

# No ground truth? No problem: Improving administrative data linking using active learning and a little bit of guile

Sarah Tahamont<sup>\*1</sup>, Zubin Jelveh<sup>1,2</sup>, Melissa McNeill<sup>3</sup>, Shi Yan<sup>4</sup>, Aaron Chalfin<sup>5</sup>,  
and Benjamin Hansen<sup>6</sup>

<sup>1</sup>Department of Criminology and Criminal Justice, University of Maryland

<sup>2</sup>College of Information Studies, University of Maryland

<sup>3</sup>University of Chicago Crime Lab

<sup>4</sup>School of Criminology & Criminal Justice, Arizona State University

<sup>5</sup>Department of Criminology, University of Pennsylvania

<sup>6</sup>Department of Economics, University of Oregon and NBER

March 10, 2022

## Abstract

While linking records across large administrative datasets [“big data”] has the potential to revolutionize empirical social science research, many administrative data files do not have common identifiers and are thus not designed to be linked to others. To address this problem, researchers have developed probabilistic record linkage algorithms which use statistical patterns in identifying characteristics to perform linking tasks. Naturally, the accuracy of a candidate linking algorithm can be substantially improved when an algorithm has access to “ground-truth” examples — matches which can be validated using institutional knowledge or auxiliary data. Unfortunately, the cost of obtaining these examples is typically high, often requiring a researcher to manually review pairs of records in order to make an informed judgement about whether they are a match. When a pool of ground-truth information is unavailable, researchers can use “active learning” algorithms for linking, which ask the user to provide ground-truth information for select candidate pairs. In this paper, we investigate the value of providing ground-truth examples via active learning for linking performance. We confirm popular intuition that data linking can be dramatically improved with the availability of ground truth examples. But critically, in many real-world applications, only a relatively small number of tactically-selected ground-truth examples are needed to obtain most of the achievable gains. With a modest investment in ground truth, researchers can approximate the performance of a supervised learning algorithm that has access to a large database of ground truth examples using a readily available off-the-shelf tool.

---

\*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

One of the more exciting recent developments in empirical social science research is the increasing availability of large administrative databases and the ability to link across them to generate new insights. Indeed linked datasets have allowed researchers working in descriptive, predictive, and causal modalities to generate systematic inferences about large numbers of individuals (1; 2; 3; 4; 5; 6; 7). Unfortunately, most administrative datasets are not designed to be linked to others and thus have no common and reliable unique identifiers. Most often, researchers must attempt to link individuals using their demographic characteristics, such as name, date of birth, race, gender, and address. Typically, only a subset of this information is available and the data are subject to missingness, transcription errors and other systematic imperfections (8; 9). Accordingly, the linking process is often difficult and fraught with error, particularly in domains like criminal justice and social service provision where the manner in which the data are collected makes them especially susceptible to recording errors (10; 11; 12).

An old adage in data analysis and statistical modeling highlights the self-evident but foundational principle that empirical estimates from a model are only as good as the underlying data used to generate them. In the data linking context, as researchers make errors in linking between administrative datasets, the analytic samples used to generate estimates of interest will also contain errors. Unfortunately, the consequences of even modest linking errors can be severe. Random linking errors lead to attenuated estimates of a treatment effect of interest (9). While attenuation bias may not lead to important qualitative distortions in reasonably large datasets, in the many applications in which statistical power is constrained by the cost or availability of data — including, notably, the vast majority of randomized field experiments — the consequence will be an increase in the probability of a false negative or a Type II error. Given the tendency for underpowered studies to end up in the “desk drawer,” Type II errors can have devastating consequences not only for individual research studies but also for entire literatures (13; 14; 15).

Recognizing the fundamental importance of data linking to applied research, statisticians and computer scientists have, for many years, worked to develop probabilistic record linkage algorithms which use statistical patterns in identifying characteristics to perform linking tasks (16). This research has established that probabilistic linking algorithms in which cases are considered to be a match when they exceed some threshold probability of

---

being a true link outperform the seemingly “conservative” approach of exact matching (9). Consequently, a great deal of work aims to improve the performance of such algorithms.

While, in principle, many of these algorithms do not require prior information on which two records refer to the same person, *supervised learning* algorithms — the most powerful approach to data linking — require a pool of “ground-truth” data in order to train the algorithm. With a large enough training sample, the algorithm can learn from patterns in the data and deduce which records are likely to match one another in unseen data. Unfortunately, ground truth data is often unavailable and the cost of obtaining ground truth examples can be high, often requiring a researcher to manually review pairs of records to make an informed judgement about match status. The time-consuming nature of the task makes it prohibitively costly to generate a large volume of ground-truth examples to improve algorithmic performance, if none already exist.

