Tahamont, S., Jelveh, Z., Chalfin, A., Yan, S., & Hansen, B. (2020). Dude, where's my treatment effect? Errors in administrative data linking and the destruction of statistical power in randomized experiments. *Journal of Quantitative Criminology*. Advance online publication. https://doi.org/10.1007/s10940-020-09461-x

Dude, Where's My Treatment Effect? Errors in Administrative Data Linking and the Destruction of Statistical Power in Randomized Experiments

Sarah Tahamont*[1], Zubin Jelveh[2], Aaron Chalfin[3], Shi Yan[4], and Benjamin Hansen[5]

[1]Department of Criminology and Criminal Justice, University of Maryland
[2]University of Chicago Crime Lab
[3]Department of Criminology, University of Pennsylvania
[4]School of Criminology & Criminal Justice, Arizona State University
[5]Department of Economics, University of Oregon and NBER

April 28, 2020

Abstract

Objective: The increasing availability of large administrative datasets has led to an exciting innovation in criminal justice research — using administrative data to measure experimental outcomes in lieu of costly primary data collection. We demonstrate that this type of randomized experiment can have an unfortunate consequence: the destruction of statistical power. Combining experimental data with administrative records to track outcomes of interest typically requires linking datasets without a common identifier. In order to minimize mistaken linkages, researchers often use stringent linking rules like "exact matching" to ensure that speculative matches do not lead to errors in an analytic dataset. We show that this, seemingly conservative, approach leads to underpowered experiments, leaves real treatment effects undetected, and can therefore have profound implications for entire experimental literatures.

Methods: We derive an analytic result for the consequences of linking errors on statistical power and show how the problem varies across combinations of relevant inputs, including linking error rate, outcome density and sample size.

Results: Given that few experiments are overly well-powered, even small amounts of linking error can have considerable impact on Type II error rates. In contrast to exact matching, machine learning-based probabilistic matching algorithms allow researchers to recover a considerable share of the statistical power lost under stringent data-linking rules.

Conclusions: Our results demonstrate that probabilistic linking substantially outperforms stringent linking criteria. Failure to implement linking procedures designed to reduce linking errors can have dire consequences for subsequent analyses and, more broadly, for the viability of this type of experimental research.

Keywords: Randomized Experiments, Administrative Data, Record Linking, Machine Learning

*We extend our sincere thanks to Melissa McNeill at the University of Chicago Crime Lab for her work in developing the records matching algorithm employed in this paper. We would also like to thank Leslie Kellam, Ryang Hui Kim, Srivatsa Kothapally, Jens Ludwig, Jim Lynch, Mike Mueller-Smith, Aurelie Ouss, Greg Ridgeway, Jesse Rothstein and Greg Stoddard for helpful comments on this project. We thank Arnold Ventures for its generous support of the University of Chicago Crime Lab New York. Points of view or opinions contained within this document are those of the authors. They do not necessarily represent those of Arnold Ventures. Of course, all remaining errors are our own. Corresponding Author: Sarah Tahamont, Contact Information: Department of Criminology & Criminal Justice, 2220J Samuel J. LeFrak Hall, 7251 Preinkert Drive, College Park, MD 20742; Ph: 301-405-6474; Email: tahamont@umd.edu

1 Introduction

Despite several important limitations, experimental evidence continues to serve as the gold — or at the very least the bronze (Berk, 2005) — standard on the evidentiary hierarchy in the social and behavioral sciences (Banerjee and Duflo, 2009; Imbens, 2010; Weisburd, 2010).[1] The fact that random assignment, in expectation, creates comparable treatment and control groups and allows us to credibly ascribe causality to group-based differences is a feature that has been appreciated since at least the 18th century (Dunn, 1997) but which was popularized in the early 20th century by the work of Jerzy Neyman and R.A. Fisher (Fisher, 1936; Splawa-Neyman et al., 1923). Recently, the foundational importance of experiments has been bolstered by concerns about the limitations of observational research, with an array of evidence from econometrics and statistics suggesting that experimental treatment effects cannot be reliably recovered using some of the most ubiquitous types of observational research designs (LaLonde, 1986; Smith and Todd, 2001, 2005; Rubin, 2008; Gordon et al., 2019) even when analysts have a rich set of covariates upon which to condition (DiNardo and Pischke, 1997; Arceneaux et al., 2010). Reflecting these concerns, Weisburd (2010) has, channeling a sentiment found in McCord (2003) suggested that "whenever possible, evaluation studies should employ random assignment."

Within criminology, a number of reviews by Farrington and colleagues (Farrington, 1983, 2003; Farrington and Welsh, 2006) and Lum and Mazerolle (2014) among others have shown a general growth in experimental research since the 1950s. Over the last two decades, an appreciation for the importance of randomization has led to the formation of the Campbell Collaboration, the Academy of Experimental Criminology and ultimately the creation of the *Journal of Experimental Criminology*, a peer-reviewed journal dedicated to the publication of experimental evidence. That said, as Sherman et al. (1997) noted in the late 1990s, experimental evidence accounts for only a small share of the evidence base in empirical criminal justice research. To this day, experimental evidence remains uncommon in our field, making up just 3 percent of of the published studies in two leading journals, *Criminology* and *Justice Quarterly* (Dezember et al., 2020). Nevertheless, an appreciation

[1] For important reviews of the limitations of experimental research especially with respect to external validity, see Berk (2005); Deaton (2010); Heckman and Smith (1995) and Sampson (2010). Also see Nagin and Sampson (2019) for a wonderfully nuanced and equally important discussion of the inherent challenges in identifying a policy-relevant counterfactual in an experimental design. For reviews of the ethical and legal considerations that are attendant in randomized experiments, we refer readers to thoughtful reviews by Boruch et al. (2000) and Weisburd (2003).

for the power of randomization to generate valid causal estimates has inspired a host of highly influential randomized experiments in areas such as hot spots policing (Sherman and Weisburd, 1995; Braga et al., 1999; Braga and Bond, 2008), physical and social disorder reduction (Keizer et al., 2008; Branas et al., 2018; Ridgeway et al., 2019) and in evaluating social programs in domains such as youth delinquency (Powers and Witmer, 1951; Tremblay et al., 2003; Petrosino et al., 2003; Heller, 2014; Heller et al., 2017), prisoner re-entry (Duwe, 2012, 2014; Farabee et al., 2014; Cook et al., 2015), drug courts (Gottfredson et al., 2006; MacDonald et al., 2007; Rossman et al., 2011) and domestic violence (Sherman and Berk, 1984; Sherman et al., 1992; Davis and Taylor, 1997), among others. The appreciation for experimental evidence has likewise led to a prioritization of randomization among key funders of empirical research in criminology including the National Institute of Justice and Arnold Ventures.[2]

A central challenge to planning and successfully carrying out a randomized experiment is statistical power (Britt and Weisburd, 2010). Primary data collection is a costly endeavor and, as such, resource constraints typically limit the size of experiments and, as a result, the amount of data that are available to analyze. Since the power to detect a given treatment effect is, in principle, a function of the sample size, it is not uncommon for randomized experiments to be marginally- or even under-powered (Moher et al., 1994; Sedlmeier and Gigerenzer, 1989; Sherman, 2007).[3] Indeed, a common rule of thumb when planning an experiment is to fix statistical power at 80 percent, meaning that a researcher is willing to make a Type II error and fail to detect a real treatment effect 20 percent of the time (Cohen, 1992).

Given the the time and risk involved in planning an experiment, Type II errors tend to be extremely costly for researchers as well as partners in the policy world. Moreover, since null findings may be less likely to be published (Braga and Apel, 2016; Gerber and Malhotra, 2008; Rothstein, 2008), Type II errors can have dire consequences not only for individual papers but also for the entire research literature. Indeed this form of publication bias (filing papers in the desk drawer) is thought to have contributed to the "replication crisis" in empirical social science (Camerer et al., 2016; Gilbert et al., 2016; Pridemore

[2]Formerly known as the Laura and John Arnold Foundation.

[3]Statistical power is, in large part a function of the available sample size but also depends on the amount of variation in the treatment and outcome variables. As was noted by Weisburd et al. (1993) some twenty-five years ago and noted recently by Nelson et al. (2015), small N studies are not necessarily more poorly powered than larger N studies empirically though, other things equal, this will be the case.

et al., 2018; Vivalt, 2017).[4] Since experimental evaluations inform the evidence base about the effectiveness of programs and interventions, the impact of Type II errors can extend beyond the ivory tower, jeopardizing the viability of a diverse set of effective interventions (Doleac et al., 2020). Therefore, it is critical to preserve statistical power to the greatest extent possible.

One of the more exciting recent developments in empirical social science research is the increasing availability of large administrative databases, often referred to colloquially as "big data" (Lane, 2018; Lynch, 2018; O'Brien and Sampson, 2015; Smith et al., 2017) With respect to experimental research, big data has enabled a new (and oncoming) wave of "low-cost randomized trials," in which observations from an experimental intervention are linked to administrative data in order to minimize the need for costly primary data collection and keep the costs of experimentation tolerably low (Hyatt and Andersen, 2019).[5]

While this type of experiment is still relatively rare in criminology journals[6], administrative data linking has, nonetheless, been a feature of a number of recent and highly influential empirical criminology and criminal justice papers. Indeed administrative data linking is a mainstay of any experimental — or quasi-experimental — research that studies criminal justice contact as an outcome including recent research on the effects of pre-trial detention (Mueller-Smith, 2016; Dobbie et al., 2018), the deterrent effects of sanctions (Hansen, 2015; Lattimore et al., 2016), life-course criminology studies (Farrington, 2006; Liberman et al., 2014; Loeffler, 2013; Sampson and Laub, 2003; Stewart et al., 2015; Van Schellen et al., 2012), research on the perpetrator-victim relationship (Broidy et al., 2006), prisoner re-entry programs (Cook et al., 2015; Duwe, 2012, 2014; Farabee et al., 2014) and interventions that are designed to affect the behavior of delinquent or at-risk youth (Tremblay et al., 2003; Petrosino et al., 2003; Heller, 2014; Heller et al., 2017). Administrative data linking is particularly valuable when researchers are interested in studying the impact of a program on multiple domains — each of which draw on data from different administrative

[4]Concerns over the misuse of researcher degrees of freedom and specification searching have likewise spurred recommendations which include the use of very small α levels (Benjamin et al., 2018), which increases the probability of Type II errors even more.

[5]A second advantage of administrative data is that it avoids the inherent challenges involved in working with self reported data (Bertrand and Mullainathan, 2001), which is not to minimize the fact that there are certainly trade-offs to using administrative data relative to self-reports. For empirical evaluations of the validity of self-reported data see: Lauritsen (1999), Morris and Slocum (2010) and Roberts and Wells (2010) among others.

[6]From 2017-present, approximately 16% of the field experiments published in *Criminology, Journal of Quantitative Criminology, Journal of Experimental Criminology, Journal of Research in Crime and Delinquency* and *Justice Quarterly* have taken the general form of a "low-cost" RCT.

agencies. For example, data linking is used by Heller et al. (2017) and Heller (2014) to link experimental subjects across multiple domains — in order to study the effects of the experimental intervention on both criminal justice contacts and education outcomes. Similarly, Gelber et al. (2016) use administrative data linking to jointly study the impact of summer employment for youth on imprisonment as well as mortality outcomes. Administrative data linking is likewise an issue that looms large in studying the intergenerational effects of delinquency from parents to children (Chalfin and Deza, 2017; Comfort et al., 2011; Dobbie et al., 2018; Hjalmarsson and Lindquist, 2012; Laub and Sampson, 1988; Wildeman and Andersen, 2017).

In this paper, we point out that linking a sample of interest to administrative data, which greatly reduces the cost of experimentation by relying on existing data sources in lieu of primary data collection can have an unfortunate consequence: the destruction of statistical power. Most administrative datasets are not designed to be linked to others, and therefore do not have a common and reliable identifier. Likewise, many individual-level identifiers, such as fingerprints in law enforcement or patient identifiers in hospitals, are internal, system-specific and, in many cases, unknown to outside researchers and agencies. Even when unique identifiers are, in principle, available across systems (e.g. a social security number) they can have reliability problems due to errors in recording or non-reporting (Curb et al., 1985; Sampson and Laub, 2003).[7] As a consequence, researchers often have to link individuals across datasets in the absence of a unique identifier (Hovde Lyngstad and Skardhamar, 2011; Taxman and Caudy, 2015). In the vast majority of cases, researchers have to try to link individuals using their demographic characteristics, such as name, date of birth, race, gender, and address. This process is often difficult and fraught with error; particularly with respect to criminal justice, because the nature of the data makes them especially susceptible to recording errors (Ferrante, 1993; Geerken, 1994; Orchowsky and Iwama, 2009). In this paper, we show that data linking errors can have dire consequences for statistical power and the validity of downstream estimates from analyses from linked data, because bad data linkages introduce noise that obscures the relationship between the outcome variable and the treatment variable (Enamorado et al., 2019). In the presence of noisy data linking, even a perfectly-executed randomized controlled trial will fail to deliver

[7] When a unique identifier is available in all of the datasets that require linking and the data are of sufficient quality, linking can, in some cases, be fairly trivial. These types of cross-system unique identifiers are frequently available in Scandinavian countries (e.g. Black et al., 2005; Dahl et al., 2014; Hovde Lyngstad and Skardhamar, 2011; Wildeman and Andersen, 2017) and occationally in South America.

an unbiased estimate of the effectiveness of an intervention.

We show analytically that linking errors will attenuate estimates of the treatment effect in a randomized experiment — or, notably, in a quasi-experiment of equivalent sample size. Critically though, while attenuation is an unwanted outcome, the most insidious result of linking errors are the effects on ex-ante statistical power and, therefore, on a researcher's ability to detect a true treatment effect when one exists (Type II errors). While modest linking errors will lead to modest attenuation, given that few experiments are *overly* well-powered, even small amounts of linking error can have large effects on Type II error rates.

