The Long-Term Effects of Housing and Criminal Justice Policy: Evidence and Methods

by

Matthew B. Gross

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2021

Doctoral Committee:

Assistant Professor Michael Mueller-Smith, Chair
Professor Charlie Brown
Assistant Professor Sara Heller
Professor Brian A. Jacob

Matthew B. Gross

mbgross@umich.edu

ORCID ID 0000-0003-4721-421X

To my grandparents - Pearl, Mel, Florence, and Elliot - who instilled in their families the value of education and the joy of exploring the world.

ACKNOWLEDGEMENTS

To rephrase the saying, it takes a village to complete a dissertation. I am indebted to so many people for their support over the past 7 years. I have been fortunate to work with some exceptionally smart and committed professors who helped me drag my dissertation over the finish line. I am constantly amazed by the dedication and intellectual firepower exhibited by my dissertation committee. Mike Mueller-Smith, my dissertation chair, boss, and coauthor, was a constant advocate and incredible adviser throughout the process. Mike taught me the skills required to be an empirical economist, and I would not have been able to complete this journey without his mentorship and help. Charlie Brown was unrelentingly enthusiastic while also displaying an uncanny ability to simultaneously ask the right questions and suggest constructive solutions. Sara Heller's brilliance as both a researcher and communicator helped me to improve my dissertation and the way I presented it in countless ways. Brian Jacob has been a research adviser going back to my third-year paper. In particular, his suggestions and comments were instrumental in pushing me to improve the matching methodology in my first chapter.

Other faculty who engaged with my research and contributed helpful suggestions include Mel Stephens, John Bound, Joel Slemrod, Ash Craig, and Gabriel Ehrlich. I would also like to thank the participants of the Labor Lunch and Labor Seminar for their attention, humor, and intelligent comments and questions at various stages of my dissertation.

The second and third chapters of my dissertation were coauthored with very talented researchers. Mike Mueller-Smith is a coauthor on the second chapter while Mike, Keith Finlay, and Elizabeth Luh are coauthors on the third chapter. I would like to thank Elizabeth for moving this paper forward when I was focused on different projects. On top of being a valued coauthor, Keith has been very gracious with his time to help shepherd this project through the sometimes maze-like Census Bureau disclosure process.

For the second and third chapters of my dissertation, I am indebted to the incredibly dedicated past and present employees working at the Criminal Justice Administrative Records System (CJARS). Matt Van Eseltine helped the project grow in its early stages and assisted with preliminary versions of the entity resolution model. Jay Choi had a hand in nearly every aspect of the project.

For my second dissertation chapter, Jay Choi, Madeleine Danes, Francis Fiore, Jordan Papp, Benjamin Pyle, Lyllian Simerly, David Smith, Brittany Street, and Peixin Yang helped out with an important handcoding exercise. Peixin also worked on the database of research papers that discuss administrative data. For the second chapter, Martha Bailey, John Bound, Charlie Brown, Aaron Chalfin, Connor Cole, James Feigenbaum, Keith Finlay, Sara Heller, Kirabo Jackson, Emily Nix, Joseph Price, Mel Stephens, Jesse Rothstein, Sarah Tahamont, and Chelsea Temple all provided substantive comments to improve the paper.

For the third chapter, many of the same CJARS employees worked together to transform millions of rows of raw criminal justice records into data that could be used for research. Diana Sutton provided important institutional knowledge when trying to understand the charges associated with driver responsibility fees. In addition, Katie Genadek was extremely helpful and generous with her time in getting a large disclosure review package released.

There are a number of teachers and mentors who have guided me throughout my academic and professional career, and without whose assistance I would not have been able to reach this milestone. Jeanne Hogarth, Ellen Merry, and Max Schmeiser took a chance on me and offered me the opportunity to conduct research professionally while at the Federal Reserve Board. Paula Malone, David Albouy, Martha Bailey, David Lam, and Mel Stephens were influential undergraduate teachers who ignited my passion for economics and research.

I am very grateful to my colleagues and friends, both within the economics department and in various other departments, for their support, help and friendship. It would be impossible to name everyone but there are a few in particular that I would like to acknowledge. Ting Lan and Huayu Xu were patient study group partners in my first year. Ellen Stuart was a gracious host on holidays and Friday evenings. Steph Owen was a committed Blank Slate buddy. Ari Binder was a gracious BBQ host and a game bike rider. Anirudh Jayanti was a great neighbor, movie buff and friend. Dhiren Patki is an all-around wonderful friend. The residents of 809 Lawrence, including Sam Haltenhof and Chad Milando, added much-needed levity and fun. George Fenton and Max Gross, the other two initials in GMM, are incredible friends who provided constant support and were with me every step of the way over the past 7 years.

I would like to thank my family for making all of this possible. My in-laws Bob, Jackie and Elizabeth welcomed me into their family with open arms despite my less than flawless table manners. My brother Adam and sister-in-law Reanna helped me generate research ideas and always supported me, even when they did not know my research topic. Most importantly, they had Mason and Brody, who made it hard to leave NYC every time I had to go back to Ann Arbor, but who also provided excellent motivation to finish my dissertation.

My parents are more responsible for my achievements and success than anyone in the world.

For treating education as an expectation and making sure that I had the best opportunities to learn. For allowing me to see the the world and stoking my intellectual curiosity. For supporting me in so many different ways throughout my time in college, Washington DC and graduate school. For nursing me through multiple shoulder surgeries while in graduate school. For knowing when not to ask me "how is your dissertation research going?" For these and countless other reasons, I have the best parents in the world, and I am so thankful for their presence in my life.

Last and most importantly, I would like to thank the brains behind the operation: my wife Rachel. Rachel encouraged me through every twist and turn of this research project from its initial conception to its current state. One silver lining of the Covid-19 pandemic was that I was able to type every word of my dissertation sitting within a few feet of Rachel. My completed dissertation is a testament to her love and support, and I could not have done this without her.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

This dissertation combines research from multiple areas of applied economics and is mostly focused on estimating the long-term impacts of housing and criminal justice policy. In addition, this dissertation covers an important methodological tool that is increasingly necessary for empirical researchers when linking multiple data sets to estimate causal treatment effects.

In the first chapter, I study the effects of rent control on the long-term outcomes of children. Rent control is a common policy enacted to limit the growth of rents and allow tenants to remain in their homes for longer. Prior empirical research has mainly focused on rent control's impact on neighborhoods and housing markets while ignoring the potential long-term impacts of rent control for the people directly affected by the policy, particularly children. Using nearest neighbor matching at the census tract level, I estimate the effects of rent control on average long-term outcomes for children, measured at the childhood census tract level. I find weakly suggestive evidence that rent control can improve the long-term labor market outcomes for children while also creating negative spillovers for children who do not directly benefit from the policy.

In the second chapter, coauthored with Michael Mueller-Smith, we develop a record linkage algorithm that is trained using a large, novel data set that includes fingerprint identifiers. Record linkage is a crucial empirical tool for contemporary applied researchers who are interested in linking data sets that do not contain unique identifiers. We show that this large training data substantially improves model performance compared to the smaller training samples frequently reported in the literature. We also show evidence that training data based on human coding can be overly conservative when identifying matches on a target sample with different characteristics than the human coder. This research has major implications for empirical researchers who wish to link data sets and estimate heterogeneous treatment effects on subpopulations.

In the last chapter, coauthored with Keith Finlay, Elizabeth Luh, and Michael Mueller-Smith, we study the long-term impacts of criminal financial sanctions on labor market outcomes and criminal recidivism. The rising use of financial sanctions in the criminal justice system in the United States necessitates a rigorous test of their impacts on criminal defendants and their families. We use data that has been processed and linked together using the record linkage algorithm detailed in my second chapter and utilize the implementation of a 2003 Michigan law that sharply increased fines associated with certain driving crimes. After carefully accounting for how the long-run behavioral effects of the policy could undermine the integrity of the research design, we find null to slightly positive effects of the policy on labor outcomes, minimal deterrent effects, and suggestive evidence of a financial burden on romantic partners.

CHAPTER I

The Long-Term Impacts of Rent Control

1.1 Introduction

There is an ongoing national conversation about inequality of opportunity and the fact that economic mobility has become increasingly difficult for those born at the bottom of the income and wealth distribution. Research by Chetty et al. (2017) confirms that economic mobility in the United States decreased for birth cohorts between 1940 and 1980, and suggests that the socioeconomic status of parents is a growing predictor of a child's life outcomes. Researchers have identified early childhood as an especially important development period when well-timed interventions can mitigate some of the gaps in achievement between children of different income or wealth levels. Because of the cumulative nature of childhood development, relatively small childhood investments can have large long-run impacts, particularly when targeting children from marginalized populations (Cunha et al., 2006). As a result, there is a large body of literature studying and measuring the effects of these interventions.[1]

Although not directly targeting families with children, rent control is an example of a policy intervention that could improve childhood development for impacted children. Defined as government regulation of allowable rent, rent control is a policy tool that is intended to transfer resources from landlords to renters. The goal of the policy is to make it easier for low-income tenants and families to remain in their current housing. In most cities with rent control, the policies were enacted in response to low rental vacancy rates, rising rental prices and the fear that without regulations, many tenants would face a heightened threat of eviction and homelessness. In the ideal scenario, rent control operates as a transfer from landlords to renters, both in the form of below-market rents and insurance against rent increases in the future.

While the short-term effect of rent control on impacted renters is likely positive, the

[1] See Almond et al. (2018) for a recent review of the literature.

long-term effects on children have never been studied. The literature on childhood interventions suggests a number of channels through which the child beneficiaries of rent control could have improved long-run outcomes as a result of growing up in rent controlled housing. Rent control can be thought of as a form of government mandated housing assistance, which has been shown in other contexts to increase earnings and decrease future incarceration rates of impacted children (Andersson et al., 2016). The benefits of rent control also include an income effect component, allowing families to shift expenditures from housing to other goods such as improved healthcare or education. Hoynes et al. (2016) shows that cash assistance to families with young children can improve child health while Carneiro et al. (2021) show how the timing of income shocks during childhood plays an important role in education outcomes. Lastly, the added housing security associated with rent control has the potential to reduce parental stress and the number of housing moves a child makes during childhood. Research shows that a mother's exposure to distressing news, such as a threat of eviction, can impact a newborn's birth weight (which is itself correlated with long-term health and achievement outcomes), while frequent childhood moves are known to be harmful to a child's academic performance and long-term development (Carlson, 2015; Wood et al., 1993; South et al., 2007).

Research also shows that one's childhood neighborhood has a causal exposure effect on long-run labor market and social outcomes, suggesting that the direction of the effect of rent control on outcomes may depend on the neighborhood (Chetty and Hendren, 2018a,c; Chyn, 2018, for example). If rent control leads children to have longer tenant durations in neighborhoods that provide negative exposure effects, then rent control could even lead to declines in long-term outcomes when compared with similar children who grew up in non-rent controlled cities. The implication from this strand of the literature is that the impacts of rent control on long-term outcomes is an empirical question in that the sign of the likely effect is ambiguous.

This paper is organized around a central research question: how does rent control affect the long-term outcomes of children? In addition to the main question, I also seek to understand how this long-term impact varies by family income. Does rent control provide benefits to children growing up at the bottom of the income distribution?

Despite the potential effects of rent control on tenants, the economics literature has mainly focused on quantifying rent control's effect on the housing market and negative spillovers on neighborhoods. Rent control is often associated with a decrease in the quality of controlled housing (Gyourko and Linneman, 1990; Sims, 2007, for example) and a misallocation of tenants and apartments (Glaeser and Luttmer, 2003; Krol and Svorny, 2005). More recently, Autor et al. (2014, 2017) show that rent control has potential negative spillover effects, not only on the value of rent controlled units, but also on the value of neighboring properties that are not rent controlled. One of the large drivers of this negative capitalization is increased crime in areas close to rent controlled units suggesting that rent control suppresses gentrification. Diamond et al. (2019) shows that rent control in San Francisco leads landlords to remove their units from the rental market, thereby decreasing the supply of rental housing and ultimately leading to a more segregated and unequal housing landscape. Asquith (2018) confirms that landlords are more likely to convert rentals to owner occupied housing as the local price of housing increases.

While the costs of rent control are well-established, the benefits are much harder to quantify. A major reason for this is that it is very difficult to define a suitable counterfactual for individuals that live in rent controlled units. Traditionally, economists used hedonic price regressions to quantify the benefit of rent control to renters by estimating the rent that would prevail in the absence of controls (Gyourko and Linneman, 1989). The difference between the controlled rent and the estimated market-rate rent is a measure of the compensating variation of the policy; however, a static measure of rent control benefits ignores the potential long-term benefits that the policy confers. From a policy perspective, these long-term benefits are fundamental to determining whether rent control passes a cost-benefit analysis.

To answer the main research questions, I use a matching method to compare tract-level average outcomes of areas that received rent control to counterfactual tracts that did not. This corresponds to the average treatment effect on the treated (ATT) of rent control on long-term tract outcomes. I measure these long-term outcomes using the publicly available Opportunity Insights data described in Chetty et al. (2018). This data set is constructed by linking children born between 1978 and 1983 to their childhood census tracts using restricted federal tax data, allowing the authors to measure the average exposure effects of neighborhoods on long-term outcomes such as economic mobility, employment, marriage, teen pregnancy and incarceration. The benefit of this data is that it is able to follow children over time regardless of where they live as adults when measuring tract-level outcomes.

For each tract and outcome variable, the data reports the unconditional mean over all children from the analysis cohort that lived in the tract between ages 6 and 23. As an example, the data on economic mobility reports the tract-level probability that a child will reach the top 20% of the income distribution in 2015. The realization of this variable gives the proportion of all children linked to a tract who have income in the top 20% of their birth cohort. Throughout this paper, I use the term “tract-level outcomes” to refer to the average outcomes of children growing up in a particular census tract.

From 1970 to 1985, a number of municipalities enacted rent control legislation in response to rising inflation and low rental vacancy rates. I identify 116 cities in California, Massachusetts and New Jersey that codify new rent control laws during this time period. For each rent controlled city, I determine the census tracts that comprise the city, enabling me to map rent control laws to the outcome data set which is measured at the tract level. Rent control laws are passed by cities according to an unknown function of economic, housing, demographic, political and other local characteristics. Many of the predictors of rent control are also likely to be correlated with the long-term outcomes of children, implying that the difference in average outcomes between places with and without rent control is a biased estimate for the effect of rent control on long-term outcomes. Unfortunately, the direction of the bias is not immediately clear based on observable traits. While tracts that receive rent control have higher unemployment rates, minority population and single parent rates, for example, they also have higher average income, college attendance and property values.

To recover a causal estimate of rent control on the average long-term outcomes measured at the tract-level, I utilize a Mahalanobis distance nearest neighbor matching procedure to pair each treated census tract with a similar comparison tract that did not receive rent control. The Mahalanobis distance is a common metric used to measure distance between two points based on underlying covariate values. I estimate the Mahalanobis distance between tracts using data from multiple sources including the 1970 decennial census, the 1972 Census of Governments and county-level voting preferences in the 1968 presidential election. This data allows me to account for observable differences between census tracts and cities that enact rent control. The census data is reported at the tract-level and includes controls for demographic, housing, income and other characteristics that are predictive of receipt of rent control and correlated with the potential outcomes of children in a given tract. The Census of Governments data includes detailed municipal-level information on government expenditures and revenue. Under the strong ignorability assumption that conditioning on the observed covariates removes all confounding variation in the assignment of rent control, I am able to interpret the estimated average treatment effect as a causal parameter. To minimize imbalance on observed covariates in the matched sample, I implement a caliper match to prune treated observations that do not have comparison tracts with similar underlying covariate values. After pruning slightly more than half of all rent controlled tracts, I show evidence that the matching strategy does an adequate job of balancing the covariates for the remaining rent controlled tracts.[2] When estimating the average effects of rent control on treated tracts, I also include bias adjustments for all covariates as proposed by Abadie and Imbens (2011).

My baseline estimates show that rent control leads to a 3.6% increase in the average time

[2] By pruning 50% of census tracts, my analysis sample is no longer representative of the baseline sample of tracts treated with rent control. Despite this, matching models with more permissive calipers yield quantitatively similar results, suggesting that the tradeoff between balance and external validity is fairly small.

that children spend in a given census tract, which implies that rent control laws achieve their primary policy goal of allowing families to stay in their housing for longer. I also show that rent control leads to slight decreases in average tract-level economic mobility. Rent control is associated with a 1.3 and 0.9 percentage point decrease in the average probability of reaching the top 20% of the family and individual income distributions, respectively. These estimates are both statistically insignificant (95% confidence interval of $[\ -0.052, 0.026]$ on a baseline mean of 23.7%). The estimates also show that rent control has a negligible effect on teen pregnancy, incarceration and employment.

The sample of tracts used to estimate the baseline results includes tracts with a low proportion of rental housing that are unlikely to have large direct effects resulting from rent control. I also generate matching estimators while limiting the sample to tracts where rental units represent at least 30% of all housing units. This removes approximately one half of the non-rent controlled tracts and 25% of the treated tracts from the sample. I find that when limiting the sample to high rental tracts, rent control increases the average time that a child spends in a given tract by 12%. This provides even stronger evidence that rent control leads families to remain in rental housing, since the tract-level effect is magnified in areas where we expect there to be more rent control.

Using the high rental sample, I also show that rent control increases the average probability of reaching the top 20% of the family (individual) income distribution by 5.9% (3.9%). Rent control also increases the tract-level average employment rate by 2.7%. Alternatively, rent control has a minimal effect on the average teen pregnancy rate while increasing the tract-level probability of being incarcerated during the 2010 census by 16.8%.

Assessing the results from the baseline and high rent samples, there is suggestive evidence that rent control does improve the average tract-level economic mobility in areas with a high percentage of renters, while also negatively impacting the economic mobility of non-rent controlled children living in cities with rent control. This result is consistent with the literature on the impact of government transfers on child outcomes as well as the literature showing that rent control is associated with negative spillovers on non-controlled housing. Furthermore, I use data from the 1980 to 2000 censuses to show that rent control leads to a decrease in the tract-level percentage of college attendance and an increase in the tract-level poverty rate and unemployment rate. According to Chetty et al. (2018), each of these demographic variables is associated with declines in the long-term outcomes of children.

An important shortcoming of the Opportunity Insights data is the fact that it is reported at the census tract level. Tract-level averages will include many children who did not grow up with rent control when aggregating over all children in a tract, making it more difficult to credibly measure small treatment effects. Assuming 20% of all children in a tract receive rent control, and rent control improves the average probability of reaching the top income quintile by 10% from a baseline of 0.1, the average tract level economic mobility rate would be $(0.8 \times 0.1) + (0.2 \times 0.11) = 0.102 \approx 0.1$. In this hypothetical example, the tract-level outcome is approximately unaffected by the existence of rent control despite the large benefit it provides to children who live in rent controlled units.

Lastly, I utilize the Opportunity Insights data on predicted outcomes for children at the 25th and 75th percentiles of parental income to determine how rent control affects children at the bottom and top of the parent income distribution. I generate estimates using both the baseline sample and the high-rent sample. In the baseline sample, rent control has a negative effect on the predicted economic mobility of children with parents at the 25th percentile of income distribution. By contrast, rent control has a minimal effect on the economic mobility of children with parents at the 75th percentile of the income distribution. In the high-rent sample, rent control leads to small and statistically significant improvements in the predicted economic mobility for children at the 25th percentile of the parent income distribution. For children at the 75th percentile of the family income distribution, rent control has a significant positive effect on predicted rates of reaching the top income quintile as adults. The results from the high rent sample suggest that rent control helps individuals at the bottom of the income distribution, though the effects are substantially stronger for children growing up at the top of the parent income distribution. These results are consistent with previous findings showing that the benefits of rent control may be larger for higher income families; however, I view the results from this exercise as merely suggestive and warranting future study with individual-level data to better grasp the heterogeneity of the effect of rent control on future outcomes by income levels.

This research adds to the economics literature on rent control by tracking the outcomes of people that are affected by the policy and quantifying the long-term benefits. This article is the first to estimate these benefits in a causal framework and will be of immediate interest to policymakers deciding whether rent control policies pass a cost benefit analysis. While the results from the high-rent sample suggest that there are positive long-term benefits for children growing up with rent control, future research can build on this work by generating more precise estimates of these effects and potentially utilizing alternative data sources to leverage individual-level variation in the assignment of rent control.

1.2 Relevant Literature

Rent control is a commonly studied topic in the economics literature going back to Grampp (1950) who argued strongly in favor of removing rent regulations to help avoid housing shortages and improve economic efficiency. This view is consistent with the implications of a simple supply and demand model which predicts that rent control leads to over-consumption and deadweight loss. For many years, a lack of natural experiments and suitable data prevented economists from estimating well-identified causal effects of rent control. As a result, there is a significant body of theoretical work exploring the implications of various rent control regimes (Fallis and Smith, 1984; McFarlane, 2003; Suen, 1989, for example). In addition, the standard economic model's clear predictions of efficiency costs due to price controls may have led some economists to think that empirical research on this topic would be superfluous (Gyourko and Linneman, 1990).

There are a number of papers that attempt to quantify the costs of rent control on housing quality (Moon and Stotsky, 1993; Gyourko and Linneman, 1990; Sims, 2007, for example), generally showing that rent controlled units are maintained at a lower quality than they would be in the absence of price controls. Other research shows how rent control negatively affects housing prices for the controlled (Autor et al., 2014) and uncontrolled stock (Fallis and Smith, 1984; Early, 2000). Lastly, there is a body of literature that attempts to characterize and quantify the costs of rent control that result from inefficiently long tenant durations (Krol and Svorny, 2005; Ault and Saba, 1990; Ault et al., 1994) and the misallocation of tenants and apartments (Glaeser and Luttmer, 2003).

Measuring the benefits of rent control to renters can be challenging without longitudinal data. There is ample evidence that rent control is associated with increased tenant durations which implies that renters with rent control receive some benefit from the policy (Olsen, 1972; Ault et al., 1994; Nagy, 1995; Munch and Svarer, 2002, for example). One common method used to estimate the size of the benefit in the absence of exogenous variation is to measure the difference between controlled rent and the predicted rent that would occur in the absence of controls. This can be done using the two-step method proposed by Gyourko and Linneman (1989) to estimate hedonic rent regressions of the uncontrolled rental stock on housing characteristics. These regressions are then used to predict what the rent would be at controlled units, conditional on observable characteristics. The difference between the predicted and actual rent can be thought of as the compensating variation or monetary value of rent control to the renter. In the second step, one can regress the compensating variation on tenant characteristics to determine how the benefits of rent control are distributed to different groups. Other papers that use this methodology include Gyourko and Linneman (1990), Ault and Saba (1990), Munch and Svarer (2002), and Early (2000). In general, these papers find that the benefits are not particularly well-targeted to the lower-income groups that price controls are intended to help.

In 1995, Massachusetts voters banned rent control in a closely contested statewide referendum, providing economists with a natural experiment to measure the effect of the end of rent control. Sims (2007) was the first to utilize this policy variation and found that rent control in Boston was associated with both decreases in the price of housing as well as housing quality. While his results indicate that rent control had no impact on the construction of new housing, he presents evidence that rent control decreased the value of neighboring, unaffected housing stock, though by a relatively small amount. Autor et al. (2014) focus on Cambridge, Massachusetts and study both the direct effect on home values of rent decontrol, and the effects on housing values of homes that were never regulated. They find that rent control suppresses the value of controlled homes, and also find substantial neighborhood effects implying that rent controlled units also suppress the value of nearby unregulated units. The end of rent control in Cambridge caused nearly $2 billion in housing value appreciation. Using a similar methodology in a follow-up paper, Autor et al. (2017) utilize detailed crime data from 1992 to 2005 to measure the effect that the end of rent control had on local crime rates. They conclude that areas with more rent controlled housing prior to decontrol saw larger decreases in crime rates than otherwise similar areas. This implies that the end of rent control had a significant effect on decreasing crime rates in Cambridge and accounts for 15% of the home value appreciation as a result of rent control found by Autor et al. (2014).

Building on the research using the 1995 Massachusetts rent decontrol natural experiment, Diamond et al. (2019) estimate a well-identified causal effect of rent control on tenants, landlords and inequality using a 1994 change in the San Francisco rent control regime. Prior to 1994, all buildings built before 1980 were subject to rent control except those that contained four or fewer units. In 1994, the small building exemption was removed such that all rental buildings with four or fewer units were now subject to rent control. Buildings with four or fewer units built after 1980 continued to be exempt from the rent control ordinance, providing a natural control group. Using a novel linkage, they collect address histories for San Francisco residents in addition to building and landlord information. The authors find that receipt of rent control led to a 15% increase in the duration of rental stays; however, they also find that rent control incentivizes landlords to remove units from the market, thereby decreasing the number of rental units. Unlike previous studies, they conclude that the benefits of rent control were well targeted to minorities, but that rent controlled units were more likely to be in neighborhoods with lower amenities (where the benefits of rent control are lower). Lastly, because landlords were more likely to remove rent controlled units in neighborhoods with more amenities, the authors argue that rent control has accelerated gentrification, inequality and rental prices in San Francisco. Asquith (2018) shows a similar result that San Francisco landlords react to increasing land values by removing tenants through no-fault evictions, allowing them to convert to non-rental uses such as condominiums.

The paper by Diamond et al. (2019) is the only research that attempts to track renters over time to measure the effect of rent control on the mobility and location decisions of tenants; however, we still do not know how rent control affects important long-term economic, labor market and social outcomes for tenants. Measuring these outcomes is key to a more complete understanding of the costs and benefits of rent control. I fill a critical gap in the literature to date with evidence from a new intergenerational perspective on the consequences of rent control.

1.3 Rent Control Sites and Institutional Background

The rent control policies that I utilize in this paper were passed in the 1970s and early 1980s, and are considered part of the second generation of rent control in the United States; however, rent control policies in the U.S. date back to the end of World War I when housing shortages in a number of cities caused states to restrict evictions of soldiers and workers involved with the war effort.[3] During World War II, the Federal Emergency Price Controls Act (EPCA) subjected many aspects of the economy to price regulation including rental housing. These federal controls continued in modified form until 1952, though some areas and units that had been initially controlled by the EPCA were decontrolled before the end of federal rent regulation. The 1947 Housing and Rent Act gave states additional authority to either extend rent regulations or decontrol rents on their own. By 1948, 10 states had some form of rent control legislation, though by the mid 1950s, New York was the only remaining state with rent controlled housing (Lett, 1976).

In the 1970s, high levels of inflation and low rental vacancy rates in cities around the country led renter advocates to push for new laws to regulate the level and growth of rents. As a result, a number of states and municipalities began to implement new rent control regimes. States that added rent control during this wave include California, New Jersey, Maryland, Massachusetts, the District of Columbia and Alaska. I focus on the laws passed in cities in California, Massachusetts and New Jersey.[4] [5] In most of these places, the justification for passing rent control was that low rental vacancy rates coupled with large rent increases

[3] See Fogelson (2013) for a detailed historical account of New York City's experience with rent control in the post World War I period.

[4] In Maryland, Takoma Park enacted rent control in 1981. Lett (1976) claims that a number of other counties implemented rent control in the early 1970s, though I have been unable to independently verify these laws. As a result, I drop Maryland from the analysis sample to ensure that I do not have measurement error in the treatment group.

[5] Washington D.C. implemented rent control in 1975 in the middle of a major, unrelated demographic shift. The population in Washington D.C. dropped 15% between 1970 and 1980 and declined from 800 thousand residents in 1950 to 570 thousand in 2000. These unrelated changes might add additional noise to the measured effects of rent control leading me to drop Washington D.C. from the analysis.

constituted an emergency that incentivized “rent gouging” and placed many residents at risk of eviction (Lett, 1976). Furthermore, evicted residents were more likely to end up homeless given inadequate supplies of rental housing. Many other cities and states considered implementing rent control during this time but had proposals fail to garner sufficient support.[6] In Maine, the state passed legislation enabling localities to implement rent control though none ended up being passed (Lett, 1976).

Rent control laws regulate the legal terms of rental agreements as well as the rent that a landlord can charge a tenant. In practice, there are many different ways that governments implement rent regulation. In the most restrictive cases, governments determine an exact price for rental housing or place a freeze on rents to prevent them from increasing for any reason.[7] In other cases, governments may place limits on the maximum possible rent increase, either through arbitrarily defined price ceilings or by tying rent increases to inflation. In these cases, rent control is only binding if the landlord would be able to raise rents above the government imposed limit in a competitive market. Another common aspect of rent control legislation is vacancy decontrol, which determines the rent that landlords are allowed to charge the next tenant after the previous tenant voluntarily vacates the rental. Depending on the law, landlords may be allowed to raise rents to market rates under full vacancy decontrol while in other cases, landlords may only be allowed to raise rents by a fixed percentage. Lastly, rent control legislation often places additional limits on evictions though the implementation of eviction restrictions varies widely by location.

In New Jersey, most rent control laws are based on the legislation enacted in 1972 by the municipal government of Fort Lee, which was the first city in New Jersey to implement rent control. Landlords quickly challenged the legality of the legislation, but in 1973, the State Supreme Court of New Jersey ruled that local governments were allowed to regulate local rent. After the court decision affirming the legality of local rent control, many New Jersey municipal governments followed Fort Lee’s example and instituted their own regulations. By 1976, nearly 100 cities and townships had rent control laws. Although the laws in New Jersey are not identical, many are based on the original law from Fort Lee (Lett, 1976). In general, the laws set base rents at current (as of the date of enactment) levels and then tied allowable rent increases to inflation. The laws generally exempted small-scale landlords (usually owners of buildings with fewer than three rentals) from the law. Also, rental units constructed after rent control enactment were often exempt from the legislation to incentivize new housing, and landlords were given permission to raise rents by more than inflation in the event that

[6] For example, municipal rent control proposals in Colorado, Pennsylvania and Wisconsin were all considered and ultimately not approved during the early 1970s.

[7] For example, see Washington D.C.’s temporary rent freeze, Regulation 74 - 13 passed in 1974.

they invested in capital improvements or if operating costs increased.

In Massachusetts, the state passed rent control enabling legislation in 1969 which allowed certain cities to pass rent control laws. Following this legislation, Boston, Somerville, Cambridge, Brookline and Lynn passed rent control laws which went into effect in 1970. Lynn and Somerville repealed rent control in 1974 and 1979 respectively. Under the laws, base rents were set at current levels and rents were allowed to increase to return a reasonable net operating income. Rent increases were also allowed for capital improvements and changes in operating expenses. New buildings and units in owner occupied houses were exempt, while all other extant rentals were subject to the law (Lett, 1976). In 1995, the voters of Massachusetts narrowly approved a referendum which made it illegal for cities to enact rent control legislation. Although Boston and Brookline loosened rent control restrictions prior to 1995, both still had a substantial percentage of units subject to control (Autor et al., 2014). Cambridge still had a heavily controlled housing market at the time of the ballot initiative in 1995.

In California, Berkeley was among the first cities to pass rent control legislation in 1972; however, this law was ruled unconstitutional by the California Supreme Court. Starting in 1979, a number of large cities began passing rent control laws including San Francisco, Los Angeles and Oakland. By 1985, 12 cities in California had implemented rent control legislation. Rent control laws varied by the city; however, all cities were forced to adhere to both the Ellis Act passed in 1985 and the Costa-Hawkins act passed in 1995. The former allowed landlords to evict tenants if they wished to remove their rental housing from the market. This was passed by the state legislature in response to a State Supreme Court ruling which stated that cities could prevent landlords from evicting tenants even when the landlord wanted to occupy the house. This forced cities with strong restrictions to allow landlords the ability to exit the rental market, though the administration of this law varied by city. The Costa-Hawkins act forced cities to allow for vacancy decontrol after a renter leaves a rent controlled unit. This allowed base rents to rise to reflect market conditions after a tenant leaves, regardless of the rent that the previous tenant paid. In terms of exemptions, most new construction and small-scale rental buildings (1-3 units) were not subject to rent regulation.

In summary, rent control laws rolled out across California, Massachusetts and New Jersey in the 1970s and early 1980s. While these policies were not identical, each placed new restrictions on landlords that could have long-run consequences for tenants, especially children. Next I discuss the data I use to estimate these effects before presenting the key results.

1.4 Data

In the states that passed rent control legislation beginning in the 1970s, the power to implement rent control devolved to local political units, meaning that rent control existed in some cities but not others. This was particularly true in California, Massachusetts and New Jersey, which are the three states I focus on to estimate the effect that rent control has on long-term outcomes of children. I follow Krol and Svorny (2005) in using Lett (1976) to collect information on local rent control laws passed in the 1970s, particularly in New Jersey and Massachusetts. This book includes a comprehensive list of cities that passed rent control by 1976, covering nearly all of the New Jersey and Massachusetts cities that added rent control. I supplement this resource with internet searches of legislative histories for large municipalities, particularly in California, to determine which cities added rent control legislation in the years following 1976.

Throughout the paper, I measure rent control as a binary variable and do not distinguish between rent control policies in Massachusetts, California and New Jersey. Though there are differences in the regulations across municipalities and states, the number of different cities with unique regulations makes it difficult to account for policy variation. Future research should attempt to study specific dimensions of rent control and the heterogeneity of treatment effects by rent control policy type. Despite this, there are reasons to believe that policies within states are fairly similar to each other. Most laws at this time were in reaction to low vacancy rates which allowed landlords to raise rents quickly. In New Jersey, rent control laws are enacted around the same time and are based on a law passed by the municipal government of Fort Lee. In Massachusetts, despite some differences between Boston, Brookline and Cambridge, all three cities had a substantial number of controlled units until 1995 when all rent control laws were invalidated by the statewide ballot initiative. Lastly, in California, the laws across cities had similar exemptions and rent increase mechanisms and were subject to statewide legislation that standardized vacancy decontrol and landlord exit.

I utilize geographic and shapefile data provided by the Census Bureau to identify census tracts that were subject to rent control. First, I merge a Census shapefile of incorporated cities from the 1990 census with a shapefile of the 2010 census tracts to create an overlap layer. Using this intersection, I determine the 2010 census tracts that comprise every city in the country. I then merge in the list of cities passing rent control between 1970 and 1985 to generate a binary variable for rent control status for each census tract in the United States. In New Jersey, there are a number of municipalities that passed rent control that do not appear in the list of census places. For these remaining locations, I use the more detailed county maps provided at `www2.census.gov` to manually identify the tracts associated with each rent controlled city. This leaves me with a database of census tracts for California, Massachusetts and New Jersey along with a binary variable indicating whether the tract had rent control established between 1970 and 1985.