An alternative approach to data linking is to use an *active learning* algorithm. Like supervised learning, the goal is to generate ground truth labels upon which the linking algorithm can be trained. However, the approach is far more surgical. Whereas supervised learning requires that the research team have access to a relatively large ground truth database, active learning algorithms request labels from the researcher for high leverage cases and learn from a much smaller number of ground truth labels. Typically, the algorithm presents a researcher with examples to label in an optimized way, starting with the example whose label would yield the most information gain and working down the list of unlabeled examples from there. The researcher can label as many or as few examples as they want, using either their own time constraints or the diminished utility of labeling an additional example as a stopping criteria. While this approach is thought to be an improvement over linking methods that do not use ground truth at all, the consequences on linking quality are unknown. In particular, we may be concerned that the number of ground-truth examples needed to maintain linking performance below a particular maximal error rate might grow as function of the size of the data sets and the relative frequency of mismatches between the data sets involved. Simply put, the concern is that the potential benefits of an active learning approach are necessarily limited by the paucity of ground truth data that active learning algorithms can access.

In this paper, we investigate the value of ground-truth examples for linking performance. Using a popular open-source record linkage algorithm, *dedupe* (17), we explore how the

---

payoff to investing in ground truth labels varies based on linking context, such as the difficulty of the linking task, the size of the training dataset, and the size of the datasets involved in the linking task itself. Specifically, we fix the share of true matches that do not match exactly on name and date of birth at 50 percent (which is likely at the “messier” end of data sets we have observed in the field) and vary the following parameters to explore the responsiveness of the linking error rate to the difficulty of the matching task: the size of data set that was used to train an algorithm, and the size of the pool of potential records.

We confirm the common sense finding that data linking can be dramatically improved with the availability ground truth examples. But critically, in many applications, only a relatively small number of tactically-selected ground-truth examples are sufficient to obtain most of the achievable gains. We show, that in the context of a common linking task, with a modest investment in obtaining ground truth examples, researchers can approximate the performance of a sophisticated supervised learning algorithm that has access to a large database of ground truth examples using simple and readily available off-the-shelf tools. In many applications, including those which are especially common in empirical social science, a small amount of effort can go a long way towards improving data linking performance.

## 2 Data and Methods

In the present study, we have ground-truth information on true matches, which allows us to report true error rates in order to assess linking performance. Since there may be certain true matches that will be very difficult for any algorithm to find (e.g. if the identifying information is completely different across records), we also estimate a proxy for the “achievable” error rate via an algorithm that has access to all ground-truth examples. To do so, we employ the Name Match Python package and compare the error rate from Name Match to the rates from dedupe.

Our test linking data is a set of identified administrative records on three million charges filed in Oregon courts between 1990-2012, maintained in the Oregon Judicial Information Network (OJIN). The individual records in the OJIN data contain a number of relevant data points: name, date of birth, race, and gender of the defendant; the date the case was filed, a description of the charge, and a unique identification number that links the same defendants across rows in the data set. As shown in Table ??, around 77% of the sample

---

is White, while Hispanic and Black individuals each makes up around 10% of the sample. Nearly 80% of the individuals are male, and the average age is around 32.6 years (SD = 11.1 years). We use the unique identifier as our indicator of ground truth to assess whether the algorithms are able to correctly identify links at the person level. We collapse the file to the case level, because there is no variation in a person’s identifying information across charges within a case. This leaves us with 1.3 million cases.

## 2.1 The Linking Task

The record linkage task we are simulating using these data is a link between a small data set from a research experiment, which we’ll call  $E$  (one record per person), and a larger administrative data set,  $D$  (potentially multiple records per person). Our set up mirrors a common linking task in a number of settings, such as linking the names of youth employment program participants to the list of arrestees to examine the crime-reduction effect of employment, or linking the list of election candidates to the list of corporate directors to study if politically connected business had easier access to loans (18; 19). We further subdivide  $E$  and  $D$  into training  $E_t$  and  $D_t$  and evaluation  $E_e$  and  $D_e$  portions, since we do not want to evaluate results of a learning algorithm on in-sample data. We also ensure that there is no overlap between the training and evaluation data sets. In the specifications to follow, the size of the training and evaluation sets are the same. Experimental data sets  $E_t$  and  $E_e$  are always set to 1,000 unique individuals. The administrative data sets vary between 5,000, 50,000, and 200,000 records, where  $D_t$  and  $D_e$  are the same number of records in each specification. In other words, in every specification, we can think of our objective as attempting to link a person-level experimental data set of 1,000 observations to a case-level data set of either 5,000, 50,000 or 200,000 records.