Linking errors are particularly problematic for rare outcomes (e.g. Gelber et al., 2016) and small sample sizes (e.g. Fischbacher et al., 2001), both of which are common and even typical in randomized experiments. Researchers who conduct ex-ante power analyses, and subsequently link to administrative data to detect the presence of their outcome of interest, will therefore overestimate their power to detect effects, often by a considerable margin. A particularly salient example of this problem can be found in a replication paper by Doleac et al. (2020) who note that a large number of experiments in the area of prisoner re-entry programs are underpowered, thus leading to lost opportunities as well as to potentially misleading conclusions about the viability of interventions for formerly-incarcerated or justice involved individuals.

How can this destructive problem be abated? Naturally, the solution lies in minimizing linking errors. Despite the existence of a number of proposed data linking methods (Enamorado et al., 2019; Lahiri and Larsen, 2005; Scheuren and Winkler, 1993, 1997), we observe that researchers in practice often use "exact matching" criteria when linking datasets, considering an individual to be a match only if his or her demographic identifiers (name and date of birth, etc.) match *exactly* across datasets (e.g., Cesarini et al., 2016; Hill, 2017; Khwaja and Mian, 2005; Mueller-Smith, 2016). The rationale for researchers to impose such a stringent linking criterion is to minimize false positive links and to ensure that speculative links do not lead to errors in the analytic data set. This popular approach to administrative data linking is therefore regarded as conservative, as it keeps the analytic data as "pure" as possible.[8] Although such an argument appears logical at first glance,

[8] A large literature considers the implications that measurement error can have for econometric models but, to our knowledge, there is considerably less formal guidance with respect to how bad data linking can confound randomized experiments. It is also worth noting that when scholars need to link datasets without a common identifier there is no "ground truth" to assess the quality of the match. Likewise, there is often no prior about what the match rate should be, rendering it difficult to diagnose whether the matching

stringent character match requirements will, by definition, increase false negative link rates (Enamorado et al., 2019). Therefore, "exact matching" can lead to a higher *total error rate*—the sum of false positive and false negative links—and, critically, to a reduction in statistical power.

The paper proceeds as follows. First, we briefly review the literature on linking errors in empirical social science. Next, we derive an analytic result for the consequences of linking error on treatment effect estimation and show that the sum of false positive and false negative error rates is what matters, not either of these error rates in isolation. We then present numerical estimates to demonstrate how the attenuation of treatment effects and the corresponding erosion of statistical power varies across different combinations of relevant inputs including the 1) the sum of false positive and false negative error rates, 2) the outcome density, and 3) the sample size. We proceed with an empirical example that shows the difference between exact matching strategies and data linking using probabilistic matching algorithms augmented by machine learning.

Our results suggest that the problem of bad data linkage is not trivial in many empirical applications. Even in a relatively large experiment ($n = 750$) with a relatively dense outcome ($\bar{y} = 0.5$), the Type II error rate — the failure to detect a true effect of an experimental intervention — would be twice as high (0.4 instead of 0.2) in the presence of a realistic amount of linking error. In other words, researchers who believe that they will fail to detect a false null hypothesis only 20 percent of the time will, in fact, fail to detect it 40 percent of the time. Given the generally high costs of experimentation, a Type II error rate of this magnitude is a clear and present danger to the feasibility of a great deal of experimental investigations, because the beneficial effects of promising interventions may go undetected. Not limited to randomized experiments, these results apply just as strongly for quasi-experimental research designs which seek to mimic randomization by using a natural experiment or a control function approach.

That said, we conclude on an optimistic note, observing that researchers can mitigate the consequences of linking errors in most cases using a simple machine-based probabilistic linking algorithm — including those that are available in most commercial software packages. In a majority of applications, more than half of linking errors introduced by using exact matching can be "overturned" with the use of a simple probabilistic linking algorithm,

procedure employed is sufficient or not.

thus preserving a large share of statistical power that is lost due to data linking issues and thus maintaining the long-term viability of experimental and quasi-experimental evidence that capitalize on administrative data linking.

2 Motivation and Context

2.1 Methodological Context for Administrative Data Linking

In this section, we provide a high-level conceptual overview of the essentials of record linking in social sciences. Our purpose is not to recommend the best algorithm or package, though we note that recent scholarship offers helpful recommendations (Enamorado et al., 2019; Karr et al., 2019). Scholarly consideration of the implications of data linking dates back to the early days of computerized record linkage itself (Fellegi and Sunter, 1969; Neter et al., 1965; Newcombe et al., 1959). Researchers from multiple disciplines have recognized that linking errors have implications for subsequent analyses (Berent et al., 2016; Campbell, 2009; Khwaja and Mian, 2005; Lahiri and Larsen, 2005; Neter et al., 1965; Scheuren and Winkler, 1993, 1997). At the inception of computerized data linkage, the key conclusion from Neter and colleagues was that "the consequences of even small mismatch rates can be considerable" (Neter et al., 1965, p. 1021). Moreover, researchers have recognized that linking errors, as a special case of classical measurement error, can lead to downward bias in effect size estimates (Aigner, 1973; Campbell, 2009; Khwaja and Mian, 2005). For example, Khwaja and Mian (2005, p. 1379, emphasis original) cautioned the readers that "when [their] algorithm matches a firm to a politician, but the match is incorrect ... estimates of political corruption are likely to be *underestimates* of the true effect." However, until recently, scholars working with empirical applications have devoted relatively little attention to describing the techniques used to link data, to evaluating the quality of the matches in linked data, or to determining how their study conclusions might have varied based on the use of different data linking strategies.

Record linking refers broadly to the practice of identifying records from different datasets that correspond to the same individual.[9] In some circumstances, linkages can be generated using biometric indicators, such as fingerprints or fingerprint-associated ID numbers (see

[9]For narrative clarity, we limit our discussion to the linking of data containing records on persons. This discussion would extend to groups or firms, but the characteristics available for linking might be different.

Freudenberg et al., 1998; Taxman and Caudy, 2015).[10] However, in the vast majority of cases, databases to which experimental subjects are matched often do not share a unique identifier. As a result, researchers have to rely on available demographic information, such as name (including first and last names, and sometimes a middle name or middle initial), date of birth, social security number, gender, race, and ethnicity. The linking enterprise is necessarily imperfect for a variety of reasons, including typographical errors, nicknames, changes in names, the commonness of certain names, geographic mobility and sometimes due to the intentional provision of inaccurate information (Christen, 2012; Harron et al., 2017).

There are two primary approaches to automated record linkage (Enamorado et al., 2019; Harron et al., 2017; Winkler, 2006). The first, deterministic matching, refers to the practice of developing a set of criteria a priori, and considering two records to be a match if and only if the set criteria are met. The strictest approach of this kind is "exact matching," under which a link is recognized only if all variables used are identical across the two records (which often includes letter-to-letter matches of names). Less stringent deterministic linking rules require only a subset of letters or digits to match, such as the first three letters of names, or the month and date digits of dates of birth.

A second approach is probabilistic linking. Instead of directly defining deterministic rules manually which lead to binary "link/no link" determinations, probabilistic linking seeks to estimate the probability that two records belong to the same individual (sometimes known as "fuzzy linking," Bowers and Johnson, 2005; Sampson and Winter, 2018).[11] The canonical approach to probabilistic linking — developed by Newcombe et al. (1959) and formalized by Fellegi and Sunter (1969) — relies on an underlying theoretical model to generate the probability of observing the set of paired characteristics given that two records refer to the same person (and vice versa).[12] The ratio of these two probabilities can then be used to define whether a pair of records is a link, nonlink, or requires further manual review.

[10] We acknowledge that biometric data are susceptible to misidentification as well. However, the literature generally considers linking using biometric indicators as more accurate than the text-based demographic identifiers that we discuss below (Watson et al., 2014).

[11] Operationally, however, the end result of most probabilistic linking processes requires the imposition of a deterministic threshold to define potential pairs as links, non-links or, in some cases, potential links.

[12] In the Fellegi-Sunter framework, pairs of records are compared across their identifying characteristics (name, date of birth, gender, etc) and a comparison vector is computed which encodes whether, for example, the names in the two records are the same, the name in one record is missing, the date of births are the same, and so on. Other extensions to this framework include string distance calculations between names (e.g. levenshtein, jarowinkler, etc.) or phonetic computation (e.g., Soundex, Double Metaphone, etc.).

In general, comparisons of matching algorithms have found that probabilistic matching algorithms have higher overall accuracy rates than deterministic rules (Campbell, 2009; Campbell et al., 2008; Gomatam et al., 2002; Tromp et al., 2011), or, at a minimum, have similar accuracy rates to deterministic rules (Clark and Hahn, 1995).

There are two types of linking errors: 1) false positives (i.e., treating records as the same person when they actually belong to different people) and 2) false negatives (i.e., treating records as different people when they actually belong to the same person). Not surprisingly, there is an inherent trade-off between false positives and false negatives (Christen and Goiser, 2007; Zingmond et al., 2004; Moore et al., 2014): raising the stringency of the criteria used to delineate a link will reduce the number of false positive links at the expense of increasing the number of false negative links (and vice versa with less stringent criteria).

Perhaps the sole existing measure of the quality of a data linking procedure is the link rate, commonly referred to as the match rate, which we define here as the percentage of records which are matched after the linking process. Link rates are most relevant when there is a prior expectation of the proportion of the records that should be linked. In general, higher link rates are seen as indicating higher quality matches (Bailey et al., 2017; Ferrante, 1993), especially when the expectation is that all records in one datset should be linked to records in another dataset.

2.2 Current Setup

Linking errors can take on many forms depending on the record linkage problem at hand. In this paper, we study a common but underexplored scenario: when the goal is to estimate the effect of an intervention by linking datasets to measure outcomes such as whether an individual was arrested, failed to graduate high school, or visited an emergency room.[13] For ease of exposition, we concentrate on the scenario where the outcome variable is binary. Our findings are particularly relevant to criminal justice contacts, they also apply to a vast array of settings throughout the social sciences because binary outcomes are extremely common.[14] We further note that in the context of planning for a randomized experiment,

[13] In particular, the aim is to estimate the difference, possibly conditioned on covariates, in means between treatment and control groups in a randomized control trial. In a related paper, Moore et al. (2014) explore the impact of matching errors on the relative risk ratio. Matching errors bias these two quantities in different ways. As we show below, false positive and false negative rates have equal impact on bias in our scenario. Moore et al. show that false positive rates are more influential on the bias in relative risk ratio estimates.

[14] Recent specific examples from throughout the social sciences include program participation in Supplemental Nutritional Assistance Program (Courtemanche et al., 2018), employment prevalence measured

conducting pre-hoc power analyses on a binary dependent variable is common practice as, in the absence of available microdata upon which statistical power can be simulated, the variance of a binary variable can be recovered simply by knowing the mean of the variable.

Another important feature of our scenario where the linking process determines the observed value of the outcome is that link rates are no longer informative for assessing linkage quality. To see why, consider a scenario in which a researcher is interested in evaluating whether a job training program for those individuals recently released from prison reduces the likelihood of an arrest. In constructing an analytic data set, individuals whose records link to the post-intervention arrest file are considered "arrested" and individuals whose records do not link to the arrest file are considered "not arrested." There is no prior expectation that all individuals will be arrested. Additionally, while a researcher could in principle use historical statistics on recidivism rates to benchmark the expected link rate, this would only serve as a rough guide since the goal of the experiment is detect *changes* in that rate. Not limited to arrest, this issue is applicable to any context in which the goal of the linking process is to determine the presence of an outcome and there is no prior prediction for how many records *should* match, or that changes relative to the prior prediction are the quantity of interest. These points distinguish this linking case from other kinds of longitudinal record linking cases (Abramitzky et al., 2019; Bailey et al., 2017; Feigenbaum, 2016) but, to our knowledge, this distinction has not been discussed in the prior literature on administrative data linking.

3 Derivation of Estimated Treatment Effects, Standard Errors and Statistical Power

In this section we derive the effects of matching errors on the estimated treatment effect, $\hat{\tau}$, as well as its standard error, $se(\hat{\tau})$, in a randomized experiment with a binary treatment condition. For the sake of simplicity, we assume, for now, that the effect of treatment is homogeneous though all subsequent results will hold under heterogeneous treatment effects, so long as false linkage rates and treatment effects do not both vary across subgroups in

through unemployment insurance wage records (Johnston and Mas, forthcoming), injuries measured through hospitalization data (Powell and Seabury, forthcoming), or financial health measured through bankruptcy or liens (Dobkin et al., 2018).

the data.[15] The results derived in this section will also hold if false linkage rates vary by sub-group under homogeneous treatment effects. While there is quite a bit of technical detail in this section, we hope the benefit of including these details is the satisfaction that comes from a closed form analytic solution. We show that the effects of linking errors on both quantities have a closed form solution. The estimated $\hat{\tau}$ will be attenuated and the degree of attenuation will be proportional to the sum of the false positive and false negative linking error rates. The effect of linking errors on $se(\hat{\tau})$ is more complicated and may result in an increase or decrease in the estimate of the standard error relative to the no linking error scenario. In general, relative to the effect on coefficient estimates, standard errors are not very sensitive to linking errors and, as a result, linking errors will *always* lead to a higher rate of failing to reject a false null hypothesis, and in so doing, lead to a higher likelihood of failing to detect a true treatment effect. As we show, linking errors can be very detrimental to statistical power in all but the largest randomized experiments. It is worth noting that we focus on experiments here, not because they are the only design affected by linking errors, but because we are able to analytically demonstrate the consequences of linking error in the unconfounded case. This is not to suggest that linking errors will not have ramifications for analyses which are not cleanly identified, just that the precise form of those ramifications is currently unknown.