Since rent control is implemented by local elected representatives, the decision to enact these policies is likely dependent on underlying observable and unobservable city characteristics. In other words, rent control is not assigned randomly throughout the country, so comparing outcomes of rent controlled and uncontrolled cities is likely to be a biased measure of the effect of rent control. Instead, I utilize data from the 1970 Decennial Census as balancing covariates in a matching framework to estimate the effect of rent control on long-term outcomes. The Census data is pulled from the SocialExplorer website, which aggregates individual responses from the 1970 Census up to the census tract level. Census tract borders change over time, so I use the 1970 Census data that is reported at the 2010 census tract level to maintain a consistent measure of geography.[8]

The matching procedure also includes city-level data on municipal spending and revenues. The municipal tax and revenue data comes from the Government Finance Database described by Pierson et al. (2015).[9] The database compiles information from the Census of Governments beginning in 1967. In years ending in either a 2 or 7, the U.S. Census Bureau collects information on the finances of every incorporated government in the United States. Unfortunately, this full census did not begin until 1972, so I use the data collected from the 1972 census to ensure that I have maximum coverage of all cities in my sample.

The enactment of rent control is a local political decision. As a result, it is necessary to control for local political views when comparing places that did or did not have rent control. To account for local political differences, I use data on the county-level partisan vote shares for the presidential election of 1968. This data is collected by Clubb et al. (2006) and distributed by the ICPSR.

The data for long-term outcomes is described in Chetty et al. (2018) and is available for public download on the Opportunity Insights website.[10] This data combines multiple sources of restricted government data to measure a series of financial, social, educational and other outcomes for children born between 1978 and 1983. Using federal income tax returns from

[8] SocialExplorer uses the area interpolation method described in Logan et al. (2014) to convert 1970-1990 tracts to the 2010 tracts. Area interpolation assigns populations from one area to another based on area overlap and does not account for the distribution of population density within tracts. This is a potential source of error, particularly in tracts with changing borders and unequal distribution of population. To the best of my knowledge, no research has theorized the direction of the expected bias.

[9] The data is publicly available at https://willamette.edu/mba/research-impact/public-datasets/; however, I downloaded the data through the Inter-university Consortium for Political and Social Research (ICPSR) website (study number 37641): https://www.icpsr.umich.edu/web/pages/ICPSR/index.html.

[10] https://opportunityinsights.org

1989 to 2000, the authors identify all children who are listed as tax dependents and were born between 1978 and 1983. Next, they utilize the Census Bureau's Protected Identification Key (PIK) to link these children to the 2000 and 2010 Decennial Census waves, 2000-2015 American Community Surveys and IRS income tax returns from 1989-2015.[11] The sample selected is representative of all children in the 1978 to 1983 birth cohort that were born in the United States or authorized immigrants and whose parents were either born in the U.S. or authorized immigrants.[12]

Once the sample is selected, the authors map children to the census tracts they grow up in through their age 23 year. A child born in 1983 can be linked to a particular tract through 12 distinct years of tax returns (1989, 1994-1995 and 1998-2006) between ages 6 and 23. Children born in 1978 are only linked to 7 years of tract data (1989, 1994-1995 and 1998-2001) between ages 11 and 23. For each 2010 census tract, the authors report the unconditional mean outcome value for all children linked to the tract. Children that appear in multiple tracts due to childhood moves are weighted to represent the relative time spent in each tract. As an example, a child born in 1983 who is linked to tract A in 6 years of tax returns and tract B for the other 6, would receive 0.5 weight in both tracts A and B when calculating tract level outcomes.

I focus on six census tract outcomes reported in the Opportunity Insights data: (1) the probability of reaching the top income quintile,[13] (2) the average income percentile,[14] (3) the probability of having a teen birth, (4) the probability of being in a correctional facility at the time of the 2010 Decennial Census, (5) the probability of having positive W2 earnings in 2015, and (6) the probability of living in a low poverty neighborhood as an adult. These variables allow me to measure how rent control impacts long-term economic mobility and other important social outcomes.

In addition to the unconditional mean outcomes, Chetty et al. also report average predicted outcomes at the tract level for children at 5 different levels of the parent income distribution. These fitted values are generated from a regression of individual outcomes on parental income level at the tract level. The regressions are estimated using all children linked to a particular tract (weighted for the number of linked years). I utilize the estimates at the 25th and

[11] The PIK is created using a probabilistic matching algorithm that is based on an individual's Social Security Number, as well as name, date of birth and address. The PIK can be used to follow an individual across a number of Census Bureau, IRS and other governmental data sets.

[12] External validity is a common concern when using data limited to those who file tax returns. According to Chetty et al., the sample used to create the public Opportunity Insights data is representative of the overall population covered by the American Community Survey and the Current Population Survey.

[13] The income quintile is measured relative to all other people born in the same year to account for rising expected earnings with age.

[14] Income percentiles are reported based on either the distribution of family income or individual income.

75th percentile of the parent income distribution to investigate whether rent control has a differential effect on children at the bottom or top of the distribution; however, it is important to emphasize that these estimates represent fitted values of a regression and may not reflect true outcomes of children. For example, a very wealthy tract may have relatively few parents at the bottom of the national income distribution. Running a regression of child outcomes on parental income percentile might predict that children with parents at the 25th percentile have positive outcomes in this hypothetical tract; however, this prediction is based entirely on a projection of children at the top of the parent income distribution. I only focus on predicted values from the 25th and 75th percentiles, ignoring estimates from the 1st, 50th and 100th percentiles. See appendix A.1 for a slightly more technical description of the Opportunity Insight data.

1.5 Methodology

Rent control legislation is not randomly assigned to cities, but is instead implemented according to an unknown function of local housing, demographic and political characteristics. Tables 1.1 and 1.2 show summary statistics broken down by rent control status at the tract and municipal level respectively. In both tables, the treated group are the tracts and cities located in California, Massachusetts and New Jersey that receive rent control between 1970 and 1985, while the control groups represent tracts and cities from the remainder of the country that are located in incorporated cities with a population greater than 5,000 people and that are represented in the 1972 Census of Governments. I also remove observations from New York, Maryland and Washington D.C. due to a mix of timing issues (New York), rent control measurement ambiguity (Maryland), and likely confounding variation (Washington D.C.).

Out of 31,261 total census tracts, 2,444 are treated with rent control. These 2,444 treated tracts comprise the 99 cities that received rent control between 1970 and 1985, have populations over 5,000 and responded to the 1972 Census of Governments.[15] In California, 26% of the population lived in cities with rent control. In Massachusetts, 15% of the population lived in a rent controlled city while in New Jersey, over 48% of the state population lived in cities with rent control.

From Table 1.1, it is clear that tracts that receive rent control are fundamentally different than those that do not. The difference in average value between treatment and control groups is statistically significant for most of the covariates suggesting that rent control is

[15] There are 16 small cities that enacted rent control between 1970 and 1985 but are not included in the Census of Governments. These are mostly located in New Jersey.

not distributed quasi-randomly. Instead, places with rent control have a higher single parent rate, minority population and higher unemployment rate, which are all variables that are negatively correlated with long-term outcomes for children. On the other hand, tracts with rent control have higher college attendance rates, average incomes and home values which are correlated with improved child outcomes.

In Table 1.2, I report summary statistics at the city level comparing cities that enacted rent control to those that never implemented a rent control law. Not surprisingly, cities with rent control had lower vacancy rates and a higher percentage of rentals as a share of total units. These cities with rent control were more likely to be in counties that had higher share of votes for Hubert Humphrey, the Democratic candidate, in the 1968 presidential election. In addition, the cities with rent control have higher municipal revenue per capita, and spend a higher fraction of total expenditures on education, police and welfare compared to non-rent controlled cities. The results from Tables 1.1 and 1.2 show that assignment of rent control is correlated with observable characteristics that are also likely correlated with the long-term outcomes of children.

Each row of Table 1.3 represents an outcome variable of interest. We can see that the average fraction of years spent in a rent controlled tract is only slightly (0.4 percentage points) higher than the non-controlled tracts; however, on average, tracts with rent control are much more likely to report a higher probability of children living with their parents as adults and staying in the same tract or commuting zone as an adult. Rows 5 through 12 show that tracts with rent control report greater average economic mobility, lower teen pregnancy and incarceration rates and higher rates of employment and living in tracts with low poverty rates as an adult. In general, the naive treatment effects suggest that rent control improves long-term outcomes; however, the differences in pre-rent control characteristics, particularly at the tract-level, imply that a simple difference in outcomes between treated and control tracts is likely to be a biased measure of the effect of rent control. Unfortunately, it is not immediately clear which direction this bias shifts the naive estimates given the countervailing effects of the individual variable imbalances.

In many observational studies, treatment is not assigned randomly but is instead assigned according to some (unknown) function of observable and unobservable characteristics. I lean on Rubin's model of causal inference to formalize the analysis and provide theoretical justification for the estimand of interest (Holland, 1986). In my setting, there is a rent control treatment $T \in \{0, 1\}$ that is assigned to the population of census tracts of size N. I hypothesize that rent control treatment T has a causal effect on long-term average outcomes at the tract level, denoted $Y(T)$. The causal effect of T on Y_i for tract i can be measured as $Y_i(T = 1) - Y_i(T = 0)$. Aggregating up to the full sample of census tracts, the average treatment effect (ATE) can be written as $\frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$. Alternatively, the average treatment effect on the treated (ATT) which measures the effect of a treatment only on the treated tracts, is written as $\frac{1}{N_1}\sum_{i=1}^{N_1}([Y_i(1) - Y_i(0)]|T = 1)$. Since we cannot observe the treated and untreated potential outcome at the same time, measurement of the causal effect of interest is reduced to a missing data problem. Unless otherwise noted, I estimate the average treatment effect on the treated throughout my analysis.

To bypass the missing data issue, I utilize techniques that balance the treatment and control groups on observable characteristics to find a suitable counterfactual and allow for improved estimates of the effect of rent control on outcomes. Rosenbaum and Rubin (1983) were among the first to formalize the framework for achieving causal estimates in observational studies by conditioning on a vector of control variables to facilitate matching. The assumptions required to identify the average treatment effect on the treated of rent control can be written as:[16]

$$(Y_0) \perp T|X, \;\; pr(T = 1|X = x) < 1$$

where X is a vector of covariates and $pr(T = 1|X = x)$ is the probability that a tract is rent controlled conditional on any realization of the covariate vector. The second part of the assumption states that the covariates cannot perfectly predict treatment. Under this strong ignorability assumption, Rosenbaum and Rubin show that treatment effects can be recovered by matching observations of different treatment levels with the same value of the conditioning function based on X. In the original case, Rosenbaum and Rubin use an estimated propensity score; however, any function of X can be used. The intuition behind this result is that controlling for the covariates X is sufficient to make the treatment assignment random. For this to be true, there must not be any unobserved variables that predict treatment and are correlated with the outcome after controlling for X. The assumption that controlling for X removes all confounding variation is quite strong and hard to prove.

I use the Mahalanobis distance metric (MDM) to determine the “closest” counterfactual match for each rent controlled tract. The Mahalanobis distance for any two tracts i, j is calculated as:

$$MDM_{i,j} = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$$

where X is a vector of covariates and S is the covariance matrix of X. The intuition behind the Mahalanobis metric is that it calculates the distance between two points in a way that is

[16] Note that the identification assumption is slightly less restrictive than the one needed to identify the ATE (Abadie and Imbens, 2006).

independent of the scale of each component of X. Recent work by King and Nielsen (2019) suggests that matching on the MDM is preferable to matching on estimated propensity scores, since the matched pairs come closer to mimicking a fully blocked experiment compared to propensity score matches which mimic a fully randomized experiment.[17] Fully blocked experiments are more efficient and should have substantially less noise in the estimated treatment effects. I match tracts with replacement, which allows a control tract to serve as the counterfactual for multiple observations.

Throughout the paper, I use 38 main covariates to calculate the MDM as well as test for balance in the subsequent matched or weighted samples. In Chetty et al. (2018), the authors report tract level variables that correlate with long-term outcomes. In particular, they show that education, poverty rates, single parenthood, income, unemployment, minority population and proxies for social capital are all correlated with economic mobility at the tract level. Therefore, it is crucially important to include these variables when calculating the distance between tracts and to ensure that they are balanced in the post-match sample. In addition to the variables highlighted by Chetty et al., I also include various housing and population variables that are likely to predict the imposition of rent control at the municipal level. These variables include per capita municipal revenue, expenditures and county level data on voting behavior from the 1968 presidential election. The municipal and county level data allows me to control for city-level differences that are not accounted for at the tract level and that could be correlated with the outcomes of interest. The full set of covariates are the same tract and city-level variables that are listed in Tables 1.1 and 1.2.

As Ho et al. (2007) suggests, the main goal of a matching strategy is to reduce covariate imbalance across treatment status (with the hope that the covariates remove all confounding variation). Despite the large body of research on propensity score and distance metric matching, there is limited consensus on the best way to estimate matching metrics (Hainmueller, 2012). This is particularly true in the context of estimating propensity scores, though still relevant for the MDM when deciding which variables to include and whether to use higher order terms when assessing the distance between two observations. Given the large number of tract and city-level covariates I control for, I do not include any higher order terms when calculating the Mahalanobis distance.[18]

Another common strategy in the matching literature is to limit matches to observations

[17] In a fully randomized experiment, treatment is assigned at random across a given population. In a fully blocked experiment, the sample is stratified based on pre-treatment characteristics, and treatment is assigned randomly within strata.

[18] There is some older research showing that the Mahalanobis distance metric performs poorly with many covariates (Gu and Rosenbaum, 1993, for example); however, in my context, the Mahalanobis distance provides the best matches resulting in the lowest residual imbalance on observables despite the large number of covariates.

that are within a given distance caliper or radius. Conducting a nearest neighbor match without a caliper will find the closest match for each treated observation. As one decreases the size of the matching caliper, the matches that are farthest in measured Mahalanobis distance are pruned, leaving the better matches with closer covariate realizations. This process of determining the match caliper is another way of describing the bias - variance trade off common to many empirical approaches. In addition, as treated observations are pruned, the estimated treatment effect from the reduced sample may not be relevant for the target sample. As a result, it is up to the researcher to determine the caliper that minimizes covariate imbalance while also maintaining a sufficient sample of observations to estimate treatment effects.

I use an iterative process to determine the optimal caliper by instituting different caliper values and checking the number of treated observations that are pruned and the resulting covariate imbalance. In all iterations of the model, I check covariate balance by comparing the standardized mean difference between the treated and control group before and after implementing the matching procedure. The standardized mean difference is given by $\frac{\bar{Y_1}-\bar{Y_0}}{\sqrt{\frac{V_1+V_0}{2}}}$, and is thus measured in units of the pooled standard deviation of the treatment and control means (Austin, 2009). Its use allows us to standardize the measure of divergence regardless of the units of each covariate. In addition, it satisfies the conditions of a good balance check statistic suggested by Imai et al. (2008) in that it is a characteristic of the sample (and not a hypothetical population) and that the value is unaffected by sample size. It is also important to note that the iterative caliper selection process is estimated without estimating treatment effects to avoid data mining particular results while searching for the optimal match radius.

Once I have finalized the caliper selection, I estimate the average treatment effect on the treated for all outcome variables by averaging the difference between the treated and matched counterfactual across all treated tracts. I also account for the bias that results when matching on continuous covariates (Abadie and Imbens, 2006) by including post matching regression adjustment on all covariates. This is equivalent to the bias-adjusted matching estimators proposed in Abadie and Imbens (2011). I report clustered standard errors at the city level using the influence function proposed by Jann (2019). Note that this standard error is likely conservative, compared to the consistent standard errors proposed by Abadie and Imbens (2006) which are harder to calculate while including clusters. In addition Abadie and Imbens (2008) show that bootstrapped standard errors are not consistent for matching estimators despite their use in the literature.

The ideal post-matching sample will have a standardized mean difference of 0 though researchers sometimes use simple thresholds such as 0.1 or 0.25 to signify when a covariate has a large enough difference to warrant returning to the matching procedure (Stuart et al., 2013; Rubin, 2001; Cohen, 1977; Normand et al., 2001; Austin, 2009). Ultimately it is up to the practitioner to identify whether the imbalanced covariate could plausibly lead to biased estimation of treatment effects. Slight imbalances of covariates that are highly correlated with the outcome will be more problematic than larger differences in a variable which has a weak correlation with the outcome of interest.

I also measure the ratio of the mean variance for the treated and control groups for each covariate. Austin (2009) shows that propensity scores that are balanced on means still may be incorrectly specified. Covariates that are fully balanced should have a variance ratio of 1, though the literature gives little direction when it comes to determining what variance ratio threshold implies an imbalanced covariate. I view variance imbalances as subordinate to imbalances of the mean when determining the match caliper.

Table 1.4 show the standardized mean differences (SMD) and variance ratios for the raw and matched data after selecting the optimal caliper. Given the large number of covariates, the resulting imbalances are expected though larger than ideal. The caliper drops slightly more than half of the original treated tracts, leaving a sample of 1,174 tracts which will be used to estimate the relevant ATTs. Out of 38 covariates, 18 have SMD of less than the desired 0.1, while 13 have SMD values of between 0.1 and 0.25. This leaves 7 covariates that have post matching SMD values of greater than 0.25. The post-matching variance ratios show that the majority of covariates have variance ratios within the $[0.7, 1.3]$ interval; however, there are still 13 covariates with variance ratios outside of this interval, further suggesting some residual imbalance. Given the resulting imbalance, the post matching regression adjustment is especially important when estimating treatment effects.

Across the distribution of included treated tracts, there are a total of 549 unique control tracts that are selected as matches. This means that each counterfactual observation is selected as the closest neighbor match for an average of 2.1 treated observations. The maximum number of times a control observation is matched is 24 while the median is 1. The caliper selection process dropped 1,290 treated tracts including most of the tracts from rent controlled cities in Massachusetts. Of the remaining treated tracts, approximately 35.4% are located in New Jersey (416 tracts), 63.1% are located in California (741 tracts) and 1.4% are located in Massachusetts (17 tracts). These tracts represent 88 out of the potential 99 cities that are included in the final analysis sample.

The observation pruning suggests that the estimated average treatment effect on the treated is not equivalent to the overall sample average treatment effect on the treated. I view the change in underlying sample to be a worthy tradeoff to achieve better covariate balance. It is also worth noting that matching models with more permissive calipers yield very similar results to this more restrictive matching procedure.

Lastly, I also estimate a matching model that limits treated and control tracts to those that have a high proportion of rental units. This removes over one half of the total tracts in the original sample, including 25% of the treated tracts. I use the same baseline covariates and matching caliper to construct these matches, which are used to determine the effect of rent control on a sample of tracts that are especially likely to be affected by the policy. The balance results from this matching procedure are reported in Table A.1. Of the 840 tracts included in the high rent sample, 72% are located in California and 28% are located in New Jersey.

1.5.1 Additional Methodological Assumptions

The publicly available outcome data has shortcomings in the context of my paper that require me to make a strong assumption before interpreting my results. The sample of children included in each tract's outcome is determined by links made starting in 1989, when the analysis cohort is between 6 and 11 years old. Figure 1.1 shows a timeline of the relevant dates when states began implementing rent control, as well as the dates when children were eligible to be linked to a given census tract. Given the timing, I must assume that rent control did not cause any selective immigration to or from census tracts in the years after it is implemented and before 1989. Under this assumption, the people living in rent controlled tracts in 1989 should be roughly equivalent to those living in the matched counterfactual tracts in 1989. This assumption is stronger for the rent controlled cities in New Jersey and Massachusetts than it is in California due to the longer duration between rent control implementation and address linkage.

Prior work suggests that violations of this assumption could bias the estimated treatment effects in either direction. Research by Autor et al. (2014, 2017) shows that the removal of rent control in Cambridge, Massachusetts led to significant property appreciation and decreased crime, suggesting that rent control is associated with reverse gentrification. If this is true, rent controlled areas may be less likely to receive an influx of new residents than similar areas without rent control. Under this scenario, the estimated ATT of rent control on long-term outcomes would likely be biased downward. Alternatively, research by Diamond et al. (2019) suggests differences in the short and long run effects of rent control on neighborhood immigration. While rent control allows some lower income individuals the ability to stay in their homes for longer, it also incentivizes landlords to exit the rental market. In the long run, they suggest that rent control increases gentrification by substituting rental housing for more expensive private uses. Under this latter scenario, we might expect the children living in rent controlled neighborhoods in 1989 and later to be relatively wealthier than the children comprising the counterfactual tracts. This would likely create an upward bias on the estimated effect of rent control on long-term outcomes, since the outcome sample would include wealthier people that select into areas with rent control.

In Figure 1.2, I test this immigration assumption by looking at the effect that rent control has on the tract level proportions of immigrants from other counties, states and countries. These variables are based on the census question asking respondents over 5 years old where they lived 5 years ago and are aggregated up to the census tract level. For each year of the decennial census, I estimate the average treatment effect on the treated on these immigration outcomes using the baseline Mahalanobis distance matching estimator. The top left panel shows the effect of rent control on the tract level percentage of people who migrated from a different census tract. The estimated ATT is negative in 1980 and positive in 1990, suggesting that rent control does not have a consistent effect on cross county migration. The top right panel shows that rent controlled tracts have fewer residents migrate from other states than the counterfactual tracts. The bottom panel shows that rent controlled tracts have a slightly higher proportion of international immigrants than non-controlled tracts in 1980 and 1990, though this effect reverses signs in 2000. These findings are very similar when using the high-rent sample and are shown in Figure A.1. In general, the balance of the evidence from this figure suggests that the Autor et al. prediction is better supported by the evidence, that rent control should lead to fewer new immigrants. Given this fact, I expect any violation of the selective immigration assumption to bias downwards the estimated benefits of rent control.

1.6 Results

I first show evidence that rent control affects the location decisions of the households included in the Opportunity Insights sample. The central finding is that rent control allows families to stay in their housing for longer. Specifically, the first column of Table 1.5 shows that rent control leads to a 2.1 percentage point (3.6%) increase in the tract-level average duration of tenancy compared to counterfactual tracts. Each child included in the Opportunity Insights data can be linked to a particular tract in up to 12 distinct tax years, but these years are not continuous. As a result, a 3.6% increase in the tract-level average of the fraction of years that a child is linked understates the increased tenant durations resulting from rent control. As an illustration of this fact, a child that lives in a given tract from 1989 to 1995 would only be linked to that tract in 3 separate years (1989, 1994, 1995) despite living there for a total of 6 years. I interpret the result as evidence that rent control increases rental durations without focusing on the exact magnitude of the increase.

The second column reports that rent control leads to a 3.1 percentage point (14.6%)

increase in the tract-level probability that a child will live with their parents as an adult. The third column shows that rent control leads to a 1.6 percentage point (7.1%) increase in the average probability that a child will live in their childhood tract as an adult. The final column shows that rent control has minimal effect on the tract-level probability that a child will live in the same commuting zone as an adult. The main takeaway is that rent control allows families to stay in their housing for longer which is exactly what we would expect if rent control laws were binding and provided benefits to families in the form of lower than market rents.

The long-term outcome results are reported in Table 1.6. The columns represent different outcome variables while the top row represents the estimated ATT of rent control on the relevant tract-level average outcome. This row is measured in percentage point differences. Standard errors are reported in parenthesis beneath the ATT estimates.[19] The first two columns report the tract-level probability that a child will reach the top family or individual income quintile in 2015. The third and fourth columns report the tract-level average income percentile in the family or individual income distribution in 2015. The fifth column reports the tract-level probability that someone linked to a given track will have a teenage birth (females only). The sixth and seventh columns report the tract-level probability that someone linked to a given tract will be in jail or prison in April, 2010 or will have positive earnings in 2015. Lastly, the eighth column reports the tract-level probability that someone linked to a given tract will live in a low poverty neighborhood as an adult, defined as a tract with a poverty rate of less than 10% according to 2000 Census data. The baseline row reports the average counterfactual value for the relevant outcome while the %Δ row calculates the percent change the ATT row represents from the baseline value.

Columns 1 and 2 show that rent control has a -1.3 and -0.9 percentage point (-5.4 and -3.5%) effect on the tract-level probability that a child will reach the top family and individual income quintile respectively, though neither result is statistically significant. For the probability of reaching the top 20% of the family income distribution, I can rule out percentage point effects outside the interval $[-0.052, 0.026]$. Column 3 shows that rent control has a small and insignificant (-1.3%) effect on the tract-level average family income percentile and a negligible effect on the average individual income percentile. For the family income percentile, I can rule out percentage point effects outside the interval $[-0.017, 0.031]$. Columns 5 through 7 show that rent control has a minimal effect on the tract-level probability that a child will have a teenage pregnancy, be incarcerated during the 2010 census or report positive employment

[19] As a reminder, the standard errors in Table 1.6 are calculated using the influence function method proposed by Jann (2019) and account for clustering at the city level. These standard errors are likely to be overly conservative compared to the consistent matching standard errors proposed by Abadie and Imbens (2006) which are more difficult to calculate including clusters.

earnings in 2015. Lastly, column 8 shows that rent control is associated with an insignificant 2.6 percentage point (5.2%) reduction in the tract-level probability that children will live in low poverty neighborhoods as an adult.

The unit of observation for the results in Table 1.6 is the tract level. Although tracts are constructed to have roughly the same population, it is possible that tracts with more children may have different estimated treatment effects. Table 1.7 reports the same matching ATT results but weights each tract by the number of children that are used to calculate the outcome variable.[20] While this does not allow for an individual-level interpretation of the results, it does scale the estimated ATT to better reflect differences in the size of tracts and, more importantly, the number of children used to measure the outcome variable. On balance, the point estimates all fail to reject the null hypothesis of no effect and are very similar to the unweighted baseline results.

In the last part of my baseline analysis, I limit the sample of tracts to include only those tracts that have a high proportion of renters, defined as any tract with greater than a 30% share of rentals of the total housing units. This sample modification drops slightly more than half of the non-rent controlled tracts and one quarter of all rent controlled tracts. The theory underpinning this adjustment is that tracts with a high percentage of rentals are more likely to exhibit effects directly related to rent control. Indeed, Table 1.8 shows that rent control leads to an estimated 6.1 percentage point (12%) increase in the fraction of years spent in a given tract and a 2.8 percentage point (13.7%) increase in the probability of living in the same tract as an adult. These results provide further proof that the rent control laws I study allow families to stay in their homes for longer.

In Table 1.9, I show the effect of rent control on long-term outcomes for children growing up in tracts with a high proportion of rentals. Column 1 shows that rent control leads to a 1.1 and 0.9 percentage point (5.9 and 3.9%) increase in the average tract-level probability of reaching the top 20% of the family and individual income distribution. Rent control is also associated with a 1.2 and 1.6 percentage point (2.7 and 3.3%) increase in the tract-level average family and individual income percentile. The coefficient on the individual income percentile is significant at the 10% reporting level. Based on the crosswalk converting income percentiles to income in 2015 dollars, the increase in average individual income attributed to rent control is between $850 and $1,709. In the high rent sample, the average tract is approximately 66% rental housing. If we assume that half of the rental units are rent controlled, and there are no spillover effects on the non-rent-controlled children, then the direct impact of rent control on the roughly 30% of children who receive the benefit would be $2,800 - $5,700

[20] The Opportunity Insights data reports the underlying sample size for each outcome variable.

in 2015 income.[21]

The results in columns 5 and 6 show that rent control has minimal impact on the average teen-pregnancy rate, and leads to a 0.2 percentage point (16.8%) increase in the tract-level probability of being in jail on April 1, 2010. Column 7 shows that rent control leads to a 2 percentage point (2.7%) increase in the tract-level employment average. This estimate is also significant at the 10% reporting level. Lastly, rent control leads to a 1.8 percentage point (4.2%) decrease in the probability of living in a low poverty neighborhood as an adult. One thing to note is that the sample of tracts used for the high-rental estimates is fairly small. By definition, these tracts are not representative of the full sample of rent controlled tracts. Despite this, the change in estimated effects between the full sample results and the high rental results shows that rent control leads to greater economic and employment outcomes in areas where there is a higher probability of residents living in rent controlled housing.

These results are suggestive though not conclusive that rent control leads to long-term benefits for children that grow up in rent controlled housing. The difference in estimated effects between the baseline and high rent sample is also consistent with prior work showing that rent control has negative spillover effects, and these negative effects may harm the long-term outcomes of children that live in close proximity to rent controlled housing. In the next subsection, I try to determine how these long-term impacts are distributed across the income distribution.

1.6.1 Rent Control at Different Points of the Income Distribution

In Table 1.10, I show the estimated average treatment effect on the treated tracts of rent control on the predicted outcomes at the 25th and 75th percentile of the parent income distribution. For these results, the outcome variable is a predicted value of a regression performed for each census tract in the data, regressing the relevant outcome on parent income rank. The 25th (75th) percentile outcome is thus the fitted value from this regression at the relevant parent income level. This means that the predicted outcomes are a projection based on the outcomes of all children that are linked to a given tract.

Panel A of the table shows the effect of rent control on predicted tract-level economic mobility for children at the 25th percentile of parent income. In general, the estimated tract-level effects of rent control on economic mobility for children at the 25th income percentile are similar to those reported in Table 1.6. Column 1 shows that rent control is associated with a 0.8 percentage point (4.9%) decrease in the predicted tract-level probability

[21] The estimate of 50% rent control exposure is a hypothetical based on summary statistics reported by Autor et al. (2014), which shows that roughly 50% of Cambridge condominiums were rent controlled in 1994. Many of the non-controlled condominiums were likely owner occupied, so the true rate of rent control among rental units was likely higher than 50%.

of reaching the top family income quintile, but this estimate has a standard error over 2 times larger than the point estimate. Columns 3 and 4 show that rent control has little effect on the tract-level predicted family and individual income percentile. Column 5 shows that rent control is associated with a 0.9 percentage point (4.5%) decrease in the tract-level probability of having a teen pregnancy. Column 7 shows that rent control is associated with a 1.3 percentage point (1.9%) increase in the predicted tract-level probability of being employed as an adult. This estimate is significant at the 10% reporting level. Lastly, column 8 shows that rent control decreases the average probability of living in a low poverty neighborhood by 1.5 percentage points (3.2%).

Panel B of the table shows the effect of rent control on predicted economic mobility for children at the 75th percentile of parent income. The reported effects of rent control on economic mobility and average income are quite small. Rent control is associated with a slight increase in the average probability of having a teen pregnancy, and a large (though statistically insignificant) decrease in the average probability of being incarcerated.

In Table 1.11, I show the results of this same exercise but limiting the sample to the high rental tracts. In panel A, I show that rent control has a small and slightly positive effect on the average predicted economic mobility of children at the 25th percentile of parent income, though none of the estimates are statistically significant. This contrasts with the effect shown in panel B for children at the 75th percentile of the parent income distribution, where rent control is associated with a 2.9 and 2.2 percentage point (13.2 and 8.2%) increase in the predicted probability of reaching the top family and individual income quintile. The estimates in columns 1-4 are all significant at the 5% significance level, further suggesting that rent control leads to better economic outcomes for children at higher income levels.

Comparing the effect of rent control on average predicted teen pregnancy between panels A and B, the estimates indicate that rent control leads to a 1.2 percentage point (5%) decrease in average teen pregnancy for lower income children, and a 1.6 percentage point (15.5%) increase in teen pregnancy for children towards the top of the parent income distribution; however, neither result is statistically significant.

The results reported in Table 1.11 suggest that there is variation in the benefits of rent control based on family income and that higher income families seem to derive a higher benefit; however, since the outcomes are projections, it is entirely possible that the results are driven by people towards the top of the parent income distribution. As such, these results should be interpreted with care and certainly do not provide conclusive proof of the effect that rent control has on children at the bottom or top of the income distribution.

1.6.2 Rent Control Effects on Housing and Demographics

In this section, I use census data from the 1980-2000 surveys to see how rent control affects the demographic composition of tracts, as well as the housing market. For each census year between 1980 and 2000, I estimate the average treatment effect on the treated tracts of rent control on a host of demographic and housing variables. Note that the census data I use in this section is comprised of people who live in the tracts in the year of the survey as opposed to the Opportunity Insights data which is comprised of people who grew up in a given census tract. Part of the goal of this section is to examine possible mechanisms to explain the baseline ATT results. Prior research has shown that rent control can suppress gentrification in a neighborhood or city by making it easier for tenants to remain in their current housing. This might explain why rent control has a slight negative effect on economic mobility when estimating the baseline ATT model. People who do not receive rent control but live in a city with rent control may receive a negative spillover from the policy. I use the same nearest neighbor matching methodology with bias adjustments to generate these estimates. As a result, the analysis sample is the same, meaning that the results are based on 1,174 rent controlled tracts spread over California, Massachusetts and New Jersey. I also estimate these effects using the high rental sample, though the results are largely the same. These supplemental figures are reported in appendix A.2.

Figure 1.3 displays estimated average treatment effects on the treated for housing market conditions. The first panel on the left column shows the average tract-level effect of rent control on the probability that a resident has been living in their current house for at least 5 years. The estimated effect of rent control on the probability of living in one's house for more than 5 years is statistically significant in all three census years reported. This indicates that rent control increases tenancy duration in tracts with rent control and provides corroborating evidence that rent control allows renters to stay in housing for longer. The top panel on the right shows that rent control leads to a small but negative effect on the average tract-level vacancy rate, though these estimates are not statistically significant.

The middle panels show the effect that rent control has on the tract-level proportion of people that have long commutes, and the tract-level percentage of rentals as a share of total housing units. The panel on the left provides weak evidence that rent control leads to decreased commutes of over 1 hour. This contrasts with Krol and Svorny (2005) who find that rent control in New Jersey lead to spatial mismatch between where people live and work further resulting in longer average commutes; however, none of the point estimates in my analysis are statistically significant.[22] The panel on the right shows that compared to

[22] The result from Krol and Svorny is not causally identified, so the comparison of results is difficult. The results from Table A.2 show estimates for the high-rental samples which are more in line with the results

counterfactual tracts, rent control has a negative effect on the proportion of rentals in 1980, and zero effect on the proportion of total units that are rentals in 1990 and 2000. The panel on the bottom shows that rent control has a statistically significant (though small) effect on the average total number of units in a tract. Taken together, this partially corroborates evidence reported in Diamond et al. (2019) and Sims (2007) that rent control leads landlords to remove rental units from the rental market.