Despite all the special sampling we have to do to generate each of these data sets, we want to hold the distribution of records per person constant at values reasonably similar to the full data set’s natural distribution. The distribution of records per person in the administrative data  $D$  is as follows:

---

Records Per Person	Proportion
1	.50
2	.40
3	.06
4	.03
5	.01

We fix the true match rate, meaning the rate of individuals in the experimental data  $E$  who have a true match in the administrative data at 50% to facilitate performance evaluation. In other words, 50% of the people in the experimental data set have one or more records in the administrative data set. If our linking algorithm performs perfectly, we would expect to find a link in  $D$  for the 500 people from  $E$  that have one. Among these true matches, the share of matches where the records in the experimental data  $E$  and the administrative data  $D$  match exactly on first name, last name and date of birth is approximately 50% for both the training and evaluation sets. Meaning that of the 1,000 individuals in  $E$ , 500 have a record in  $D$  and 250 of those have a record in  $D$  that matches the information in  $E$  exactly on first name, last name, and date of birth.

This particular type of linking task is becoming increasingly common in the social sciences, particularly as it relates to the proliferation of the “low-cost” randomized controlled trial, in which a data set of participants from an intervention is linked to existing administrative databases to capture outcome data in one or more domains (20).

## 2.2 Running dedupe

The dedupe tool can perform record linkage tasks using either a traditional record linkage framework, in which the administrative database  $D$  is searched for each of the records in the experimental data  $E$ , or a de-duplication linkage framework in which the records in the experimental data  $E$  are appended to the administrative data set  $D$  making one large database with the number of records equal to the sum of the records from  $E$  and  $D$ . From there, links are identified via de-duplication. To streamline the presentation of results here, we present results for the de-duplication linkage framework, because it tends to have better performance in this context than the traditional record linkage framework. Although the patterns we discuss hold for both kinds of linkage frameworks (see Appendix ?? for the traditional record linkage framework results).

Evaluating all possible pairs of records as potential links is often too computationally

---

intensive to be feasible in real-world linking tasks. As a result, record linkage processes begin with a blocking step designed to reduce the number of comparisons required during linking by eliminating those that have low likelihood of being true matches. During blocking, records are organized into groups that consist of broadly similar records according to particular criteria. For example, blocking on birth year would mean that only records with the same birth year will be considered as potential links. The goal is to choose blocking criteria to discard as many nonlinks as possible without discarding true links. Once the data is organized into blocks, the linking task proceeds by assessing pairs of records within blocks to find potential links. This restriction of the comparison space can greatly reduce runtime of the linkage task, but often forces the researcher to make an a priori decision about which blocking rules to impose. Fortunately, with dedupe the optimal blocking rules are part of what is learned via the active learning process, with dedupe learning blocking rules to maximize the share of true pairs that make it past the blocking step. Accordingly there are no a priori decisions that the user has to make about what defines a block before the linking process begins.

This active learning component, which typically requires users to label some number of pairs as either a match, non-match, or unknown, is a critical component of dedupe’s linking process. We simulate human input by feeding dedupe the labels it requests from ground truth data in order to ensure consistency. By automating the active learning process using ground truth data, we are evaluating dedupe’s performance under optimal conditions, in which there are no errors in the labeling process. In other words, we are assessing dedupe’s performance under the condition that what it learns from the labeling process is entirely correct. The active learning component of dedupe asks the user to label training pairs until it reaches a minimum of 10 positive and 10 negative labels. While dedupe does not limit the training process, the website notes that it should be rare to need more than 50 positive and 50 negative cases to train the algorithm on even the most difficult and messy data sets (21).

There are parameters in dedupe that control the set of pairs that are available during the active learning step of the linking process. The “sample size” parameter determines the number of label-eligible record pairs and the “blocked proportion” parameter influences what share of label-eligible record pairs are similar pairs vs. random pairs. We will show that in the context of our linking task, performance was highly sensitive to the choice of the

---

sample-size parameter. In contrast, our results were not sensitive to variations in blocked proportion. As a result, we fixed blocked proportion at the deduplication framework’s default value of 0.90. A value of 0.90 means that 90% of all the record pairs that the user could be asked to label are records that were similar enough to appear in the same block at least once. The remaining 10% are random pairs of records that, most likely, look nothing alike. Finally, we also make a minor adjustment to the train function. The “recall” parameter is set to 1 by default, meaning that the algorithm must find blocking rules such that 100% of labeled matches are blocked together. We adjusted the recall parameter to .99, because occasionally there are real world cases where an individual actually has two very different looking records. We run dedupe four times for each specification of our linking task in order to observe any variation in linking performance that may occur due to random sampling performed by the algorithm.

We ran dedupe with access to 8 Intel Xeon Platinum 8180 CPUs and 700 gigabytes of RAM. While we did not conduct an extensive run-time analysis, the relatively short duration of even the longest dedupe run (an average of 134 minutes for a specification with a budget of 1,000, an administrative dataset size of 200,000, and a sample size setting of 150,000), suggests that dedupe’s runtime is a very reasonable investment for the purposes of constructing an analytic data set.