3.1 Estimated Treatment Effect

We begin by showing that, in a randomized experiment, linking errors lead to attenuated estimates of an average causal effect in absolute terms. Consider a randomized control trial with a study sample of n units of which a fraction, p, are assigned to treatment and the remaining $(1-p)$ are assigned to a control condition. Information on this experimental sample is stored in a dataset, E. We are interested in estimating the average treatment effect of our intervention on an outcome, y. In this case, the outcome measure is stored in an administrative dataset, D, where the number of individuals in D is much larger than n. In order to estimate the average treatment effect, we need to match our experimental sample to the outcomes stored in administrative dataset. If a record in E links to a record in D then observed $y = 1$ and otherwise observed $y = 0$. The realized outcome for an individual

[15] One scenario where this assumption would not hold is if *both* linking errors *and* treatment effects vary by one or more subgroups. In the event that both treatment effects and false linkage rates both vary by subgroup, the solution is slightly more complex and is explored in Appendix E.

in E is given by the potential outcomes corresponding to the treatment condition:

$$y_i(T_i) \begin{cases} y_i(0) & \text{if } T_i = 0, \\ y_i(1) & \text{if } T_i = 1 \end{cases} \tag{1}$$

As such, the average treatment effect of the intervention, τ can be computed as:

$$\tau = \mathbb{E}[y_i(1) - y_i(0)] = P(y_i = 1|T_i = 1) - P(y_i = 1|T_i = 0) \tag{2}$$

where T is a treatment indicator. In practice, once we introduce the idea of error we have to consider both the true outcome which we denote using y^* and the observed outcome which we refer to using y. The process of linking the experimental data to the outcome data can lead to two types of errors:

- False positive link (FP): An instance in which an individual i has the true outcome $y_i^* = 0$, but was incorrectly linked to some other individual's record in D with $y_j^* = 1$ where $i \neq j$, such that in this case, the observed value of y after the linking process is equal to 1.

- False negative link (FN): An instance in which an individual i has the true outcome $y_i^* = 1$, but was not linked to a record in D. In this case, the observed value after the linking process is $y = 0$.

It is important to note that since we are trying to match observations in E to a given outcome in D, linking errors are only driven by whether the correct outcome is observed, and not that the link refers to the same person in both datasets. Specifically, this means that it would not be considered an error with respect to measuring the outcome if a record in E is linked to the wrong person in D, provided the linked record had the same outcome value as the true match. When records are matched to erroneous outcomes this will lead to the following biased estimate of τ:

$$\hat{\tau} = P(y_i = 1|T_i = 1) - P(y_i = 1|T_i = 0) \tag{3}$$

To further characterize the nature of the bias in τ we introduce the following four definitions:

- True Positive Rate (TPR): $P(y_i = 1|y_i^* = 1)$, or the probability that an individual

with outcome $y = 1$ will be linked to an individual in D yielding an observed outcome $y^* = y = 1$.

- True Negative Rate (TNR): $P(y_i = 0|y_i^* = 0)$, or the probability that an individual with outcome $y = 0$ will not be linked to an individual in D yielding an observed outcome $y^* = y = 0$.

- False Negative Rate (FNR): $P(y_i = 0|y_i^* = 1)$, or the probability that an individual with outcome $y^* = 1$ will not be linked to an individual in D, yielding an observed outcome $y^* \neq y$. FNR is equivalent to 1 - TPR.

- False Positive Rate (FPR): $P(y_i = 1|y_i^* = 0)$, or the probability that an individual with outcome $y^* = 0$ will be incorrectly linked to an individual in D, resulting in an observed outcome $y^* \neq y$. FPR is equivalent to 1 - TNR.

Then the observed, and potentially biased, treatment effect can be written as:

$$\begin{aligned}
\hat{\tau} &= P(y_i = 1|T_i = 1) - P(y_i = 1|T_i = 0) \\
&= \sum_{j \in \{0,1\}} P(y_i = 1, y_i^* = j|T_i = 1) - \sum_{j \in \{0,1\}} P(y_i = 1, y_i^* = j|T_i = 0) \\
&= \sum_{j \in \{0,1\}} P(y_i = 1|y_i^* = j, T_i = 1) P(y_i^* = j|T_i = 1) \\
&- \sum_{j \in \{0,1\}} P(y_i = 1|y_i^* = j, T_i = 0) P(y_i^* = j|T_i = 0) \\
&= TPR_T \; P(y_i^* = 1|T_i = 1) \; - \; TPR_C \; P(y_i^* = 1|T_i = 0) \\
&+ FPR_T \; P(y_i^* = 0|T_i = 1) \; - \; FPR_C \; P(y_i^* = 0|T_i = 0)
\end{aligned} \tag{4}$$

TPR_T and TPR_C are the true positive rates for the treatment and control groups, respectively. Similarly, the false positive rate for the treatment and control groups are FPR_T and FPR_C. In the case where linking error rates are equivalent for both treatment and control groups, as is expected under random assignment, we let $TPR_T = TPR_C$ and $FPR_T = FPR_C$. We can then re-write the expression more compactly:

$$\begin{aligned}
\hat{\tau} = TPR\,[P(y_i^* = 1|T_i = 1) - P(y_i^* = 1|T_i = 0)] \\
+ FPR\,[P(y_i^* = 0|T_i = 1) - P(y_i^* = 0|T_i = 0)]
\end{aligned} \tag{5}$$

which can, in turn, be written as:

$$\begin{aligned}\hat{\tau} = & \, TPR\,[P(y_i^* = 1|T_i = 1) - P(y_i^* = 1|T_i = 0)] \\ & - FPR\,[P(y_i^* = 1|T_i = 1) - P(y_i^* = 1|T_i = 0)]\end{aligned} \tag{6}$$

The bracketed term in (6) is simply τ, the true treatment effect, which leads to the following final form in the case of equivalent matching error across treatment and control:

$$\hat{\tau} = (TPR - FPR)\ \tau \tag{7}$$

We note that if the error rates were known, the true treatment effect could be recaptured:

$$\tau = \frac{\hat{\tau}}{TPR - FPR} \tag{8}$$

Non-zero linking error will always attenuate the absolute value of the true treatment effect.[16] Finally, we can re-write the denominator as $1 - (FNR + FPR)$ and generate two critical insights. First, bias will be proportional to the total matching error rate. The finding that the sum of false positive and false negative error rates drives the bias is particularly important given the tendency toward "exact matching," which is thought to minimize error, but, in fact, reduces the number of false positive links while increasing the number of false negative links. Second, under reasonable assumptions on the magnitude of the error rates (i.e. when $FNR + FPR < 1$), $\hat{\tau}$ will be attenuated towards zero — that is, the estimated treatment effect will be too small.

3.2 Estimated Standard Errors

In Section 3.1 we showed that linking error leads to an attenuated estimate of the average treatment effect and we further posited that bias introduced by linking errors will reduce statistical power. However, in order to draw conclusions about the effect of linking errors on statistical power, we must also consider the effect of linking error on estimated standard errors. To see how linking error affects σ_τ, note that the variance of τ is given by:

$$\sigma_\tau^2 = \frac{1}{p(1-p)}\frac{\sigma^2}{N} \tag{9}$$

[16] If $TPR = FPR$ then the previous equation is undefined and the observed treatment effect will equal zero, but that situation is unlikely to occur in practice as it implies a random match.

where p is the proportion of the study sample enrolled in treatment, N is the sample size, and σ^2 is the residual outcome variance from a regression of y on the treatment indicator, T. Taking the square root of the quantity on the right-hand side of (9) yields the estimated standard error around τ.

The only remaining step is to estimate the residual variance. We note that in the case of linear regression, σ^2 can be defined via the residual sum of squares, and, with a binary outcome and a binary treatment, results in the following form where y_T and y_C are the number of individuals in the experimental group linked to records in the administrative data for the treatment and control group, respectively (see derivation in Appendix B).

$$\sum_i (y_i - \hat{y}_i)^2 = y_T \left(1 - \frac{y_T}{N_T}\right) + y_C \left(1 - \frac{y_C}{N_C}\right) \tag{10}$$

While attenuation in the treatment effect depends only on the the sum of false positive and false negative error rates, linking error affects the standard errors through the control group mean, the treatment effect, and the distribution of false positive and false negative links. To see how the distribution of linking error types affects the standard errors, consider a scenario where there is no treatment effect. When the false positive rate is greater than the false negative rate, the number of instances where $y = 1$ will increase and the outcome density will also increase. Conversely, when the false negative rate is higher the number of instances where $y = 0$ will increase and the outcome density will decrease. The outcome density that maximizes the variance is $\bar{y} = 0.5$. Whether the standard errors increase or decrease depends on the extent to which the errors move the outcome density toward or away from 0.5. For example, if the overall mean is 0.4, but the matching algorithm produces more false negatives than false positives, then the observed treatment group mean will be less than 0.4 and the resulting standard error will shrink. The situation is slightly more complicated when there is a treatment effect, but we show in Appendix C that Equation 10 is maximized when the control group mean plus the treatment effect equal 0.5.

The interplay between these factors means that there will be scenarios in which matching error will produce *smaller* standard errors when compared with the no error case. But in the next section we show that even in these situations, matching error compromises a researcher's ability to detect a true treatment effect.

3.3 Implications for Statistical Power

While attenuation of coefficients can be troublesome, the effect of matching errors on statistical power is a far greater concern. Due to resource constraints, few randomized experiments are overpowered, so modest matching errors can have an outsize effect on statistical power.[17] We begin by noting that since there is a closed form solution for the effect of matching errors on the estimated average treatment effect and its standard error, there is also a closed form solution for the effect of matching errors on statistical power $(1-\beta)$. To see this, consider that, for a given Type I error rate (α) and a standard error around the average treatment effect, the probability of a Type II error is given by:

$$\beta = \Phi\left[-\Phi^{-1}\left(\frac{\alpha}{2}\right) - \frac{\tau_h}{\sigma_{\tau_h}}\right] \tag{11}$$

where τ_h is the hypothesized treatment effect, σ_{τ_h} is the standard error, and Φ is the cumulative distribution function for the normal distribution.[18] One minus this quantity is statistical power. If the following condition holds, then power will always be lower under matching error:

$$\frac{\tau_h}{\sigma_{\tau_h}} > \frac{\hat{\tau}}{\sigma_{\hat{\tau}}} \tag{12}$$

Since the true treatment effect will be adjusted according to $1-(FNR+FPR)$, Equation 12 can be re-written as:

$$\begin{aligned} \frac{\tau_h}{\sigma_{\tau_h}} &> \frac{\big[1-(FNR+FPR)\big]\tau_h}{\sigma_{\hat{\tau}}} \\ \sigma_{\hat{\tau}} &> \big[1-(FNR+FPR)\big]\ \sigma_{\tau_h} \end{aligned}$$

As we discussed in the previous section, there will be situations in which $\sigma_{\hat{\tau}} < \sigma_{\tau_h}$, but in Appendix D we show that even in these situations the shrinkage in the standard errors is never enough to offset the consequences of coefficient attenuation, and, therefore, statistical power *always* decreases under linking error.

In the next section we show that even with modest linking errors, there can be large

[17] Ioannidis et al. (2017) show that the median statistical power for a large body of studies in economics, most of them observational, is just 18%.

[18] Here, τ_h refers to the candidate treatment effect for which statistical power will be computed. For smaller samples, Φ would be replaced by the cumulative distribution function for the t distribution.

declines in statistical power. Since, in the context of a randomized experiment, researchers tend to set $1-\beta$ on the basis of their relative tolerance for the risk of an underpowered finding, the result is that researchers will undertake randomized experiments that are underpowered relative to their desired power thresholds.

4 Analytic Results

In order to provide a sense for the degree to which linking errors lead to attenuation in experimental estimates, incorrect standard errors, and corresponding declines in statistical power, we compute the Type II error rate over a range of reasonable parameter values. We focus specifically on the outcome density for the control group $\bar{y}_C$, the hypothesized treatment effect τ_h, the sample size N and the linking error rates. Our goal here is to demonstrate the dynamics of this problem and the contexts in applied research for which it is likely to be especially pernicious.

4.1 Setup

In order to explore the effect of linking errors under a range of different parameterizations, using the analytic results in Section 3, we derive closed form solutions for τ, $se(\tau)$ and, ultimately, the Type II error rate, β, in two potential scenarios: one in which there are no linking errors and another in which linking errors are present. While it is the sum of false positive and false negative error rates (FPR+FNR) that dictates the degree of attenuation in $\hat{\tau}$, as we have shown, the extent to which linking errors affect the standard error around this estimate and, relatedly statistical power, will also depend on the ratio of false positive to false negative match rates. We motivate our setup using a dichotomous outcome, y and a binary treatment, T where, as before, p is the proportion of the sample that is treated and the remaining $1-p$ are untreated.[19]

[Insert Figure 1 About Here]

4.2 Main Results

Figure 1 contains four panels that report ex ante power calculations with and without matching errors, corresponding to four control mean-effect size combinations ($\bar{y}_C^* = 0.3$,

[19] The computational details of this exercise are described in Appendix A to this paper.

0.5 and $\tau_h = 15\%, 25\%$) that are typical of power calculations in planning a randomized experiment. In each panel, the total linking error rate — that is the sum of the false negative and false positive link rates — is plotted on the X-axis while the Type II error rate (β) is plotted on the Y-axis. The lines plot Type II error rates for a given sample size, N.

We begin our discussion with Panel (a) which corresponds with $\bar{y}_C^* = 0.5$ and $\tau_h = 25\%$, the parameterization which is best powered for a given sample size. Consider, for example, a very large experiment in which $N = 2{,}000$. In such an experiment, a Type II error will be extraordinarily rare — approximately zero — in the absence of linking errors. Even when the linking error rate is as high as 30%, the probability of a Type II error will be approximately 3%, meaning that such an experiment will have a 97% probability of detecting a treatment effect of 25%. This is sensible as linking errors have little effect on statistical power when an experiment is extremely overpowered. However, due to resource constraints, overpowered experiments are rare. A more realistic scenario is an experiment in which $N = 500$. This sample size corresponds with the solid, red line in Panel (a). In the absence of linking errors, this study has a Type II error rate of approximately 20% which is considered by many researchers to be a reasonable default rule in conducting ex ante statistical power calculations (Moher et al., 1994). Under even a relatively modest linking error rate of 10%, Type II error rates rise to approximately 28%; with a 20% linking error rate, the probability of a Type II error nearly doubles to 39%.

Another way to understand the impact of linking errors is to consider how much larger the study would have to be to maintain a given Type II error rate, β. This too can be seen in Figure 1. Referring to Panel (a), consider a study of size, $N = 500$ which has a Type II error rate of approximately 20% in the absence of linking errors but a 38% Type II error rate under 20% linking error. Here, it would take a 50% increase in the size of the study (from $N = 500$ to 750) to return to the desired Type II error rate of 20%. As resource constraints are often binding, increasing the size of a study by 50% is most often infeasible.