Figure 1.4 displays estimated average treatment effects on the treated tracts for demographic variables that could play a role in changing the long-term outcomes of residents. Chetty et al. (2018) show that long-term outcomes are correlated with tract level characteristics such as the poverty rate, education, unemployment and rate of single parent families. In the top left panel, I show that rent control is associated with a small increase in the proportion of single parent families. In the top right panel, I show that rent control is associated with a consistent decrease in the tract-level share of college graduates. In the bottom panels, I show that rent control is associated with a slight increase in the tract-level poverty rate and increases in the unemployment rate in the 1990 and 2000 decennial census. Although the measured effects are small, the results suggest that rent control leads to a reversal or hold on neighborhood gentrification. This is mostly consistent with results reported by Autor et al. (2014, 2017). In addition, this suggests that the benefits of rent control might be offset by negative spillover effects on long-term outcomes for those that do not live in rent controlled housing.

1.7 Discussion and Limitations

The results from the high rental sample provide suggestive though not convincing evidence that rent control improves economic mobility and employment outcomes in areas where we should see rent controlled tenants exert a greater effect on the tract-level average outcome. I interpret these latter results to indicate that rent control provides a long-term benefit to children that grow up in rent-controlled housing, though the exact magnitude of this benefit is difficult to quantify with the data in this study. While few of the estimated ATT effects in the high rent sample are statistically significant, the data limitations may make it difficult to identify small effects of rent control on renters. In addition, conservative standard errors are likely to overstate the variance of my estimates, making it more difficult to reject the null hypothesis of 0 effect.

The methodology I use is only capable of estimating the net effect of rent control in a census tract and cannot disentangle the effects on rent-controlled tenants and everyone else. As the figures in section 1.6.2 show, rent control is associated with a slight demographic shift

from Krol and Svorny, particularly in 1990 and 2000.

that leads to lower college attendance, higher unemployment rates and higher poverty rates. Chetty et al. (2018) reports that the long-term outcomes of children are negatively correlated with these demographic shifts, implying that rent control could harm children growing up in neighborhoods with rent control but who do not live in rent controlled housing. If this is the case, the estimated benefit of rent control on tract-averages would include both the effect on those who receive rent control and the spillover effects on those children who do not but are affected by the changing neighborhood demographics. This would indicate that the effects I measure understate the true benefit of rent control to children who grow up in rent controlled housing.

Even if we suppose that these spillover effects are negligible, it would still require a large magnitude effect to detect tract-level changes in outcomes as a result of rent control. Suppose there are 100 children in a tract, and the true effect of rent control is that it improves the probability of reaching the top income quintile by 10% (improving the baseline probability from 20 to 22%). If we also assume that half of the children live in rental housing (which is roughly equal to the share of rental housing in rent controlled tracts), and half of these renter children live in rent controlled housing, then 25 children will have improved chances of reaching the top income quintile; however, the tract-level economic mobility probability will be $((0.75) \times (0.2)) + ((0.25) \times (0.22)) = 0.205$, which is roughly equal to the baseline probability of 0.2, even though rent control has a large effect on those that receive the benefit.

From the high rental sample, the estimated effect of rent control on the average tract-level individual income is 1.6 percentiles. The difference in annual income between percentile 49.2 and percentile 50.8 is roughly equivalent to $1,111, measured in 2015 dollars. This represents a 4% increase (from a baseline of $28,500) that is attributable to rent control. This effect is similar though slightly smaller than the one reported byAndersson et al. (2016), who find that an additional year of public housing is associated with a 5% increase in annual earnings.

Another related issue that can attenuate the estimated benefits of rent control is the fact that I do not have tract-level measures of rent control prevalence. By treating rent control as a binary variable, I do not account for the fact that different tracts are likely to have different levels of rent control intensity. Rent control intensity can be thought of as a combination of the number of households that receive rent control as well as the degree that the rent regulations are binding. Given the complicated rules that govern exemptions to rent control laws, it would be very challenging to construct a measure of the true rate of rent control intensity in each tract over all cities in California, Massachusetts and New Jersey that enacted rent control between 1970 and 1985. Despite this, the methodological decision to measure rent control as binary will likely lead to an attenuation bias of the target average treatment effect on the treated. This bias is likely mitigated in the high rental sample analysis, though there is still the probability that there is differential rent control treatment within these tracts leading to measurement error and attenuation bias.

Lastly, the causal interpretation of the matching results rests on the assumption that conditioning the receipt of rent control on the matching covariates removes all confounding variation. I attempt to use a wide variety of tract and city-level covariates when matching to account for as many observable characteristics as possible. The large number of covariates makes it more difficult to balance the full covariate vector after matching, which further leads to possible bias of the estimated ATT on the outcomes of interests. I mitigate this concern by using the bias corrected estimators proposed by Abadie and Imbens (2011); however, this is a suboptimal method for dealing with residual imbalance and likely cannot remove all potential sources of confounding variation.

Future research can avoid these problems using a number of different strategies. Most importantly, access to better data that is reported at the individual child level would allow for one to disentangle the direct and spillover long-term effects of rent control. Additionally, studies that include more detail on the level of rent control intensity would have a better chance of avoiding the attenuation bias that is likely to plague my results. Lastly, researchers can utilize random or quasi-random variation in the assignment of rent control to more confidently avoid potential confounding of the effect of interest.

1.8 Conclusion

This paper is the first to provide estimates of the long-term benefits to children of growing up with rent controlled housing. Rent control is a commonly implemented policy that currently exists in 5 states including some of the largest cities in the country. Despite its ubiquity, previous research on rent control has mainly focused on its impact to housing markets and neighborhoods while ignoring the potential benefits to renters. There has been some attempt to quantify the benefits of rent control by calculating the difference between controlled rent and the rent that would be charged in a competitive market, but these static measures ignore the potential long-term benefits that the policy confers.

I show that rent control leads to statistically insignificant decreases in economic mobility and minimal changes in tract-level teen pregnancy rates, incarceration and employment; however, when limiting the sample of tracts to those that have at least a 30% rental share, I show that the estimated effect of rent control on economic mobility, income and employment is marginally positive. These findings are weakly consistent with prior research showing that rent control leads to negative spillovers on surrounding properties, while also providing benefits to those who live in controlled housing. To further contextualize these results, I use census data from the 1980 to 2000 census to show that rent control leads to tract-level changes in demographics and housing variables that are consistent with rent control suppressing gentrification.

In addition, I show some evidence that rent control can lead to minor improvements in economic mobility for people at the bottom of the income distribution; however, the economic mobility effects of rent control appear significantly stronger for children growing up at the top of the parent income distribution.

These results represent an important first step in measuring the long-term benefits of rent control for children that live in rent controlled housing, as well as the potential negative spillovers that worsen long-run outcomes for those that live in non-rent controlled housing. This research suggests that the long-term direct benefits may be positive, though future research should explore this topic using alternative data to determine whether the estimated effects can be reproduced using individual-level variation instead of aggregated data. Additional inquiry can also help to determine how the direct benefits of rent control compare to the potential negative long-term spillovers which can help policymakers measure the net effect of rent control on residents.

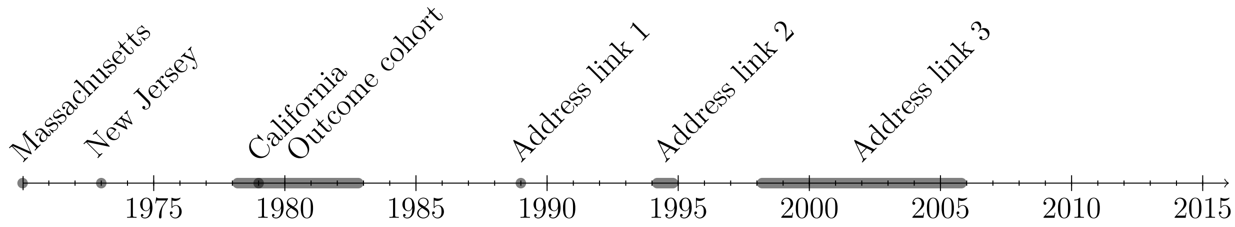


Figure 1.1: Timeline of Relevant Rent Control Laws and Data Construction Dates

Figure shows the years that each state began to implement rent control, as well as the years that are relevant to the construction of the Opportunity Insights data. There is a sizeable gap between the implementation of rent control and the first address link, particularly for New Jersey and Massachusetts. To recover a causal estimate of the impact of rent control, I must assume that there is no selective inmigration to areas with rent control between the date of implementation and the address link.

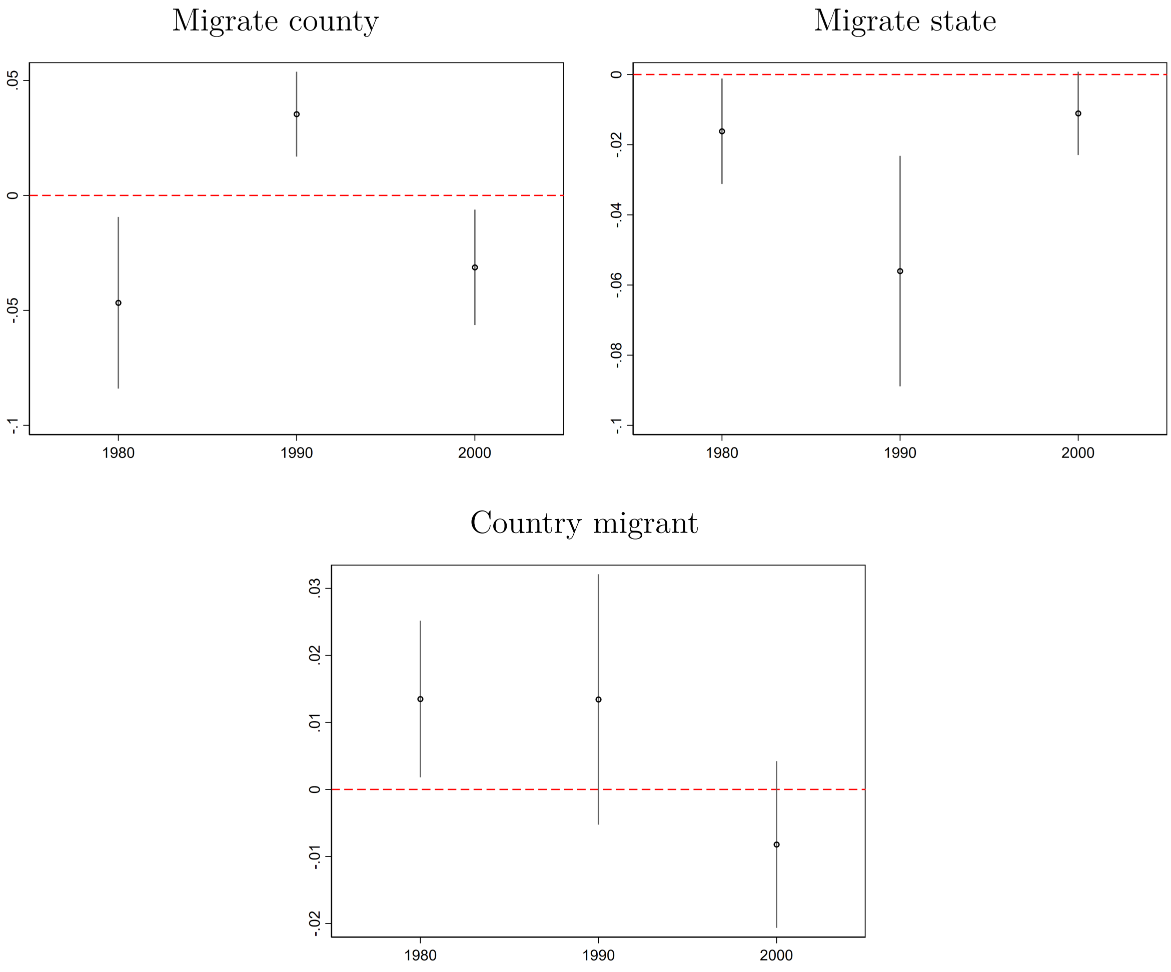


Figure 1.2: Estimated ATT of Rent Control on Immigration Outcomes by Year

Source: 1980 - 2000 decennial census data reported at the tract level by SocialExplorer.

Notes: Figures show the average treatment effect on the treated tracts of rent control on immigration outcomes. Each outcome is the percentage of tract inhabitants that have moved from a given location in the last 5 years. The error bars represent 95% confidence intervals from standard errors that are clustered at the city level.

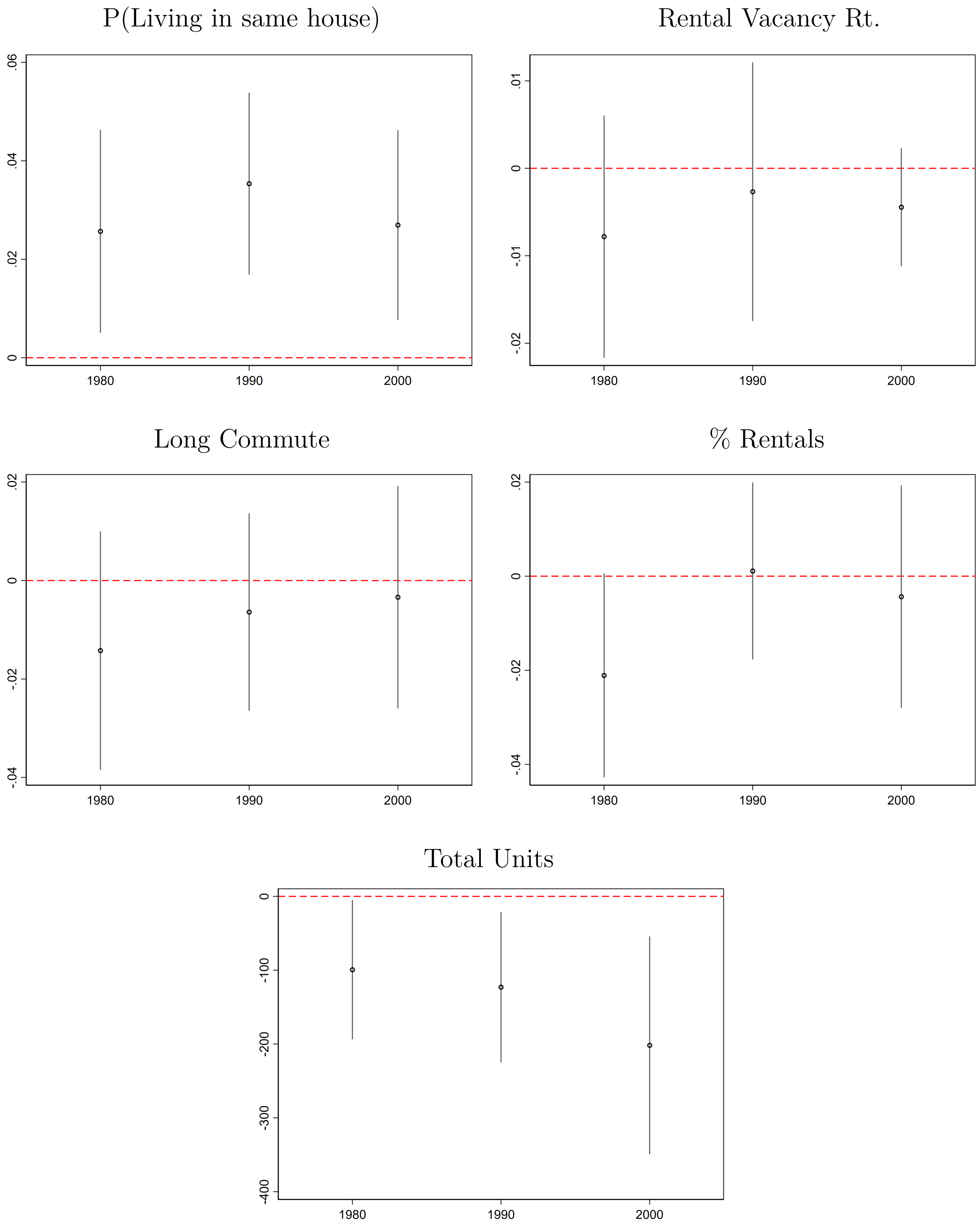


Figure 1.3: Estimated ATT of Rent Control on Housing Outcomes by Year

Source: 1980 - 2000 decennial census data reported at the tract level by SocialExplorer.

Notes: Figures show the average treatment effect on the treated tracts of rent control on housing outcomes. The average outcomes are generated using the baseline nearest neighbor match model. The error bars represent 95% confidence intervals from standard errors that are clustered at the city level.

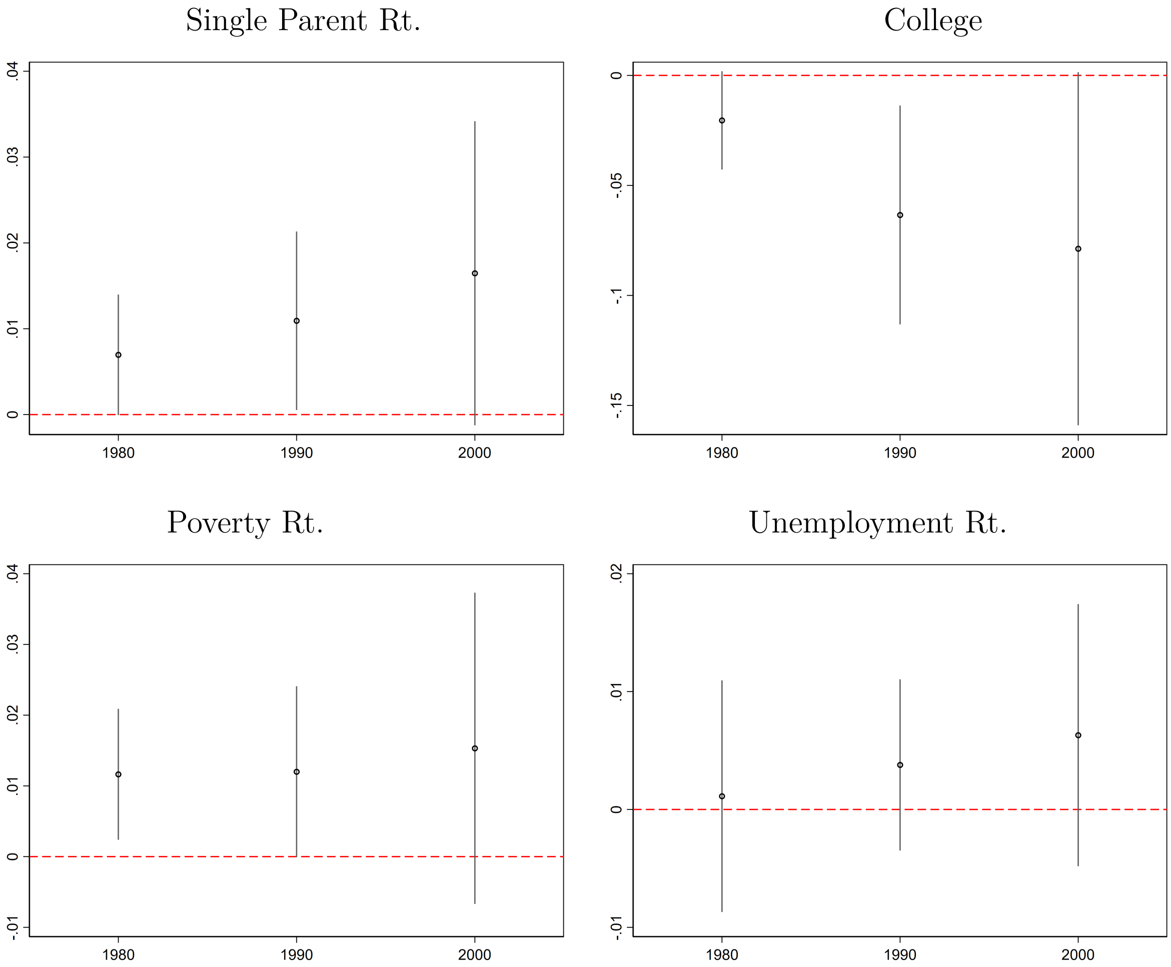


Figure 1.4: Estimated ATT of Rent Control on Demographic Outcomes

Source: 1980 - 2000 decennial census data reported at the tract level by SocialExplorer.

Notes: Figures show the average treatment effect on the treated tracts of rent control on employment and demograhic outcomes. The average outcomes are generated using the baseline nearest neighbor match model. The error bars represent 95% confidence intervals from standard errors that are clustered at the city level.

Table 1.1: T-Test of Means to Compare Characteristics of Treated and Controlled Census Tracts

	Average					
	Control	Treat	Difference	CA Diff.	MA Diff.	NJ Diff.
Population	3,065.197	3,422.419	357.222***	348.446***	-327.377**	243.072**
Male (%)	0.485	0.478	-0.007***	-0.010***	-0.011***	-0.002*
Pop./sq. mile	4,607.690	13,303.566	8,695.876***	6,557.932***	14,391.206***	11,001.167***
Age median	28.322	30.856	2.534***	2.323***	-0.748	1.762***
White (%)	0.898	0.818	-0.081***	-0.141***	-0.132***	-0.051***
Black (%)	0.092	0.144	0.052***	0.113***	0.121***	0.049***
Married (%)	0.634	0.561	-0.074***	-0.082***	-0.125***	-0.032***
Single parent fam. (%)	0.071	0.100	0.029***	0.026***	0.047***	0.016***
Educ. Less than HS (%)	0.437	0.434	-0.002	0.012**	-0.039***	0.073***
Educ. HS (%)	0.320	0.308	-0.012***	-0.025***	-0.012	-0.026***
LFP rate	0.598	0.603	0.005***	0.013***	-0.012**	0.013***
Unemployment rate	0.043	0.060	0.017***	0.007***	0.003*	0.004***
Avg. inc.	4,557.440	4,914.329	356.889***	272.167***	-9.041	-346.089***
Family poverty rt. (%)	0.089	0.093	0.003**	0.016***	0.041***	0.015***
Current addr. 5 years (%)	0.478	0.474	-0.004	0.015***	-0.096***	-0.026***
Housing units	1,026.231	1,237.823	211.592***	232.245***	17.667	154.410***
Rental (% of total units)	0.334	0.532	0.197***	0.142***	0.249***	0.233***
Rental vacancy rate (%)	0.052	0.035	-0.017***	-0.000	0.025***	0.000
Rent vacancy x Rental %	0.020	0.021	0.001*	0.006***	0.018***	0.004***
Avg. rent	127.908	143.727	15.819***	6.621***	18.171***	-3.224
Avg. home value	20,213.463	26,258.793	6,045.330***	3,713.865***	2,200.790**	-343.242
N	28,813	2,444				
N by State:						
California	4,052	1,563				
Massachusetts	538	184				
New Jersey	506	697				
Other States	23,717					

Source: Author calculations of 1970 decennial census data reported at the tract level by SocialExplorer.

Sample includes tracts in all states except Washington DC, Maryland and New York. These excluded states had cities with rent control and cannot be used as possible control tracts. * = p < 0.1, ** = p < 0.05, *** = p < 0.01.

Table 1.2: T-Test of Means to Compare Characteristics of Treated and Controlled Cities

	Average					
	Control	Treat	Difference	CA Diff.	MA Diff.	NJ Diff.
City rental (% of total units)	0.298	0.450	0.152***	0.130**	0.311***	0.166***
City rental vacancy rate (%)	0.066	0.023	-0.043***	0.003	-0.004	-0.001
City white (%)	0.914	0.904	-0.010	-0.085*	-0.087	-0.020
City black (%)	0.054	0.069	0.015	0.077*	0.063	0.020
City unemployment rate	0.041	0.038	-0.003	0.001	-0.003	0.002
City avg. rent	118.028	140.819	22.790***	23.756**	50.587	11.518***
County Dem. vote share 1968	41.408	45.876	4.468***	5.802**	3.040	3.296***
County Wallace vote share 1968	14.154	8.005	-6.149***	-0.927**	0.264	-0.766**
City population	33,718.552	88,277.263	54,558.711*	435,493.354	204,199.588	26,450.147***
City family poverty rt. (%)	0.076	0.054	-0.022***	-0.003	0.018	0.006
City revenue per capita	0.173	0.228	0.055***	0.189**	0.214*	0.031*
City tax revenue per capita	0.075	0.146	0.072***	0.086**	0.151**	0.029**
Property tax share of rev.	0.293	0.511	0.218***	0.033	0.022	0.029
Other gov. sources share of rev.	0.142	0.155	0.013	-0.069***	-0.080**	0.008
Educ. share of expenditure	0.032	0.133	0.101***	0.000	-0.115*	0.040
Police share of expenditure	0.155	0.178	0.023***	-0.045***	0.015	0.003
Welfare share of expenditure	0.002	0.010	0.008***	0.027	-0.002	0.003***
N	2,373	99				
N by State:						
California	274	10				
Massachusetts	34	3				
New Jersey	149	86				
Other States	1,916					

Source: Author calculations of 1972 Census of Governments data and publicly available county level vote shares.

Sample includes cities from all states except Washington DC, Maryland and New York. These excluded states had cities with rent control and cannot be used as possible control cities. * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$.

Table 1.3: T-Test of Means to Compare Outcomes of Treated and Controlled Census Tracts

	Average					
	Control	Treat	Difference	CA Diff.	MA Diff.	NJ Diff.
Fract. years in tract	0.594	0.597	0.004*	0.008***	-0.067***	-0.089***
Live with parents	0.164	0.248	0.083***	0.045***	0.029***	0.037***
Stay tract	0.189	0.235	0.045***	0.027***	-0.002	0.000
Stay comm. zone	0.698	0.741	0.043***	0.041***	-0.005	0.042***
Top 20% fam. inc.	0.194	0.224	0.030***	-0.004	-0.010	-0.064***
Top 20% ind. inc.	0.197	0.254	0.057***	0.011***	0.017**	-0.037***
Percentile fam. inc.	0.491	0.503	0.012***	-0.013***	-0.023***	-0.052***
Percentile ind. inc.	0.498	0.527	0.029***	0.005**	-0.003	-0.029***
Teen birth	0.210	0.164	-0.046***	-0.001	0.002	0.048***
Jail	0.018	0.012	-0.006***	0.000	0.004***	0.003***
Employed	0.766	0.754	-0.012***	-0.009***	-0.025***	-0.016***
Low pov. nbhd.	0.466	0.475	0.009**	-0.058***	-0.110***	-0.112***

Source: Author calculations of Opportunity Insights census tract level outcome data. All other outcome tables are generated from this source. Sample includes tracts in all states except Washington DC, Maryland and New York. These excluded states had cities with rent control and cannot be used as possible control tracts. * = p < 0.1, ** = p < 0.05, *** = p < 0.01.

Table 1.4: Comparing Means and Variances of the Raw and Weighted Samples Using the Two-Step Nearest Neighbor Matching Algorithm

	Std. Mean Diff.		Var. Ratio	
	Raw	Matched	Raw	Matched
Population	0.185	-0.203	0.655	1.080
Male (%)	-0.226	0.026	1.101	1.259
Pop./sq. mile	0.910	-0.046	3.533	0.928
Age median	0.357	-0.068	1.139	1.220
White (%)	-0.322	-0.080	1.616	0.951
Black (%)	0.211	0.041	1.548	0.931
Married (%)	-0.694	-0.053	1.210	1.433
Single parent fam. (%)	0.447	0.191	2.121	1.301
Educ. Less than HS (%)	-0.012	-0.214	0.962	1.075
Educ. HS (%)	-0.157	-0.009	0.695	1.096
LFP rate	0.076	-0.052	0.683	1.175
Unemployment rate	0.583	0.422	1.494	1.311
Avg. inc.	0.216	0.043	1.630	1.238
Family poverty rt. (%)	0.044	0.104	0.894	1.165
Current addr. 5 years (%)	-0.026	-0.362	0.938	1.175
Housing units	0.315	-0.164	0.739	1.107
Rental (% of total units)	0.843	0.116	1.616	1.183
Rental vacancy rate (%)	-0.258	0.066	0.470	1.407
Rent vacancy x Rental %	0.037	0.082	0.762	1.265
Avg. rent	0.367	0.153	1.334	1.124
Avg. home value	0.683	0.266	1.348	1.044
City rental (% of total units)	1.448	0.167	1.399	0.828
City rental vacancy rate (%)	-0.586	-0.167	0.240	0.739
City white (%)	-0.528	0.134	0.960	0.537
City black (%)	0.341	-0.254	0.822	0.416
City unemployment rate	0.912	0.638	0.976	1.659
City avg. rent	0.190	0.041	0.012	0.580
County Dem. vote share 1968	0.875	-0.010	0.536	0.521
County Wallace vote share 1968	-0.904	-0.120	0.021	0.870
City population	1.015	-0.021	4.548	0.681
City family poverty rt. (%)	-0.057	-0.017	0.395	0.880
City revenue per capita	0.882	0.257	2.282	1.454
City tax revenue per capita	0.935	0.048	2.677	1.052
Property tax share of rev.	0.394	-0.221	1.464	1.444
Other gov. sources share of rev.	0.135	-0.131	0.677	1.011
Educ. share of expenditure	0.375	-0.043	2.296	1.128
Police share of expenditure	-0.022	-0.382	0.419	0.687
Welfare share of expenditure	0.293	-0.058	5.377	0.879
N	31,443			
N Treat	1,174			
N unique control	549			

Source: Author calculations of 1970 census, 1972 Census of Governments, and 1968 county-level vote data.
Table shows the standardized differences in means and variances between the raw and weighted sample. The unit of observation is a census tract. The treated tracts are matched to a control tract using a nearest neighbor Mahalanobis distance matching procedure. The Mahalanobis distance metric includes linear terms for tract-level, city-level and county-level characteristics.

Table 1.5: Mahalanobis Distance Nearest Neighbor Match Estimates of Average Treatment Effect on the Treated of Rent Control on Location as Child and Adult

	Frac. years in tract	Stay home	Stay Tract	Stay Comm. Zone
ATT	0.021**	0.031***	0.016	0.004
	(0.010)	(0.011)	(0.010)	(0.025)
Baseline	0.576	0.216	0.219	0.738
%Δ	0.036	0.146	0.071	0.005
N Treat	1,156	1,155	1,156	1,156
N Control	541	539	540	540

The ATT row reports the average treatment effect on the treated of rent control on the average tract outcome. The ATT estimates are generated using a nearest neighbor Mahalanobis distance metric matching estimator. The baseline row represents the average outcome of the matched counterfactual tracts. * = p < 0.1, ** = p < 0.05, *** = p < 0.01.

Table 1.6: Mahalanobis Distance Nearest Neighbor Match Estimates of Average Treatment Effect on the Treated of Rent Control on Long-Term Outcomes

	Top 20% Fam.	Top 20% Ind.	Percentile Fam. Inc.	Percentile Ind. Inc.	Teen birth	Jail	Employed	Low Pov. nbhd.
ATT	-0.013	-0.009	-0.007	0.000	-0.000	-0.000	0.005	-0.026
	(0.020)	(0.018)	(0.012)	(0.009)	(0.013)	(0.002)	(0.004)	(0.022)
Baseline	0.237	0.263	0.510	0.526	0.164	0.012	0.749	0.501
%Δ	-0.054	-0.035	-0.013	0.001	-0.000	-0.001	0.007	-0.052
N Treat	1,172	1,172	1,172	1,172	1,171	1,172	1,172	1,172
N Control	549	549	549	549	549	549	549	549

The ATT row reports the average treatment effect on the treated of rent control on the average tract outcome. The ATT estimates are generated using a nearest neighbor Mahalanobis distance matching estimator that accounts for tract, city and county-level traits. The baseline row represents the average outcome of the matched counterfactual tracts. * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$.

Table 1.7: Mahalanobis Distance Nearest Neighbor Match Estimates of Average Treatment Effect on the Treated of Rent Control on Long-Term Outcomes, Weighted by Children Linked to a Tract

	Top 20% Fam.	Top 20% Ind.	Percentile Fam. Inc.	Percentile Ind. Inc.	Teen birth	Jail	Employed	Low Pov. nbhd.
ATT	-0.009	-0.007	-0.004	0.002	-0.003	-0.001	0.005	-0.012
	(0.013)	(0.013)	(0.009)	(0.007)	(0.016)	(0.002)	(0.005)	(0.020)
Baseline	0.219	0.248	0.499	0.518	0.183	0.014	0.752	0.477
%Δ	-0.042	-0.026	-0.008	0.005	-0.017	-0.091	0.006	-0.024
N Treat	1,172	1,172	1,172	1,172	1,171	1,172	1,172	1,172
N Control	549	549	549	549	549	549	549	549

The ATT row reports the average treatment effect on the treated of rent control on the average tract outcome, weighted by the number of children linked to each tract. The ATT estimates are generated using a nearest neighbor Mahalanobis distance matching estimator that accounts for tract, city and county-level traits. The baseline row represents the average outcome of the matched counterfactual tracts. * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$.

Table 1.8: Mahalanobis Distance Nearest Neighbor Match Estimates of Average Treatment Effect on the Treated of Rent Control on Location as Child and Adult: High Rental Tract Sample

	Frac. years in tract	Stay home	Stay Tract	Stay Comm. Zone
ATT	0.061**	0.047**	0.028*	0.019
	(0.027)	(0.020)	(0.017)	(0.021)
Baseline	0.507	0.204	0.207	0.737
%Δ	0.121	0.230	0.137	0.026
N Treat	827	826	826	826
N Control	372	373	372	372

The ATT row reports the average treatment effect on the treated of rent control on the average tract outcome. The unit of observation is a census tract with at least 30% rental share. The ATT estimates are generated using a nearest neighbor Mahalanobis distance metric matching estimator. The baseline row represents the average outcome of the matched counterfactual tracts. * = p < 0.1, ** = p < 0.05, *** = p < 0.01.

Table 1.9: Mahalanobis Distance Nearest Neighbor Match Estimates of Average Treatment Effect on the Treated of Rent Control on Long-Term Outcomes: High Rental Tract Sample

	Top 20% Fam.	Top 20% Ind.	Percentile Fam. Inc.	Percentile Ind. Inc.	Teen birth	Jail	Employed	Low Pov. nbhd.
ATT	0.011	0.009	0.012	0.016*	-0.001	0.002	0.020*	-0.018
	(0.012)	(0.012)	(0.010)	(0.008)	(0.015)	(0.002)	(0.010)	(0.021)
Baseline	0.181	0.218	0.466	0.492	0.188	0.012	0.726	0.440
%Δ	0.059	0.039	0.027	0.033	-0.005	0.168	0.027	-0.042
N Treat	840	840	840	840	839	840	840	840
N Control	373	373	373	373	373	373	373	373

The ATT row reports the average treatment effect on the treated of rent control on the average tract outcome. Only tracts with more than 30% rental share are included in the sample. The ATT estimates are generated using a nearest neighbor Mahalanobis distance matching estimator that accounts for tract, city and county-level traits. The baseline row represents the average outcome of the matched counterfactual tracts. * = p < 0.1, ** = p < 0.05, *** = p < 0.01.