### 2.3 Benchmarking Using a Supervised Learning Algorithm

Since administrative data linking is a noisy task under the best conditions, solely comparing dedupe’s performance to the true link rate is a stringent standard. Accordingly, we also conducted a benchmarking exercise so that we had a point of comparison for dedupe’s performance that was more indicative of what might be possible for a linking algorithm that had access to all the ground truth information, which is likely the best possible performance under real world conditions. For each combination of experimental data set  $E$  and administrative data set  $D$  we also used a supervised learning algorithm, Name Match, to link the data in order to demonstrate what linking performance we might reasonably expect with access to all of the ground truth information available, as opposed to the limited number of ground truth labels that dedupe uses to learn the linking task.



---

## 2.4 Evaluation Metrics

There are a number of candidate performance metrics we can use to assess the performance of record linking tools. We focus on three evaluation metrics: precision, recall, and total error. Precision and recall are commonly used to evaluate the performance of information retrieval and prediction tasks, because they each focus on a different type of error. In the context of data linkage, high precision indicates fewer false positives. In other words, the algorithm is not linking records that belong to two different individuals. Whereas, recall focuses on false negatives. High recall means that the algorithm is not missing the true links that do, in fact, belong to the same individual. While, ideally, an algorithm has both - high precision and high recall - there is an inherent trade off between the two: fewer false positives typically results in more false negatives, and vice versa. In some cases, the relative costs of false positive and false negative links varies, so it is useful to assess these two measures separately. The total error metric – which is a combination of the false positive rate (FPR) and the false negative rate (FNR) – is also key because the bias introduced by linkage errors is a function of the total linking error (9). Naturally, to minimize the bias introduced by the linking process, we want to minimize the total error metric.

## 3 Results

### 3.1 Off the Shelf Performance

We begin by assessing dedupe’s “off the shelf” performance in the task of linking the experimental data set,  $E$ , to administrative data sets,  $D$ , of varying sizes. Figure 1 presents the results of this exercise using dedupe’s deduplication framework to link  $E$  to  $D$ . We perform the task with dedupe’s default settings: the “sample size” parameter set to 1,500 label-eligible pairs and blocked proportion set to 0.90. We compare dedupe’s performance to that of a sophisticated supervised learning algorithm (Name Match).

In each of the plots, the dashed line represents the performance of the supervised learning algorithm (Name Match). Each of the points represents a separate run of dedupe. (Note that some of the runs had nearly identical performance, such that they are not individually distinguishable in the figure.) The plots in Columns 2 ( $n=50,000$ ) & 3 ( $n=200,000$ ) of Figure 1 had extraordinarily high total error rates, in some cases approaching 100 percent. For example, the best performance for 50,000 admin rows was 33 percent total linking error

---

even after feeding the algorithm 1,000 accurate ground truth labels. The results in Column 3 reveal even larger error rates. Here the best total error rate was 54 percent — for the task of linking an experimental data  $E$  of 1,000 individuals to an administrative data set  $D$  containing 200,000 rows with 1,000 accurate ground truth labels. This is a concerning result given that many administrative data sets that might be targeted for linking tasks are quite large. In order to investigate the reason for such high error rates in Figure 1, we next examine the sensitivity of the results to varying each of dedupe’s hyperparameters (i.e. the set of adjustable parameters).

### 3.2 “Sample Size” Implications

We found that linking performance using dedupe is highly sensitive to the “sample size” parameter as the size of the administrative data set  $D$  increases. In other words, for larger administrative data set sizes, dedupe needs a larger pool of candidate pairs from which it can select pairs to be labeled during the active learning component of the linking process. Figure 2 shows the performance of the linking task for “sample sizes” of 1,500 (default), 15,000, 150,000 and 300,000 candidate pairs. With respect to performance, all of the sample sizes performed similarly with respect to precision (Figure 2, Panel B). While precision performance is not very sensitive to sample size, increasing the number of candidate pairs to which dedupe has access when it is looking for candidates to label has large effects on recall performance (Figure 2, Panel C). As Panel C, Column 3 shows, recall never exceeded 40 percent on average even when we provided accurate labels for 1,000 of the 1,500 candidate pairs. By contrast, for all three specifications, a sample size of 300,000 candidate pairs performed well on both precision and recall and, by extension, total error. However, there are costs to the speed of the process by substantially increasing sample size. To balance run time costs and performance, we shifted to a sample size of 150,000 for all specifications.

