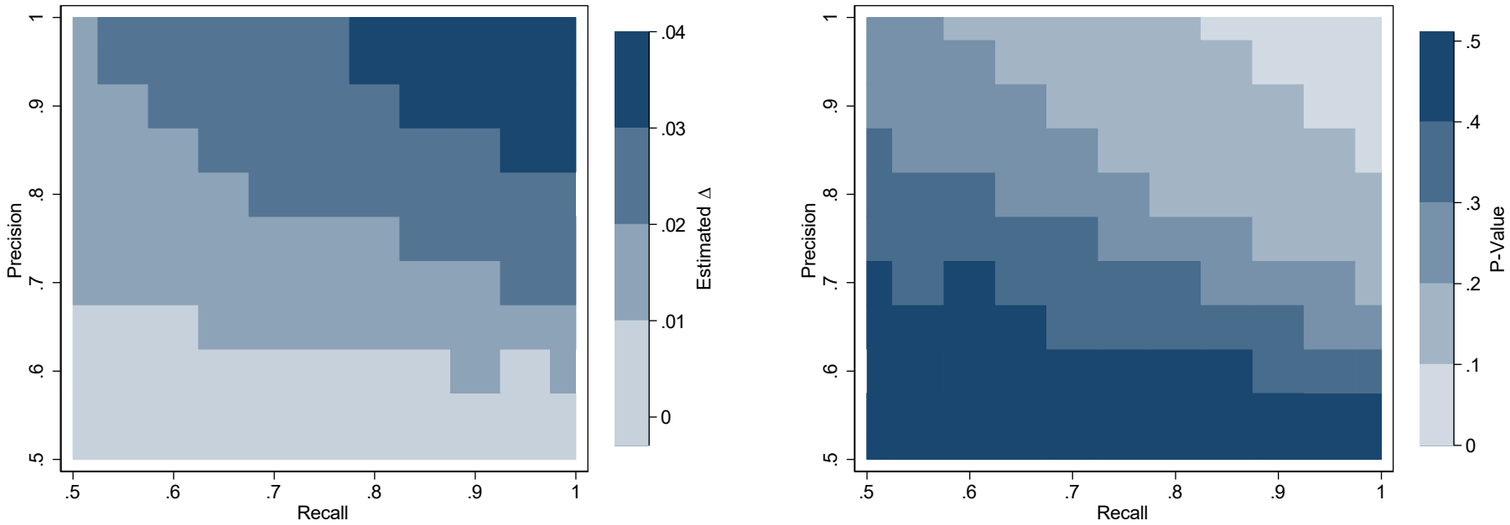
$$y_i = \Delta d_i + v_i$$

# Example of simulation exercise in scenario 1

- Sample size of 5,000
- 50% of the sample receives the treatment
- Baseline rate of outcome is 50%
- $\beta$ is 10% (or $\sim$4.5 percentage point treatment in linear model)

We estimate the treatment effect in the presence of both recall and precision errors

# Example of simulation exercise




- Left panel shows estimated treatment effect at different combinations of precision and recall
  - True effect is 0.045, in the top right corner
  - Errors in precision and recall both lead to attenuation bias
- Right panel shows p values of the hypothesis test that the treatment effect is 0
  - Changes to precision and recall lead us to incorrectly fail to reject the null hypothesis

# Main Takeaways

- Large training sets ($\geq$ 250,000 obs.) yield better and more stable results when estimating matching models
- Fair degree of heterogeneity across different matching algorithms conditional on training data
- Training data produced by clerical review may be weighted towards precision at the expense of recall
  - Training data with biometric IDs has better overall performance
  - Performance varies by demographic subgroup
- Errors in precision and recall may lead to biased treatment effect estimates and statistical imprecision
  - Especially concerning given prior result on failures of traditional matching approaches for women and minority communities

# Thank You For Your Comments

Mike Mueller-Smith - mgms@umich.edu

# Illustration of Matching Algorithm Sequence of Events