Table 1.10: Mahalanobis Distance Nearest Neighbor Match Estimates of Average Treatment Effect on the Treated of Rent Control on Long-Term Predicted Outcomes

	Top 20%	Top 20%	Percentile	Percentile	Teen			Low Pov.
	Fam.	Ind.	Fam. Inc.	Ind. Inc.	birth	Jail	Employed	nbhd.
Panel A: ATT estimates for children at P25 of parent income distribution								
Estimate	-0.008	-0.006	-0.002	0.003	-0.009	0.001	0.013**	-0.015
	(0.018)	(0.017)	(0.010)	(0.008)	(0.017)	(0.002)	(0.007)	(0.019)
Baseline	0.172	0.206	0.452	0.480	0.210	0.014	0.714	0.459
%Δ	-0.049	-0.027	-0.003	0.007	-0.045	0.039	0.019	-0.032
Panel B: ATT estimates for children at P75 of parent income distribution								
Estimate	-0.002	0.000	-0.003	0.006	0.003	-0.001	0.006	-0.014
	(0.012)	(0.011)	(0.008)	(0.005)	(0.010)	(0.001)	(0.007)	(0.018)
Baseline	0.271	0.303	0.554	0.566	0.108	0.008	0.792	0.537
%Δ	-0.008	0.001	-0.005	0.010	0.024	-0.081	0.007	-0.026
N Treat	1,172	1,172	1,172	1,172	1,171	1,172	1,172	1,172
N Control	549	549	549	549	549	549	549	549

Estimates are reported for the predicted outcomes of children growing up in tracts at the 25th and 75th percentile of parent income. The predicted outcomes are generated by regressing individual outcomes on a transformation of parent income rank and recovering the fitted values at these two points of the parent income rank. The ATT is estimated using the baseline Mahalanobis distance matching estimator. * = p < 0.1, ** = p < 0.05, *** = p < 0.01.

Table 1.11: Mahalanobis Distance Nearest Neighbor Match Estimates of Average Treatment Effect on the Treated of Rent Control on Long-Term Predicted Outcomes: High Rental Tract Sample

	Top 20% Fam.	Top 20% Ind.	Percentile Fam. Inc.	Percentile Ind. Inc.	Teen birth	Jail	Employed	Low Pov. nbhd.
Panel A: ATT estimates for children at P25 of parent income distribution								
Estimate	0.003	0.004	0.004	0.007	-0.012	0.001	0.012	-0.014
	(0.011)	(0.011)	(0.009)	(0.007)	(0.016)	(0.003)	(0.010)	(0.022)
Baseline	0.140	0.180	0.432	0.467	0.231	0.014	0.715	0.408
%Δ	0.019	0.021	0.009	0.015	-0.050	0.094	0.016	-0.034
Panel B: ATT estimates for children at P75 of parent income distribution								
Estimate	0.029**	0.022**	0.022**	0.021**	0.016	0.000	0.016*	0.006
	(0.012)	(0.011)	(0.010)	(0.009)	(0.017)	(0.002)	(0.009)	(0.021)
Baseline	0.222	0.269	0.515	0.543	0.106	0.008	0.778	0.474
%Δ	0.132	0.082	0.043	0.039	0.155	0.004	0.020	0.012
N Treat	840	840	840	840	839	840	840	840
N Control	373	373	373	373	373	373	373	373

Estimates are reported for the predicted outcomes of children growing up in tracts at the 25th and 75th percentile of parent income. Only tracts with more than 30% rental share are included in the sample. The predicted outcomes are generated by regressing individual outcomes on a transformation of parent income rank and recovering the fitted values at these two points of the parent income rank. The ATT is estimated using the baseline Mahalanobis distance matching estimator. * = p < 0.1, ** = p < 0.05, *** = p < 0.01.

CHAPTER II

Modernizing Person-Level Entity Resolution with Biometrically Linked Records

2.1 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, 2018b,d), 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 2.1).[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 primary 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

[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.

the absence of accurate, unique identifiers, researchers must rely on similarity comparisons of plausibly identifying variables common to all data.[3] For person-level linkage, issues like data entry and optical character recognition errors, name confusion, nicknames and abbreviations, or naturally occurring name changes raise 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 2.1).[4]

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 potential 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

[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.

[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.

and women.[6] Allowing our machine learning algorithm to train on a 1 million observation sample strengthens performance on both *recall* (% of true matches correctly identified) and *precision* (% of statistical matches that are correct), 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 record linkage of registered Washington voters in 2008 and 2012 to assess performance in a more representative population, and (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.

Another useful finding from these exercises is the broad applicability of our approach to both record linkage (matching rows *across* datasets) and deduplication (matching rows *within* a dataset) problems. These two matching problems are closely related; for example, any deduplication exercise can be restated as a record linkage exercise.[9] But, specific applications generate important distinctions; a one-to-one record linkage problem does not allow for independent error terms, which is not necessarily the case for deduplication. The performance stability of the algorithm across these diverse applications broadens the relevance of our findings to a range of contexts.[10]

While details on matching often get shortchanged in academic publications,[11] the common

[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.

[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 B.3 for more details.

[9] In lieu of matching set A to set B, simply consider matching set A to set A (itself) excluding pairwise exact matches.

[10] In fact, the model developed in this paper simultaneously serves both record linkage and deduplication purposes in practice in constructing the Criminal Justice Administrative Record System's (CJARS) (see Finlay and Mueller-Smith (2021) data through identifying unique individuals both within and across criminal justice administrative datasets from jurisdictions around the United States.

[11] Recent literature has explored the concept of data matching strategies and its 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

matching performance metrics of recall and precision directly relate to concepts of internal and external validity in causal inference, which empirical researchers care about. 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 inference. 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 inflated estimates of the effect of interest. This last fact is directly 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.

This research has led to the creation of the *Biometrically Validated Entity Resolution System* ("B-VERS") tool. This trained algorithm captures much of the work described in this paper, allowing social science researchers to leverage the benefits of our unique, biometrically-linked data and improve their person-level record linkage.[12] It has been built to perform both deduplication and record linkage applications while providing options for a number of common data quality issues including missing demographic information, missing or abbreviated middle name, or omitted exact date of birth (i.e., year or age only).

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.

2.2 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 combining two or more datasets without a unique linking variable (i.e. record linkage), a range of statistical techniques and methodologies have been developed. Although the goals of deduplication and record linkage are different in practice, the underlying theory

to administrative data.

[12] B-VERS will be published publicly online through GitHub for general use in the coming months.

and methodology are closely related as either problem can be restated as the other with only modest restrictions imposed. For example, a historical linkage of two census waves can be thought of as a one-to-one match between two separate datasets; or alternatively, a deduplication using the two waves appended together and limiting to singular matches among observations from different sources. Both linkage strategies would generate the same matches in practice if using the same underlying prediction model.

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.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.[13] 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. 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.[14] Then, they attempt the link using methodologies previously

[13] We explore a simulation where this assumption does not hold in section 2.6.

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

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 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 Schnepel, 2021), 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.

Economics research utilizing linked records has become more prevalent given the increase in reliance on administrative data. Recent examples of the kind of data that can be linked include Federal Government data sources such as IRS tax records and U.S. Census Bureau Records (Chetty et al., 2016; Chetty and Hendren, 2018b,d), randomized control trial participation records, public school records and arrest records (Heller et al., 2016), health insurance records, hospital discharge records and credit bureau records (Finkelstein et al., 2012) and pollution records and grants for pollution abatement (Keiser and Shapiro, 2018). These data are generated by a mix of public and private sources and represent a small sample of the types of data that can be linked.

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 **A** and **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

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

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 ($\mathbf{U^S}$).

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

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

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

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 simple applications of deterministic algorithms, 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 $\mathbf{A}$ and $\mathbf{B}$, 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. Recent work on commonly used training data sets underscores the importance of training data accuracy by showing that errors in the training data are particularly costly when estimating non-linear or higher feature machine learning models (Northcutt et al., 2021). Elaborate models can overlearn from the improperly labeled training set, leading to situations where simpler models may actually outperform high feature models.

Usually training data is created by manually determining match status for a sample of paired observations through a process referred to as clerical review.[15] 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.[16] 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 applications 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).

[15] 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.

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

2.3 Data and Background

We utilize a novel method for identifying true matches and generating training data by leveraging finger-print based identifiers found in two criminal justice data sources. The training data is then used to estimate the model described in Section 2.4. The first training data source 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 [17] and remain relatively constant over time.[18] 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.[19]

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.[20]

While there are overlapping populations between the TDCJ and JIMS data systems and

[17] 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., 2017; Watson et al., 2014).

[18] 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.

[19] 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.

[20] 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.

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 2.2 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 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 2.5.3 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.

2.4 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.[21] The potential match space $\boldsymbol{\Delta}$ of all records $d_i \in \mathbf{D}$ contains $\frac{N\times(N-1)}{2}$ unique ordered pairings:

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

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

$$\boldsymbol{\Omega} = \Big\{(d_i, d_j); \Omega_{d_i} = \Omega_{d_j}, (d_i, d_j) \in \boldsymbol{\Delta}\Big\}$$

[21] There are many observations with perfectly matching PII in the raw data, reflecting the high recidivism rates in the criminal justice system. As is common in the matching literature, we eliminate duplicative observations with the exact same combination of PII. This ensures that measures of the algorithm's performance are not driven by observations with identical PII that are likely to be matched regardless of the matching strategy that is employed.

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.

In practice, we utilize the union of 10 blocks described in Table 2.3. 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 2 trillion observation pairs to just 17,577,515 observation pairs, with 95.2% of actual matches included in the blocked subset of paired observations.[22]

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

[22] 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.

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.[23] At this stage, we also account for the possibility that names can be flipped by calculating the distance between first-middle, first-last and middle-last names. When the flipped name comparison reveals a closer string edit distance than the original comparison, we update the string edit distance to reflect the flipped value. Lastly, we calculate a measure of uniqueness for each first, middle and last name in our data.[24] 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 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 B.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 2.5, ultimately settling on a *random forest* classifier as our preferred specification.

[23] These edit distances attempt to quantify the similarity of two text strings. The raw edit distance calculates the number of edits one would have to make to make string A equivalent to string B. The Jaro Winkler and Levenshtein Edit Distance are variations of the raw edit distance.

[24] The uniqueness variable is measured by calculating the reciprocal of the total number observations with the same name. Person A who has a first name shared by 500 other people has a first name uniqueness score of 1/500, while person B who has a unique first name has a uniqueness score of 1.

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.[25] 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.[26]

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 (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 2.2.

[25] The technique of utilizing multiple draws of a random sample is also known as bagging.

[26] See appendix B.2 for more details about the random forest classification methodology.

2.5 Evaluating Classification Performance

2.5.1 Baseline Results

Table 2.4 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.2.

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.[27] 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.

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.

[27] 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.

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. Alternatively, the random forest (Year of Birth) model is the same as the baseline random forest except we drop all comparators that are based on the day or month of birth. In many historical linkage contexts, matching is based on name and age, so we view this model as a rough comparison to the methods used in the historical linkage literature. Not surprisingly, the algorithm performs worse when date and month of birth are not included.

Figure 2.3 shows a variety of diagnostic graphs from the training data for the DE-RF model. Figure 2.3a 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 2.3b). 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 2.3c. 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-Score balances these tradeoffs through combining the concepts of recall and precision into a single measure that takes the harmonic mean of both components:

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

We utilize the F-Score for two purposes. First, in the training data, we establish our statistical match threshold using the predicted probability that maximizes the in-sample F-Score (0.42 in the case of our demographic-enhanced random forest model as seen in Figure 2.3d). Second, we identify the preferred algorithm in Table 2.4 based on the routine that delivers the best out-of-sample F-Score.

It happens that our preferred specification also performs relatively well on our second performance criteria: the duration of combined estimation and prediction routines.[28] 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 2.4 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

[28] All models are estimated on the Criminal Justice Administrative Record System's (CJARS) server which has 256 GB of RAM and 12 virtual processors.

the pair of variables, likely reflecting the fact that middles names are irregularly collected, often receiving less oversight than other components of PII.

2.5.2 Decomposing Model Performance and Assessing Demographic Heterogeneity

We compare the DE-RF model's performance to three alternatives in Table 2.5, 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 biometric identifier.[29] 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.[30]

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 biometric model deliver F-Scores 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 2.5 further investigates performance gains as training sample size is increased.

[29] See Appendix B.1 for details on our hand-coding procedure.

[30] 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 2.4 that date of birth information is more uniquely identifying that other components of PII.

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.[31] 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.

A natural follow-up question is whether out-of-sample performance can be improved by adjusting the composition rather than the size of the training sample. For example, including more marginal matches in the training data may strengthen the model's ability to correctly differentiate hard-to-classify, ambiguous matches as well as identify an appropriate match threshold. Figure 2.6 explores this concept in a bootstrapping exercise that varies the composition of the training records. First, we use OLS to predict the likelihood of being a true match using the full sample, and split the training data into three predicted match likelihood groups: low (1% TM), marginal (51% TM), and high (89% TM).[32] Then, we sample records according to various targeted sample compositions to generate 5,000 training pairs per iteration. This repeated 100 times per composition scenario, with the resulting out-of-sample precision and recall estimates plotted in panels a and b. Changing composition can substantially improve precision with only slight penalties to recall (see Panel a).[33] Taking this logic to an extreme, though, can degrade model performance: populating the training data exclusively with "marginal" pairs worsens recall and precision relative to our baseline scenario (see Panel b).

Returning to Table 2.5, 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

[31] 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.

[32] Note that this exercise is academic in nature as it presupposes the existence of true match variable to generate the "low", "marginal", and "high" predicted match likelihood variables.

[33] To help better quantify the gain from a tailored sample, average recall and precision rates in panel a which are built on 5,000 training observations approach the sample size results for 25,000 training observations see in Figure 2.5 although with greater variability.

theme in Section 2.6 where we discuss the implications of match quality for causal inference.

2.5.3 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 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[34] (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. A voter's PII may change if they move or change their name and must register again.[35] The goal of this application is to assess model performance in a non-criminal justice record linkage 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.[36] 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.

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

[35] The Washington Secretary of State elections webpage indicates that approximately 10-15% of the population moves each year while another 40,000 people change their name each year.

[36] See Appendix B.3 for a more detailed description of the corruption algorithm.

Table 2.6 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 in the blocked pair sample).[37] 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 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-Score 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-Score. 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.

2.6 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.[38] 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

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

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

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:

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

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 2.7 and 2.8 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,[39] and impair statistical precision

[39] 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.

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 2.8, 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 a covariate ($x_i \sim N(0, 1)$) into the model resulting in heterogeneous 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:

$$\tilde{y}_i = \begin{cases} missing & \text{if } \rho_i \geq \bar{\rho} \\ y_{\not{i}} & \text{if } \pi_i \geq \bar{\pi} \\ y_i & \text{otherwise} \end{cases}$$

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 $y_{\not{i}}$ 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 2.5.2, match quality does vary by key demographic traits that in many settings drive heterogeneous response to interventions, making this setup consistent with common applications.

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,[40] 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 2.9 and 2.10 report the average estimated $\hat{\Delta}$ and corresponding p-value testing the

[40] 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.

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 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 2.5), 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.

2.7 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, and to both record linkage and deduplication applications. 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.

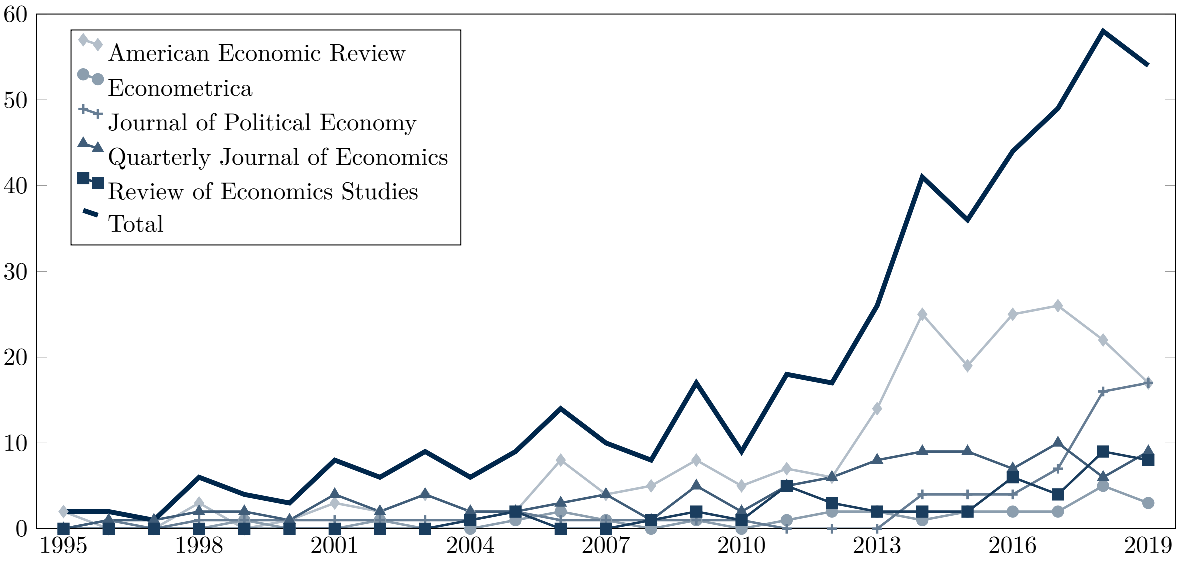


Figure 2.1: Total Publications Mentioning "administrative data" in the Top 5 Economic Journals, 1995-2019.

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.

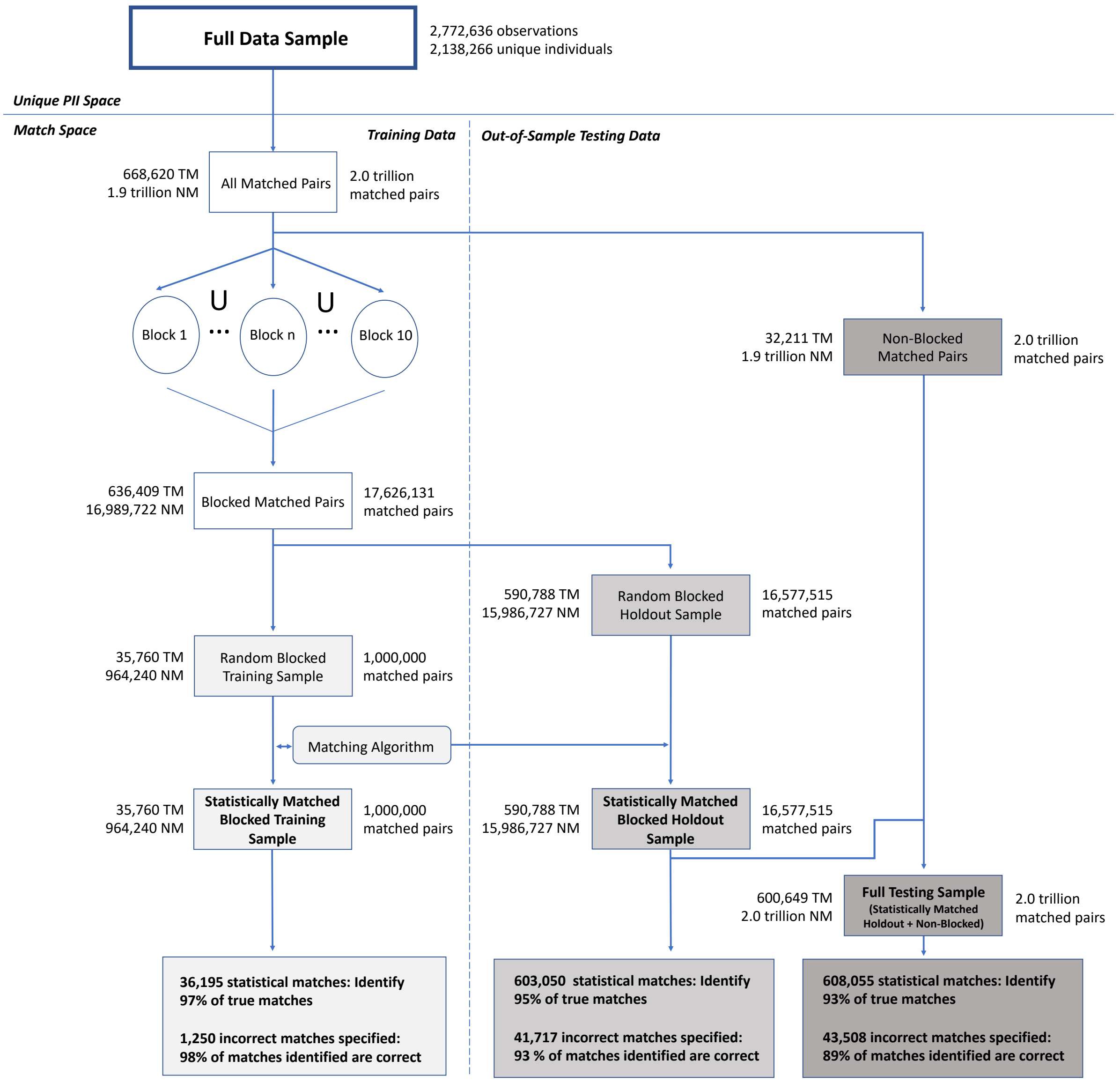
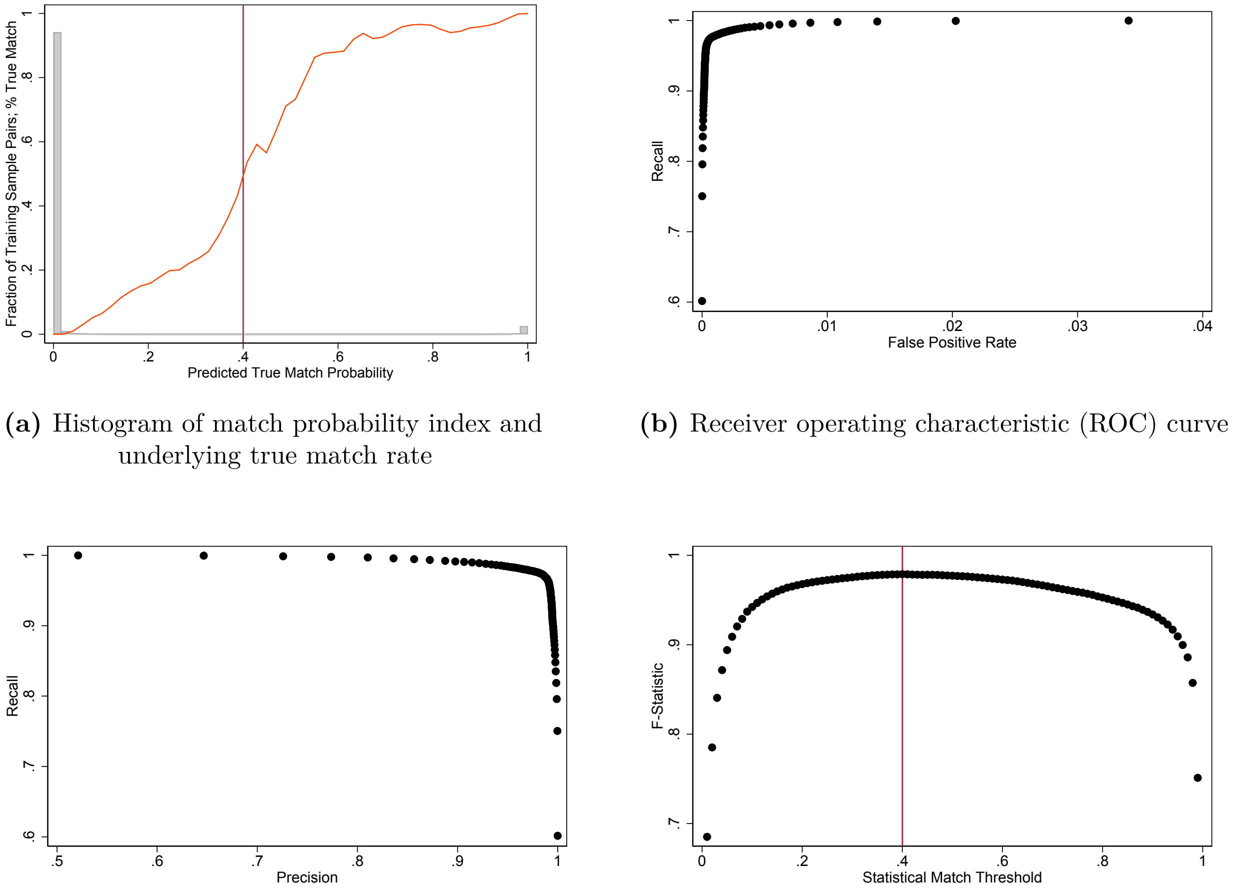


Figure 2.2: Model Training and Testing Overview

Notes: TM - True Match; NM - True Non-Match. Performance statistics based on the demographic enhanced random forest model described in Section 2.4. 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.



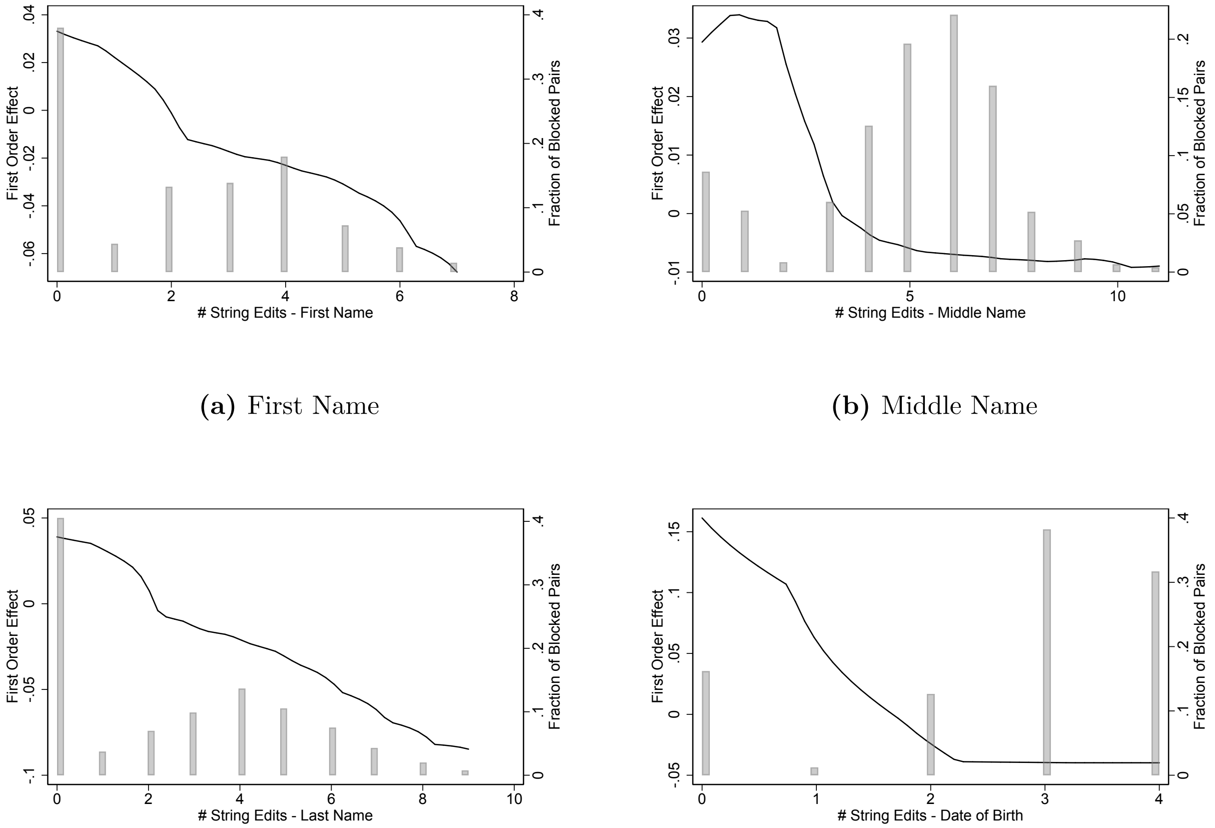
(a) Histogram of match probability index and underlying true match rate

(b) Receiver operating characteristic (ROC) curve

(c) Recall/precision tradeoff curve

(d) Implied F-Statistic at varying threshold values

Figure 2.3: 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

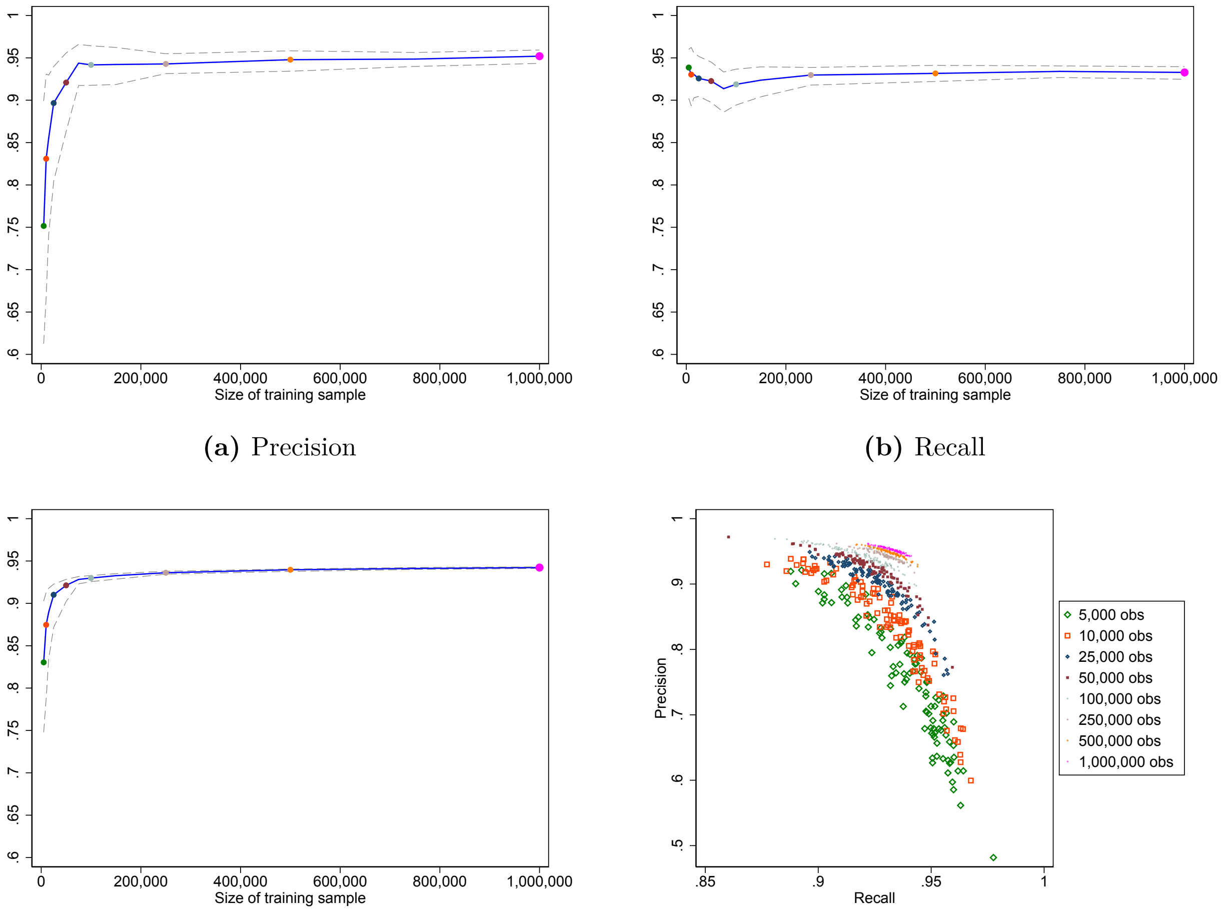
(b) Middle Name

(c) Last Name

(d) Date of Birth

Figure 2.4: 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

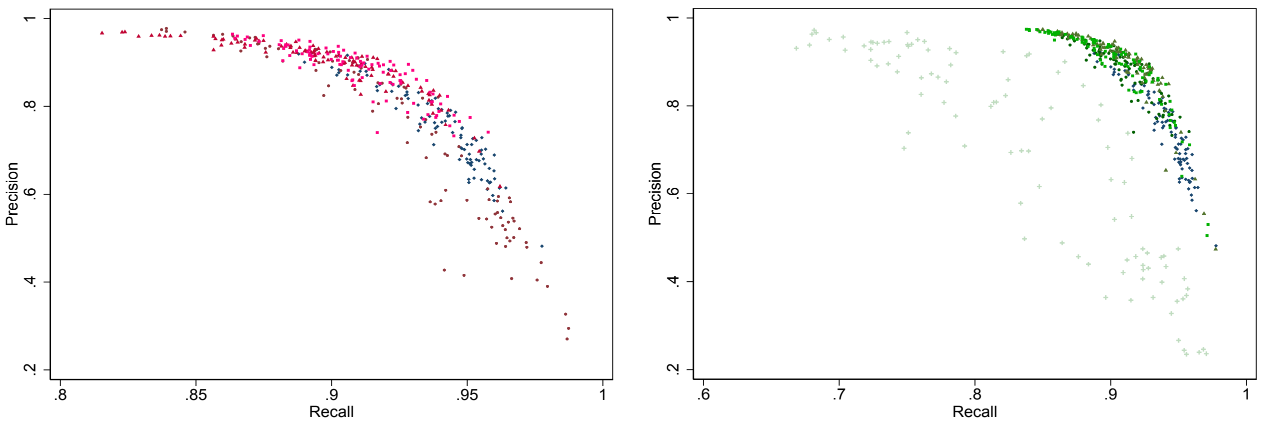
(b) Recall

(c) F-Statistic

(d) Performance Distribution

Figure 2.5: 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.