### 3.3 Main Results

Figure 3 presents the performance of our linking task with the larger “sample size.” A couple important findings emerge. First, when we look at the plots in Figure 3 we see that many more of the dots are filled in compared to the dots on the plots in Figure 1, indicating that the runs were much more likely to meet the baseline training threshold of 10 positive and 10 negative links during the labeling process when dedupe is given access to

---

a larger pool of candidate links during the labeling process. However, we also see that even conditional on meeting the training criterion of 10 positive and 10 negative labels dedupe’s performance improves when we provide more ground truth labels *up to a point* when it starts to flatten out around 200 labels. We see from the figure that there are diminishing returns to providing ground truth labels after approximately 200 labels. This is a key result because it means that in linking contexts when there are no pre-existing ground truth labels available, linking using active learning performs well with reasonable investments in ground truth labels. The performance improvements are largely driven by improvements in recall performance. We note that with at least 200 ground truth labels, the performance of dedupe is in the neighborhood of the performance of a supervised algorithm given the same task. Importantly, we are not suggesting that active learning with a modest number of labels approximates the linking performance we might expect from a supervised linking process in all linking cases. However, in this specific case, we are able to approximate the performance of a supervised learning algorithm with a small number of ground truth labels. While the case we set up is specific, it is also a very common linking task in empirical social science.

The result is even more remarkable put in context. Figure 3 shows the performance of the supervised learning algorithm, Name Match, as a benchmark for the linking performance we might expect under the best case scenario of access to a large amount of ground truth information. For all of the administrative data set sizes we tested, we are able to achieve performance similar to (or better than) a supervised learning algorithm with access to all the ground truth information available with dedupe’s active learning algorithm and an investment in providing approximately 200 ground truth labels (or fewer in some circumstances). This is a very encouraging result. With a few small adjustments to an open source package, researchers who have data linking needs but little specific linkage expertise can leverage active learning and a relatively small investment in providing ground truth labels to achieve remarkable performance even if their data is quite messy.

## 4 Discussion

In this paper, we investigate the value of ground-truth examples for linking performance. Using a popular open-source record linkage algorithm, *dedupe* (17), we explore how the

---

payoff to investing in ground truth labels varies based on the context and the difficulty of the linking task, such as the size of the training dataset and the size of the datasets involved in the linking task itself. We show that in many applications, only a relatively small number of tactically-selected ground-truth examples are sufficient to obtain most of the achievable gains. With a modest investment in ground truth, researchers can approximate the performance of a sophisticated supervised learning algorithm that has access to a large database of ground truth examples using simple and readily available off-the-shelf tools.

Of course, it is important to discuss some of the limitations of our work. First, the number of labels required to maximize achievable performance is likely to grow with the size of the administrative dataset being linked. Our results speak to linking scenarios using databases that have fewer than 200,000 observations. Next, as we show in this paper, linkage performance is related to the difficulty of the matching task. We have explored one particularly common and important type of linking task in this paper, matching individuals when we have access to full name and date of birth information. To the extent other types of linkages are more or less difficult, that would have implications for the number of ground-truth labels needed to obtain maximum achievable performance. To that end, further avenues for research would include exploring other types of linkages: not having access to name and/or dob in full, or non-person linking such as linking location or organization names. Third, we explored how many labels are required to get back to overall levels of performance. However, there may still be differences in performance by subgroups such as by race or gender.

These limitations notwithstanding, while many applied research projects could benefit from implementing processes like active learning to perform record linkage tasks, there is relatively little practical guidance available. Often applied researchers consider record linkage to be a pre-processing step, rather than an early-stage component of their analytic plan. However, as prior work has demonstrated, record linkage errors can increase bias and diminish statistical power (9). To that end, our paper serves as guide for how researchers can benefit from the power of record linkage without learning data science from scratch. An added benefit is that using a publicly available linking tool promotes transparency and replicability since the underlying code is freely available, which is particularly important for applied researchers who have little formal training in data scientific methods.