The effects of linking error on statistical power are even more dramatic with a less dense outcome and a smaller treatment effect of interest. In Panel (b), $\bar{y}_C^*$ is fixed at 0.5 but now we are interested in being able to detect a smaller treatment effect, $\tau_h = 15\%$. Now, even in the absence of linking error, we will need a larger sample size to detect a treatment effect of this magnitude (e.g. for $N = 500$, the Type II error rate at zero linking error is greater than 60%). Focusing on the sample size ($N = 1{,}500$) that roughly yields the default Type

II error rate of 20% in the absence of linking errors. In this case, we see that when the sum of false positive and false negative errors is at a reasonable level (15%) Type II error rates will increase by approximately 50%, from 20% to around 30%. We see a similar relationship when the treatment effect of interest is 25% but the outcome is less dense (Panel c). Finally, we turn to Panel (d) in which we have both a less dense outcome $\bar{y}_C^* = 0.3$ and a smaller treatment effect of interest 15%. Here, even a very large experiment will sometimes fail to detect a true treatment effect as the Type II error rate for a study of size $N = 3{,}000$ is approximately 25% in the absence of matching errors. In this case, a reasonable matching error rate of 15%, takes the Type II error rate to 40%.

4.3 Allowing for Covariate Adjustment

The results reported in Section 4.2 presume that researchers do not have access to or, at least, do not use pre-test covariates in estimating $\hat{\tau}$. While a healthy debate exists about the wisdom of controlling for covariates in a finite sample, it is common empirical practice in analyzing randomized experiments to condition on covariates and estimate an average treatment effect by regressing y on both T and a vector of covariates, X (Angrist and Pischke, 2009; Duflo et al., 2007). The wisdom behind controlling for covariates is straightforward. Given that the treatment is randomized, X will be unrelated to T but may be helpful in explaining y. The result is that residual variation will shrink and so too will estimated standard errors. Thus, controlling for covariates will increase a researcher's power to detect treatment effects and, in expectation, will not bias the estimated treatment effect. Given that the primary purpose of controlling for covariates in an experimental setup is to increase statistical power, a natural question is whether doing so has implications for the effect of linking errors on statistical power.

In order to answer this question, we generate a covariate, X, that is correlated with y_C^* but which, by construction, is uncorrelated with T. For simplicity, we generate a dichotomous X which is found in equal proportions in the treatment and control groups (though all of the analytic results will also hold in the case in which X is continuous). The setup is the same as before with the exception that we specify an imbalance parameter, r, which governs the strength of the relationship between y_C^* and X. Specifically, r is difference in the proportion of the sample for which $y^* = 1$ when $X = 0$ and when $X = 1$. In other words, r represents the amount of imbalance in the outcome density between individuals

who possess characteristic X and those who do not. For example, if $\bar{y}_C^*$ is 0.5, when $r = 0.1$, $\bar{y}_C^* = 0.4$ for the $X = 1$ group and 0.6 for $X = 0$ group, or vice versa. When r is large, y_C^* and X will be highly correlated and standard errors shrink by a relatively large amount. In the demonstration below, we fix r =0.1. However, the choice of r does not have a first order effect on the extent to which linking errors lead to Type II errors.[20] We present findings in Figure 2 in which Panels (a)-(d) correspond with the same parameterizations shown in Figure 1.

[Insert Figure 2 About Here]

Referring to Panel (a) in which $\bar{y}_C^* = 0.5$ and $\tau_h = 25\%$, we see that, compared to Figure 1, the y-intercept has shifted downwards reducing the probability of Type II error when a covariate is added to the model. Without matching error, a sample of $N = 500$ yielded a Type II error rate of approximately 20% in the absence of a covariate, conditioning on a reasonably predictive covariate reduces the Type II error rate to just over 15%. In the case of this marginally powered sample ($N = 500$), a reasonable error rate of 15% doubles the Type II error. Referring to Panel (b) where the researcher would like to detect a treatment effect of 15%, we see that the consequences of linking errors continue to be severe in the presence of a covariate with Type II error rates typically increasing by between 50% and 75% with a relatively modest linking error rate of 15%. The key takeaway is that despite the statistical power gains from covariate adjustment, linking error remains a concern for experiments with marginally palatable Type II error rates.

5 Empirical Example

Having established that matching errors can lead to a considerable number of Type II errors in empirical applications, we next consider how to mitigate this problem. In Section 3, we established that it is the sum of false positive and false negative error rates (rather than either the false positive or false negative match rates individually) that controls the

[20] The parameter r captures the strength of the relationship between X and y_C^*. Therefore, as r increases in magnitude, statistical power increases, both in the absence and the presence of matching errors. However, the *relative* gain statistical power is slightly larger when we do not condition on X. Across the parameterizations we examine, in the absence of a covariate, the average loss of power under matching errors is 8.4%. When r = 0.1, the loss of power is 8.8% when X is conditioned on. When $r = 0.3$, the average loss of power under matching errors is 11.9% when X is conditioned on. Hence while a larger r is uniformly power enhancing, it does mean that controlling for a covariate will be slightly less helpful in maximizing statistical power than it otherwise would be.

degree of attenuation of parameter estimates and therefore, statistical power. While exact matching will reduce the number of false positive links, it will, in general, not minimize the *sum* of false positive and false negative error rates since the number of false negative links grows because of the stringency of the matching criteria. There is, therefore, promise in testing the performance of more flexible strategies as an alternative to exact matching. These alternative strategies allow for linkages between records that have discrepancies in demographic attributes, a common feature of real world data. Even though allowing records to link that do not match exactly is likely to increase false positives, we show below in our empirical example that the reduction in false negatives links reduces the false negative rate by a much larger amount than the new erroneous links increase the false positive rate. Indeed, in most cases below the increase in the false positive rate is negligible.[21]

In order to explore the potential gains from probabilistic matching with machine learning, we need an empirical example. The reason for this is that while we can solve for the bias that accrues from a given error rate, sample size and effect size, the extent to which we can reduce bias via a given matching strategy requires empirical data — names, dates of birth and addresses, etc. which can be used to generate candidate matches.

5.1 Empirical Simulation

For this study we use identified administrative records on 3 million charges filed in Oregon courts during the 1990-2012 window, maintained in the Oregon Judicial Information Network (OJIN). These data have been used previously to show how legal access to alcohol affects criminality. Hansen and Waddell (2018) measured recidivism by recording whether individuals appeared in a dataset multiple times using exact matching. The individual records in the OJIN data contain the following relevant variables: name, date of birth, race, incident date, and a unique identification number that links the same individuals across

[21] The probabilistic matching approaches we deploy in this section also take advantage of the latest advancements in the field of machine learning for two primary reasons. First, administrative datasets often span hundreds of thousands and often millions of records. Probabilistic techniques involve computing similarity metrics across a number of identifying characteristics such as name and date of birth. It becomes prohibitively, computationally expensive to perform these calculations for each potential record pair as the administrative dataset size grows. Ideally, we would only perform these computations for records with high prior probability of referring to the same person. Techniques for detecting approximate nearest neighbors (Sadosky et al., 2015) allow for fast detection of likely matches that drastically reduce the number of comparisons that need to be made in the linking process. Second, the adaptivity of machine learning models for learning non-linear functions and the practice of assessing performance on out-of-sample data lead to predictive accuracy that outperforms linear models such as logistic regression.

rows in the dataset.[22]

In order to simulate the linking scenario described above, we first randomly sample 80% of the data to treat as the dataset from which we will simulate various administrative datasets, which we refer to as D. The remaining 20% will serve as the dataset from which we will sample various experimental datasets, which we refer to as E. While D is at the record level, meaning a person can appear multiple times, in our simulations we convert E to be at the person level.

While there a number of factors which dictate how well any particular matching strategy performs, in our simulation we focus on the most common in the literature by constructing different versions of D and E according to the following:

- **Size of the administrative data set**: As the number of records in a data set grows, the number of ways in which a typographic error can be introduced can also grow. The end result is that an algorithm will need to learn more patterns that distinguish true positive from true negative links. (Johndrow et al., 2018). We consider administrative data sets of 10,000, 100,000 and 1,000,000 records.

- **Size of the experiment**: The size of the experiment is crucial to understanding the effect of linking errors on subsequent analyses. To explore variations in statistical power, we construct experimental data sets of the following sizes: 500, 750, 1,000, 2,000, and 3,000 individual observations.

- **Share of true matches containing typographic errors**: Similarly, the number of true matches which contain typographic errors can increase the number of patterns that must be learned by a matching algorithm. It is clear that an exact match strategy will perform worse when there are fewer exact matches to be found. We vary the share of true matches that are not exact matches between 10% and 40% in 10% increments.

- **Overlap between data sets**: A small overlap between data sets being linked, which in our scenario would correspond to a low base rate for the outcome, can lead to poor results in the Fellegi-Sunter framework, which is the seminal model for probabilistic record linkage (Yancey, 2004). In order to account for variations in the proportion of

[22]There are situations where the two rows in the dataset will match on all relevant variables save for the unique identifier. As it is ambiguous whether these rows refer to different individuals or if there is an error in the unique identifier, we drop these records from the empirical simulation. This reduces the number of records to 2.6 million.

true matches to be found in the administrative data set we set the share of E that exists in the administrative data to be between 10% and 50% at 10% intervals.

- **Presence of ground truth labels**: The standard approach to implementing the Fellegi-Sunter model for linking does not require information on whether two records actually refer to the same person, or what we will refer to as ground truth data. Prior work in record-linkage has demonstrated the improved performance of record linkage when this data is available (Winkler, 2002). With respect to incorporating ground truth information, we deploy two types of modeling approaches. Note that the Fellegi-Sunter model does not rely on human labeling of data, an approach referred to as *unsupervised* learning. A recent augmentation of this unsupervised approach in probabilistic linking is the use of *active learning* (Enamorado, 2018; Bilenko, 2004). With this approach, the linking algorithm will identify a small number of hard to adjudicate cases and query a human reviewer to provide a match status. By incorporating just a small number of manually labeled data in this manner, the learning algorithm can greatly reduce matching error (Sariyar et al., 2012). In the simulations here, we explore the performance of active learning with the *dedupe* library for the Python programming language.[23] To do so, we use the RecordLink functionality of *Dedupe*, which means that the data used to train the underlying model is from both the experimental an administrative data set.[24] Another approach we explore in this paper is supervised learning, or the use of large quantities of ground truth data (Price et al., 2019). Our algorithm works by identifying instances in the training data where two records are known to either refer to, or not refer to, the same person. We then compute similarity measures between these known pairs. A random forest model (Breiman, 2001) run on these data produces probabilities for whether two records refer to the same person.[25] We use a cutoff threshold and we consider record pairs

[23] https://github.com/dedupeio/dedupe

[24] RecordLink works by identifying potential matches across the two data sets and asking for human labels for pairs which the algorithm is most uncertain about. This information is then incorporated into the learning algorithm to improve predictions. A user providing labels has the option to stop at any point and have *dedupe* produce predictions based on the current version of the algorithm. To simulate a human providing responses, we modified *dedupe*'s code so that ground truth labels would be provided until either of the following conditions was met: the number of labels provided was equal to 50% of the experimental data set size, or the number of labels which identified a true positive links was greater than or equal to 75% of the number of true matches. For 95% of simulations, the number of labels provided was greater than 50% of the experimental data set size, and in 15% of simulations the number of labels provided was greater than 75% of the experimental data set size.

[25] Further details of the algorithm to appear in Jelveh and McNeill (2018).

with predicted probabilities above that threshold value to be links.[26] For both the supervised and active learning algorithms, we provide the following fields with which to compute similarity metrics: first name, last name, date of birth, race, and indictment date. We report all results on heldout data that reflects the same simulation settings (administrative data sample size, share overlap, share non-exact match, experimental sample size) as was used for training the model.

[Insert Table 1 About Here]

Table 1 compares the performance of the machine learning algorithms against exact matching by name and date of birth when linking E to D. As expected, the true positive rate for exact matching is lower than that achieved by probabilistic matching, but imperceptibly so. On the other hand, exact matching is significantly more likely to introduce false negatives. Most importantly, as Table 1 shows, we substantially reduce the sum of false positive and false negative error rates by using a machine-learning strategy. For the supervised learning approach, there is about a 60% reduction in total error and for active learning there is about a 36% reduction.

[Insert Table 2 About Here]

Table 2 provides a more detailed view of how error rates varied with simulation parameters for active and supervised learning. The share of non-exact matches was strongly predictive of higher total error rates for both schemes, but was relatively more influential for active learning. A greater overlap between the experimental and administrative datasets was associated with lower error for active learning but was not related to error for supervised learning. The size of the administrative dataset was also positively associated for both approaches, but more influential for active learning. The size of the experimental dataset had a very small negative association with error rates for both approaches. Finally, the number of human labels provided had negligible impact on error rates.

[26] While ground truth data for record linkage is often hard to come by, in the context of low-cost RCTs it may actually be likely that the administrative data set being linked to will meet the conditions needed to deploy a supervised approach. In particular, the conditions that are needed for supervised learning are that the administrative data set contains a unique identifier (such as an agency identifier assigned by a police department, public hospital, or school system) and that a person can appear multiple times in the data set with the same unique identifier but with discrepancies between records in identifying characteristics.

5.2 Empirical Simulation Results

To simulate matching error bias we follow the same procedure as in Section 4, this time using empirical linkage rates from exact matching as well as probabilistic matching. We explore the comparative performance of exact matching versus two flavors of probabilistic matching: active learning as proxied by the ubiquitous *Dedupe* library for Python and supervised learning using the probabilistic matching algorithm created by the authors. For each $\bar{y}_C^*$, τ_h and N combination and each of the two probabilistic matching strategies, we compute the share of linking errors in the empirical data that are abated by using probabilistic matching instead of exact matching. In other words, we compute the share of linking errors under exact matching that are corrected using probabilistic matching. Results are presented in Figure 3 which contains two histograms: the performance of active learning is summarized in Panel 3a and the performance of supervised learning is summarized in Panel 3b. The histograms plot the distribution of the share of linking errors overturned across a range of $\bar{y}_C^*$, τ_h and N combinations, focusing exclusively on the combinations that result in power of between 0.6 and 0.8. These are the studies that are marginally powered an are precisely the studies for which linking errors are most pivotal.