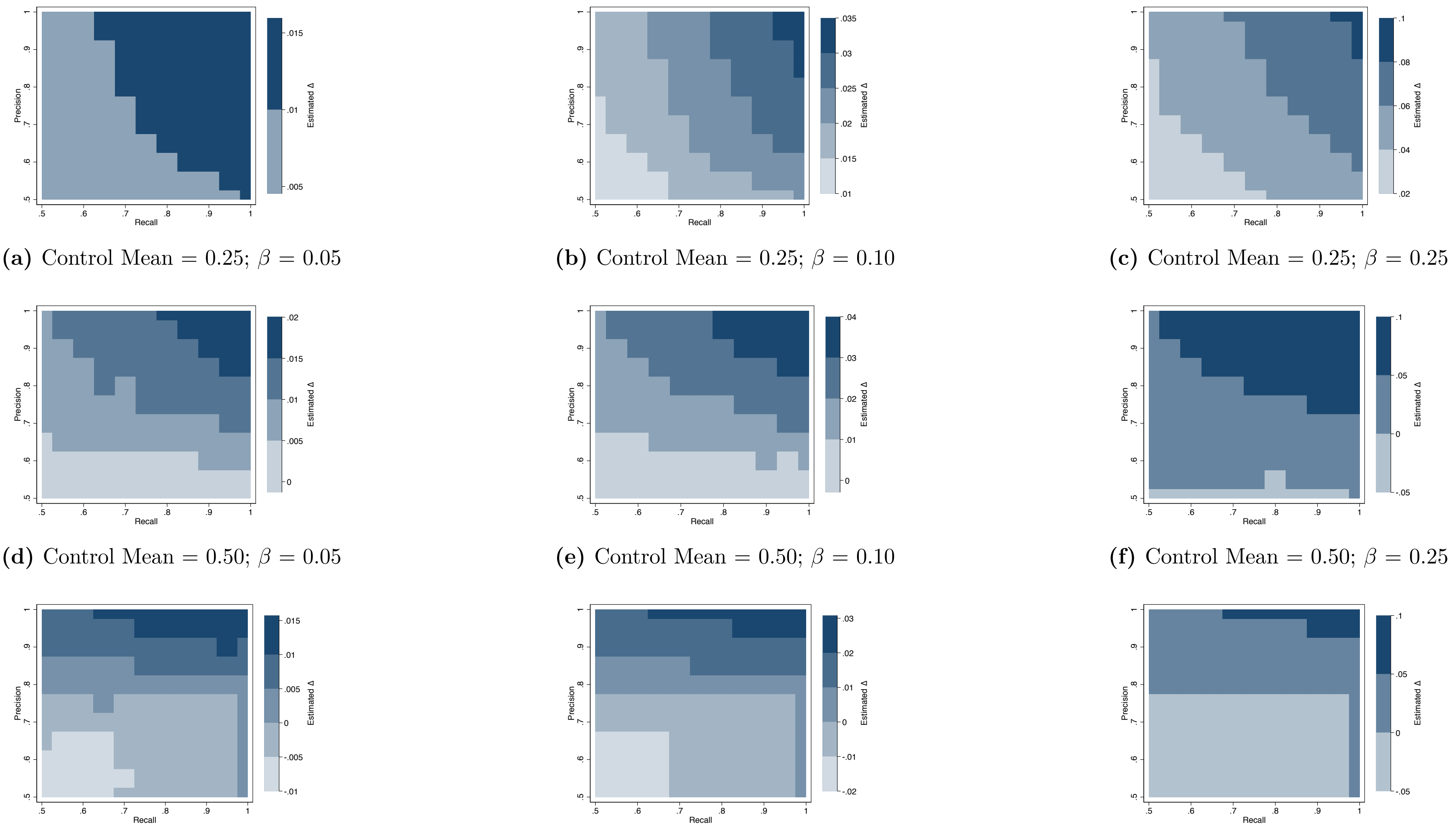


(a) Excluding low, high, and marginal matches in training sample

(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				

Figure 2.6: Model Performance Under Varying Training Sample Composition, Conditional on Training Sample Size
This figure shows the variation in out-of-sample model performance given varying composition of a fixed 5,000 observation training sample. 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. OLS was used to predict true match status using all available comparators in the data; blocked match pairs were designated then as either "low", "high", or "marginal" matches based on the predicted linear probabilities of being a true match. 100 independent models were estimated for each targeted composition, with training observations selected at random from the 16.6 million pool of eligible pairs.



(a) Control Mean = 0.25; β = 0.05

(b) Control Mean = 0.25; β = 0.10

(c) Control Mean = 0.25; β = 0.25

(d) Control Mean = 0.50; β = 0.05

(e) Control Mean = 0.50; β = 0.10

(f) Control Mean = 0.50; β = 0.25

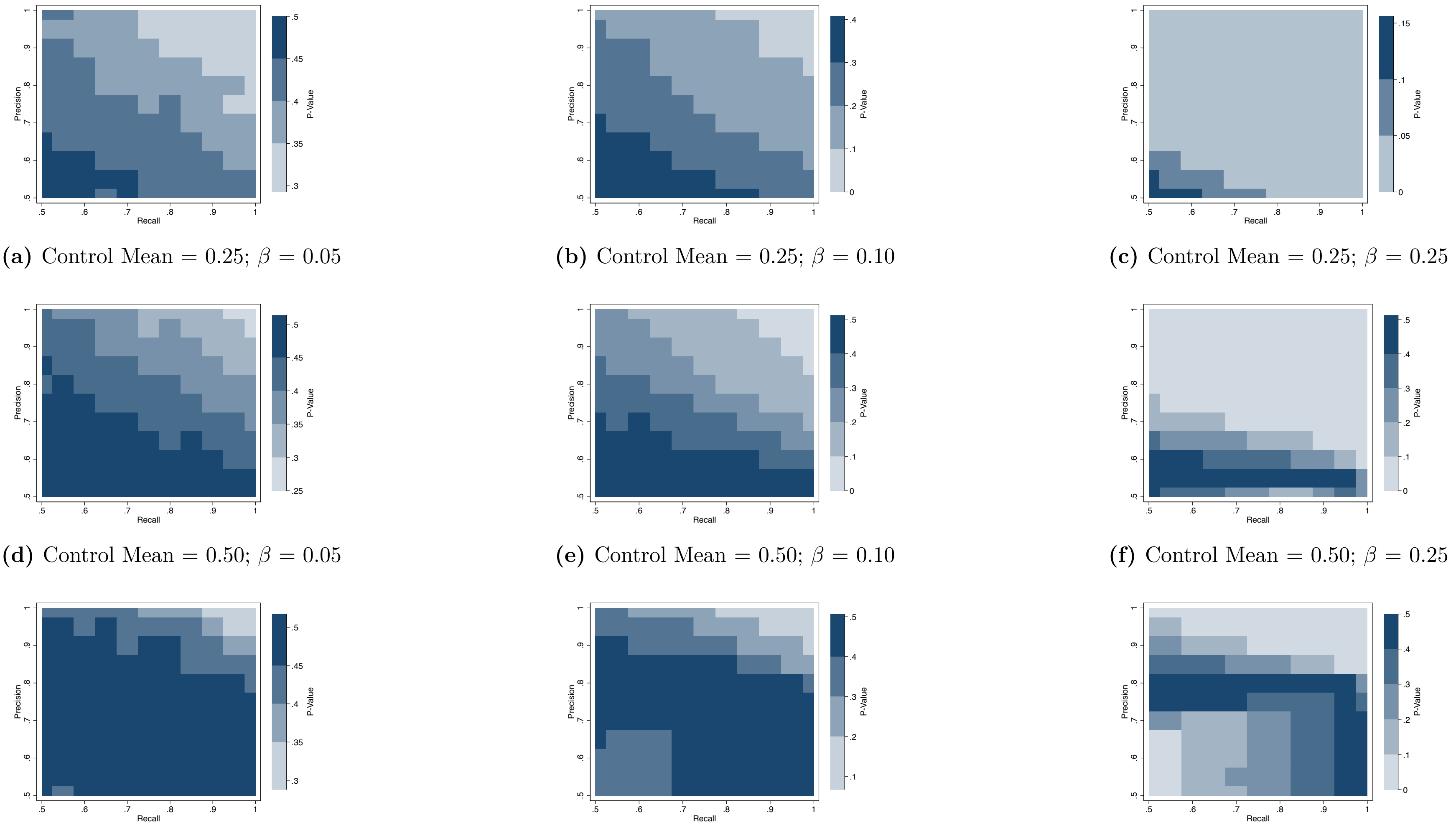
(g) Control Mean = 0.75; β = 0.05

(h) Control Mean = 0.75; β = 0.10

(i) Control Mean = 0.75; β = 0.25

Figure 2.7: 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 2.6. 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; $\beta = 0.25$

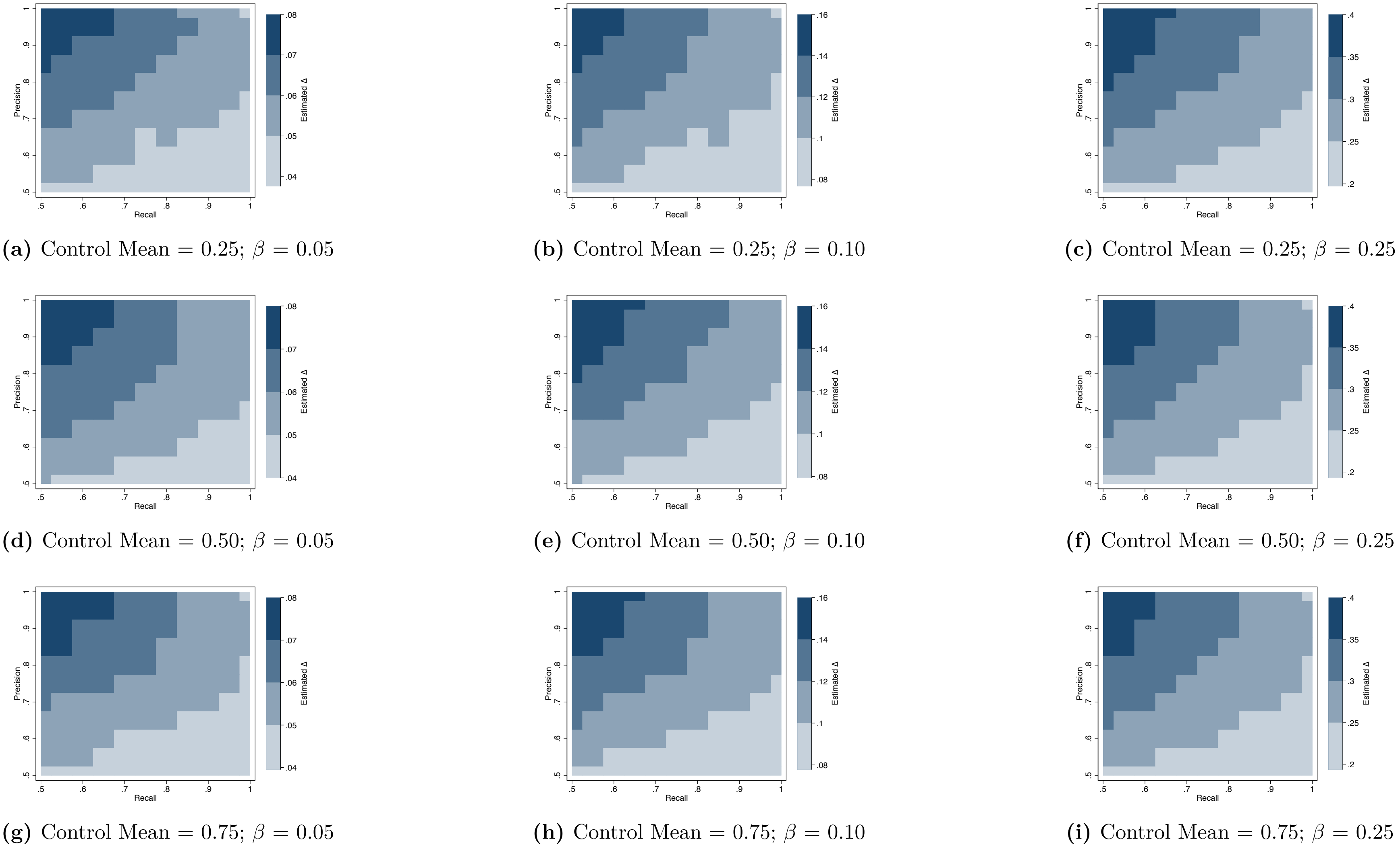
(g) Control Mean = 0.75; $\beta = 0.05$

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

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

Figure 2.8: 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 2.6. 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; β = 0.05

(b) Control Mean = 0.25; β = 0.10

(c) Control Mean = 0.25; β = 0.25

(d) Control Mean = 0.50; β = 0.05

(e) Control Mean = 0.50; β = 0.10

(f) Control Mean = 0.50; β = 0.25

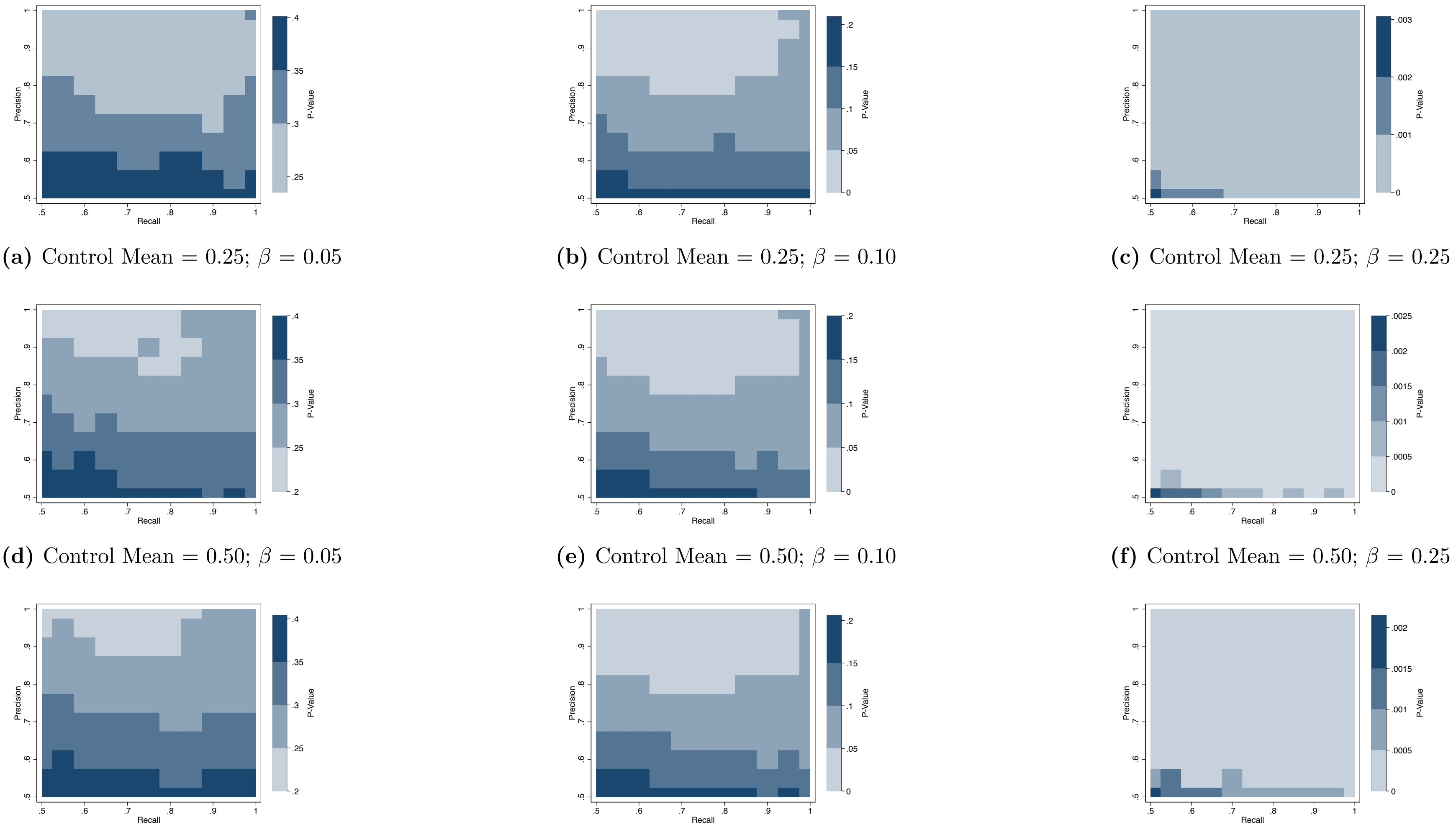
(g) Control Mean = 0.75; β = 0.05

(h) Control Mean = 0.75; β = 0.10

(i) Control Mean = 0.75; β = 0.25

Figure 2.9: 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 2.6 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; β = 0.05

(b) Control Mean = 0.25; β = 0.10

(c) Control Mean = 0.25; β = 0.25

(d) Control Mean = 0.50; β = 0.05

(e) Control Mean = 0.50; β = 0.10

(f) Control Mean = 0.50; β = 0.25

(g) Control Mean = 0.75; β = 0.05

(h) Control Mean = 0.75; β = 0.10

(i) Control Mean = 0.75; β = 0.25

Figure 2.10: 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 2.6 for the scenario with heterogeneous treatment effects. Worse precision and recall rates impact statistical precision though the effects are not systematic.

Table 2.1: 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 2.2: 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	0.880	0.800	0.500	0.930	0.480	0.490
Share White	0.360	0.270	0.720	0.440	0.820	0.640
Share Black	0.370	0.330	0.350	0.510	0.080	0.120
Share Hispanic	0.260	0.320	0.110	0.040	0.060	0.160
Average Age	42.500	40.900	37.800	44.200	78.300	37.200
Share in Texas	100.000	100.000	0	0	0.050	0.080
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 2.3: Description of Individual Blocks

Block	Fraction of True Matches	True Matches Not Included
Date of birth + last soundex	77.900	147,654
Date of birth + first soundex	81.500	123,808
Month of birth + first soundex + last soundex	72.700	182,324
Day of birth + First soundex + last soundex	72.100	186,694
Year of birth + first soundex + last soundex	72.100	186,798
Date of birth + last phonex	77.900	147,761
Date of birth + first phonex	82.100	119,861
Month of birth + first phonex + last phonex	73.200	179,241
Day of birth + First phonex + last phonex	72.500	183,624
Year of birth + first phonex + last phonex	72.500	183,720
Union of Blocks	95.200	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 2.4: Comparison of Out-of-Sample Model Performance

Model	Accuracy	Precision	Recall	F-Statistic	False Positive Rate	Estimation + Prediction Duration (Hours)
Deterministic	0.970	0.930	0.760	0.840	**0.006**	**0.00**
Naive Bayes Classifier (Discrete)	*0.97*	0.900	*0.72*	*0.80*	0.008	0.060
Naive Bayes Classifier (Kernel)	0.970	0.880	0.810	0.840	0.011	1.420
Support Vector Machine	0.980	**0.94**	0.830	0.880	0.006	*199.55*
Lasso Shrinkage Model	0.980	0.900	0.820	0.860	0.009	21.010
Random Forest	0.980	0.930	0.880	0.900	0.006	0.260
Random Forest (Demog. Enhanced)	**0.98**	0.930	**0.89**	**0.91**	0.007	0.280
Random Forest (Year of Birth)	0.970	*0.82*	0.820	0.820	*0.018*	0.180
Neural Net Perceptron	0.980	0.930	0.850	0.890	0.006	2.110
Neural Net	0.980	0.920	0.880	0.900	0.008	10.130

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 2.5: Demographic-Specific Performance Statistics

		Demographic Enhanced Random Forest		
		5,000 Hand-Coded	5,000 Biometric	1,000,000 Biometric
	Deterministic	Training Obs.	Training Obs.	Training Obs.
Panel A: Precision Rates				
Overall	0.930	**0.95**	*0.91*	0.930
Race/Ethnicity				
White	0.960	**0.97**	*0.94*	0.940
Black	0.970	**0.97**	0.960	*0.95*
Hispanic	*0.88*	**0.98**	0.950	0.930
Sex				
Male	*0.92*	**0.96**	0.950	0.940
Female	**0.97**	0.950	*0.83*	0.900
Decade of Birth				
1960s	0.930	**0.94**	*0.89*	0.920
1970s	0.940	**0.96**	*0.93*	0.930
1980s	0.970	**0.97**	*0.93*	0.940
1990s	**0.98**	0.980	*0.95*	0.960
Panel B: Recall Rates				
Overall	*0.76*	0.770	0.840	**0.89**
Race/Ethnicity				
White	0.810	*0.81*	0.890	**0.93**
Black	*0.80*	0.810	0.860	**0.91**
Hispanic	*0.73*	0.740	0.880	**0.93**
Sex				
Male	0.790	*0.77*	0.830	**0.90**
Female	*0.68*	0.760	0.860	**0.88**
Decade of Birth				
1960s	*0.72*	0.720	0.800	**0.86**
1970s	*0.80*	0.820	0.880	**0.92**
1980s	*0.86*	0.880	0.930	**0.96**
1990s	*0.88*	0.900	0.950	**0.98**
Panel C: F-Statistics				
Overall	*0.84*	0.850	0.870	**0.91**
Race/Ethnicity				
White	0.880	*0.88*	0.910	**0.94**
Black	*0.88*	0.890	0.900	**0.93**
Hispanic	*0.84*	0.840	0.920	**0.93**
Sex				
Male	*0.85*	0.870	0.890	**0.92**
Female	*0.80*	0.840	0.840	**0.89**
Decade of Birth				
1960s	*0.81*	0.810	0.840	**0.89**
1970s	*0.86*	0.880	0.900	**0.93**
1980s	*0.91*	0.920	0.930	**0.95**
1990s	*0.93*	0.940	0.950	**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 2.6: Testing Model Performance in External Applications

					False Pos.
Application	Accuracy	Precision	Recall	F-Stat.	Rate
Multi-State Inmate Snapshot (July 1, 2017)	1.000	–	–	–	0.000
Washington State Voter Records (2008 & 2012)	0.980	0.920	0.880	0.900	0.008
Corrupted Death Master File (2000-2009)	0.980	0.970	0.930	0.950	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.

CHAPTER III

Effect of Financial Sanctions: Evidence From Michigan's Driver Responsibility Fees

3.1 Introduction

Understanding the impacts of legal financial obligations (LFO) on criminal justice involved individuals has become more urgent over the past twenty years as state and municipal courts have steadily increased the number and magnitude of fines and fees owed by defendants (Bannon et al., 2010; Harris et al., 2010). Many state and local governments rely on the revenue generated from these fines and fees to fund courts and other government services. According to the Survey of Inmates in State and Federal Correctional Facilities, the percentage of inmates that had LFOs imposed by courts has increased from 25% in 1991 to 66% in 2007 Harris et al. (2010).

Descriptive research has found strong correlational evidence linking fines and fees with financial instability, criminal recidivism, and poor labor market outcomes (Harris et al., 2010; Pleggenkuhle, 2018).[1] Given the high incidence of criminal convictions in the United States, such evidence would suggest that these fees may have wide-ranging impacts on not only the most disadvantaged criminal defendants but also on the economy at large. Causal evidence remains quite limited given the lack of data availability and exogenous variation. One exception is Mello (2021) who finds that driver fees in Florida are associated with short-term financial distress, using difference-in-differences and event-study research designs applied to traffic stop data linked with high-frequency credit report data.

In this paper, we exploit a policy change in the state of Michigan as a source of exogenous variation on the magnitude of financial sanctions faced by defendants to get a more complete understanding of the ways that criminal financial sanctions impact long-term recidivism and

[1] See Martin et al. (2018) or Fernandes et al. (2019) for recent reviews of the literature on financial sanctions.

labor market outcomes. In 2003, Michigan passed Public Act 165, or the driver responsibility fee (DRF), which mandated new fines to criminal defendants who were convicted of certain driving crimes. The goal of the act was to raise revenue for the government while improving driving safety.[2] The amount of the fines varied based on the severity of the offense, ranging from $300 to $2,000. Failure to pay these fines would lead to driver's license suspension, which could lead to even more fines since driving with a suspended license was also a DRF-qualifying offense. In the first two years after the law was enacted, the state levied over $250 million in driver responsibility fines (Wild, 2008).

To estimate the long-term causal effect of the financial sanctions, we exploit the fact that the DRF program in Michigan applied only to individuals convicted of a DRF-eligible offense on or after October 1, 2003, a context well-suited for regression discontinuity analysis. Two key mechanisms potentially link DRF sanctions to behavioral responses: (1) an income effect generated from the fine itself, and (2) a license revocation in the event of non-payment. A challenge in studying this program is that a subset of the caseload commits DRF-related offenses at a regular interval, meaning that those to the left of the cutoff who originally avoided the DRF sanction pick up DRF sanctions when they reoffend in a year or two. Ignoring this issue might underestimate the true impact of the policy since the license revocation channel is partially neutralized through such behavior.

To address this challenge, we develop a prediction model based on observable demographic information, criminal history, and longitudinal earnings profiles to distinguish between those likely and unlikely to commit a new DRF-related offense in the short-run.[3] Among those with with low likelihood of DRF recidivism, defined as having a predicted likelihood of DRF recidivism below the median, which we refer to as the low contamination sample, the integrity of both mechanisms is maintained over roughly the entire 10-year follow-up period. For the high contamination sample, the extensive margin of the first stage (i.e., getting one or more DRF sanctions) declines by over 50 percent within five years, severely curtailing our ability to capture the long-term impact of the license revocation mechanism for this subpopulation.

Our results show modest behavioral responses to DRF sanctions. For the *low contamination sample*, we observe no economically meaningful or statistically significant long-term labor market responses overall. Subgroup analysis, however, suggests improved earnings outcomes at the median of the predicted income distribution. We find short-term negative impacts on earnings that are likely mitigated through the channel of incapacitation following driver's license revocation from non-payment; these effects attenuate in the long run. This subsample

[2] This program is not unique to Michigan as New Jersey, New York, Texas, and Virginia have all had similar programs.

[3] There is no imbalance across the discontinuity in being categorized as either being part of the high or low risk samples.

does not show a crime response either in the short or long-run. For the *high contamination sample*, we observe a short-run increase in earnings from the income effect that is not sustained in the long-run. We also find evidence of small long-run deterrent effects. Finally, we find spillover labor market effects onto the romantic partners of the high contamination group in the long-run. Specifically, we find a 9% increase in the cumulative earnings of partners from 2005 to 2015 relative to a mean of $202,900, suggesting other household members may be stepping in to cover the cost of the LFO.

Overall, we find DRFs to be a regressive form of funding for the government with limited benefits in terms of labor market outcomes or criminal behavior. We observe no change in the rate of DRF-related offending in the general population, suggesting no evidence of a general deterrence response, and no fall in recidivism, suggesting no evidence of a specific deterrence response in the study sample. Given the low average income of individuals in our sample, the impacts of the policy were concentrated on those less likely to pay the fees, placing them at higher risk for driver's license suspension. While unmeasured, it is possible that consumption declined in response to the fines without a concurrent change in income to compensate for the negative financial shocks. Even more concerning, the spillover impacts onto partners' earnings indicate that some people who did not commit the DRF offense bore the monetary burden of the fines.

The findings of our paper are surprising given the small but growing economics literature on the impacts of financial sanctions, which find negative effects on labor outcomes. In the paper that is most similar to ours, Mello (2021) finds that speeding ticket citations lead to financial instability and decreased labor force participation in the two years following the initial fine using data from credit reporting agencies.[4] The result is particularly strong for lower income individuals, suggesting that even small negative income shocks in the form of unexpected fines can lead to negative labor market outcomes. In contrast, our paper relies on administrative tax records, covering a wider range of individuals.

Previous empirical work on financial sanctions has mostly focused on the direct impacts on the defendants but has ignored the potential effects on other household members; however, qualitative and anecdotal work on this topic suggests that spillover effects are important to quantify, especially in the context of the criminal justice system (Kearney et al., 2014; Shapiro, 2014; Mathews and Curiel, 2019).[5] Our paper is the first to attempt to document

[4] In general, research on driving infractions finds that higher cost citations and stronger enforcement deter people from future offenses and lead them to drive more safely (Makowsky and Stratmann, 2009; Mello, 2021; Hansen, 2015; Luca, 2015). Compared to past work, our paper examines costs from administrative fees, rather than citations and also significantly increases the timeline in which we study the effect of these financial sanctions.

[5] A range of literatures in economics provide evidence on the household spillovers of negative income shocks (Page et al., 2009; Mörk et al., 2014; Liu and Zhao, 2014; Coile, 2004; Bloemen and Stancanelli, 2008;

and measure these spillover effects on romantic partners.

This paper makes several important contributions to the literature. First, we provide robust, causal estimates on the effects of financial sanctions on labor market and recidivism outcomes. Second, our unique data allows us to estimate these effects on a longer time frame and test how they might generate spillovers to other household members, which helps better characterize the full impact of the policy. Third, we develop a methodological strategy to address high DRF recidivism rates, which would otherwise undermine the use of the policy changes over time in regression discontinuity research designs. Our findings are much more modest compared with the previous literature. Future research is needed to expand the evidence base on financial sanctions and increase our understanding on the potential mechanisms that might contribute to behavioral changes.

The remainder of the paper is as follows: Section 3.2 describes the policy change and judicial system of the state of Michigan; Section 3.3 describes the data used in this analysis; Section 3.4 describes the empirical methodology and provides evidence to support the identification strategy; Section 3.5 presents the results; lastly, we conclude in Section 3.7.

3.2 Michigan's Driver Responsibility Fees Law

In an effort to promote safer driving and increase state revenue, the governor of Michigan signed Public Act 165 into law on August 11, 2003. The legislation, which became effective on October 1, 2003, mandated new fines to defendants who were convicted of certain driving crimes.[6] The DRF would be enforced by the Michigan State Treasurer as its revenue would be directed toward the state's General Fund. As a result, the DRFs were classified as administrative fines, rather than criminal penalties (Wild, 2008).

The act created two categories of fees. Category 1 was for drivers who accrued seven or more driving or traffic violations in two years. Category 2, which is the focus of this research, fined drivers for specific violations ranging from driving without a driver's license to driving under the influence. This fee was determined using three distinct tiers of driving violations, where the lowest level defendants were forced to pay a \$150 or \$200 dollar fee for two consecutive years, the middle level defendants were forced to pay a \$500 dollar fee for two consecutive years, and the highest level defendants were forced to pay a \$1000 fee for two consecutive years (Wild, 2008). Table 3.1 shows a detailed list of the Category 2 type of offenses associated with each fee level.

One particular criticism of the DRF policy relates to its onerous impact on poor defendants

Hankins and Hoekstra, 2011).

[6] The law was modeled on similar legislation in New Jersey. Since the passage of Michigan's law, Texas, New York, and Virginia have each instituted their own version of DRFs.

(Hausman, 2013). Failure to pay the fees after 60 days led to the suspension of one's driver's license and driving with a suspended license was itself a DRF-qualifying offense, so that lower income recipients may have been more at risk of receiving multiple DRFs. By the time that the law was repealed in 2018, an estimated 317,000 drivers had had their driver's licenses suspended for failure to pay DRFs (Carrasco, 2018). In order to reinstate a license, one was required to pay all outstanding DRFs along with an additional $125 fee, otherwise driving without a license put individuals at risk for new criminal charges and additional fines and fees. In fact, from 2005 to 2007, the number of citations for driving with a suspended license increased by 44% (Wild, 2008).

The fees failed to generate planned revenue or improve driver safety. The initial collection rate, from 2003 to 2008, of 52% was lower than the state's initial projected collection rate of 60%. Alcohol-related driving crimes increased by 21% after the bill went into effect, which suggested that the deterrent aims of the policy failed to materialize (Wild, 2008).

In 2018, the state of Michigan repealed the driver responsibility fee legislation and canceled all remaining debt owed under the law.[7][8] At the time of nullification, the state forgave approximately $630 million in outstanding driver responsibility payments (Carrasco, 2018). From 2004 to 2008, Michigan assessed approximately $780 million in fees but collected only 49% of them (Wild, 2008). This collection rate is lower than the one reported by the state of New Jersey for their DRF program (Wild, 2008).

Driver responsibility fees were assessed upon conviction for a qualifying offense and were administered by the Michigan Secretary of State. It is worth noting that these fees were distinct from fines, restitution payments, and other fees that could be imposed by the courts upon conviction for any crime. As such, driver responsibility fees in this paper represent a lower bound on the total financial burden imposed after a conviction for a driving offense.

3.3 Data

This project leverages several sources of rich population-level data, including criminal records from the Criminal Justice Administrative Record System (CJARS), longitudinal earnings data from IRS W-2 information returns, and romantic partner linkages compiled from a combination of survey and administrative data. All of these data were analyzed within the Census Bureau's Data Linkage Infrastructure, where data can be linked at the person level using the anonymous Protected Identification Key (PIK).

[7] Since Michigan's repeal, Texas and New Jersey have also repealed their own versions of the DRF law. Virginia repealed its law in 2009.

[8] This repeal only covered the Category 2 fees, which is the focus of this study. Category 1 fees were repealed in 2011.

Michigan DRF-eligible offenses that form the basis of our sample are identified from the adjudication records in CJARS. We use the charge offense to identify all charges that would trigger a DRF as defined in Michigan Public Act 165. Our sample consists of an individual's first conviction for a DRF-eligible offense in Michigan from April 1, 2001 to March 31, 2006, which covers the 2.5 years before and after the DRF effective date of October 1, 2003.[9]

To measure criminal outcomes, we identify all misdemeanor or felony convictions in Michigan for the set of individual in our analytic sample who received a DRF-eligible offense during the 2.5 years surrounding the effective date of the DRF law, starting in April 1, 2001 and ending in March 31, 2006.

To measure employment and earnings, we use IRS W-2 information returns from 2005, the first year we have W-2 information, to 2015. We define earnings as the sum of inflation-adjusted wages across all W-2 filings in a given period.[10] One major benefit of using W-2s is that they cover all formal employment regardless of the duration of the employment; they are not affected by the selective tax filing behavior inherent in IRS 1040 individual tax returns.[11] Furthermore, if an individual works for multiple employers in one year, each of the employers must issue a W-2 tax return. We can use the number of W-2 returns filed in a year on behalf of an individual as a measure of the number of jobs that individual worked.

In order to document spillover effects within romantic partnerships, we link individuals to their partner or spouse using a wide array of government data including the 2000 and 2010 Decennial Censuses, IRS 1040 individual tax returns, housing assistance data from the Department of Housing and Urban Development, American Community Survey responses, and other survey and administrative records that identify romantic partnerships between individuals over time.[12] Romantic relationships of interest in our sample are spousal, romantic non-cohabiting, or romantic cohabiting. Once we have linked an individual with a DRF-qualifying offense to a romantic partner whose relationship inception predates the DRF-eligible offense, we are able to draw on the same IRS and CJARS data to identify the labor market outcomes and criminal behavior of the partner. This enables us to test how pre-existing relationships and partners' outcomes are affected by the fees.[13]

Finally, we leverage Census Bureau survey and administrative records to identify demographic characteristics so that we do not have to rely on possibly mismeasured analogues in court

[9] We focus on conviction date because the DRF law affected only cases disposed after October 1, 2003.