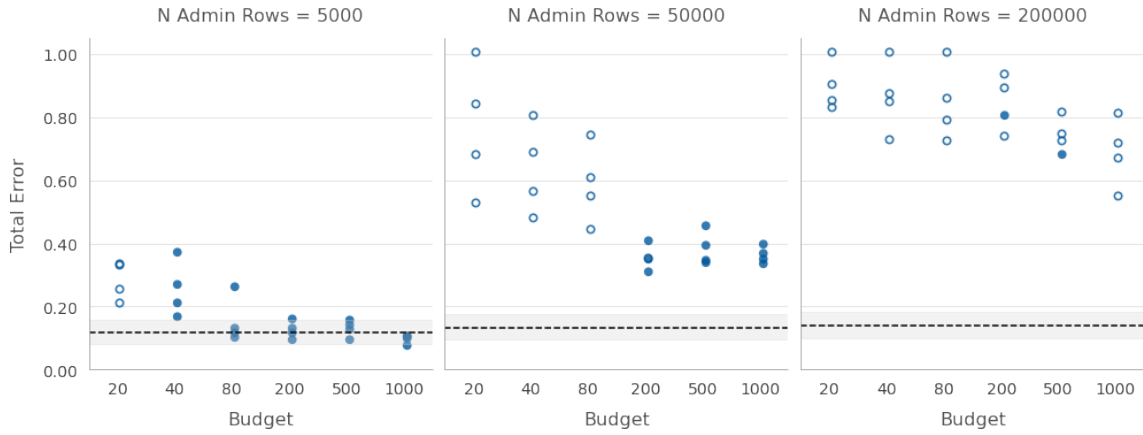
---

## References

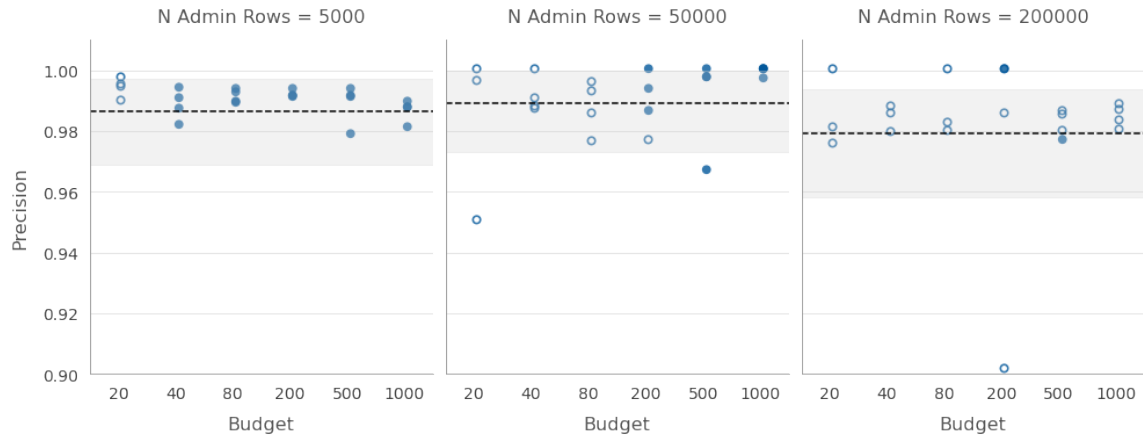
- [1] Einav L, Levin J. Economics in the age of big data. *Science*. 2014;346(6210):1243089.
- [2] Smith GJD, Bennett Moses L, Chan J. The challenges of doing criminology in the big data era: Towards a digital and data-driven approach [Journal Article]. *The British Journal of Criminology*. 2017;57(2):259–274.
- [3] Lane J. Building an infrastructure to support the use of government administrative data for program performance and social science research [Journal Article]. *The ANNALS of the American Academy of Political and Social Science*. 2018;675(1):240–252.
- [4] Lynch J. Not even our own facts: Criminology in the era of big data. *Criminology*. 2018;56(3):437–454.
- [5] O’Brien DT, Sampson RJ. Public and private spheres of neighborhood disorder: Assessing pathways to violence using large-scale digital records. *Journal of research in Crime and Delinquency*. 2015;52(4):486–510.
- [6] Finlay K, Mueller-Smith M. Justice-involved individuals in the labor market since the Great Recession. *The ANNALS of the American Academy of Political and Social Science*. 2021;695(1):107–122.
- [7] Finlay K, Mueller-Smith M, Street B. Myself as the Employer of Last Resort: Criminal Justice Involvement, Self-employment, and Barriers in Recent Public Policy. 2021;.
- [8] Gross M, Mueller-Smith M. Modernizing Person-Level Entity Resolution with Biometrically Linked Records. Working paper; 2020.
- [9] Tahamont S, Jelveh Z, Chalfin A, Yan S, Hansen B. Dude, Where’s My Treatment Effect? Errors in Administrative Data Linking and the Destruction of Statistical Power in Randomized Experiments [Journal Article]. *Journal of Quantitative Criminology*. 2021;37(3):715 – 749. Available from: [<GotoISI>://WOS:A1994NM32300001](#).
- [10] Ferrante A. Developing an offender-based tracking system: The Western Australia INOIS project [Journal Article]. *Australian and New Zealand Journal of Criminology*. 1993;26(3):232–250. Available from: [<GotoISI>://WOS:A1993MN25000005](#).