[Insert Figure 3 About Here]

Referring to the figure, we see that across all pivotal parameterizations, using an active-learning based matching algorithm overturns, on average, 37 percent of the linking errors under exact matching. Using our supervised matching algorithm, we can overturn an average of 62 percent of linking errors. In both cases, the degree to which probabilistic matching is power enhancing varies and depends on the specific $\bar{y}_C^*$, τ_h and N combination in the data. Under supervised learning, the 95 percent confidence interval runs from 50 percent to 77 percent, indicating that, in nearly all cases, probabilistic matching can overturn between one half and three quarters of linking errors. Active learning, as proxied by *Dedupe* is, on average less successful (95 percent confidence interval: 10 percent to 71 percent). However, given the relative simplicity of implementing an active learning algorithm and the wide availability of canned software packages that do so, in most empirical applications, this will be an excellent option to retain a large share of statistical power that would otherwise be lost due to exact matching.[27]

[27] We note that in a very small number of parameterizations, the share of errors overturned is negative

6 Conclusion

We have shown that linking errors, even when they are random, can have serious consequences for the evidence base in empirical criminal justice research — in particular, by creating potentially enormous challenges for developing evidence from randomized experiments. Our reading of the prior literature is that scholars sometimes favor stringent linking criteria (i.e., exact matching) in an effort to minimize false positive links with the goal of generating an analytic data set with as few errors as possible. However, a key insight from this research is that the the sum of false positive and false negative error rates is the parameter that drives the attenuation bias from linking errors, which means that stringent linking criteria will increase, rather than minimize linking error bias. This is because while stringent criteria minimize false positive links, they substantially increase false negative links. As linking error affects coefficient estimates, there are descriptive as well as inferential consequences.

In the presence of linking errors, for any sample size, coefficients will be underestimated, with degree of attenuation being proportional to the error rate in the linkage.[28] While attenuation is unwelcome, linking errors have far more destructive consequences for statistical inference. This is because researchers who plan randomized experiments rarely have an excess of the statistical power they need to detect an effect. The result is that a small degree of attenuation can easily make an effect size (that was thought a priori to be detectable) undetectable. As we show, this problem can be especially severe in experiments with small samples or with larger samples and small effect sizes. Taken study by study, this issue might be dismissed as trivial, but because studies with "null results" are plausibly less likely to be submitted and accepted for publication, as "low-cost randomized trials" gain traction, this problem stands to erode the quality of the social scientific evidence base — perhaps substantially.

While our analytic results apply specifically to randomized experiments where the treatment is allocated randomly, we note that our findings — and advice — also hold for a great deal of quasi-experimental research which implicitly seeks to mimic randomization using

indicating that exact matching leads to fewer linking errors than active learning. Common features of these parameterizations include low exact matching error rates, low overlap between the experimental and administrative datasets, and/or larger administrative datasets.

[28] It is worthwhile to note that the descriptive consequences of linking error cannot be resolved by increasing sample size.

either natural experiment or a control function. Likewise, although we specifically focus on binary outcomes, continuous variables measuring outcomes like program utilization, earnings, or duration could all suffer from similar problems when derived from administrative data. In fact, linking errors might even be more problematic if the absence of a link is a recorded as a zero, a common mass point in those types of continuous variables.

On an optimistic note, we find that probabilistic linking via machine learning algorithms vastly outperforms exact matching and, in fact, in many scenarios approximates a zero error scenario. We argue that these results provide compelling evidence that exact matching should be abandoned in favor of probabilistic linking and that applied researchers should pay greater attention to the way in which data linking is done more generally.

References

Abramitzky, R., L. P. Boustan, K. Eriksson, J. J. Feigenbaum, and S. Pérez (2019, May). Automated linking of historical data. Working Paper 25825, National Bureau of Economic Research.

Aigner, D. J. (1973). Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics 1*(1), 49–59.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Arceneaux, K., A. S. Gerber, and D. P. Green (2010). A cautionary note on the use of matching to estimate causal effects: An empirical example comparing matching estimates to an experimental benchmark. *Sociological Methods & Research 39*(2), 256–282.

Bailey, M., C. Cole, M. Henderson, and C. Massey (2017, November). How well do automated methods perform in historical samples? Evidence from new ground truth. Working Paper 24019, National Bureau of Economic Research.

Banerjee, A. V. and E. Duflo (2009). The experimental approach to development economics. *Annual Review of Economics 1*(1), 151–178.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson (2018). Redefine statistical significance. *Nature Human Behaviour 2*(1), 6–10.

Berent, M. K., J. A. Krosnick, and A. Lupia (2016). Measuring voter registration and turnout in surveys: Do official government records yield more accurate assessments? *Public Opinion Quarterly 80*(3), 597–621.

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology 1*(4), 417–433.

Bertrand, M. and S. Mullainathan (2001). Do people mean what they say? Implications for subjective survey data. *The American Economic Review 91*(2), 67–72.

Bilenko, M. (2004). Learnable similarity functions and their applications to clustering and record linkage. In *Proceedings of the Ninth AAAI/SIGART Doctoral Consortium*, pp. 981–982.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2005). Why the apple doesn't fall far: Understanding intergenerational transmission of human capital. *The American Economic Review 95*(1), 437–449.

Boruch, R. F., T. Victor, and J. S. Cecil (2000). Resolving ethical and legal problems in randomized experiments. *Crime & Delinquency 46*(3), 330–353.

Bowers, K. J. and S. D. Johnson (2005). Domestic burglary repeats and space-time clusters: The dimensions of risk. *European Journal of Criminology 2*(1), 67–92.

Braga, A. A. and R. Apel (2016). And we wonder why criminology is sometimes considered irrelevant in real-world policy conversations. *Criminology Public Policy 15*(3), 813–829.

Braga, A. A. and B. J. Bond (2008). Policing crime and disorder hot spots: A randomized controlled trial. *Criminology 46*(3), 577–607.

Braga, A. A., D. L. Weisburd, E. J. Waring, L. G. Mazerolle, W. Spelman, and F. Gajewski (1999). Problem-oriented policing in violent crime places: A randomized controlled experiment. *Criminology 37*(3), 541–580.

Branas, C. C., E. South, M. C. Kondo, B. C. Hohl, P. Bourgois, D. J. Wiebe, and J. M. MacDonald (2018). Citywide cluster randomized trial to restore blighted vacant land and its effects on violence, crime, and fear. *Proceedings of the National Academy of Sciences 115*(12), 2946–2951.

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

Britt, C. L. and D. Weisburd (2010). Statistical power. In *Handbook of Quantitative Criminology*, pp. 313–332. Springer.

Broidy, L. M., J. K. Daday, C. S. Crandall, D. P. Sklar, and P. F. Jost (2006). Exploring demographic, structural, and behavioral overlap among homicide offenders and victims. *Homicide Studies 10*(3), 155–180.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016). Evaluating replicability of laboratory experiments in economics. *Science 351*(6280), 1433–1436.

Campbell, K. M. (2009). Impact of record-linkage methodology on performance indicators and multivariate relationships. *Journal of Substance Abuse Treatment 36*(1), 110–117.

Campbell, K. M., D. Deck, and A. Krupski (2008). Record linkage software in the public domain: A comparison of Link Plus, the Link King, and a 'basic' deterministic algorithm. *Health Informatics Journal 14*(1), 5–15.

Cesarini, D., E. Lindqvist, R. Östling, and B. Wallace (2016). Wealth, health, and child development: Evidence from administrative data on swedish lottery players. *The Quarterly Journal of Economics 131*(2), 687–738.

Chalfin, A. and M. Deza (2017). The intergenerational effects of education on delinquency. *Journal of Economic Behavior & Organization*.

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. New York, NY: Springer.

Christen, P. and K. Goiser (2007). Quality and complexity measures for data linkage and deduplication. In F. J. Guillet and H. J. Hamilton (Eds.), *Quality measures in data mining*, pp. 127–151. Berlin, Germany: Springer.

Clark, D. E. and D. R. Hahn (1995). Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. *Proceedings of the Annual Symposium on Computer Application in Medical Care 1995*, 397–401.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science 1*(3), 98–101.

Comfort, M., A. M. Nurse, T. McKay, and K. Kramer (2011). Taking children into account: Addressing the intergenerational effects of parental incarceration. *Criminology & Pub. Policy 10*, 839.

Cook, P. J., S. Kang, A. A. Braga, J. Ludwig, and M. E. O'Brien (2015). An experimental evaluation of a comprehensive employment-oriented prisoner re-entry program. *Journal of Quantitative Criminology 31*(3), 355–382.

Courtemanche, C. J., A. Denteh, and R. Tchernis (2018). Estimating the associations between snap and food insecurity, obesity, and food purchases with imperfect administrative measures of participation. Technical report, National Bureau of Economic Research.

Curb, J. D., C. E. Ford, S. Pressel, M. Palmer, C. Babcock, and C. M. Hawkins (1985). Ascertainment of vital status through the national death index and the social security administration. *American Journal of Epidemiology 121*(5), 754–766.

Dahl, G. B., A. R. Kostøl, and M. Mogstad (2014). Family welfare cultures. *The Quarterly Journal of Economics 129*(4), 1711–1752.

Davis, R. C. and B. G. Taylor (1997). A proactive response to family violence: The results of a randomized experiment. *Criminology 35*(2), 307–333.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–55.

Dezember, A., M. Stoltz, L. Marmolejo, L. C. Kanewske, K. D. Feingold, S. Wire, L. Duhaime, and C. Maupin (2020). The lack of experimental research in criminology — evidence from criminology and justice quarterly. *Journal of Experimental Criminology*.

DiNardo, J. E. and J.-S. Pischke (1997). The returns to computer use revisited: Have pencils changed the wage structure too? *The Quarterly Journal of Economics 112*(1), 291–303.

Dobbie, W., J. Goldin, and C. S. Yang (2018). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *The American Economic Review 108*(2), 201–40.

Dobbie, W., H. Grönqvist, S. Niknami, M. Palme, and M. Priks (2018). The intergenerational effects of parental incarceration. Technical report, National Bureau of Economic Research.

Dobkin, C., A. Finkelstein, R. Kluender, and M. J. Notowidigdo (2018). The economic consequences of hospital admissions. *The American Economic Review 108*(2), 308–352.

Doleac, J. L., C. Temple, D. Pritchard, and A. Roberts (2020). Which prisoner reentry programs work? replicating and extending analyses of three *rct*s. *International Review of Law and Economics 62*, 105902.

Duflo, E., R. Glennerster, and M. Kremer (2007). Using randomization in development economics research: A toolkit. In T. P. Schultz and J. A. Strauss (Eds.), *Handbook of Development Economics*, Volume 4 of *Handbook of Development Economics*, pp. 3895–3962. Amsterdam, the Netherlands: North-Holland.

Dunn, P. M. (1997). James lind (1716-94) of edinburgh and the treatment of scurvy. *Archives of Disease in Childhood-Fetal and Neonatal Edition 76*(1), F64–F65.

Duwe, G. (2012). Evaluating the minnesota comprehensive offender reentry plan (mcorp): Results from a randomized experiment. *Justice Quarterly 29*(3), 347–383.

Duwe, G. (2014). A randomized experiment of a prisoner reentry program: Updated results from an evaluation of the minnesota comprehensive offender reentry plan (mcorp). *Criminal Justice Studies 27*(2), 172–190.

Enamorado, T. (2018). Active learning for probabilistic record linkage. *Available at SSRN 3257638*.

Enamorado, T., B. Fifield, and K. Imai (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review 113*(2), 353–371.

Farabee, D., S. X. Zhang, and B. Wright (2014). An experimental evaluation of a nationally recognized employment-focused offender reentry program. *Journal of Experimental Criminology 10*(3), 309–322.

Farrington, D. P. (1983). Randomized experiments on crime and justice. *Crime and Justice 4*, 257–308.

Farrington, D. P. (2003). A short history of randomized experiments in criminology: A meager feast. *Evaluation Review 27*(3), 218–227.

Farrington, D. P. (2006). Key longitudinal-experimental studies in criminology. *Journal of Experimental Criminology 2*(2), 121–141.

Farrington, D. P. and B. C. Welsh (2006). A half century of randomized experiments on crime and justice. *Crime and Justice 34*(1), 55–132.

Feigenbaum, J. J. (2016). Automated census record linking: A machine learning approach. *Working Paper*.

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

Ferrante, A. (1993). Developing an offender-based tracking system: The western australia inois project. *Australian and New Zealand Journal of Criminology 26*(3), 232–250.

Fischbacher, U., S. Gächter, and F. Ernst (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters 71*(3), 397–404.

Fisher, R. A. (1936). Design of experiments. *The British Medical Journal 1*(3923), 554–554.

Freudenberg, N., I. Wilets, M. B. Greene, and B. E. Richet (1998). Linking women in jail to community services: Factors associated with rearrest and retention of drug-using women following release from jail. *Journal of the American Medical Women's Association 53*(2), 89–93.

Geerken, M. R. (1994). Rap sheets in criminological research: Considerations and caveats. *Journal of Quantitative Criminology 10*(1), 3–21.

Gelber, A., A. Isen, and J. B. Kessler (2016). The effects of youth employment: Evidence from new york city lotteries. *The Quarterly Journal of Economics 131*(1), 423–460.

Gerber, A. S. and N. Malhotra (2008). Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science 3*(3), 313–326.

Gilbert, D. T., G. King, S. Pettigrew, and T. D. Wilson (2016). Comment on "Estimating the reproducibility of psychological science". *Science 351*(6277), 1037.

Gomatam, S., R. Carter, M. Ariet, and G. Mitchell (2002). An empirical comparison of record linkage procedures. *Statistics in Medicine 21*(10), 1485–1496.

Gordon, B. R., F. Zettelmeyer, N. Bhargava, and D. Chapsky (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science 38*(2), 193–225.