[10] All earnings are inflated to 2017 dollars using the Consumer Price Index for All Urban Consumers (CPI-All Urban). Fines and fees generated from DRF-eligible offenses are not adjusted.

[11] Employers are required to file W-2 returns if an employee earns at least $600 in a tax year.

[12] See Finlay et al. (2021) for more details on how these links were identified.

[13] A limitation of this approach is that we are less likely to observe informal relationships, such as unmarried romantic relationships that do not involve cohabitation, since they are unlikely to jointly file taxes, co-reside, or respond to household surveys together.

records.[14] We use date of birth and gender records from the 2020 Census Bureau Numident file, which is based on the Social Security Administrations Numident register. For race and ethnicity information, we use the Census Bureau 2016 Title 13 race and ethnicity file, which combines self-reported and administrative records of an individual's race and ethnicity from various sources, such as the Census Numident and the 2000 and 2010 Decennial Censuses.

3.4 Research Design and Methodology

To determine the causal effects of DRF sanctions, we exploit the discontinuous implementation of the policy on October 1, 2003. Specifically, the statute only applied to individuals convicted of a DRF eligible offense on or after October 1, 2003. Therefore, individuals convicted of the same offense prior to October 1, 2003 would not be subject to the additional fine. Given the policy design, we utilize a sharp regression discontinuity designed to compare outcomes for individuals convicted of the same crimes right before and after the policy implementation. Under standard assumptions, the difference in outcomes can be attributed to the policy change at the discontinuity. In order to have a causal interpretation, the change in policy must be the only variable correlated with the outcomes to shift. In other words, the convictions around the discontinuity must be randomly sorted, so there should be no difference in the people charged with DRF crimes in the neighborhood of the policy change.

For the formal regression discontinuity estimates, we follow the method proposed by Calonico et al. (2014) and implemented in the Stata command *rdrobust* (Calonico et al., 2017). The point estimates $\hat{\tau}$ are estimated using the following framework:

$$\tau = \mu_+ - \mu_-,$$

where μ is the estimating equation of the outcome variable and

$$\mu_+ = \lim_{x \to d^+} \mu(x), \;\; \mu_- = \lim_{x \to d^-} \mu(x), \;\; \mu(x) = E[Y_i | X_i = x].$$

In this model, the average outcome function, $E[Y_i]$ is estimated on either side of the threshold where the DRF conviction date (X_i) is equal to the date of the policy change, $X = d$. The causal effect of the driver responsibility fees, τ, is thought of as the jump in the estimating equation moving from the left (d_-) to the right side (d_+) of the conviction date threshold. Instead of taking a simple average of the outcome variable, Calonico et al. (2014) propose a parameterization of the estimating equation using first-order local polynomials on

[14] Hispanic ethnicity is especially underreported in criminal justice administrative records (Eppler-Epstein et al., 2016; Ford, 2015).

each side of the discontinuity.[15]

Throughout the analysis, we use the sharp RD design defined above with initial DRF conviction date as the running variable. We also include sample averages of the outcomes to contextualize estimate effect sizes.

The identifying assumption of this research design is that justice-involved individuals whose cases were disposed right before October 1, 2003 are observationally equivalent to individuals whose cases were disposed right after October 1, 2003. Behavioral responses to the policy, such as deterrence or delayed sentencing, would violate this assumption. This is of particular concern since Public Act 165 was signed into law six months before the driver responsibility fee policy went into effect. Thus, both government agents and drivers could change their behavior in anticipation of the adoption of the new fees. We test both of these threats to identification by providing graphical and regression evidence that the caseload and demographic composition of DRF defendants is smooth around the policy threshold.

Table 3.2 reports regression discontinuity estimates for select demographic and pre-offense income variables. Nearly all balance test estimates are statistically insignificant and close to zero supporting the causal interpretation of the proposed research design. We estimate a significant, but small increase in the likelihood of being black and small decreases in the likelihood of being white or convicted of a DRF level 1 offense, but these estimates are relatively small when compared to the overall sample averages. All other estimates are statistically indistinguishable from zero. We also show that the estimated probability of linking to a romantic partner in the year of DRF conviction is unchanged at the discontinuity. This last fact is used to justify our empirical analysis of the effects of the fees on partner labor supply in Section 3.5.5.

One of the benefits of the regression discontinuity methodology is that results are easily visualized. By plotting local polynomial estimates for outcome variables on either side of the threshold, we can visually represent the estimated jump or discontinuity.[16] Panel C of Figure 3.1 shows the smoothed case counts of all DRF convictions around the discontinuity. The difference in case load count is smooth across the threshold suggesting that there was no change in enforcement or charging behavior after the implementation of the new fees. For this local polynomial figure as well as throughout the remainder of the paper, we use the Epanechnikov kernel and a 120-day bandwidth to estimate the outcome function and display

[15] We use a uniform kernel which equally weights all observations within the bandwidth and a data-driven bandwidth selector that chooses two mean squared error optimal bandwidths, one for each side of the discontinuity. This corresponds to the *msetwo* bandwidth selector within the *rdrobust* command. We estimate first-order linear polynomials with second-order bias correction and implement heteroskedasticity-robust plug-in residuals variance estimators with HC_2 weights.

[16] For our local polynomial graphs, we residualize the variables using the same set of controls with the mean added back to aid in interpretability.

95% confidence intervals.[17] The scatter plots are binned at the monthly level with marker size scaled according to the number of convictions in each bin.

For further evidence of the robustness of the identification strategy, we calculate predicted income using cumulative income reported on annual IRS W-2 information returns from 2005 to 2015 and predicted recidivism using total future convictions 10 years after DRF conviction shown in Table 3.3. Since W-2 information is only available starting in 2005, we use different measures of time across the labor and recidivism measures. To generate both predicted variables, we use a fully interacted regression model with the following variables: age at conviction for DRF offense, gender, race/ethnicity, average annual 1040 income, average 1040 form filing rates 1–3 years prior to conviction, fixed effects for the number of previous convictions, and the full interaction of fixed effects for the DRF offense level with fixed effects for the county of adjudication. Since we do not use the cutoff in constructing these predictions, the smoothness in the predicted variables across the cutoff would imply that the identifying assumptions of the sharp RD design are met. This is indeed what we show in the bottom two rows of Table 3.3 where coefficients on the predicted outcomes are insignificant.[18]

A visualization of the balance tests of the predicted income and predicted recidivism variables are shown in Panels A and B of Figure 3.1. Both figures are smooth around the discontinuity, further underscoring the overall balance of covariates.

Given the balance in demographic characteristics and predicted variables, we quantify the first-stage relationship in Figure 3.2, which shows the changing likelihood of being subject to a driver responsibility fee conditional on being convicted of a DRF-qualifying offense for individuals in our sample over time. Unsurprisingly, we find a sharp jump of approximately 100% in the likelihood of being charged with a driver responsibility fee after the law goes into effect.

Throughout the analysis below of labor market and recidivism outcomes, all outcomes are residualized by including controls for the driver's demographic characteristics (age, gender, race/ethnicity, total convictions 1–3 years before DRF conviction, income 1–3 years before DRF conviction as reported on 1040 tax filings, and the likelihood of filing a 1040 tax return 1–3 years before DRF conviction) and fixed effects for the DRF level of offense.

[17] Census Bureau disclosure rules prevented release of the binned quantities from the uniform kernel model using for tabular output. The kernel differences result in slightly varied regression discontinuity point estimates across tables and figures.

[18] Due to limits on the number of results that can be disclosed by the Census Bureau, we do not include smoothed local polynomial figures for all covariates. As a compromise, we report the caseload density, predicted W-2, and predicted recidivism (which aggregate the individual covariates into two indices).

3.4.1 Traffic Offense Recidivism and Integrity of the Experimental Variation

One possible threat to identification of the impact of DRFs is that individuals convicted of DRF-eligible offenses before the DRF goes into effect will eventually be convicted of DRF-eligible offenses after the effective date, thereby exposing the group to DRFs and contaminating the RDD. This type of DRF-eligible recidivism is particularly likely given the role of license revocation and the fees associated with driving without a license discussed earlier.

To circumvent this issue, we identify drivers at risk of reoffending two years after their first conviction within the sample time period based on a prediction model using the full interaction of age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication and case characteristics (fixed effects for the DRF level of offense, county of adjudication). Using the median of the predicted risk of DRF recidivism, we separate our sample into "high contamination" and "low contamination" groups where the high contamination group has greater than median predicted risk of DRF recidivism and is therefore more likely to contaminate our treated group. For the rest of the paper, we refer to the high (low) contamination group as the group with more (less) contamination.

To illustrate this contamination over time and to demonstrate the stability of the low contamination group, the top panels of Figure 3.3 show the year-by-year evolution of the cumulative likelihood of being convicted of a DRF offense after the effective date separately for individuals in the high contamination group compared to individuals in the low contamination group. The DRF conviction estimates for Year 0 are equivalent to the first-stage estimates in Figure 3.2, with each additional year increasing the follow-up period by a year.

For the high contamination group, we find that the validity of the experiment erodes significantly on the extensive margin and that the first-stage estimate falls from nearly 100 percentage point to almost 60 percentage points two years after the first DRF conviction. This implies that a significant portion of observations in the untreated sample were convicted of an additional DRF-qualifying offense after the effective date, thus contaminating the sample. In contrast, the low contamination group is significantly more stable. In fact, the low contamination group is relatively stable, with the first-stage estimate falling from nearly 100 percentage points to 95 percentage points four years after the first conviction.

Since some individuals with initial DRF-eligible convictions before the law's effective date will go on to lose their licenses (i.e., the high contamination group), we cannot as easily measure the impact of DRF conviction on labor market and recidivism outcomes that may result from license suspension. We can, however, still measure the impacts of the accumulation of fines from DRF conviction that operate through the income effect.

Since the analysis will be stratified by contamination group, we should consider how these groups differ along other dimensions. Figure 3.4 plots the average demographic characteristics across 5 percentage point intervals of the distribution of the predicted DRF recidivism. We include individuals within 15 percentiles on either side of the central point. For example, at the 20th percentile, we include individuals from the 5th to 35th percentile. Statistics near the ends of the distribution will have fewer observations as they are bounded by 0 and 100. We also mark where the sample splits between the low contamination group (0–50) and the high contamination group (50–100).

Taken together, these figures show that the low contamination group is older and is composed of more females and fewer Blacks relative to the high contamination group. We also find that predicted income is lower in the high contamination group. We also find a sharp contrast in criminal histories between the low and high contamination groups. Specifically, individuals with any prior criminal convictions one to three years prior to DRF conviction are more likely to have a predicted likelihood of DRF recidivism above the median and thereby be in the high contamination group.

To ensure that splitting the sample by predicted risk of DRF recidivism does not violate the identifying assumptions for a sharp RD, we graph the local polynomial estimates on either side of the discontinuity in Figure 3.5 of the likelihood of being in the high contamination group. We also show the regression discontinuity estimate in Table 3.3. Both the figure and the estimates show smoothness across the discontinuity.

3.5 Results

We split our discussion of the effects of financial sanctions into two categories: direct effects on individuals charged with a DRF-qualifying offense in the sample period and partner spillovers. For our direct effects analysis, we also include heterogeneity analysis across predicted income, recognizing that personal financial stability may influence the impact of the DRFs. We also separate our results across the different DRF levels to assess how the monetary size of the fees affects criminal justice and labor market outcomes. As mentioned in the prior section, we show results separately for the predicted high and low contamination groups.

3.5.1 Direct Impacts on Labor Market Outcomes and Recidivism

Our main analysis examines the direct effect of the fees on the charged individual's cumulative employment outcomes from 2005 to 2007, from 2005 to 2015, and on recidivism outcomes two years and ten years after initial DRF conviction. We use different periods of measurement for recidivism and labor market outcomes because the IRS W-2 return data are only available beginning in 2005.

Table 3.4 shows the regression discontinuity estimates and standard errors as well as the underlying mean of the outcome from the sample of defendants with initial DRF conviction dates within 2.5 years of the policy threshold. Panel A reports the estimates for labor market outcomes and Panel B reports the estimates for criminal activity. All estimates in the table are estimated using a sharp RD design and should be interpreted as causal.We split our results by contamination group to separate the two potential mechanisms: the income effect from the additional fines and the incapacitation effect from license suspension due to non-payment of DRFs. For comparison purposes, we also include long-run results for the full sample.

Using W-2 tax return information from 2005–2007, we find differing effects of DRF fees on earnings in the short run across the high and low contamination groups. For the high contamination group, DRF fees are associated with increased earnings of $2,185 from 2005 to 2007. Since individuals in the high contamination group are at a higher risk of re-committing a DRF-qualifying offense, increased earnings are consistent with a labor supply income effect caused by the DRF conviction. We observe no change in the average number of W-2 returns filed during the same period, which indicates that individuals increased labor supply within existing jobs rather than taking on more jobs to pay off the fees.

For the low contamination group, DRF fees are associated in the short run with a small decrease in the average number of W-2 returns received per year (0.03 percentage point decrease from a mean of 1.18 annual returns). Since there is no statistically significant decrease in the likelihood of any W-2 return being filed or on cumulative earnings within the same period, the reduction of jobs was focused on secondary employment. Given that the income effect would generate the opposite labor supply response, the reduction in the number of jobs could be caused by driver's license revocation from non-payment of DRF fees.

For both low and high contamination groups, effects of DRF fees attenuate when we increase the period of study to 2005–2015. We find no effects of DRF conviction on labor outcomes on the long-run intensive or extensive margin for either contamination group. Specifically, our estimates show small and mostly insignificant effects on the likelihood of receiving a W-2 tax return, the average number of W-2 returns received, and on cumulative earnings measured using total earnings reported on W-2 tax return.

These results taken together indicate that the imposition of DRFs did not create barriers to employment for individuals in either group as we do not see a corresponding fall in the likelihood of receiving a W-2 return from 2005 to 2007. As mentioned previously, this was a major concern given that the punishment for DRF non-payment led to driver's license suspension, potentially eliminating a means of transportation to work. Furthermore, all significant impacts generated by DRF conviction in 2005–2007 are no longer significant from 2005–2015.

In Panel B of Table 3.4, we shift our focus to recidivism outcomes. The first row shows the impacts of the fees on the likelihood of receiving any conviction (felony or misdemeanor, all offense types), measured by average recidivism. We find insignificant effects of DRFs on recidivism two and ten years after DRF conviction for the high contamination group. The estimates and standard errors are small relative to the mean, which indicated precise null effects. On the intensive margin (total convictions), estimates show similar null effects of DRFs on recidivism for each contamination group.

For the low contamination group, we find a statistically significant but small decrease in the likelihood of felony conviction two years after the initial DRF conviction. This decreased recidivism is not sustained when measured ten years after DRF conviction. We also find null effects when we measure effects by type of crime. Overall, DRF conviction had no meaningful impact on criminal behavior.

Our results using the full sample show null results for nearly all outcomes except for increased cumulative earnings measured using income reported on 1040 tax filings and decreased total convictions. Without correcting for the contamination in our sample, we would incorrectly conclude that the policy improved market labor outcomes and increased deterrence. We observe none of these effects in either the high or low contamination results. The contrast in results for long-run outcomes between the full sample and the results by contamination group only highlights the need for careful consideration of the interaction of the research design and behavioral responses to the policy.

Figure 3.6 shows graphical evidence of selected results from Table 3.4 by reporting non-parametric smoothed estimates of long-term cumulative W-2 earnings and total convictions before and after effective date of the DRF law in October 2003. These figures confirm the findings in Table 3.4.

3.5.2 Evolution of the Effects on Labor Market Outcomes and Recidivism

Next we shift our focus to investigating the evolution of the effects over time. Figure 3.7 shows the year-by-year estimates of the discontinuity and 95% confidence intervals for the main income and recidivism outcomes and for the low and high contamination groups separately. To help quantify the estimated effect size, we also include the mean of the entire sample in the figures. The first estimate represents the effect of a DRF offense charge on cumulative income (cumulative recidivism) one year after DRF conviction (from 2005–2006) with each additional estimate increasing the follow-up an additional year up to ten years.

For the low contamination group, we observe no statistically significant effects on cumulative income or cumulative recidivism over time. Notably, the estimates for cumulative earnings are close to zero until the 2005–2011 measurement period. Similarly, estimates for cumulative recidivism are close to zero until seven years after conviction. Given the size of the sample, we are confident that these are precisely estimated null effects for both outcomes.

For the high contamination group, we observe a different pattern on the estimated impacts on cumulative earnings. From 2005–2007 and 2005–2011, we find statistically significant increases in cumulative earnings reported on W-2 information returns. These estimated impacts begin decreasing and become statistically insignificant from zero from the 2005–2012 period onward. One possible reason for the effect fade out could be the loss of labor market opportunities during Great Recession, but further research is needed to confirm this.

For total convictions, we observe a different pattern over time with no statistically significant impacts on criminal behavior across all years in the follow-up period after the first DRF conviction.

3.5.3 Heterogeneous Effects by Ability to Pay

Our findings so far show null effects of DRFs on criminal recidivism and labor market outcomes over the long run. Despite this, the fees may have real but countervailing effects in different subsets of the population. In particular, we are interested in determining whether an individual's initial ability to pay off driver responsibility fees may play a role in driving the headline null effects.

We proxy for the ability to pay driver responsibility fees using predicted W-2 income. Because we do not have access to IRS W-2 returns before 2005, we cannot condition analysis on pre-conviction income. We do have access to IRS 1040 individual tax returns before the DRF policy goes into effect, but, as discussed earlier, the 1040 tax filing rate is only around 60%, so its use would greatly reduce the power of the heterogeneity analysis. Instead we rank individuals based on their predicted cumulative 2005–2015 W-2 income. The predicted income model is estimated using the same covariates used in predicting DRF recidivism. These covariates are: age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication. Therefore, this measure of predicted W-2 income does not include any information that is affected by DRF-eligible charges. As a result, we use it as a way to rank people in the sample by their ability to pay the fees. People with higher predicted income will have greater capacity to earn money to pay for the fines. We use the predicted W-2 income as a proxy for "affordability" and estimate how the effects of the fees differ by predicted income.

After generating predicted income, we estimate regression discontinuity models at different points of the predicted income distribution—every 5 percentiles between the 5th and 95th. To reduce noise, we include individuals within 15 percentiles of the central point in either direction. As an example, our estimate at the 20th percentile will include everyone with predicted income between the 5th and 35th percentile. Note that the estimates for the 5th and 10th (90th and 95th) percentiles include fewer observations as they are bounded below (above) by 0 (100).

Panels A–D of Figure 3.8 show the effects of DRFs on earnings and recidivism across the distribution of predicted income and separately for individuals with predicted low and high risk of DRF recidivism (with 95% confidence intervals for each point estimate).

For the heterogeneous effects of the driver responsibility fees on cumulative income, we find statistically significant increases of about $20,000 in income for drivers in the 60th to 65th percentile of the income distribution.

For the low contamination group, Panel B shows that there is little variation in the effect of DRF fees on ten-year average recidivism rates. Most of the beneficial effects are concentrated on individuals in the 65th to 70th percentiles of the income distribution with a statistically significant reduction in total recidivism of around 0.1 crimes ten years after conviction. The estimated decreases in recidivism overlap with the effects on income at the 65th percentile of predicted income.

Only the income effect from a DRF conviction could generate these results. Driver's license revocation from non-payment would place the individual at higher risk for recidivism if caught driving with a suspended license, thereby increasing recidivism. Furthermore, the revocation could potentially reduce an individual's earnings through their ability to work by removing their means of transportation and thereby reducing employment or earnings.

We observe a different pattern across predicted income for individuals in the high contamination sample. Notably, there are no effects on total convictions throughout the entire distribution except for a statistically significant decrease of approximately 0.2 crimes at the 45th and 50th percentile of the distribution. We do not observe a corresponding effect on total earnings. For individuals in the 20th to 40th percentile, we observe similarly sized increases in earnings of nearly $10,000, but the estimates are not quite statistically significant.

3.5.4 Heterogeneous Effects Across DRF Fee Levels

The previous section analyzes how the effects of the fees vary by the individual's ability to pay. However, the size of the fees relative to income can also be an important determinant for the DRF impacts. Furthermore, heterogeneity in the characteristics of the subpopulations of individuals across the fee levels could influence measurement of the impact of DRFs.

Given that being charged with a DRF-eligible offense may have monetary consequences and incapacitation effects from driver's license revocation, analyzing the outcomes across the fee levels may provide insight on the main mechanisms driving the outcomes.

To test this hypothesis, we stratify the analysis of the impacts on cumulative earnings and total convictions by the three DRF levels and the two predicted risk groups. Using the same sharp regression discontinuity design, Table 3.5 reports the effect of being convicted with a DRF by the fee amount for both the high contamination and low contamination group on short-run and long-run cumulative earnings and total convictions.

In the high contamination sample, we observe effects similar to what we observe in the analysis of heterogeneity in ability to pay DRFs. Specifically, for cumulative earnings from 2005–2015, we find no significant effects for any of the fee amounts; for total recidivism ten years after conviction, we estimate a decrease of 0.3 convictions for individuals assigned the \$300 fee, which is similar to the estimate from the heterogeneity analysis at the 50th percentile. The estimated decrease is small relative to the average of 4.3 total convictions. Interestingly, for the estimated impacts on the short run, the positive increase in earnings is not concentrated at a particular fine level.

For individuals in the low contamination sample, we estimate heterogeneous impacts of DRFs on cumulative earnings from 2005–2007 across the different fees. For individuals assigned the \$300 fee, we find an increase of earnings of \$5,608, but for individuals assigned the \$2,000 fee, we estimate a decrease of \$7,830. One possible explanation for the short-term negative impacts for individuals assigned the \$2,000 is license suspension from fee non-payment. Given that average annual income is approximately \$28,000 (calculated from average total income from 2005–2007 of \$83,930 in Table 3.5), a \$2,000 fine paid over the course of two years represents a significant portion of income, which places these individuals at higher risk of driver's license suspension from fee non-payment.

These effects on earnings are sustained for those assigned the \$300 fee. We estimate that DRFs caused a statistically significant \$27,080 increase in long-run cumulative W-2 earnings, which is equivalent to an 11% increase relative to mean earnings of \$254,900. Individuals in this sample are unlikely to receive a second DRF conviction so this income growth stems from the first DRF conviction. Thus, the fee of \$300 had a sustained increase in earnings even after the fee had been paid.

Assignment of fees stemming from a DRF conviction did not impact criminal behavior for any level of fees in the low contamination sample. Across all levels and time periods, we observe no statistically significant effect on total convictions. The estimates and standard errors are also small relative to the means—indicating null effects.

3.5.5 Effects of DRFs on Romantic Partners

While we generally find that the driver responsibility fees have small or null effects on labor market and recidivism outcomes of DRF recipients, we are also interested in how potential impacts may be internalized by households or romantic partners. For example, a large fine may trigger a change in a partner's labor supply if he or she is the primary earner or in a better position to adjust labor supply on the intensive margin by increasing work hours. To measure partner spillovers, we use the household crosswalk discussed in Section 3.3 that synthesizes information from a variety of Census Bureau, IRS and other federal program data. This crosswalk allows us to link individuals convicted of DRF-eligible offenses to their partners in the year of their initial DRF conviction.

In order to identify the causal impact of DRFs on partner outcomes, we need to have balance in the likelihood of being linked to a partner across the DRF effective date. Panel A of Table 3.6 shows no effect of the fines on the likelihood an individual charged with a DRF offense is linked to a romantic partner in the year of conviction—for individuals in either low or high contamination groups.

The rest of Table 3.6 shows outcomes of interest for partners: the likelihood of remaining in a relationship and the length of the relations in Panel B; cumulative W-2 earnings in Panel C and the total number of convictions in Panel D. As before, we split our results by contamination group. We also consider both short- and long-term versions of the outcomes (relationship status in 2007 and 2015, labor market outcomes from 2005–2007 and 2005–2015, and recidivism outcomes two and ten years after the initial DRF conviction). All outcomes are estimated using the same covariates in previous analyses describing the individual charged with the offense (not the partner).

For the low contamination group, we find no evidence of spillover effects of the DRFs on partnership rates or partner outcomes. Not only are estimates statistically indistinguishable from zero, but, relative to the mean, the effect sizes and standard errors are small, which indicates null effects. These results indicate that earlier findings of null impacts of DRFs on labor market and recidivism outcomes were not confounded by secondary impacts on partnership length, partner labor supply, or partner criminal activity.

For the high contamination group, we find a statistically significant effect of DRFs on partner long-term cumulative earnings (2005–2015), but no statistically significant effects on criminal behavior or partnership outcomes. That is, for individuals at a higher risk of DRF recidivism, partners increased labor supply as a result of DRF conviction, and some of the financial burden affected someone other than the individual convicted of a DRF offense.

3.6 Relationship to Prior Work

Our paper represents the first attempt to measure the long-term impact of financial sanctions using detailed complete criminal case histories, administrative tax records on earnings, and population-level links to romantic partners. In contrast to previous empirical studies on the impacts of financial sanctions from driving offenses, our results generally show null impacts on labor market and recidivism outcomes with some small increases in labor supply of fee recipients and their partners.

These results are divergent from past research, especially that of Mello (2021), which studies the effect of Florida drivers receiving a traffic citation on future employment and economic stability. Using highly detailed credit report data, he finds that driving fines are associated with an increase in financial instability as well as a small but significant decrease in labor supply over the following two years. Below we lay out potential hypotheses as to why our results diverge from past research on the effect of financial sanctions on driving outcomes.

First, the institutional context of our study is markedly different than that of past studies, which affects the efficacy of the financial sanction. In our study, those convicted of a DRF-qualifying offense were subjected to a higher financial burden because they were required to pay the DRF on top of the traffic fine. In contrast, the focus of Mello (2021) is only on traffic citations.

Second, from the definition of Public Act 165, we are only able to study a subset of traffic violations. In contrast, Mello's sample includes all individuals with a traffic violation in Florida. Furthermore, we identify individuals in our sample using court adjudication data. Therefore, even if an individual was convicted of a DRF-qualifying offense but did not go to court, they are not in our data set. In Mello's sample, he is able to identify all individuals who received a citation in Florida.

Third, our earnings data cover a greater scope of employment relationships. Mello's primary employment data comes from monthly credit reports, which may select toward large employers who choose to report to credit agencies. We rely on W-2 information returns that cover all formally employed individuals regardless of employer firm size, as long as those employees earned at least $600 in a tax year.

Finally, there may be differences in data and sample composition between the two settings. Our sample is conditioned on individuals with a DRF-eligible offense receiving a PIK at the Census Bureau, which excludes undocumented immigrants. The sample in Mello's paper is conditioned on individuals with a driving citation data being linked to credit bureau reporting data, which may also exclude some undocumented immigrants.

3.7 Conclusion

This paper examines the effects of financial sanctions on labor market and recidivism outcomes in the state of Michigan. Despite widespread criticism of the law, especially related to driver's license suspensions, we find relatively muted impacts of DRFs. After carefully addressing research design contamination caused by DRF recidivism, we find that the fees had a small positive effect on short-term W-2 income for individuals in the high contamination sample. This effect attenuates in the long term, which may relate to the erosion of the license revocation channel in this subgroup over time. We find suggestive evidence that these individuals may be shifting part of the financial burden onto their romantic partners who have increased earnings of approximately $1,700 on average from 2005 to 2015. We also observe slight deterrent effects for individuals assigned the $300 fee ten years after DRF conviction.

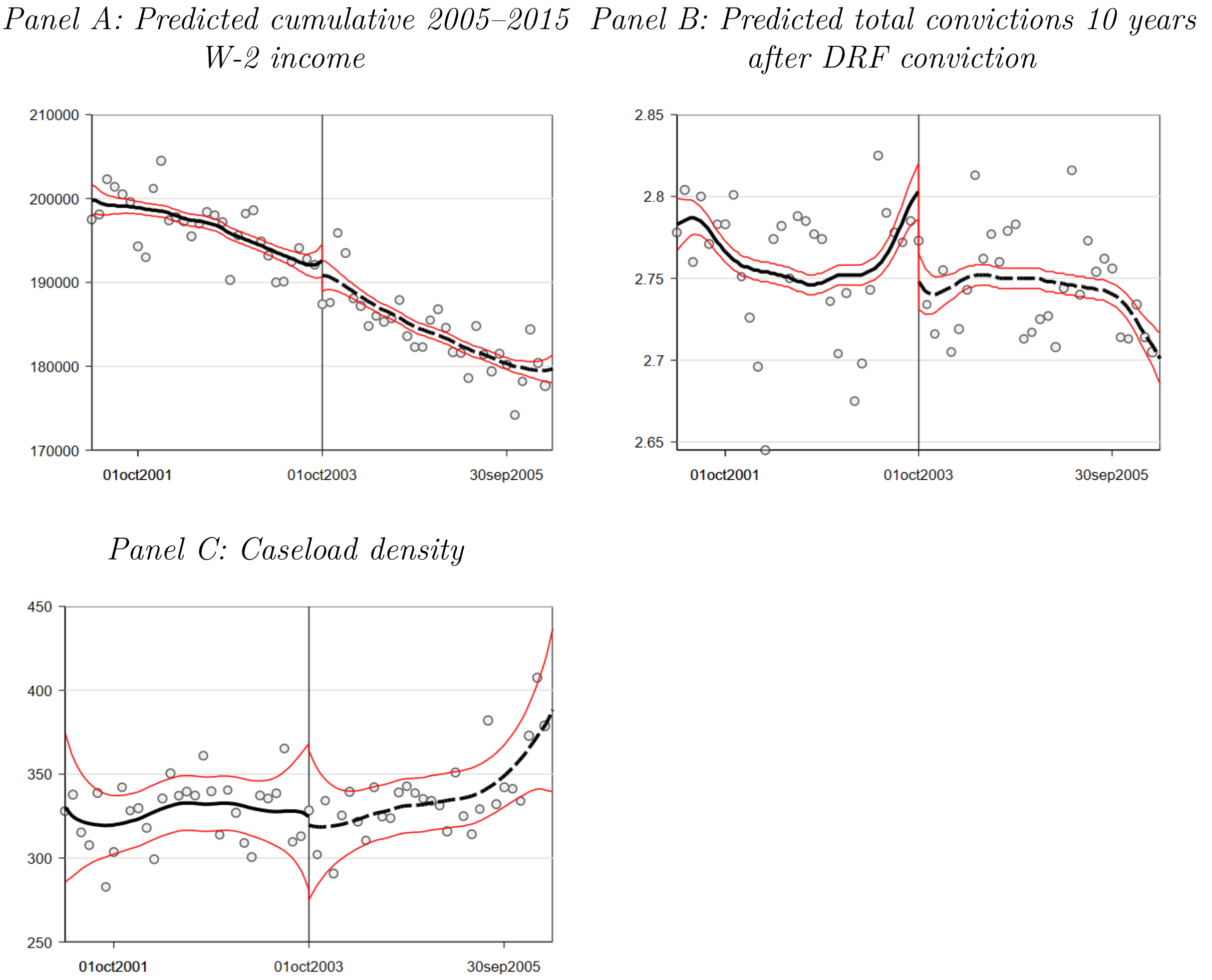
At the same time, we find predominantly null effects on both short and long-term outcomes for individuals in the overall, low contamination sample with no spillover effects onto partners. Our heterogeneity analysis reveals long-term positive impacts of the fees on individuals' cumulative earnings and mild deterrent effects for individuals at the 60th–70th percentiles. Separating individuals by the assigned fee reveals potential incapacitation effects for those with the most severe offense, who have reduced total earnings of almost $8,000 from 2005 to 2007. Since these individuals have low likelihood of re-committing a DRF offense, the negative impact on earnings is not sustained in the long-run. To the best of our knowledge, our results are the first to find such consistent and precisely estimated null to mildly positive effects of these types of financial sanctions.

Our findings contrast with prior work, in particular the causal estimates from Mello (2021). We believe four factors may contribute to these divergent findings: (1) different approaches to research design (RD versus event study); (2) measurement of labor market outcomes (W2 tax filings versus credit bureau employment data, which principally comes from large employers); (3) different study populations; and, (4) the size of the financial sanctions. Further research is warranted to better understand the contribution of each of these differences.

While we find no significant harm on individuals' labor market outcomes or criminal behavior, we also find limited evidence of benefits to justify this policy. As a source of revenue generation, the DRF was an inefficient and regressive form of taxation; if LFOs increased without a concurrent change in labor supply, consumption may have in fact been negatively impacted. Funds were being raised from individuals with lower income compared to the general population.[19] It is therefore unsurprising that DRF payment rates were quite low, reducing

[19] Average per capita, annual personal income in 2005 in Michigan was just over $30,000, unadjusted (of Economic Analysis and of St. Louis, 2021). From Table 3.4, average annual income per capita in our sample is just over $19,000 from 2005-2007 using W-2 information.

revenue and placing these individuals at higher risk of recidivism due to driver's license revocation from non-payment. Our results show that DRF conviction increased the labor supply of romantic partners, suggesting that other household members may be shouldering the monetary burden of the DRF fines. Without clear evidence of general or specific deterrence arising from the DRF policy, it still remains unlikely that the DRF regime was welfare improving even if our causal estimates are less pessimistic than prior research has found.



Panel A: Predicted cumulative 2005–2015 W-2 income

Panel B: Predicted total convictions 10 years after DRF conviction

Panel C: Caseload density

Figure 3.1: Balance Tests Showing Smoothness Around the DRF Effective Date

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage. Notes: These figures represent a series of balance tests done to ensure that there is smoothness around the discontinuity. Equivalent RD estimates shown in Table 3.2 support the figures findings that the discontinuity is not significant. Predicted total recidivism ten years after conviction and predicted cumulative 2005–2015 W-2 income are predicted using the full interaction of age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of: controls for the DRF offense level and fixed effects for the county of adjudication. Caseload density is measured using the total number of DRF convictions per day. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1st, 2001 to March 31st, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. RD figure notes: Figures show smoothed, nonparametric estimates and 95% confidence intervals of the relevant outcome variable, estimated on either side of the discontinuity. In addition, the monthly average of the outcome variable is plotted as a scatter, where the size of each point is weighted by the monthly case count. The estimates are generated using a non-parametric local polynomial with a 120-day bandwidth and weighted with an Epanechnikov kernel estimated separately across both sides of the discontinuity.