- 
- [11] Geerken MR. Rap sheets in criminological research: Considerations and caveats [Journal Article]. *Journal of Quantitative Criminology*. 1994;10(1):3–21. Available from: [<GotoISI>://WOS:A1994NM32300001](https://www.isi.com/).
- [12] Orchowsky S, Iwama J. Improving state criminal history records: Recidivism of sex offenders released in 2001. Justice Research and Statistics Association; 2009.
- [13] Easterbrook PJ, Gopalan R, Berlin J, Matthews DR. Publication bias in clinical research. *The Lancet*. 1991;337(8746):867–872.
- [14] Andrews I, Kasy M. Identification of and correction for publication bias. *The American Economic Review*. 2019;109(8):2766–94.
- [15] Brodeur A, Cook N, Heyes A. Methods matter: P-hacking and publication bias in causal analysis in economics. *The American Economic Review*. 2020;110(11):3634–60.
- [16] Fellegi IP, Sunter AB. A theory for record linkage [Journal Article]. *Journal of the American Statistical Association*. 1969;64(328):1183–1210.
- [17] Gregg F, Eder D. Dedupe; 2019. <https://github.com/dedupeio/dedupe>.
- [18] Gelber A, Isen A, Kessler JB. The effects of youth employment: Evidence from New York city lotteries. *The Quarterly Journal of Economics*. 2016;131(1):423–460.
- [19] Khwaja AI, Mian A. Do lenders favor politically connected firms? Rent provision in an emerging financial market [Journal Article]. *The Quarterly Journal of Economics*. 2005;120(4):1371–1411.
- [20] Shankar M. How Low-cost Randomized Controlled Trials Can Drive Effective Social Spending; 2014.
- [21] Dedupe. Frequently asked questions (FAQ);. Available from: <https://dedupe.io/documentation/faq.html>.

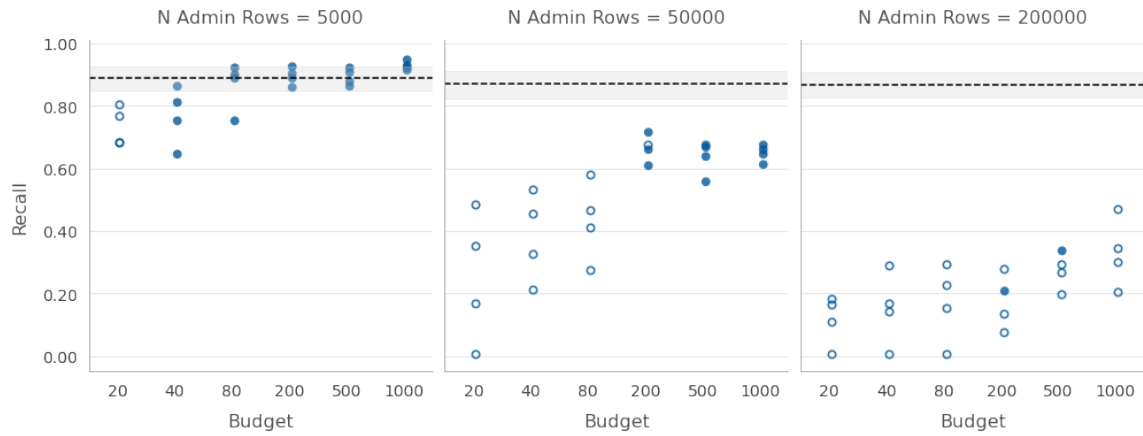
**Figure 1:** Performance of dedupe, with "off the shelf" setup



(a) Total errors using dedupe, with default



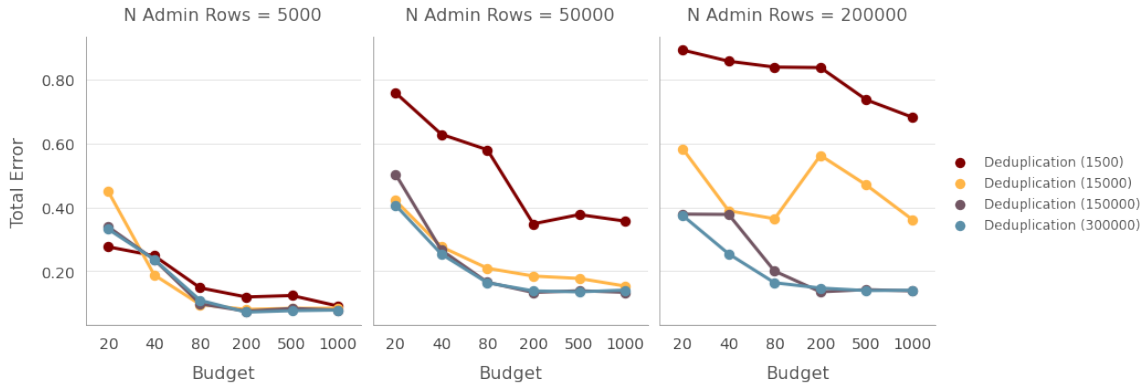
(b) Precision using dedupe, with default