Gottfredson, D. C., S. S. Najaka, B. W. Kearley, and C. M. Rocha (2006). Long-term effects of participation in the baltimore city drug treatment court: Results from an experimental study. *Journal of Experimental Criminology 2*(1), 67–98.

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

Hansen, B. and G. R. Waddell (2018). Legal access to alcohol and criminality. *Journal of Health Economics 57*, 277–289.

Harron, K., C. Dibben, J. Boyd, A. Hjern, M. Azimaee, M. L. Barreto, and H. Goldstein (2017). Challenges in administrative data linkage for research. *Big Data & Society 4*(2), 1–12.

Heckman, J. J. and J. A. Smith (1995). Assessing the case for social experiments. *Journal of Economic Perspectives 9*(2), 85–110.

Heller, S. B. (2014). Summer jobs reduce violence among disadvantaged youth. *Science 346*(6214), 1219–1223.

Heller, S. B., A. K. Shah, J. Guryan, J. Ludwig, S. Mullainathan, and H. A. Pollack (2017). Thinking, fast and slow? some field experiments to reduce crime and dropout in chicago. *The Quarterly Journal of Economics 132*(1), 1–54.

Hill, S. J. (2017). Changing votes or changing voters? How candidates and election context swing voters and mobilize the base. *Electoral Studies 48*, 131–148.

Hjalmarsson, R. and M. J. Lindquist (2012). Like godfather, like son exploring the intergenerational nature of crime. *Journal of Human Resources 47*(2), 550–582.

Hovde Lyngstad, T. and T. Skardhamar (2011). Nordic register data and their untapped potential for criminological knowledge. *Crime and Justice 40*(1), 613–645.

Hyatt, J. M. and S. N. Andersen (2019). On the potential of incorporating administrative register data into randomized experiments. *Journal of Experimental Criminology 15*(3), 469–497.

Imbens, G. W. (2010). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature 48*(2), 399–423.

Ioannidis, J. P., T. D. Stanley, and H. Doucouliagos (2017). The power of bias in economics research. *The Economic Journal 127*(605), F236–F265.

Jelveh, Z. and M. McNeill (2018). Downstream impact of probabilistic matching quality on prediction performance. *In Progress*.

Johndrow, J., K. Lum, and D. Dunson (2018). Theoretical limits of microclustering for record linkage. *Biometrika 105*(2), 431–446.

Johnston, A. and A. Mas (2018). Potential unemployment insurance duration and labor supply: The individual and market-level response to a benefit cut. *Journal of Political Economy 126*(6), 2480–2522.

Karr, A. F., M. T. Taylor, S. L. West, S. Setoguchi, T. D. Kou, T. Gerhard, and D. B. Horton (2019). Comparing record linkage software programs and algorithms using real-world data. *PloS One 14*(9), e0221459.

Keizer, K., S. Lindenberg, and L. Steg (2008). The spreading of disorder. *Science 322*(5908), 1681–1685.

Khwaja, A. I. and A. Mian (2005). Do lenders favor politically connected firms? Rent provision in an emerging financial market. *The Quarterly Journal of Economics 120*(4), 1371–1411.

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

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review 76*(4), 604–620.

Lane, J. (2018). Building an infrastructure to support the use of government administrative data for program performance and social science research. *The ANNALS of the American Academy of Political and Social Science 675*(1), 240–252.

Lattimore, P. K., D. L. MacKenzie, G. Zajac, D. Dawes, E. Arsenault, and S. Tueller (2016). Outcome findings from the hope demonstration field experiment: Is swift, certain, and fair an effective supervision strategy? *Criminology & Public Policy 15*(4), 1103–1141.

Laub, J. H. and R. J. Sampson (1988). Unraveling families and delinquency: A reanalysis of the glueks'data. *Criminology 26*(3), 355–380.

Lauritsen, J. L. (1999). Limitations in the use of longitudinal self-report data: A comment. *Criminology 37*, 687.

Liberman, A. M., D. S. Kirk, and K. Kim (2014). Labeling effects of first juvenile arrests: Secondary deviance and secondary sanctioning. *Criminology 52*(3), 345–370.

Loeffler, C. E. (2013). Does imprisonment alter the life course? evidence on crime and employment from a natural experiment. *Criminology 51*(1), 137–166.

Lum, C. and L. Mazerolle (2014). History of randomized controlled experiments in criminal justice. *The Encyclopedia of Criminology and Criminal Justice*, 2227–2239.

Lynch, J. (2018). Not even our own facts: Criminology in the era of big data. *Criminology 56*(3), 437–454.

MacDonald, J. M., A. R. Morral, B. Raymond, and C. Eibner (2007). The efficacy of the rio hondo dui court: A 2-year field experiment. *Evaluation Review 31*(1), 4–23.

McCord, J. (2003). Cures that harm: Unanticipated outcomes of crime prevention programs. *The Annals of the American Academy of Political and Social Science 587*(1), 16–30.

Moher, D., C. S. Dulberg, and G. A. Wells (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association 272*(2), 122–124.

Moore, C. L., J. Amin, H. F. Gidding, and M. G. Law (2014). A new method for assessing how sensitivity and specificity of linkage studies affects estimation. *PloS One 9*(7), e103690.

Morris, N. A. and L. A. Slocum (2010). The validity of self-reported prevalence, frequency, and timing of arrest: An evaluation of data collected using a life event calendar. *Journal of Research in Crime and Delinquency 47*(2), 210–240.

Mueller-Smith, M. (2016). The criminal and labor market impacts of incarceration. *Working paper*.

Nagin, D. S. and R. J. Sampson (2019). The real gold standard: measuring counterfactual worlds that matter most to social science and policy. *Annual Review of Criminology 2*, 123–145.

Nelson, M. S., A. Wooditch, and L. M. Dario (2015). Sample size, effect size, and statistical power: A replication study of weisburd's paradox. *Journal of Experimental Criminology 11*(1), 141–163.

Neter, J., E. S. Maynes, and R. Ramanathan (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association 60*(312), 1005–1027.

Newcombe, H. B., J. M. Kennedy, S. Axford, and A. P. James (1959). Automatic linkage of vital records. *Science 330*(3381), 954–959.

Orchowsky, S. and J. Iwama (2009). Improving state criminal history records: Recidivism of sex offenders released in 2001. Report, Justice Research and Statistics Association.

O'Brien, D. T. and R. J. Sampson (2015). Public and private spheres of neighborhood disorder: Assessing pathways to violence using large-scale digital records. *Journal of research in Crime and Delinquency 52*(4), 486–510.

Petrosino, A., C. Turpin-Petrosino, and J. Buehler (2003). Scared straight and other juvenile awareness programs for preventing juvenile delinquency: A systematic review of the randomized experimental evidence. *The Annals of the American Academy of Political and Social Science 589*(1), 41–62.

Powell, D. and S. Seabury (2018). Medical care spending and labor market outcomes: Evidence from workers' compensation reforms. *The American Economic Review 108*(10), 2995–3027.

Powers, E. and H. Witmer (1951). *An experiment in the prevention of delinquency; the Cambridge-Somerville Youth Study.* Columbia University Press.

Price, J., K. Buckles, J. Van Leeuwen, and I. Riley (2019, September). Combining family history and machine learning to link historical records. Working Paper 26227, National Bureau of Economic Research.

Pridemore, W. A., M. C. Makel, and J. A. Plucker (2018). Replication in criminology and the social sciences. *Annual Review of Criminology 1*(1), 19–38.

Ridgeway, G., J. Grogger, R. A. Moyer, and J. M. MacDonald (2019). Effect of gang injunctions on crime: A study of los angeles from 1988–2014. *Journal of Quantitative Criminology 35*(3), 517–541.

Roberts, J. and W. Wells (2010). The validity of criminal justice contacts reported by inmates: A comparison of self-reported data with official prison records. *Journal of Criminal Justice 38*(5), 1031–1037.

Rossman, S. B., J. K. Roman, J. M. Zweig, M. Rempel, C. H. Lindquist, et al. (2011). *The multi-site adult drug court evaluation: Executive summary.* Urban Institute.

Rothstein, H. R. (2008). Publication bias as a threat to the validity of meta-analytic results. *Journal of Experimental Criminology 4*(1), 61–81.

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association 103*(484), 1350–1353.

Sadosky, P., A. Shrivastava, M. Price, and R. C. Steorts (2015). Blocking methods applied to casualty records from the syrian conflict. *arXiv preprint arXiv:1510.07714*.

Sampson, R. J. (2010). Gold standard myths: Observations on the experimental turn in quantitative criminology. *Journal of Quantitative Criminology 26*(4), 489–500.

Sampson, R. J. and J. H. Laub (2003). Life-course desisters? trajectories of crime among delinquent boys followed to age 70. *Criminology 41*(3), 555–592.

Sampson, R. J. and A. S. Winter (2018). Poisoned development: Assessing childhood lead exposure as a cause of crime in a birth cohort followed through adolescence. *Criminology 56*(2), 269–301.

Sariyar, M., A. Borg, and K. Pommerening (2012). Active learning strategies for the deduplication of electronic patient data using classification trees. *Journal of Biomedical Informatics 45*(5), 893–900.

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

Scheuren, F. and W. E. Winkler (1997). Regression analysis of data files that are computer matched, part ii. *Survey Methodology 23*(2), 157–165.

Sedlmeier, P. and G. Gigerenzer (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin 105*(2), 309–316.

Sherman, L. W. (2007). The power few: experimental criminology and the reduction of harm. *Journal of Experimental Criminology 3*(4), 299–321.

Sherman, L. W., D. C. Gottfredson, D. L. MacKenzie, J. Eck, P. Reuter, S. Bushway, et al. (1997). *Preventing crime: What works, what doesn't, what's promising: A report to the United States Congress*. US Department of Justice, Office of Justice Programs Washington, DC.

Sherman, L. W., J. D. Schmidt, D. P. Rogan, and D. A. Smith (1992). The variable effects of arrest on criminal careers: The milwaukee domestic violence experiment. *Journal of Criminal Law & Criminology 83*, 137.

Sherman, L. W. and D. Weisburd (1995). General deterrent effects of police patrol in crime "hot spots": A randomized, controlled trial. *Justice Quarterly 12*(4), 625–648.

Sherman, W. and R. A. Berk (1984). The minneapolis domestic violence experiment.

Smith, G. J. D., L. Bennett Moses, and J. Chan (2017). The challenges of doing criminology in the big data era: Towards a digital and data-driven approach. *The British Journal of Criminology 57*(2), 259–274.

Smith, J. A. and P. E. Todd (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. *The American Economic Review 91*(2), 112–118.

Smith, J. A. and P. E. Todd (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics 125*(1-2), 305–353.

Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed (1990[1923]). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472.

Stewart, A., S. Dennison, T. Allard, C. Thompson, L. Broidy, and A. Chrzanowski (2015). Administrative data linkage as a tool for developmental and life-course criminology: The queensland linkage project. *Australian & New Zealand Journal of Criminology 48*(3), 409–428.

Taxman, F. S. and M. S. Caudy (2015). Risk tells us who, but not what or how. *Criminology Public Policy 14*(1), 71–103.

Tremblay, R. E., F. Vitaro, D. Nagin, L. Pagani, and J. R. Seguin (2003). The montreal longitudinal and experimental study. In *Taking Stock of Delinquency*, pp. 205–254. Springer.

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

Van Schellen, M., R. Apel, and P. Nieuwbeerta (2012). The impact of military service on criminal offending over the life course: evidence from a dutch conviction cohort. *Journal of Experimental Criminology 8*(2), 135–164.

Vivalt, E. (2017). The trajectory of specification searching and publication bias across methods and disciplines. *Working paper*.

Watson, C. I., G. P. Fiumara, E. Tabassi, W. J. Salamon, and P. A. Flanagan (2014). Fingerprint vendor technology evaluation. Report, NIST.

Weisburd, D. (2003). Ethical practice and evaluation of interventions in crime and justice: The moral imperative for randomized trials. *Evaluation Review 27*(3), 336–354.

Weisburd, D. (2010). Justifying the use of non-experimental methods and disqualifying the use of randomized controlled trials: Challenging folklore in evaluation research in crime and justice. *Journal of Experimental Criminology 6*(2), 209–227.

Weisburd, D., A. Petrosino, and G. Mason (1993). Design sensitivity in criminal justice experiments. *Crime and Justice 17*, 337–379.

Wildeman, C. and S. H. Andersen (2017). Paternal incarceration and children's risk of being charged by early adulthood: Evidence from a danish policy shock. *Criminology 55*(1), 32–58.