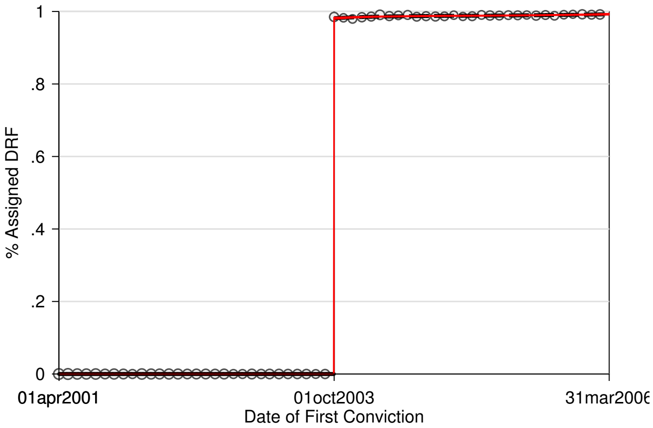
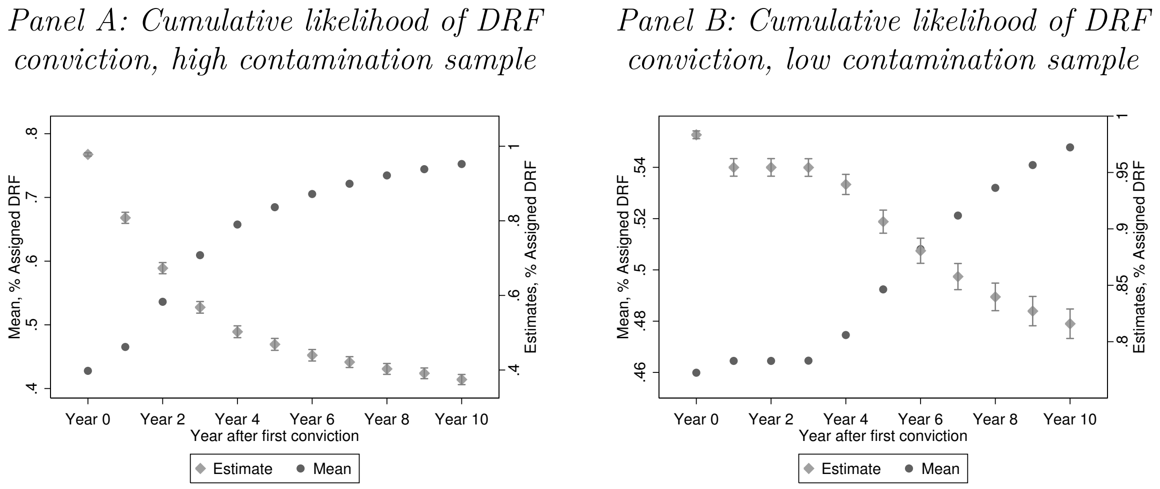


Figure 3.2: First Stage: Likelihood of Receiving Driver Responsibility Fee Before and After DRF Effective Date

Source: Author's calculations from Michigan criminal justice histories from the CJARS 2020Q1 vintage. Notes: This figure shows the smoothed non-parametric estimates and 95% confidence intervals of the first stage estimated on either side of the discontinuity. The outcome is likelihood of DRF conviction after the policy enactment. Equivalent RD estimates support the figures findings that the discontinuity is not significant. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1st, 2001 to March 31st, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. RD figure notes: Figures show smoothed, nonparametric estimates and 95% confidence intervals of the relevant outcome variable, estimated on either side of the discontinuity. In addition, the monthly average of the outcome variable is plotted as a scatter, where the size of each point is weighted by the monthly case count. The estimates are generated using a non-parametric local polynomial with a 120-day bandwidth and weighted with an Epanechnikov kernel estimated separately across both sides of the discontinuity.

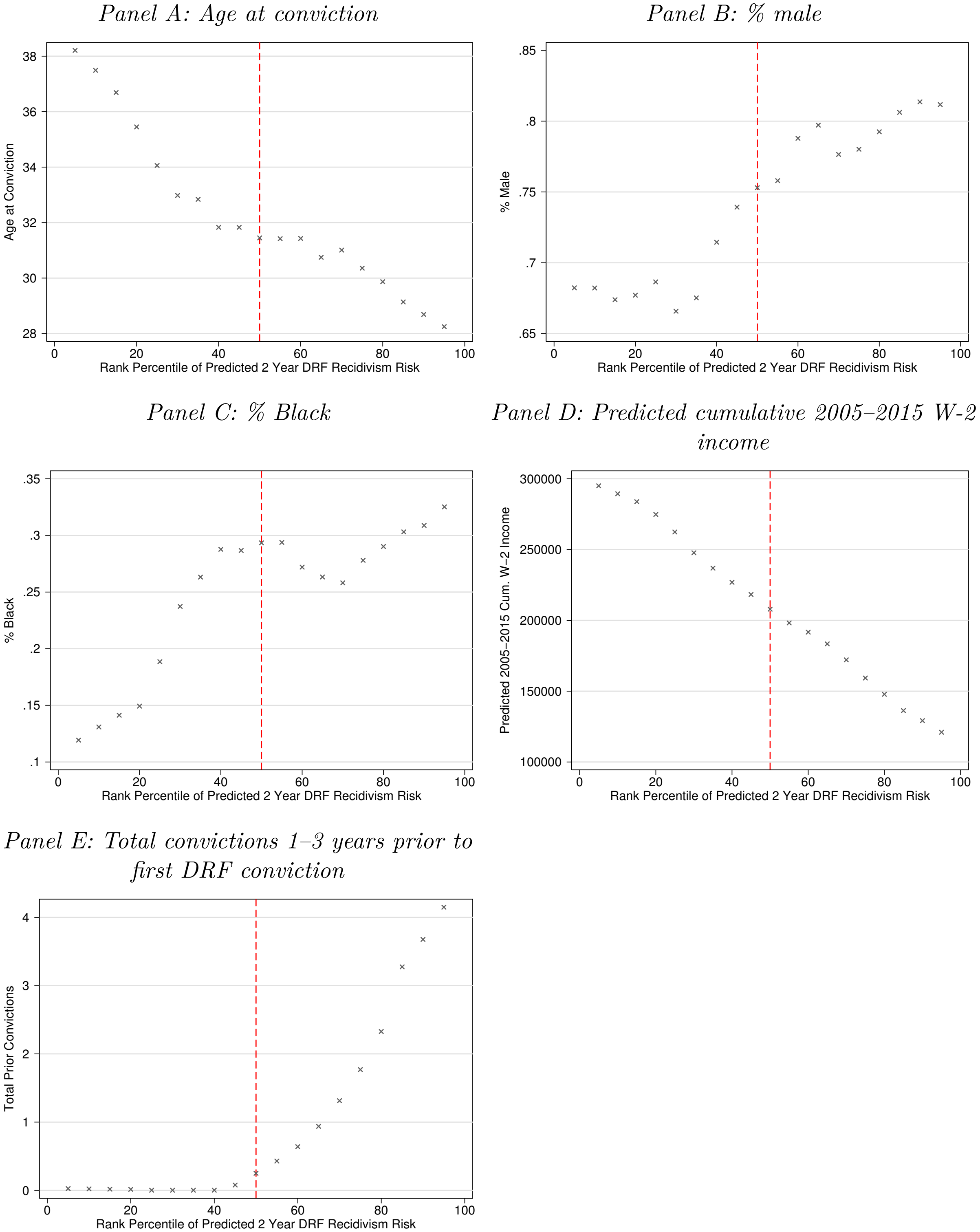


Panel A: Cumulative likelihood of DRF conviction, high contamination sample

Panel B: Cumulative likelihood of DRF conviction, low contamination sample

Figure 3.3: Evolution of Cumulative Likelihood of DRF Conviction Regression-Discontinuity Estimates Over Time and by Contamination Group

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage. Note: These figures plot regression discontinuity estimates measuring likelihood of being assigned a DRF over a cumulative time period starting from the first conviction and the ten years after. The left graphs are for the subsample of individuals in the high contamination group and the right is for the low contamination group. Low (high) contamination is defined as having below (above) median risk for predicted DRF recidivism 2 years after conviction. Two-year DRF recidivism is predicted using the full interaction of age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication. The sample means are included for each outcome variable (dark grey circles). All estimates are shown with 95% confidence intervals. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1st, 2001 to March 31st, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. RD estimate notes: We present estimates using a local-polynomial regression discontinuity. Unless otherwise noted, all regressions include covariates for individual characteristics (age at conviction, gender, race and ethnicity, average income reported on 1040 tax filings 1–3 years pre-conviction, average 1040 filing rate 1–3 years pre-conviction) and the full interaction of fixed effects for county of adjudication with fixed effects for the level of DRF offense.



Panel A: Age at conviction

Panel B: % male

Panel C: % Black

Panel D: Predicted cumulative 2005–2015 W-2 income

Panel E: Total convictions 1–3 years prior to first DRF conviction

Figure 3.4: Means of Selected Characteristics Across the Distribution of Predicted DRF Recidivism

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage.

Note: This figure shows the average of select individual characteristics at different points on the distribution of predicted DRF recidivism two years after original conviction date. Two-year DRF recidivism and 2005–2015 cumulative W-2 income is predicted using the full interaction of age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication. Pre-conviction criminal history is total convictions 1–3 years prior to DRF conviction. Wages and income are adjusted to 2017 dollars using the CPI-All Urban. The means are calculated at every 5th percentile from the 5th to 95th percentile. To reduce noise, we include individuals in the 15 percentiles above and below the central point. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1st, 2001 to March 31st, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023.

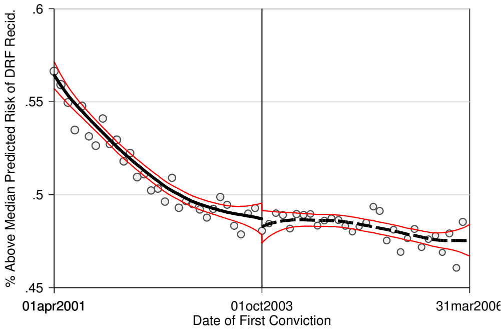
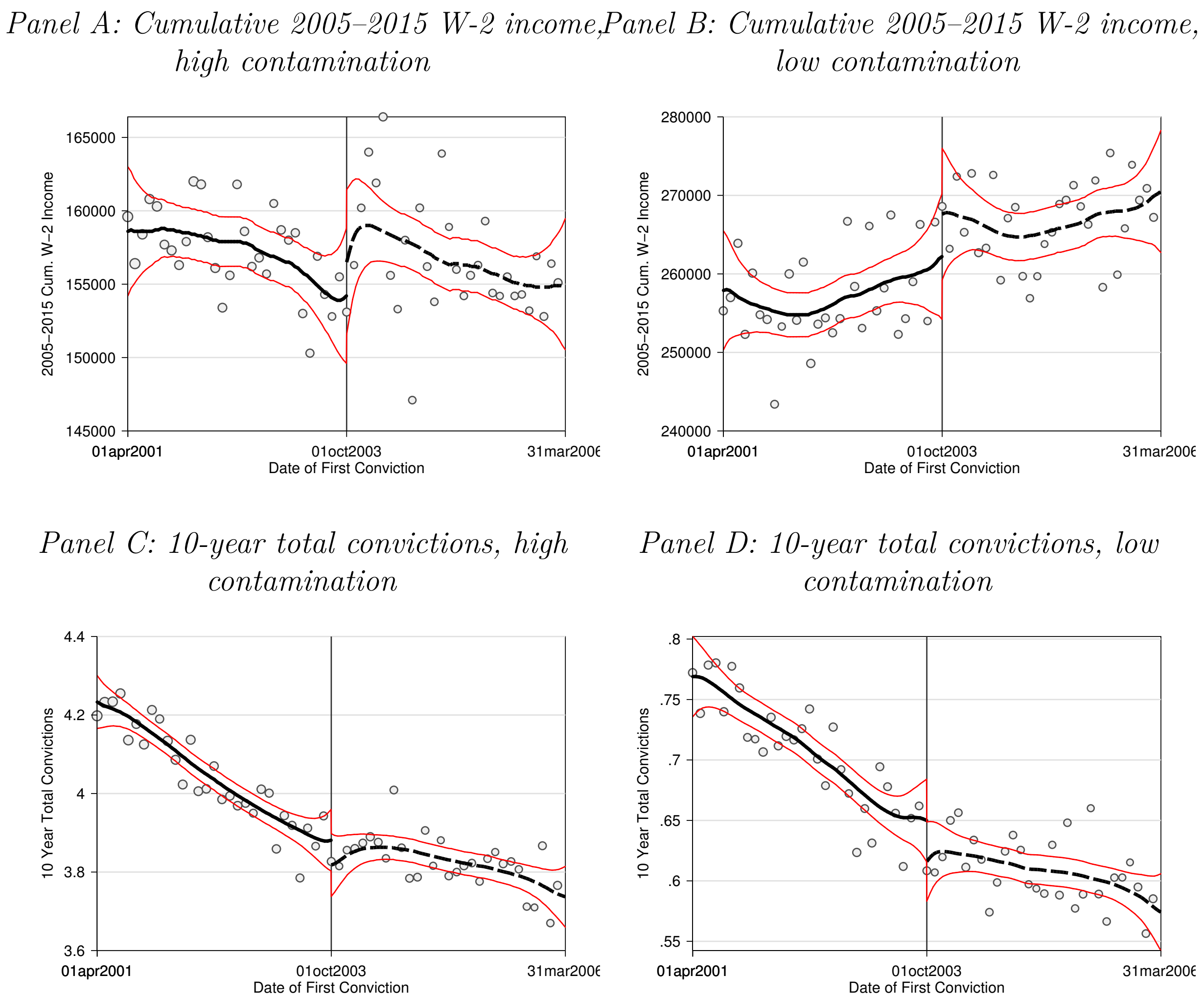


Figure 3.5: Balance Test for Likelihood of Above-Median DRF Recidivism Two Years After Initial DRF Conviction

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage. Note: This figure shows the smoothed non-parametric estimates and 95% confidence intervals of the likelihood of having above median predicted risk of DRF recidivism 2 years after original conviction date on either side of the discontinuity. Two-year DRF recidivism is predicted using the full interaction of: age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication. Equivalent RD estimates in Table 3.2 support the figure's findings that the discontinuity is not significant. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1st, 2001 to March 31st, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. RD figure notes: Figures show smoothed, nonparametric estimates and 95% confidence intervals of the relevant outcome variable, estimated on either side of the discontinuity. In addition, the monthly average of the outcome variable is plotted as a scatter, where the size of each point is weighted by the monthly case count. The estimates are generated using a non-parametric local polynomial with a 120-day bandwidth and weighted with an Epanechnikov kernel estimated separately across both sides of the discontinuity.



Panel A: Cumulative 2005–2015 W-2 income, high contamination

Panel B: Cumulative 2005–2015 W-2 income, low contamination

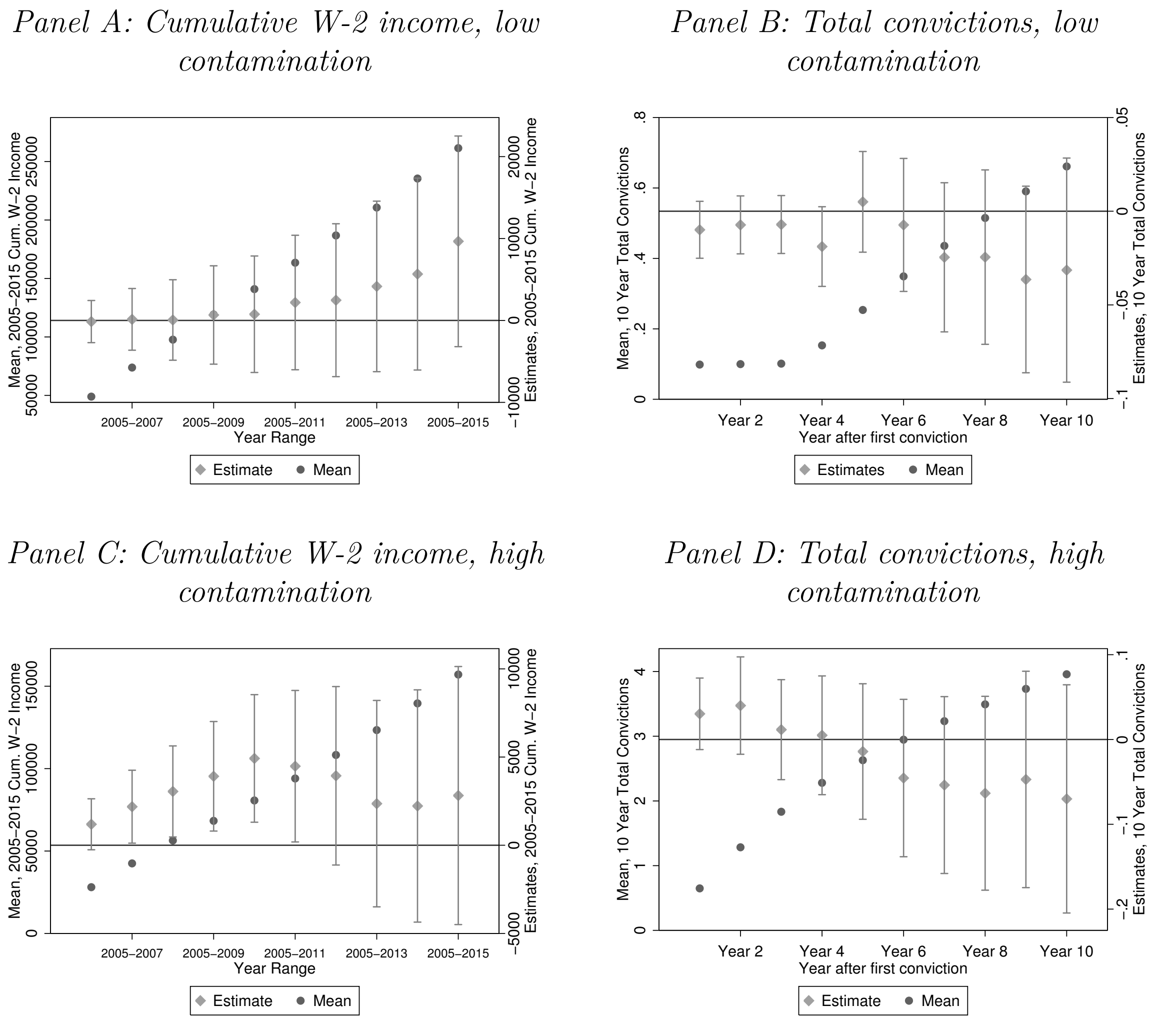
Panel C: 10-year total convictions, high contamination

Panel D: 10-year total convictions, low contamination

Figure 3.6: Effects of DRF Conviction on Long-Term Labor Market Outcomes and Criminal Behavior, by Contamination Group

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage.

Note: This figure shows smoothed nonparametric estimates and 95% confidence intervals of the effect of DRF conviction on total recidivism 10 years after and 2005–2015 cumulative W-2 income (adjusted to 2017 dollars using the CPI-All Urban), estimated on either side of the discontinuity for the low and high contamination groups separately. Low (high) contamination is defined as having below (above) median risk for predicted DRF recidivism 2 years after conviction. 2 year DRF recidivism is predicted using the full interaction of: age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication. For all subfigures, the outcome variable is residualized using the full interaction of: age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level and fixed effects for the county of adjudication. The mean of the outcome variable is added back to the residuals. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1st, 2001 to March 31st, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. RD figure notes: Figures show smoothed, nonparametric estimates and 95% confidence intervals of the relevant outcome variable, estimated on either side of the discontinuity. In addition, the monthly average of the outcome variable is plotted as a scatter, where the size of each point is weighted by the monthly case count. The estimates are generated using a non-parametric local polynomial with a 120-day bandwidth and weighted with an Epanechnikov kernel estimated separately across both sides of the discontinuity.



Panel A: Cumulative W-2 income, low contamination

Panel B: Total convictions, low contamination

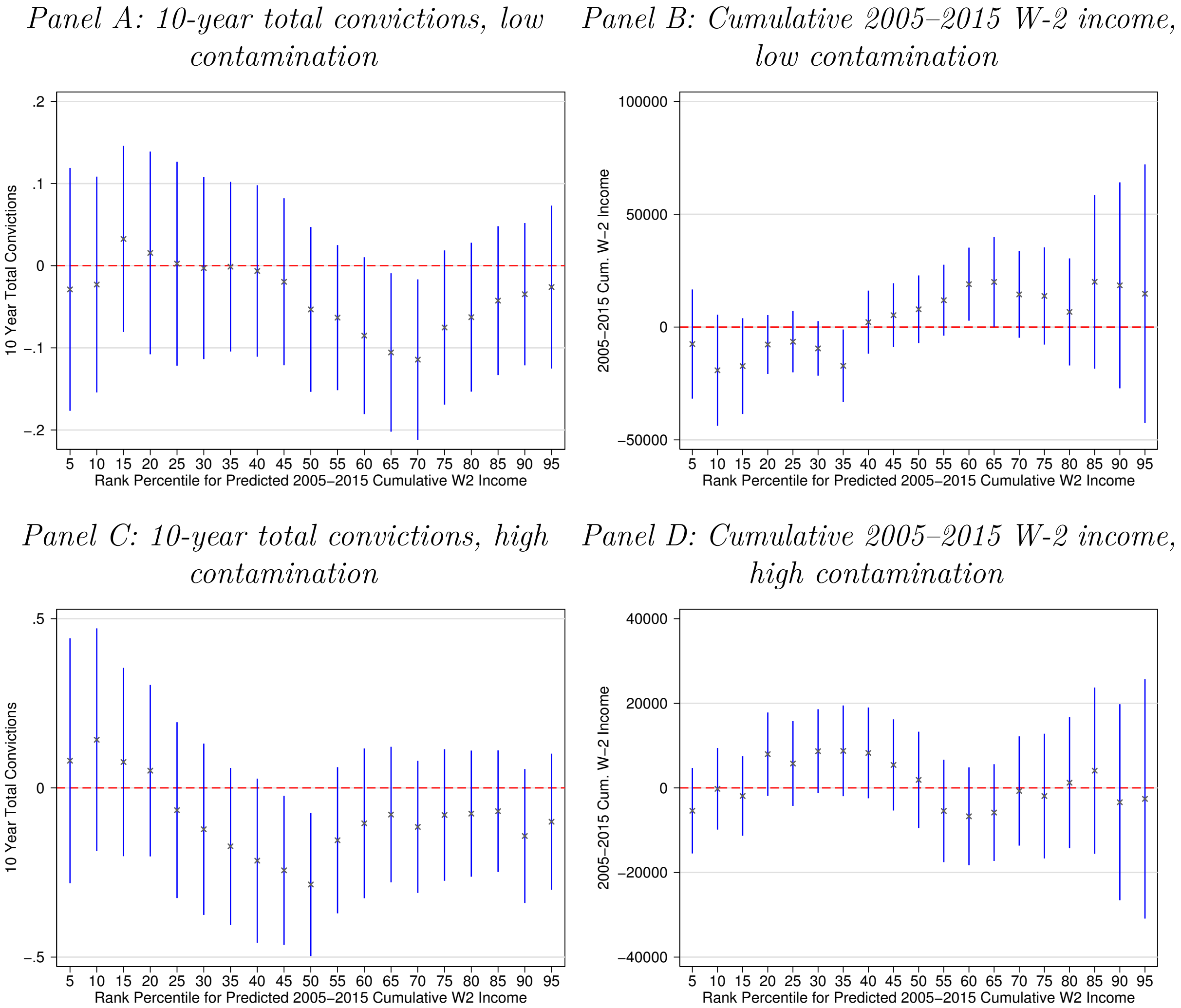
Panel C: Cumulative W-2 income, high contamination

Panel D: Total convictions, high contamination

Figure 3.7: Evolution of DRF Effects on Labor and Recidivism Outcomes Over Time, by Contamination Group

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage.

Note: This figure plots regression discontinuity estimates measuring the effects of DRFs on labor and recidivism outcomes over a cumulative time period that varies by graph. For the labor market outcomes (Adjusted to 2017 dollars using the CPI-All Urban) (subgraphs (a) and (c), the time frame covered is from 2005–2006 to 2005–2015. For the recidivism outcomes (subgraphs (b) and (d)) the time frame is between 1 and 10 years following conviction of first DRF offense. The full sample means are also included for each outcome variable (dark grey circles). All estimates are generated for individuals in the low and high contamination group separately. Low (high) contamination is defined as having below (above) median risk for predicted DRF recidivism two years after conviction. Two-year DRF recidivism is predicted using the full interaction of: age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication. All estimates are shown with 95% confidence intervals. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1st, 2001 to March 31st, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. RD estimate notes: We present estimates using a local-polynomial regression discontinuity. Unless otherwise noted, all regressions include covariates for individual characteristics (age at conviction, gender, race and ethnicity, average income reported on 1040 tax filings 1–3 years pre-conviction, average 1040 filing rate 1–3 years pre-conviction) and the full interaction of fixed effects for county of adjudication with fixed effects for the level of DRF offense.


Panel A: 10-year total convictions, low contamination

Panel B: Cumulative 2005–2015 W-2 income, low contamination

Panel C: 10-year total convictions, high contamination

Panel D: Cumulative 2005–2015 W-2 income, high contamination

Figure 3.8: Heterogeneity Analysis of Effects of DRF Conviction on Labor Market Outcomes and Criminal Behavior, by Predicted Income and by Contamination Group

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage.

Notes: This figure presents the sharp RD estimates for the effect of the DRFs on labor and criminal outcomes at different points in the percentile rank of predicted 2005-2015 cumulative W-2 income distribution for individuals in the low and high contamination group separately. Low (high) contamination is defined as having below (above) median risk for predicted DRF recidivism 2 years after conviction. Predicted 2005–2015 cumulative W-2 income (Adjusted to 2017 dollars using the CPI-All Urban) and two-year DRF recidivism are predicted using the full interaction of: age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication. Estimates are generated separately at every 5 percentiles from the 5th through 95th percentiles. To improve stability, we include individuals in the 15 percentiles above and below the central point. We also include robust 95% confidence intervals of the estimate. The left panels measure the 10 year total recidivism, the top right panels measure the 2005-2015 cumulative W-2 income. The low and high contamination groups are presented separately in the top and bottom panel respectively. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1st, 2001 to March 31st, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. RD estimate notes: We present estimates using a local-polynomial regression discontinuity. Unless otherwise noted, all regressions include covariates for individual characteristics (age at conviction, gender, race and ethnicity, average income reported on 1040 tax filings 1–3 years pre-conviction, average 1040 filing rate 1–3 years pre-conviction) and the full interaction of fixed effects for county of adjudication with fixed effects for the level of DRF offense.

Table 3.1: Michigan Driver Responsibility Fee Amounts and Eligible Offenses

Fee amounts	Eligible offenses
$300–400	• No proof of insurance at the time of the stop • Driving with an expired/invalid license
$1,000	• Driving on a suspended/revoked/denied license • No insurance • Operating with presence of drugs • Operating while impaired by liquor/controlled substance • Reckless driving
$2,000	• Vehicular manslaughter • Felony with an auto • Unlawful/felonious driving • Failing to stop after accident causing injury • Operating while intoxicated • Fleeing or eluding an officer

Source: Michigan Department of State

Notes: This table presents the list of offenses associated with the driver responsibility fee (DRF) assigned upon conviction. This list includes offenses enumerated under Michigan Public Act 165, Category 2, which was in effect from October 1, 2003 to October 1, 2018.

Table 3.2: Balance Tests of Selected Characteristics at Date of Conviction Across DRF Effective Date

	RD estimate		RD estimate
	(standard error)		(standard error)
Variable	[sample mean]	Variable	[sample mean]
Average daily DRF caseload	-16	DRF level 1	-0.012**
	(32)		(0.006)
	[265]		[0.195]
Age at conviction	0.250	DRF level 2	-0.006
	(0.170)		(0.008)
	[32.370]		[0.369]
Male	0.009	DRF level 3	0.004
	(0.008)		(0.009)
	[0.741]		[0.343]
Hispanic	0.002	Pre-conviction average 1040 filing rate	-0.001
	(0.003)		(0.007)
	[0.028]		[0.617]
Black	0.022**	Pre-conviction average 1040 income	-8
	(0.008)		(723)
	[0.247]		[23,890]
White	-0.026**	Pre-conviction average total convictions	-0.004
	(0.009)		(0.033)
	[0.686]		[1.117]
Matched to romantic partner in conviction year	-0.006		
	(0.005)		
	[0.200]		
Observations	423,000	Observations	423,000

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage. Individuals are linked to their romantic partners using the universe of 1040 filings and survey responses to the Decennial (2000) and American Community Survey (ACS) (2005–2018).

Note: This table presents the sharp RD estimates for select characteristics describing the individual at the time of conviction. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1, 2001 to March 31, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. Standard errors are enclosed in parentheses and sample means are enclosed in brackets. $*$ p<0.1, $**$ p<0.05, $***$ p<0.01.

Table 3.3: Balance Tests of Predicted Variables Across DRF Effective Date

Variable	RD estimate (standard error) [sample mean]
Predicted cumulative 2005-2015 W-2 income	914 (2,589) [209,300]
Predicted total convictions 10 years after DRF disposition	-0.011 (0.038) [2.310]
Predicted above-median risk for two-year DRF recidivism	-0.007 (0.007) [0.500]
Observations	423,000

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage. Individuals are linked to their romantic partners using the universe of 1040 filings and survey responses to the Decennial (2000) and American Community Survey (ACS) (2005–2018).

Note: This table presents the sharp RD estimates for select predicted variables. All predicted variables are predicted using the full interaction of: age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication. Wages and income are adjusted to 2017 dollars using the CPI-All Urban. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1, 2001 to March 31, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. Standard errors are enclosed in parentheses and sample means are enclosed in brackets. $*$ p<0.1, $**$ p<0.05, $***$ p<0.01.

Table 3.4: Effects of DRF Conviction on Labor Market Outcomes and Criminal Behavior

Panel A: Labor market outcomes					
Sample→	High contamination		Low contamination		Full sample
Outcome Period→	2005–2007	2005–2015	2005–2007	2005–2015	2005–2015
Average W-2 return rate	0.005	0.010	-0.011	-0.000	0.004
	(0.007)	(0.006)	(0.009)	(0.007)	(0.005)
	[0.675]	[0.597]	[0.725]	[0.650]	[0.624]
Average number of annual W-2 returns	-0.001	0.013	-0.033*	-0.015	-0.001
	(0.021)	(0.014)	(0.020)	(0.014)	(0.010)
	[1.211]	[0.976]	[1.181]	[0.980]	[0.978]
Cumulative W-2 earnings	2,185**	2,818	134	9,660	5,171
	(1,055)	(3,732)	(1,922)	(6,561)	(3,959)
	[42,510]	[157,100]	[73,810]	[261,500]	[209,300]
Cumulative 1040 earnings	1,291	4,699	1,217	12,980	11,430*
	(1,422)	(5,494)	(3,511)	(11,900)	(6,522)
	[48,590]	[196,700]	[106,200]	[401,100]	[298,900]
Panel B: Criminal activity					
Sample→	High contamination		Low contamination		Full sample
Outcome Period→	2 years	10 years	2 years	10 years	10 years
Any conviction	0.014	0.008	-0.003	-0.011	-0.003
	(0.010)	(0.007)	(0.005)	(0.009)	(0.006)
	[0.582]	[0.866]	[0.070]	[0.292]	[0.579]
Any felony conviction	-0.008	-0.004	-0.004*	-0.002	-0.002
	(0.006)	(0.008)	(0.002)	(0.004)	(0.005)
	[0.131]	[0.319]	[0.011]	[0.062]	[0.191]
Total convictions	0.040	-0.070	-0.007	-0.031	-0.071*
	(0.029)	(0.069)	(0.008)	(0.031)	(0.038)
	[1.284]	[3.958]	[0.100]	[0.661]	[2.310]
Total drug convictions	-0.006	-0.039	-0.002	0.002	-0.020
	(0.011)	(0.024)	(0.003)	(0.007)	(0.013)
	[0.131]	[0.449]	[0.009]	[0.065]	[0.257]
Total property convictions	0.006	-0.012	-0.000	-0.002	-0.009
	(0.011)	(0.026)	(0.002)	(0.009)	(0.014)
	[0.163]	[0.529]	[0.011]	[0.075]	[0.302]
Total violent convictions	0.002	-0.016	-0.002	-0.007	-0.010
	(0.008)	(0.014)	(0.002)	(0.006)	(0.008)
	[0.098]	[0.322]	[0.008]	[0.051]	[0.187]
Observations	211,500		211,500		423,000

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage. Note: This table presents the sharp RD estimates for the effects of DRF conviction on labor and criminal outcomes across varying time periods noted in the column names. Wages and income are CPI adjusted to 2017 dollars using the CPI-All Urban. Estimates are generated for individuals in the low and high contamination group and the full sample separately. Low (high) contamination is defined as having below (above) median risk for predicted DRF recidivism two years after conviction. Two-year DRF recidivism is predicted using the full interaction of age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. Standard errors are enclosed in parentheses and sample means are enclosed in brackets. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1, 2001 to March 31, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. Standard errors are enclosed in parentheses and sample means are enclosed in brackets. RD estimate notes: We present estimates using a local-polynomial regression discontinuity. Unless otherwise noted, all regressions include covariates for individual characteristics (age at conviction, gender, race and ethnicity, average income reported on 1040 tax filings 1–3 years pre-conviction, average 1040 filing rate 1–3 years pre-conviction) and the full interaction of fixed effects for county of adjudication with fixed effects for the level of DRF offense. $*$ p<0.1, $**$ p<0.05, $* * *$ p<0.01.