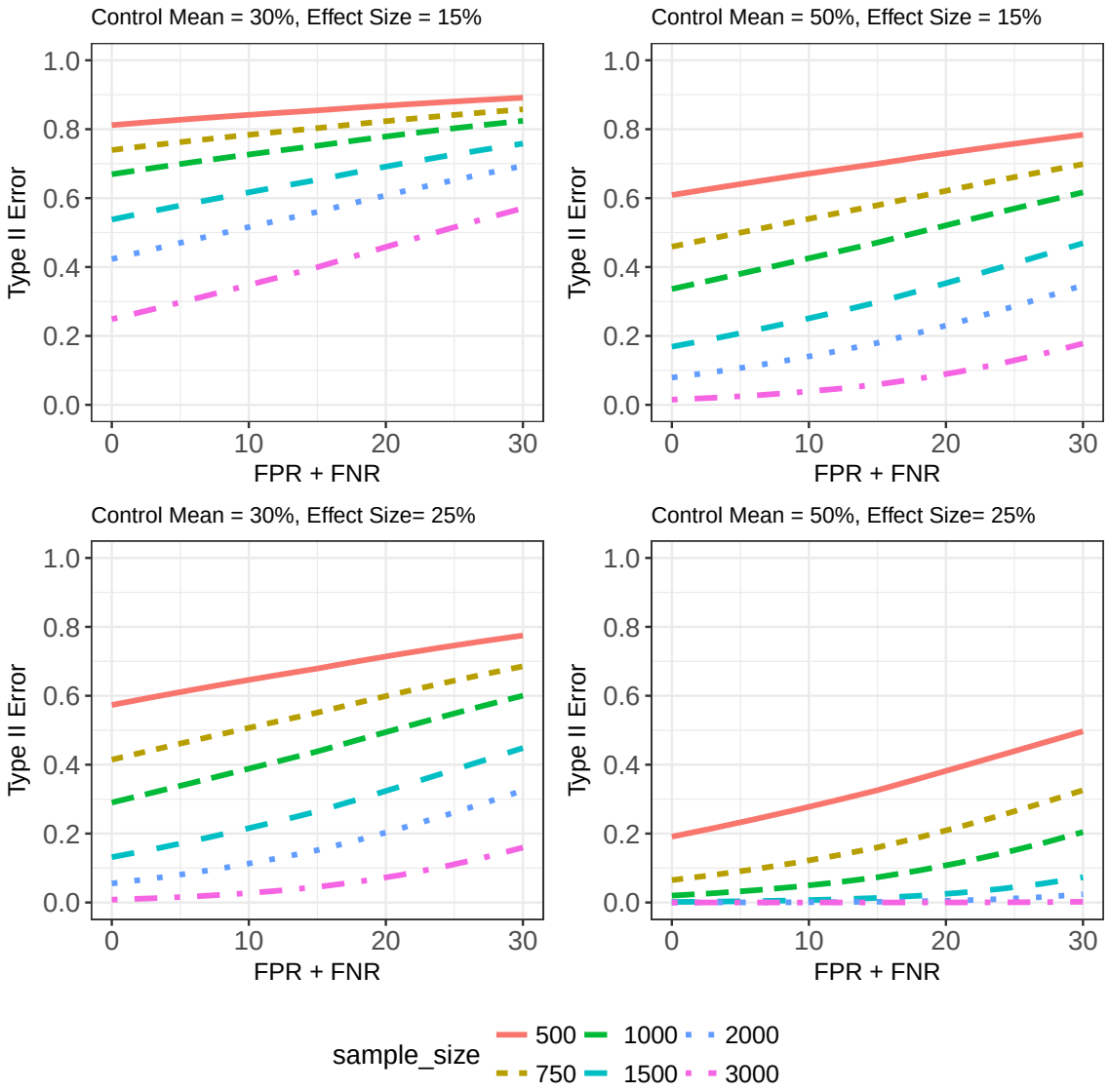
Winkler, W. E. (2002). Methods for record linkage and bayesian networks. Report, Statistical Research Division, U.S. Bureau of the Census.

Winkler, W. E. (2006). Overview of record linkage and current research directions. Report, U.S. Bureau of the Census.

Yancey, W. E. (2004). Improving EM algorithm estimates for record linkage parameters. Report, U.S. Bureau of the Census.

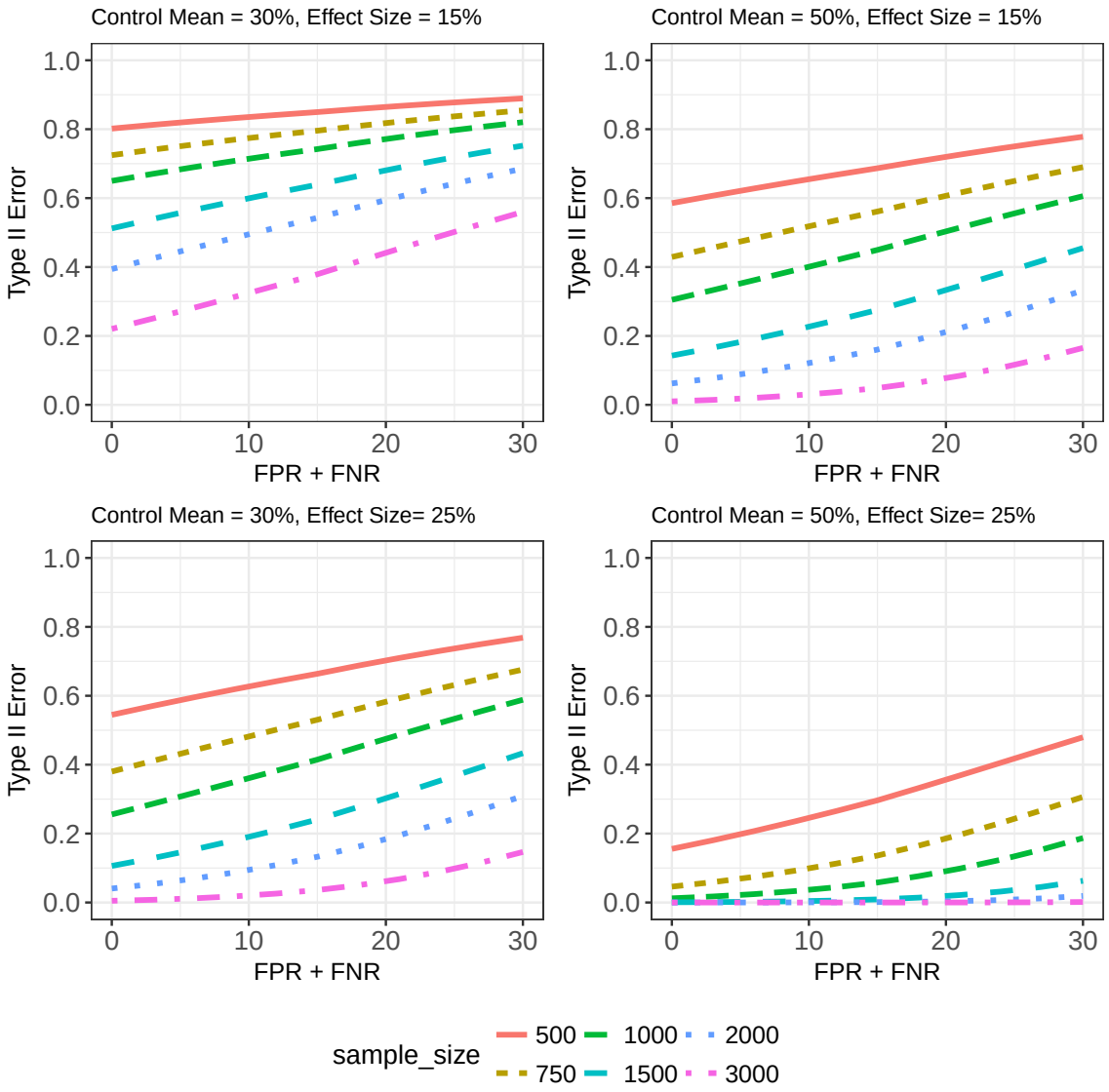
Zingmond, D. S., Z. Ye, S. L. Ettner, and H. Liu (2004). Linking hospital discharge and death records—accuracy and sources of bias. *Journal of Clinical Epidemiology 57*(1), 21–29.

Figure 1: Matching Error and Type II Error Rates By Outcome Density and Treatment Effect Size



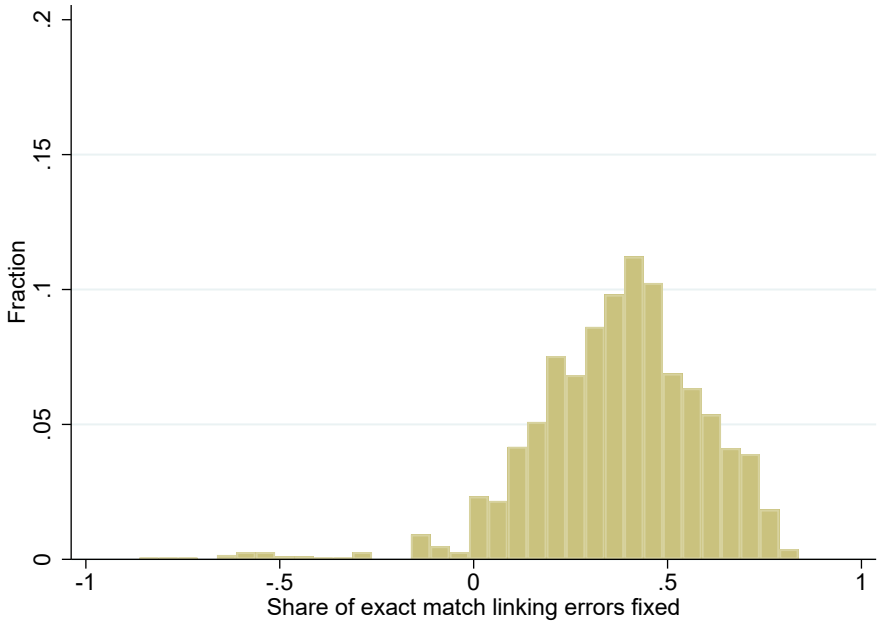
Note: Figures plot the Type II error rate (β) as a function of the total matching error rate for a given hypothesized effect size and control mean. In each plot, Type II error rates are plotted for sample sizes that range from $N = 500$ to $N = 3{,}000$.

Figure 2: Matching Error and Type II Error Rates w/ Covariate Adjustment By Outcome Density and Treatment Effect Size

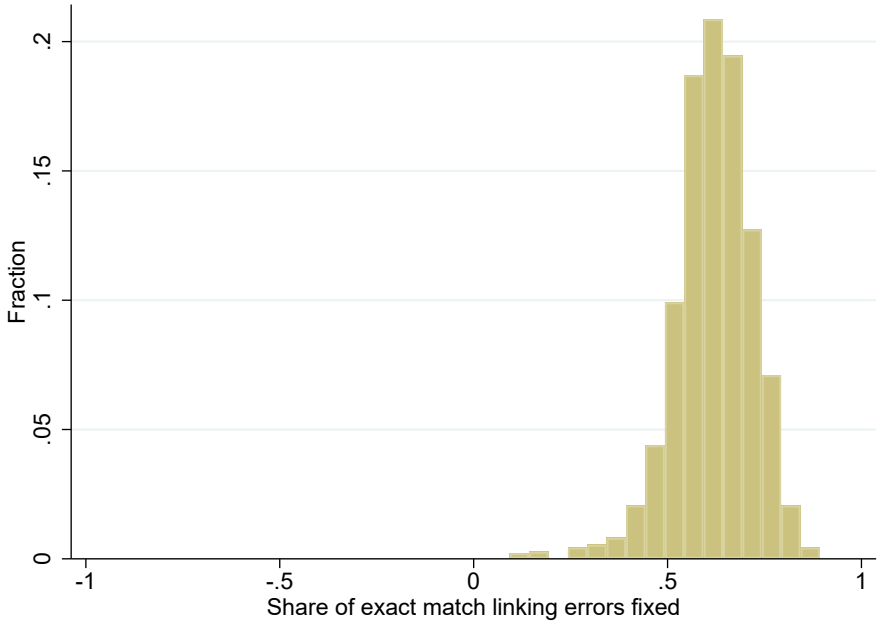


Note: Figures plot the Type II error rate (β) as a function of the total matching error rate for a given hypothesized effect size, control mean, and correlated covariate. In each plot, Type II error rates are plotted for sample sizes that range from $N = 500$ to $N = 3{,}000$.

Figure 3: Distributions of exact match linking errors fixed by active learning and supervised learning



(a) Distribution of exact match linking errors fixed by active learning



(b) Distribution of exact match linking errors fixed by supervised learning

Note: The histograms plot the distribution of the share of linking errors that are overturned using probabalistic matching across the $\bar{y}_C^*$, τ_h and N combinations where the ground truth power lies between 0.6 and 0.8.

Table 1: Performance metrics across matching schemes

Share Non-Exact	Error	(1) Exact Matching	(2) Active Learning	(3) Supervised Learning
0.1	FPR	0.000	0.002	0.002
	FNR	0.100	0.072	0.038
	FPR+FNR	0.100	0.074	0.040
0.2	FPR	0.000	0.004	0.002
	FNR	0.200	0.125	0.074
	FPR+FNR	0.200	0.129	0.076
0.3	FPR	0.000	0.005	0.002
	FNR	0.300	0.178	0.114
	FPR+FNR	0.300	0.183	0.116
0.4	FPR	0.000	0.006	0.002
	FNR	0.400	0.219	0.147
	FPR+FNR	0.400	0.225	0.148

Note: This table displays three different error rates (false positive rate, false negative rate, and the sum of the two error rates) from simulated matches for each share of non-exact matches between the input experimental data set E and the administrative data set D. We simulate matches for each of the matching schemes, estimating an error rate for each of the combinations of the following parameters: administrative data set (10,000, 100,000 and 1,000,000); experimental data set (500, 750, 1,000, 2,000, and 3,000); and share non-exact matches (.1, .2, .3, and .4). The error rates presented in Columns (1)-(3) are averaged across simulations for each matching scheme.

Table 2: Relationship between total error rate and simulation parameters

	Dependent variable:	
	FPR + FNR	
	Active Learning	Supervised Learning
Share Non-Exact	0.509***	0.372***
	(0.013)	(0.006)
Share Overlap	−0.127***	−0.006
	(0.016)	(0.006)
Admin Sample Size = 100,000	0.062***	0.027***
	(0.005)	(0.002)
Admin Sample Size = 1,000,000	0.107***	0.030***
	(0.006)	(0.002)
Experimental Sample Size (in 100s)	−0.001***	−0.0002**
	(0.0004)	(0.0001)
Number of Human Labels Provided	0.0001***	
	(0.00003)	
Constant	−0.004	−0.012***
	(0.007)	(0.003)
Observations	450	450
Adjusted R^2	0.813	0.901

Note: This table presents the coefficient estimates from ordinary least squares regressions of the sum of the false positive and false negative rates from the simulated matches on the attributes of the simulation (N=450). Standard errors are in parentheses. The results are shown separately for active learning and supevised learning. The goal here is to decompose the sum of false positive and false negative error rates to demonstrate the relative contributions of the different components of the match. Statistical significance is indicated by asterisks according to the following: *p<0.1; **p<0.05; ***p<0.01.

Appendices

A Computational Details

In this appendix we provide additional details for how statistical power can be computed under two possible states of the world: 1) in the absence of linking errors and 2) in the presence of linking errors. We use the derivations in this appendix to empirically demonstrate the effect of linking errors on statistical power in a hypothetical experiment in Section 5 of the paper.

For illustrative purposes, we will assume that a roster of individuals involved in a treatment program is being linked to arrest data to measure whether the program reduced the likelihood of arrest. Additionally, we will assume a record-linkage algorithm was run on the arrest data and that there existed a unique identifier allowing us to measure when predicted links between two records represented true and false matches and when predicted non-links represtented true and false non-matches.

We motivate the derivation by introducing a framework — a confusion matrix — that governs the incidence of linking errors in the arrest. Each row of the confusion matrix represents the incidence of an actual class (true non-match and true match) while each column represents the instances in a predicted class (predicted non-link and predicted link). The matrix thus allows us to understand the extent to which the algorithm is successful in classifying that two records belong to the same person.

In the following confusion matrix, y^* represents the true state of the world and y represents the observed state of the world after linking. The cells provide counts of the number of true negatives, false negatives, false positives and true positives, respectively in linking the data.

	$y^* = 0$	$y^* = 1$
$y = 0$	TN	FN
$y = 1$	FP	TP

The diagonal entries of the matrix correspond to an alignment of the true and observed states of the world — observations for which $y^* = y = 0$ are true negatives and observations for which $y^* = y = 1$ are true positives. The off-diagonal entries provide us with the number of linking errors. In particular, the 2,1 element of the matrix provides the number of false positive links — this is the number of times in which an observation which is truly $y^* = 0$ is mistakenly linked to $y = 1$. Similarly, the 1,2 element of the matrix provides the number of false negative links where an observation that is truly $y^* = 1$ is mistakenly linked to a record for which $y = 0$.

The matrix allows us to compute four different rates capturing the success of a given linking strategy: the true positive and true negative rate and the false positive and false

negative rate.

$$TPR = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

$$FPR = 1 - TNR$$

$$FNR = 1 - TPR$$

The true positive rate (TPR) is defined as the number of linked positives divided by the number of true positives (TP+FN). Likewise the true negative rate (TNR) is the number of linked negatives divided by the number of true negatives (TN+FP). The corresponding false positive and false negative link rates are obtained by subtracting each of these quantities from 1. As we show in Section 3 of the paper, estimated treatment effects will be attenuated under linking errors and the attenuation will be proportional to 1-FPR-FNR. So long as FPR+$FNR < 1$, this will be strict attenuation towards zero but if FPR+FNR exceeds 1 then there can be a change in the sign of the bias.

To appreciate how this works, assume that the arrest dataset contained N = 10,000 records and that after running the matching algorithm, the following confusion matrix was generated:

	$y^* = 0$	$y^* = 1$
$y = 0$	TN=3,000	FN=1,000
$y = 1$	FP=2,000	TP=4,000

The error rates for the matching algorithm can be computed as:

- FNR = $\frac{1{,}000}{4{,}000+1{,}000} = 0.20$
- FPR = $\frac{2{,}000}{3{,}000+2{,}000} = 0.40$

Assume that to test the effectiveness of the treatment program, 1500 individuals where randomized, with one-half in the treatment group ($p = 0.5$), the control group mean was $\frac{1}{3}$, and that the treatment effect was $\tau = \frac{-1}{15}$.

In the true state of the world, there are 250 individuals arrested in the control group and 200 in the treatment group, reflecting the fact that $\tau = \frac{-1}{15}$. We next apply the error rates from the matching algorithm to the control and treatment groups respectively to generate the following confusion matrices:

Control Group

	$y^* = 0$	$y^* = 1$
$y = 0$	TN=300	FN=50
$y = 1$	FP=200	TP=200

	Treatment Group	
	$y^* = 0$	$y^* = 1$
$y = 0$	TN=330	FN=40
$y = 1$	FP=220	TP=160

Let $y^*_{T=0}$ be the true number of individuals arrested in the control group and $y_{T=0}$ be the observed number of individuals arrested in the control group. We see that $y^*_{T=0} = 50 + 200 = 250$ and $y_{T=0} = 200 + 200 = 400$.

Let $y^*_{T=1}$ be the true number of individuals arrested in the treatment group and $y_{T=1}$ be the observed number of individuals arrested in the treatment group. We see that $y^*_{T=1} = 40 + 160 = 200$ and $y_{T=1} = 220 + 160 = 380$.

The observed treatment effect can be computed as $\bar{y}_{T=1} - \bar{y}_{T=0} = \frac{400}{750} - \frac{380}{750} = \frac{-2}{75}$, which is equivalent to $\tau * (TPR - FPR) = \frac{-1}{15} * (0.8 - 0.4) = \frac{-2}{75}$

In order to compute statistical power to detect a given potential treatment effect, we need to compute a standard error which is computed according to:

$$var(\hat{\tau}) = \frac{1}{p(1-p)} \frac{\sigma^2}{N}$$

The square root of this quantity is the standard error around the estimated treatment effect. N and p are simply the sample size and the proportion treated but we will need to compute ς which is the mean square error from a regression of either y^* or y on D, depending on which state of the world we are in. We show how to compute σ^2 in absence and presence of linking errors in Appendix B.

We can then compute statistical power according to:

$$\beta = \Phi\left[- \Phi^{-1}\left(\frac{\alpha}{2}\right) - \frac{\tau_h}{\sigma_{\tau_h}} \right]$$

Carrying through the numerical example from our confusion table, power to detect a treatment effect of 20% in these data is 81 percent in the true state of the world and just 72 percent in the state of the world with linking errors. What would have been an adequately well-powered experiment is no longer well-powered in the presence of modest linking errors.

B Deriving Outcome Variance

In this section we show how to compute the residual sum of squares with a binary outcome and binary treatment in order to compute the σ^2. Let $\bar{y}_C$ equal the control group mean and τ the treatment effect:

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \bar{y}_C - \tau T_i)^2$$

We can decompose the above equation into four mutually exclusive groups determined by whether an individual is in the treatment or control group, and whether their associated outcome is $y = 0$ or $y = 1$.

$$\begin{aligned}
&\sum_{i \in \{i|T_i=0, y_i=0\}} (-\bar{y}_C)^2 + \sum_{i \in \{i|T_i=0, y_i=1\}} (1 - \bar{y}_C)^2 + \sum_{i \in \{i|T_i=1, y_i=0\}} (-\bar{y}_C - \tau)^2 + \sum_{i \in \{i|T_i=1, y_i=1\}} (1 - \bar{y}_C - \tau)^2 \\
&= N_{C,0}\ \bar{y}_C^2 + N_{C,1} + N_{C,1}\ \bar{y}_C^2 - N_{C,1}\ 2\bar{y}_C + N_{T,0}\bar{y}_T^2 + N_{T,1} + N_{T,1}\ \bar{y}_T^2 - N_{T,1}\ 2\bar{y}_T \\
&= n_C\ \bar{y}_C^2 + N_{C,1} - N_{C,1}\ 2\bar{y}_C + n_T\bar{y}_T^2 + N_{T,1} - N_{T,1}\ 2\bar{y}_T \\
&= N_{C,1}\ \bar{y}_C + N_{C,1} - N_{C,1}\ 2\bar{y}_C + N_{T,1}\bar{y}_T + N_{T,1} - N_{T,1}\ 2\bar{y}_T \\
&= N_{C,1}(\bar{y}_C + 1 - 2\bar{y}_C) + N_{T,1}(\bar{y}_T + 1 - 2\bar{y}_T) \\
&= N_{C,1}(1 - \bar{y}_C) + N_{T,1}(1 - \bar{y}_T)
\end{aligned}$$

C Maximizing RSS

We now show why Equation 10 is maximized when the control group mean, $\bar{y}_C$, plus the treatment effect, τ, equal 0.5. Let $N_{T,1}$ equal the number of individuals in the treatment group with $y = 1$ and $N_{T,0}$ equal the number of individuals in the treatment group with $y = 0$. Note that $N_{T,1} = (\bar{y}_C + \tau)N_T$ and $N_{T,0} = N_T(1 - (\bar{y}_C + \tau))$.

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (\bar{y}_C + \tau T_i))^2$$

For a given control group mean we will take derivatives with respect to τ, which means we will only consider individuals in the treatment group. We can decompose the previous equation into:

$$\begin{aligned}
\sum_{i \in T} (y_i - \hat{y}_i)^2 &= N_{T,0}(-\bar{y}_C - \tau)^2 + N_{T,1}(1 - \bar{y}_C - \tau)^2 \\
&= N_T(1 - (\bar{y}_C + \tau))(-\bar{y}_C - \tau)^2 + N_T(\bar{y}_C + \tau)(1 - \bar{y}_C - \tau)^2 \\
&= N_T(1 - \bar{y}_C - \tau)(\bar{y}_C + \tau)^2 + N_T(\bar{y}_C + \tau)(1 - \bar{y}_C - \tau)^2 \\
&= N_T(1 - \bar{y}_C - \tau)(\bar{y}_C + \tau)(\bar{y}_C + \tau + 1 - \alpha - \tau) \\
&= N_T(1 - \bar{y}_C - \tau)(\bar{y}_C + \tau)
\end{aligned}$$

Let $\kappa = N_T(1 - \bar{y}_C - \tau)(\bar{y}_C + \tau)$, then taking derivatives with respect to τ:

$$\frac{d\kappa}{d\tau} = N_T(1 - 2\bar{y}_C - 2\tau)$$

Setting the previous equation to zero and solving for τ leads to

$$\bar{y}_C + \tau = 0.5$$

D Proof for Power Attenuation

In this section we show that even when the standard error estimated under linking error is smaller than the standard error estimated under no error, statistical power will still be larger for the latter scenario. Let η be True Positive Rate, ω the False Positive Rate, tau^* the true treatment effect, σ_{τ^*} the true standard error, $\hat{\tau}$ the observed treatment effect, and $\sigma_{\hat{\tau}}$ the observed standard error. Note that $0 \leq \eta \leq 1$ and $0 \leq \omega \leq 1$. In order to show that

$$\frac{\tau^*}{\sigma_{\tau^*}} > \frac{\hat{\tau}}{\sigma_{\hat{\tau}}}$$

we use the result from Equation 3.1 to substitute for the observed treatment effect to get

$$\frac{\tau^*}{\sigma_{\tau^*}} > \frac{(\eta - \omega)\tau^*}{\sigma_{\hat{\tau}}}$$

and then rearrange terms to get the following:

$$\sigma_{\hat{\tau}} > (\eta - \omega)\sigma_{\tau_h}$$

It is straightforward to show that this is equivalent to

$$\sqrt{\widehat{RSS}} > (\eta - \omega)\sqrt{RSS^*}$$

Or that

$$\widehat{RSS} - (\eta - \omega)^2 RSS^* > 0$$

where $\widehat{RSS}$ is the residual sum of squares with linking error and RSS^* is the residual sum of squares without linking error.

In the following, $N^*_{j,1}$ represents the number of observations for which the true value of y, $y^* = 1$ and $N^*_{j,0}$ represents the number of observations for which the true value of y, $y^* = 0$. This allows us to write the last inequality above as

$$\sum_{j \in \{T,C\}} (\eta\ N^*_{j,1} + \omega\ N^*_{j,0}) \left(1 - \frac{\eta\ N^*_{j,1} + \omega\ N^*_{j,0}}{N^*_j}\right) - (\eta - \omega)^2 \sum_{j \in \{T,C\}} N^*_{j,1} \left(1 - \frac{N^*_{j,1}}{N^*_j}\right) > 0 \implies$$

$$\sum_{j \in \{T,C\}} (\eta\ N^*_{j,1} + \omega\ N^*_{j,0}) \left(\frac{(1-\eta)N^*_{j,1} + (1-\omega)N^*_{j,0}}{N^*_{j,1} + N^*_{j,0}}\right) - (\eta - \omega)^2 \left(\frac{N_{j,1^*}\ N^*_{j,0}}{N^*_{j,1} + N^*_{j,0}}\right) > 0 \implies$$

$$\sum_{j \in \{T,C\}} \frac{\eta(1-\eta)N^{*2}_{j,1} + \omega(1-\omega)N^{*2}_{j,0} + N^*_{j,1}N^*_{j,0}\left[\eta(1-\omega) + (1-\eta)\omega - (\eta-\omega)^2\right]}{N^*_{j,1} + N^*_{j,0}} > 0 \implies$$

$$\sum_{j \in \{T,C\}} \frac{\eta(1-\eta)N^{*2}_{j,1} + \omega(1-\omega)N^{*2}_{j,0} + N^*_{j,1}N^*_{j,0}\left[\eta + \omega - \eta^2 - \omega^2\right]}{N^*_{j,1} + N^*_{j,0}} > 0$$

All terms in the numerator of the last inequality are greater than zero, satisfying the condition.

E Treatment Heterogeneity Correlated With Matching Error

In this section we demonstrate how treatment effect heterogeneity that's correlated with linking error can impact coefficient estimates. Consider a dichotomous covariate G which takes on two values M and F. We assume that linking error rates within group are equal across treatment and control but that $TPR_M \neq TPR_F$ and $FPR_M \neq FPR_F$. Further, assume that $\tau_M \neq \tau_F$. We rewrite Equation 4 as:

$$
\begin{aligned}
\hat{\tau} = & \sum_{j \in \{0,1\}} \sum_{g \in \{M,F\}} P(y_i = 1, y_i^* = j, G_i = g | T_i = 1) \\
& - P(y_i = 1, y_i^* = j, G_i = g | T_i = 0) \\
= & \sum_{j \in \{0,1\}} \sum_{g \in \{M,F\}} P(y_i = 1 | y_i^* = j, G_i = g, T_i = 1) P(y_i^* = j | G_i = g, T_i = 1) P(G_i = g | T_i = 1) \\
& - P(y_i = 1 | y_i^* = j, G_i = g, T_i = 0) P(y_i^* = j | G_i = g, T_i = 0) P(G_i = g | T_i = 0) \\
= & P(M)(TPR_M - FPR_M)\tau_M + P(F)(TPR_F - FPR_F)\tau_F
\end{aligned}
$$

When both τ_M and τ_F are in the same direction, then linking error will only attenuate the pooled treatment effect in absolute value. However, if the signs of τ_M and τ_F are different, then the observed treatment effect may be greater than the true average treatment effect in absolute value.