Table 3.5: Effects of DRF Conviction on Labor Market Outcomes and Criminal Behavior by Fee Level

Panel A: Cumulative W-2 earnings

	Sample→	High contamination		Low contamination	
Fine	Period→	2005–2007	2005–2015	2005–2007	2005–2015
$300		2,574	3,376	5,608*	27,080**
		(2,843)	(8,060)	(3,330)	(12,320)
		[37,990]	[147,000]	[68,310]	[254,900]
$1000		1,400	3,941	4,199	15,160
		(1,365)	(5,086)	(4,083)	(15,020)
		[41,550]	[154,200]	[61,250]	[208,700]
$2000		2,113	4,907	-7,830**	-16,090
		(2,522)	(8,266)	(3,668)	(12,160)
		[49,140]	[173,600]	[83,930]	[291,400]

Panel B: Total convictions

	Sample→	High contamination		Low contamination	
Fine	Period→	2 years	10 years	2 years	10 years
$300		0.086	-0.263**	0.002	-0.021
		(0.065)	(0.134)	(0.014)	(0.044)
		[1.433]	[4.341]	[0.091]	[0.639]
$1000		0.016	-0.043	-0.019	-0.083
		(0.038)	(0.094)	(0.019)	(0.068)
		[1.176]	[3.780]	[0.128]	[0.858]
$2000		0.028	0.050	-0.006	-0.031
		(0.062)	(0.137)	(0.012)	(0.045)
		[1.373]	[3.972]	[0.094]	[0.585]

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage. Note: This table presents the sharp RD estimates for the effects of DRF conviction on cumulative income (Adjusted to 2017 dollars using the CPI-All Urban) measured using W-2 tax returns and total convictions by the fee amount across varying time periods noted in the columns. Estimates are generated for individuals in the low and high contamination group. Low (high) contamination is defined as having below (above) median risk for predicted DRF recidivism two years after conviction. Two-year DRF recidivism is predicted using the full interaction of: age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication. The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1, 2001 to March 31, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. Standard errors are enclosed in parentheses and sample means are enclosed in brackets. RD estimate notes: We present estimates using a local-polynomial regression discontinuity. Unless otherwise noted, all regressions include covariates for individual characteristics (age at conviction, gender, race and ethnicity, average income reported on 1040 tax filings 1–3 years pre-conviction, average 1040 filing rate 1–3 years pre-conviction) and the full interaction of fixed effects for county of adjudication with fixed effects for the level of DRF offense. $*$ p<0.1, $**$ p<0.05, $***$ p<0.01.

Table 3.6: Effects of DRF Conviction on Partnership Outcomes and Partner's Labor Market Outcomes and Criminal Behavior

Panel A: Balance tests					
Outcome	Sample→	High contamination		Low contamination	
Matched to romantic partner in year of conviction		-0.005		-0.008	
		(0.007)		(0.008)	
		[0.1756]		[0.2151]	
Observations		211,500		211,500	
Panel B: Relationship outcomes					
	Sample→	High contamination		Low contamination	
Outcome	Period→	2007	2015	2007	2015
Survival rate		-0.020	-0.002	0.005	-0.013
		(0.026)	(0.024)	(0.022)	(0.023)
		[0.627]	[0.314]	[0.739]	[0.474]
Years together		-0.103	-0.153	0.009	-0.072
		(0.068)	(0.204)	(0.061)	(0.205)
		[2.047]	[5.139]	[2.341]	[6.450]
Observations		36,000		44,500	
Panel C: Partner labor market outcomes					
	Sample→	High contamination		Low contamination	
Outcome	Period→	2005–2007	2005–2015	2005–2007	2005–2015
Cum. W-2 earnings		4,040	18,670*	-1,400	-3,214
		(3,601)	(11,310)	(4,489)	(17,680)
		[58,690]	[202,900]	[100,900]	[361,900]
Observations		36,000		44,500	
Panel D: Partner criminal behavior					
	Sample→	High contamination		Low contamination	
Outcome	Period→	2 years	10 years	2 years	10 years
Total convictions		0.007	0.037	-0.004	0.022
		(0.036)	(0.112)	(0.025)	(0.068)
		[0.284]	[1.051]	[0.139]	[0.529]
Observations		36,000		44,500	

Source: Authors' calculations from 1998–2015 IRS 1040 individual tax returns, 2005–2015 IRS W-2 information returns, the 2020 Census Numident (to measure year of birth, state of birth, and gender), the 2020 Census Bureau Title 13 race/ethnicity file, and Michigan criminal justice histories from the CJARS 2020Q1 vintage. Individuals are linked to their romantic partners using the universe of IRS 1040 returns and survey responses to the Decennial (2000) and American Community Survey (ACS) (2005–2018). Note: This table presents the sharp RD estimates for the effects of DRF conviction on partnership and partner outcomes. Panel A is a balance test for the likelihood an individual is in a relationship in the year of conviction. Panel B presents results on partnership outcomes, cumulative W-2 earnings (Adjusted to 2017 dollars using the CPI-All Urban), and total convictions for the partners of individuals conditional on observing a relationship during the year of conviction. Estimates are generated separately for individuals in the low and high contamination group. Low (high) contamination is defined as having below (above) median risk for predicted DRF recidivism two years after conviction. Two-year DRF recidivism is predicted using the full interaction of age at conviction for DRF offense, gender, race/ethnicity, average 1040 income and filing rates 1–3 years prior to conviction, and fixed effects for number of previous convictions and the full interaction of controls for the DRF offense level with fixed effects for the county of adjudication.The estimates are based off of a sample of all individuals convicted of a driver responsibility fee qualifying offense from April 1, 2001 to March 31, 2006 in Michigan. The sample contains only the first DRF-qualifying offense of the individuals within the relevant time period. Estimates and sample sizes have been rounded according to Census Bureau DRB rules. All results were approved for release by the Census Bureau, authorization number CBDRB-FY21-ERD002-023. Standard errors are enclosed in parentheses and sample means are enclosed in brackets. RD estimate notes: We present estimates using a local-polynomial regression discontinuity. Unless otherwise noted, all regressions include covariates for individual characteristics (age at conviction, gender, race and ethnicity, average income reported on 1040 tax filings 1–3 years pre-conviction, average 1040 filing rate 1–3 years pre-conviction) and the full interaction of fixed effects for county of adjudication with fixed effects for the level of DRF offense. $*$ $p<0.1$, $**$ $p<0.05$, $***$ $p<0.01$.

APPENDICES

APPENDIX A

Appendix to The Long-Term Impacts of Rent Control

A.1 Opportunity Insight Data Description

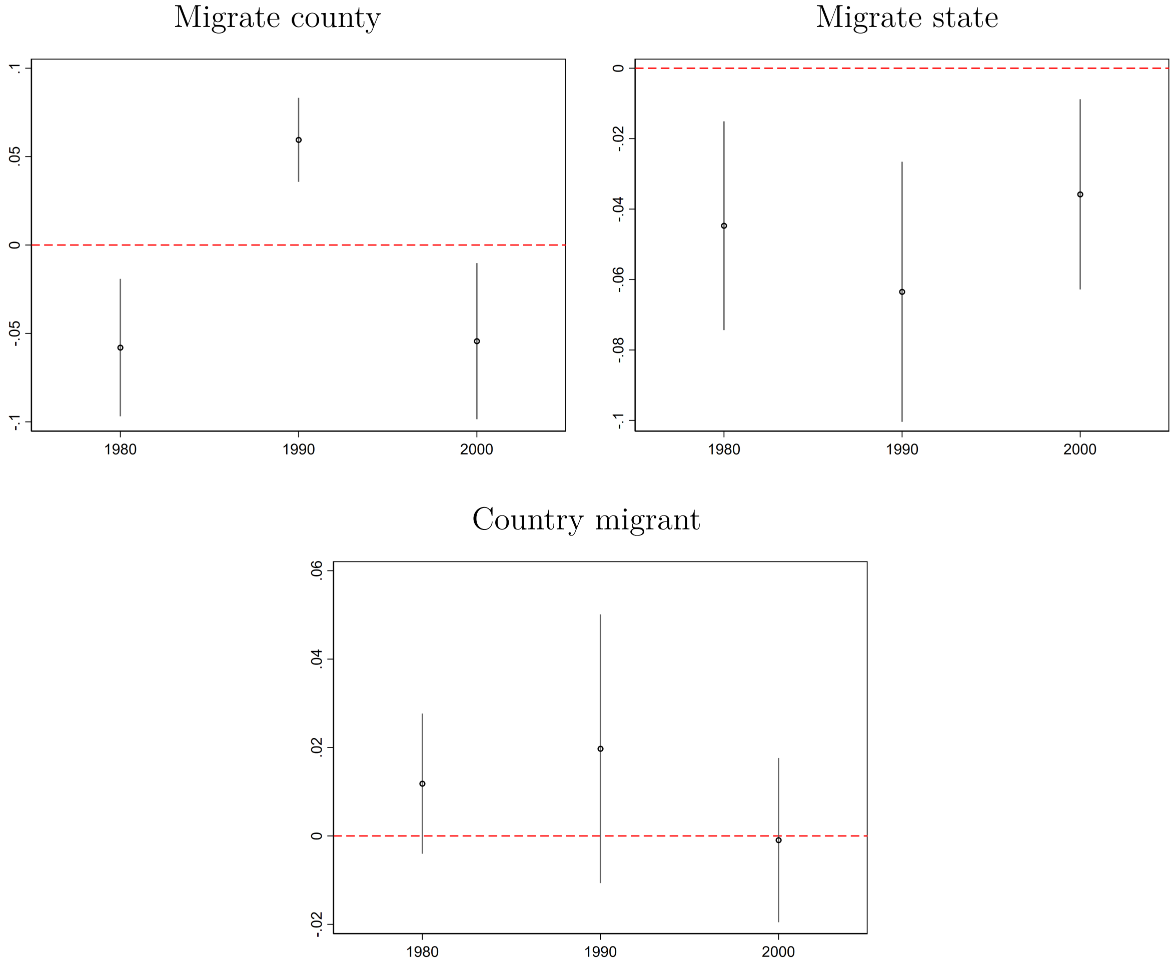
As discussed in the body of the paper, the Opportunity Insights includes unconditional mean outcome estimates for all children linked to a given census tract. In addition, they provide predicted outcomes for children growing up at 5 specific points of the parent income distribution. Using data on the income level of parents, as well as the race, gender and census tract that children grew up in, they run the following regression:

$$y_i = \alpha_{crg} + \beta_{crg} \times f_{rg}(p_i) + \epsilon_i$$

where y_i is an outcome for child i, α_{crg} is a fixed effect for the combination of census tract (c), race (r) and gender (g). $f_{rg}(p_i)$ refers to a transformation of parental income rank that is estimated at the national level using local polynomial regression. This is done because the true relationship between parental income and the outcome of children is often non-linear, so the transformed variable allows for the capture of these non-linearities within a simple regression framework. Once the individual regression is estimated, the authors use predicted values at a given level of parental income for each census tract-race-gender combination to create outcome predictions at the census tract level. The outcome data released to the public is the average fitted value over the entire census tract for the given population. Lastly, the authors add a small amount of noise to the estimates to avoid involuntary disclosure before releasing the data to the public.

The noise infusion procedure is detailed in Chetty and Friedman (2019), though the main intuition behind the algorithm is that the amount of noise added is proportional to the sensitivity of the statistic to one observation. For each tract, they determine the observation that has the largest impact on the result by estimating the result in the absence of each observation in the tract. Once they determine the maximum observed sensitivity, they add noise to the final result with a mean zero, normally distributed term that has a standard deviation proportional to the maximum observed sensitivity. Chetty et al. (2018) report that the added noise is generally smaller than the sampling variance. This will be especially true in tracts located in cities that have higher populations since the influence of any one observation has a smaller effect on the final output.

A.2 Appendix Tables and Figures



Migrate county

Migrate state

Country migrant

Figure A.1: Estimated ATT of Rent Control on Immigration Outcomes by Year: High Rental Tract Sample

Source: 1980 - 2000 decennial census data reported at the tract level by SocialExplorer.

Notes: Figures show the average treatment effect on the treated tracts of rent control on immigration outcomes. The unit of observation is a census tract with at least 30% rental share. Each outcome is the percentage of tract inhabitants that have moved from a given location in the last 5 years. The error bars represent 95% confidence intervals from standard errors that are clustered at the city level.

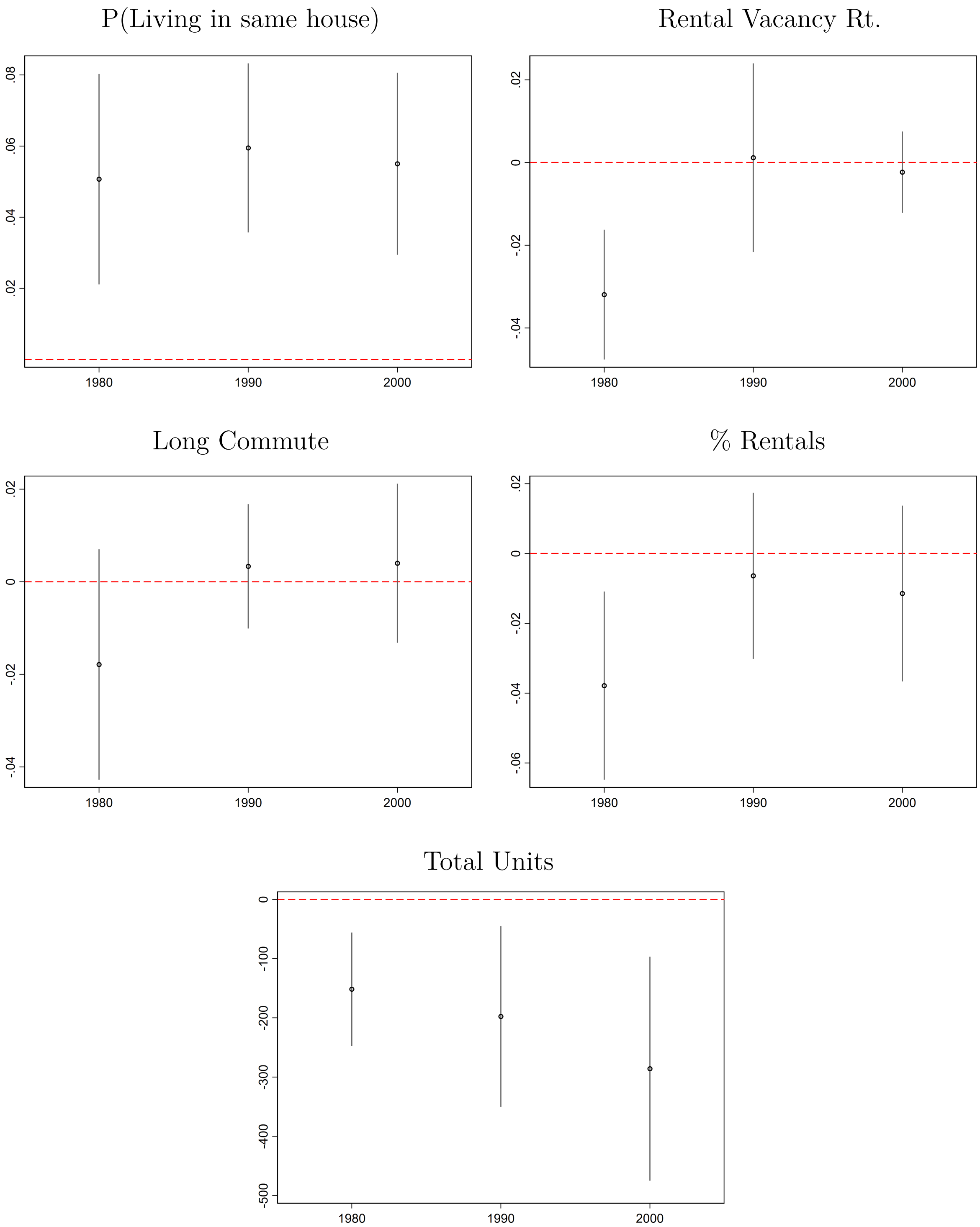


Figure A.2: Estimated ATT of Rent Control on Housing Outcomes by Year: High Rental Tract Sample

Source: 1980 - 2000 decennial census data reported at the tract level by SocialExplorer.

Notes: Figures show the average treatment effect on the treated tracts of rent control on housing outcomes. The unit of observation is a census tract with at least 30% rental share. The average outcomes are generated using the baseline nearest neighbor match model. The error bars represent 95% confidence intervals from standard errors that are clustered at the city level.

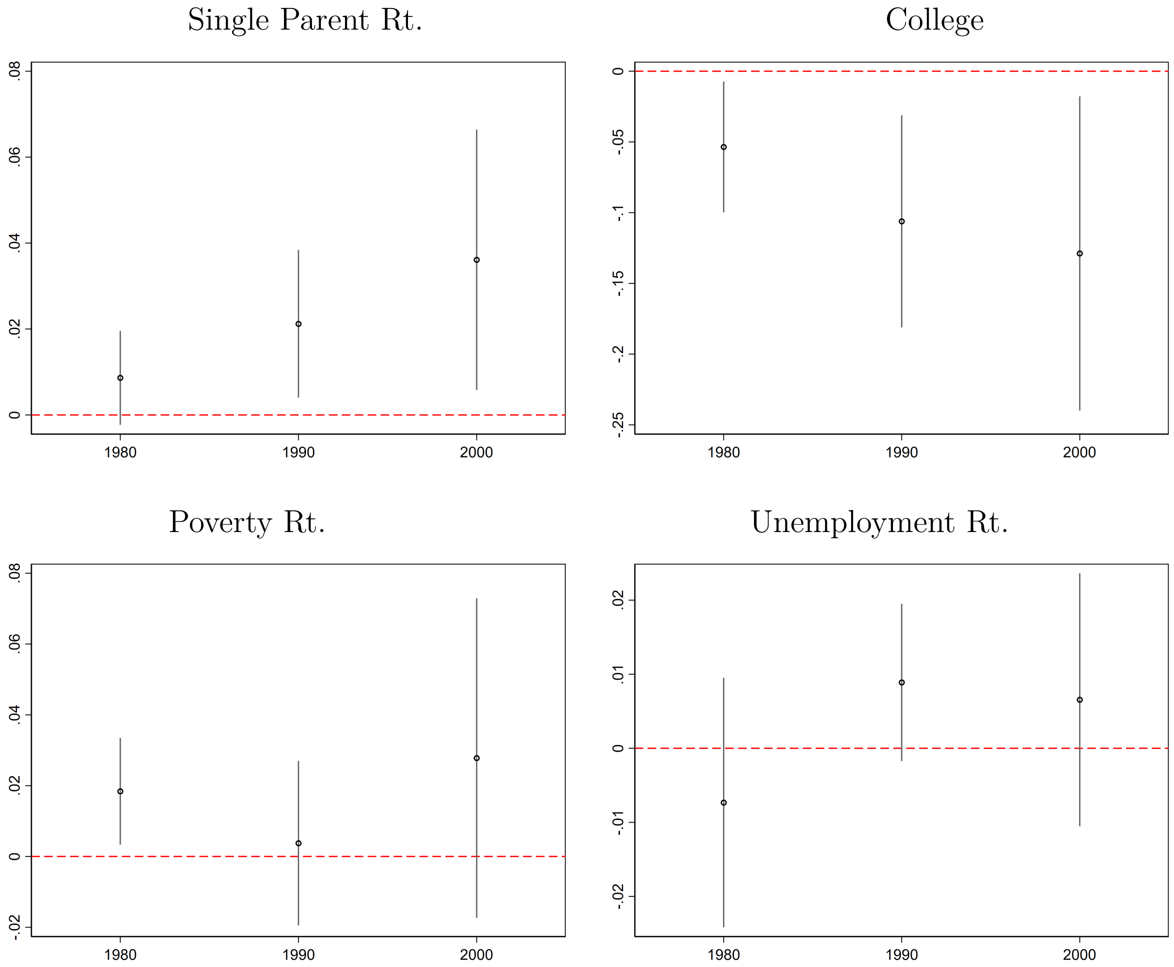


Figure A.3: Estimated ATT of Rent Control on Demographic Outcomes

Source: 1980 - 2000 decennial census data reported at the tract level by SocialExplorer.

Notes: Figures show the average treatment effect on the treated tracts of rent control on employment and demograhic outcomes. The unit of observation is a census tract with at least 30% rental share. The average outcomes are generated using the baseline nearest neighbor match model. The error bars represent 95% confidence intervals from standard errors that are clustered at the city level.

Table A.1: Comparing Means and Variances of the Raw and Weighted Samples Using the Two-Step Nearest Neighbor Matching Algorithm: High Rental Tract Sample

	Std. Mean Diff.		Var. Ratio	
	Raw	Matched	Raw	Matched
Population	-0.033	-0.199	0.567	1.128
Male (%)	-0.101	0.008	1.028	1.225
Pop./sq. mile	0.868	-0.006	2.552	0.882
Age median	0.347	-0.079	1.060	1.150
white (%)	-0.189	-0.104	1.167	1.083
black (%)	0.086	0.064	1.126	1.099
Married (%)	-0.492	-0.135	0.910	1.652
Single parent fam. (%)	0.282	0.220	1.537	1.127
Educ. Less than HS (%)	-0.149	-0.198	0.861	1.340
Educ. HS (%)	0.113	0.050	0.679	1.330
LFP rate	0.145	-0.046	0.696	1.189
Unemployment rate	0.443	0.481	1.245	1.002
Avg. inc.	0.309	0.004	1.269	1.352
Family poverty rt. (%)	-0.142	0.114	0.683	0.941
Current addr. 5 years (%)	-0.019	-0.346	0.844	1.098
Housing units	0.072	-0.160	0.645	1.254
Rental (% of total units)	0.807	0.216	1.128	1.111
Rental vacancy rate (%)	-0.478	0.051	0.358	1.130
Rent vacancy x Rental %	-0.223	0.088	0.572	1.191
Avg. rent	0.478	0.100	0.962	1.305
Avg. home value	0.782	0.315	1.404	1.221
City rental (% of total units)	1.560	0.248	1.249	0.933
City rental vacancy rate (%)	-0.720	-0.298	0.311	0.768
City white (%)	-0.444	0.060	0.816	0.536
City black (%)	0.256	-0.188	0.717	0.390
City unemployment rate	0.842	0.682	0.825	1.412
City avg. rent	0.994	0.165	0.417	0.752
County Dem. vote share 1968	0.792	0.024	0.641	0.232
County Wallace vote share 1968	-0.884	-0.136	0.019	0.735
City population	0.935	0.218	3.001	0.753
City family poverty rt. (%)	-0.136	-0.029	0.387	0.888
City revenue per capita	0.952	0.328	2.169	1.264
City tax revenue per capita	0.959	0.061	2.667	0.760
Property tax share of rev.	0.406	-0.272	1.536	1.308
Other gov. sources share of rev.	0.180	-0.111	0.863	1.354
Educ. share of expenditure	0.426	-0.026	2.233	1.206
Police share of expenditure	-0.121	-0.406	0.377	0.465
Welfare share of expenditure	0.270	-0.077	4.426	0.571
N	15,199			
N Treat	840			
N unique control	373			

Table shows the standardized differences in means and variances between the raw and weighted sample. The unit of observation is a census tract with at least 30% rental share. The treated tracts are matched to a control tract using a nearest neighbor Mahalanobis distance matching procedure. The Mahalanobis distance metric includes linear terms for tract-level, city-level and county-level characteristics.

APPENDIX B

Appendix to Modernizing Person-Level Entity Resolution with Biometrically Linked Records

Appendix Tables

Table B.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.

B.1 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.[1] 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.2 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

[1] Another possible strategy could take a demographically stratified sample to ensure that specific demographic groups are sufficiently represented in the training sample. While this may improve model performance within smaller demographic groups, it could come at the expense of worse performance in larger demographic groups and consequentially worse overall performance given a fixed budget constraint on the total number of feasible hand-coded training observations.

same person by the deterministic algorithm.

Naive 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-Score 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 continuous 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-Score 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-Score. 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-Score 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. [2] 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-Score across the score space.

[2] See Appendix B.2 for details on sample construction.

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.

B.3 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, 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. [3]

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 character recognition

[3] 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.

(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.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

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.

Abramitzky, R., Boustan, L. P., and Eriksson, K. (2014). A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy*, 122(3):467–506.

Abramitzky, R., Boustan, L. P., Eriksson, K., Feigenbaum, J. J., and Pérez, 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.

Almond, D., Currie, J., and Duque, V. (2018). Childhood circumstances and adult outcomes: Act ii. *Journal of Economic Literature*, 56(4):1360–1446.

Andersson, F., Haltiwanger, J. C., Kutzbach, M. J., Palloni, G. E., Pollakowski, H. O., and Weinberg, D. H. (2016). Childhood housing and adult earnings: A between-siblings analysis of housing vouchers and public housing. Working Paper 22721, National Bureau of Economic Research.

Asquith, B. J. (2018). Do rent increases reduce the housing supply under rent control? evidence from evictions in san francisco. *Working Paper*.

Ault, R. and Saba, R. (1990). The economic effects of long-term rent control: The case of new york city. *The Journal of Real Estate Finance and Economics*, 3(1):25–41.

Ault, R. W., Jackson, J. D., and Saba, R. P. (1994). The effect of long-term rent control on tenant mobility. *Journal of Urban Economics*, 35(2):140 – 158.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25):3083–3107.

Autor, D. H., Palmer, C. J., and Pathak, P. A. (2014). Housing market spillovers: Evidence from the end of rent control in cambridge, massachusetts. *Journal of Political Economy*, 122(3):661–717.

Autor, D. H., Palmer, C. J., and Pathak, P. A. (2017). Gentrification and the amenity value of crime reductions: Evidence from rent deregulation. Working Paper 23914, National Bureau of Economic Research.

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.

Bannon, A., Mitali, N., and Rebekah, D. (2010). Criminal justice debt: A barrier to reentry. Brennan Center for Justice.

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.

Bloemen, H. and Stancanelli, E. (2008). How do parents allocate time? Tinbergen Institute Discussion Paper 08-079/3.

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

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *Stata Journal*, 17(2):372–404.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression discontinuity designs. *Econometrica*, 82(6):2295–2326.

Carlson, K. (2015). Fear itself: The effects of distressing economic news on birth outcomes. *Journal of Health Economics*, 41:117 – 132.

Carneiro, P., García, I. L., Salvanes, K. G., and Tominey, E. (2021). Intergenerational mobility and the timing of parental income. *The Journal of political economy*.

Carrasco, Jr., J. (2018). Slamming the brakes on driver responsibility fees. *State Notes: Topics of Legislative Interest*, Fall 2018.

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

Chetty, R., Friedman, J., Hendren, N., Jones, M., and Porter, S. (2018). The opportunity atlas: Mapping the childhood roots of social mobility. Working Paper 25147, National Bureau of Economic Research.

Chetty, R. and Friedman, J. N. (2019). A practical method to reduce privacy loss when disclosing statistics based on small samples. *Journal of Privacy and Confidentiality*, 9(2).

Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R., and Narang, J. (2017). The fading american dream: Trends in absolute income mobility since 1940. *Science*, 356(6336):398–406.

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.

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

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

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

Chetty, R., Hendren, N., 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.

Christen, P. and Churches, T. (2002). Febrl - freely extensible biomedical record linkage. Technical report.

Chyn, E. (2018). Moved to opportunity: The long-run effects of public housing demolition on children. *American Economic Review*, 108(10):3028–56.

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.

Clubb, J. M., Flanigan, W. H., and Zingale, N. H. (2006). Electoral data for counties in the united states: Presidential and congressional races, 1840-1972.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences.

Coile, C. (2004). Retirement incentives and couples' retirement decisions. *Topics in Economic Analysis and Policy*, 4(1):1–30.

Cunha, F., Heckman, J. J., Lochner, L., and Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. volume 1 of *Handbook of the Economics of Education*, pages 697 – 812. Elsevier.

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.

Diamond, R., McQuade, T., and Qian, F. (2019). The effects of rent control expansion on tenants, landlords, and inequality: Evidence from san francisco. *American Economic Review*, 109(9):3365–94.

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. Technical Report 14-EHC033-EF.

Early, D. W. (2000). Rent control, rental housing supply, and the distribution of tenant benefits. *Journal of Urban Economics*, 48(2):185 – 204.

Eppler-Epstein, S., Gurvis, A., and King, R. (2016). The alarming lack of data on latinos in the criminal justice system. Washington, DC: Urban Institute.

Fallis, G. and Smith, L. B. (1984). Uncontrolled prices in a controlled market: The case of rent controls. *The American Economic Review*, 74(1):193–200.

Feigenbaum, J. J. (2016). A machine learning approach to census record linking. Technical report.

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

Fernandes, A. D., Cadigan, M., Edwards, F., and Harris, A. (2019). Monetary sanctions: A review of revenue generation, legal challenges, and reform. *Annual Review of Law and Social Science*, 15(1):397–413.

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.

Finlay, K. and Mueller-Smith, M. (2021). *Criminal Justice Administrative Records System (CJARS) [dataset]*. `https://cjars.isr.umich.edu`.

Finlay, K., Mueller-Smith, M., and Street, B. (2021). Inequalities in child exposure to the us criminal justice system and implications of changing household structure. Working Paper.

Fogelson, R. M. (2013). *The Great Rent Wars : New York, 1917-1929*. Yale University Press.

Ford, M. (2015). The missing statistics of criminal justice. *The Atlantic*.

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.

Glaeser, E. L. and Luttmer, E. F. P. (2003). The misallocation of housing under rent control. *The American Economic Review*, 93(4):1027–1046.

Grampp, W. D. (1950). Some effects of rent control. *Southern Economic Journal*, 16(4):425–447.

Grogger, J., Ivandic, R., and Kirchmaier, T. (2020). Comparing conventional and machine-learning approaches to risk assessment in domestic abuse cases. Technical Report Discussion Paper No 1676, Centre for Economic Performance.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420.

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

Gyourko, J. and Linneman, P. (1989). Equity and efficiency aspects of rent control: An empirical study of new york city. *Journal of Urban Economics*, 26(1):54 – 74.

Gyourko, J. and Linneman, P. (1990). Rent controls and rental housing quality: A note on the effects of new york city's old controls. *Journal of Urban Economics*, 27(3):398 – 409.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hankins, S. and Hoekstra, M. (2011). Lucky in life, unlucky in love? the effect of random income shocks on marriage and divorce. *Journal of Human Resources*, 46(2):403–26.

Hansen, B. (2015). Punishment and deterrence: Evidence from drunk driving. *American Economic Review*, 105(4):1581–1617.

Harris, A., Evans, H., and Beckett, K. (2010). Drawing blood from stones: Legal debt and social inequality in the contemporary united states. *American Journal of Sociology*, 115(6):1753–99.

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

Hausman, J. S. (2013). Driving up fees: Muskegon court officials bemoan michigan's driver responsibility fees' effects on poor. *Michigan Live*.

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.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hoynes, H., Schanzenbach, D. W., and Almond, D. (2016). Long-run impacts of childhood access to the safety net. *The American Economic Review*, 106(4):903–934.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 171(2):481–502.

Jann, B. (2019). Influence functions for linear regression (with an application to regression adjustment). Working Paper 32, University of Bern.

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.

Kearney, M. S., Harris, B. H., Jácome, E., and Parker, L. (2014). Ten economic facts about crime and incarceration in the united states. Washington, DC: Brookings Institution Hamilton Project Policy Memo.

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.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454.

Krol, R. and Svorny, S. (2005). The effect of rent control on commute times. *Journal of Urban Economics*, 58(3):421 – 436.

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.

Lett, M. R. (1976). *Rent Control: Concepts, Realities, and Mechanisms*. Center for Urban Policy Research.

Liu, H. and Zhao, Z. (2014). Parental job loss and children's health: Ten years after the massive layoff of the soes' workers in china. *China Economic Review*, 31:303–19.

Logan, J., Xu, Z., and Stults, B. (2014). Interpolating u.s. decennial census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer*, 66.

Luca, D. L. (2015). Do traffic tickets reduce motor vehicle accidents? evidence from a natural experiment. *Journal of Policy Analysis and Management*, 34(1):85–106.

Makowsky, M. D. and Stratmann, T. (2009). Political economy at any speed: What determines traffic citations? *American Economic Review*, 99(1):509–27.

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

Martin, K. D., Sykes, B. L., Shannon, S., Edwards, F., and Harris, A. (2018). Monetary sanctions: Legal financial obligations in us systems of justice. *Annual Review of Criminology*, 1(1):471–95.

Mathews, John, I. and Curiel, F. (2019). Criminal justice debt problems. *Human Rights Magazine*, 44(3).

McFarlane, A. (2003). Rent stabilization and the long-run supply of housing. *Regional Science and Urban Economics*, 33(3):305 – 333.

Mello, S. (2021). Fines and financial wellbeing. Working Paper.

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.

Moon, C.-G. and Stotsky, J. G. (1993). The effect of rent control on housing quality change: A longitudinal analysis. *Journal of Political Economy*, 101(6):1114–1148.

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.

Mörk, E., Sjögren, A., and Svaleryd, H. (2014). Parental unemployment and child health. *CESifo Economic Studies*, 60(2):366–401.

Mueller-Smith, M. and Schnepel, K. (2021). Diversion in the Criminal Justice System. *The Review of Economic Studies*, 88(2).

Munch, J. R. and Svarer, M. (2002). Rent control and tenancy duration. *Journal of Urban Economics*, 52(3):542 – 560.

Nagy, J. (1995). Increased duration and sample attrition in new york city's rent controlled sector. *Journal of Urban Economics*, 38(2):127 – 137.

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., and McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54(4):387 – 398.

Northcutt, C. G., Athalye, A., and Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks.

of Economic Analysis, U. B. and of St. Louis, F. R. B. (2021). Per Capita Personal Income in Michigan [MIPCPI].

Olsen, E. O. (1972). An econometric analysis of rent control. *Journal of Political Economy*, 80(6):1081–1100.

Page, M., Stevens, A. H., and Lindo, J. (2009). *Parental Income Shocks and Outcomes of Disadvantaged Youth in the United States*, pages 213–36. University of Chicago Press.

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).

Pierson, K., Hand, M. L., and Thompson, F. (2015). The government finance database: A common resource for quantitative research in public financial analysis.

Pleggenkuhle, B. (2018). The financial cost of a criminal conviction: Context and consequences. *Criminal Justice and Behavior*, 45(1):121–45.

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.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188.

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.

Scheuren, F. 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.

Shapiro, J. (2014). As court fees rise, the poor are paying the price. *NPR*.

Sims, D. P. (2007). Out of control: What can we learn from the end of massachusetts rent control? *Journal of Urban Economics*, 61(1):129 – 151.

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

South, S. J., Haynie, D. L., and Bose, S. (2007). Student mobility and school dropout. *Social Science Research*, 36(1):68 – 94.

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

Steorts, R. C., 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.

Stuart, E. A., Lee, B. K., and Leacy, F. P. (2013). Prognostic score based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, 66(8, Supplement):S84 – S90.

Suen, W. (1989). Rationing and rent dissipation in the presence of heterogeneous individuals. *Journal of Political Economy*, 97(6):1384–1394.

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, pages 2473–2476, New York, NY, USA. ACM.

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. Technical Report NISTIT 8034, National Institute of Standards and Technology.

Wild, E. (2008). Driver responsibility fees: A five-year checkup. *State Notes: Topics of Legislative Interest*, July/August 2008.

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.

Wood, D., Halfon, N., Scarlata, D., Newacheck, P., and Nessim, S. (1993). Impact of Family Relocation on Children's Growth, Development, School Function, and Behavior. *JAMA*, 270(11):1334–1338.

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.