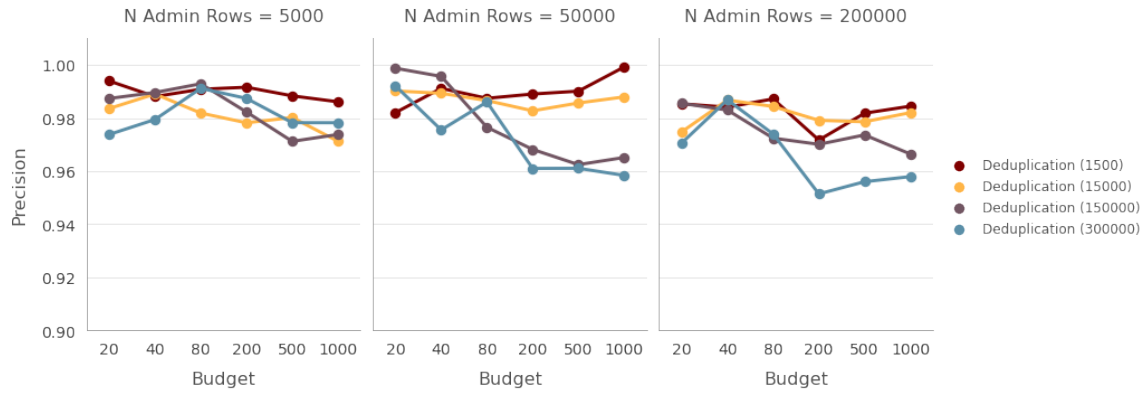
(c) Recall using dedupe, with default

*Note:* Dashed line represents the performance of Name Match. Each dot represents a dedupe run. Each filled dot represents a run where 10 positive and 10 negative links were obtained.

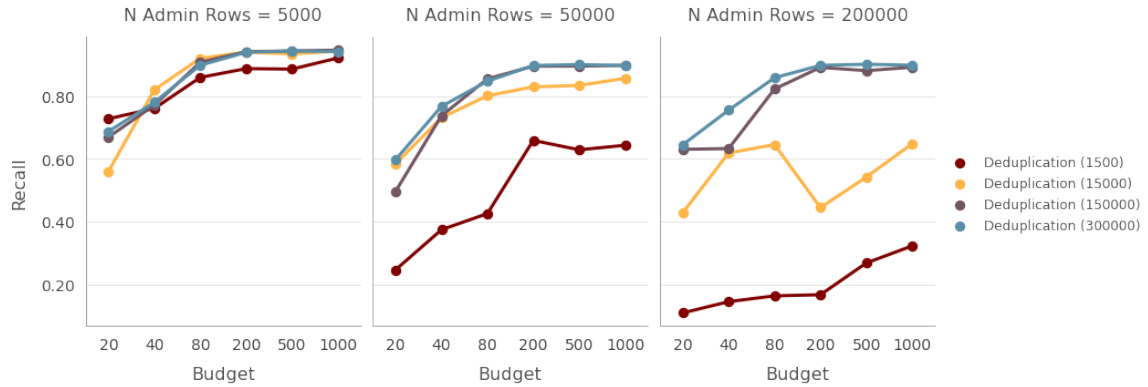
**Figure 2:** Performance for different sample sizes



**(a)** Total errors for different sample sizes



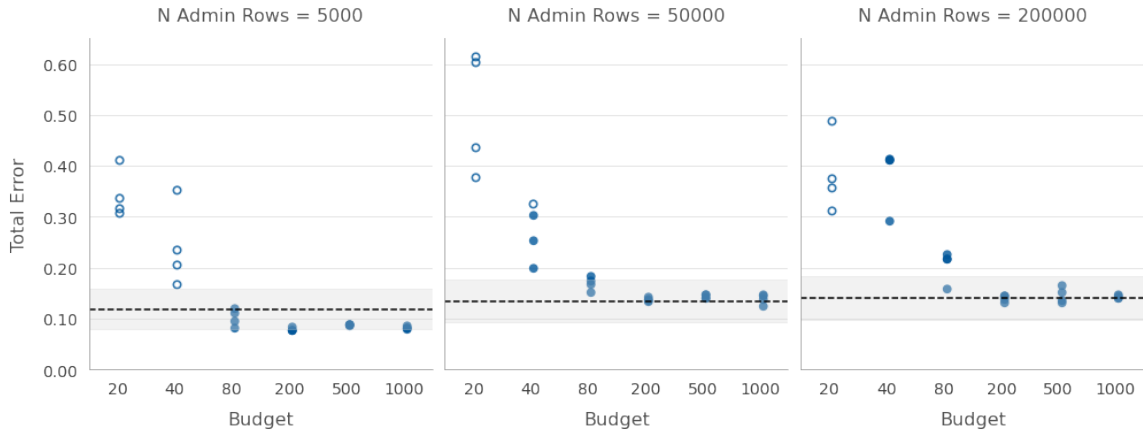
**(b)** Precision for different sample sizes



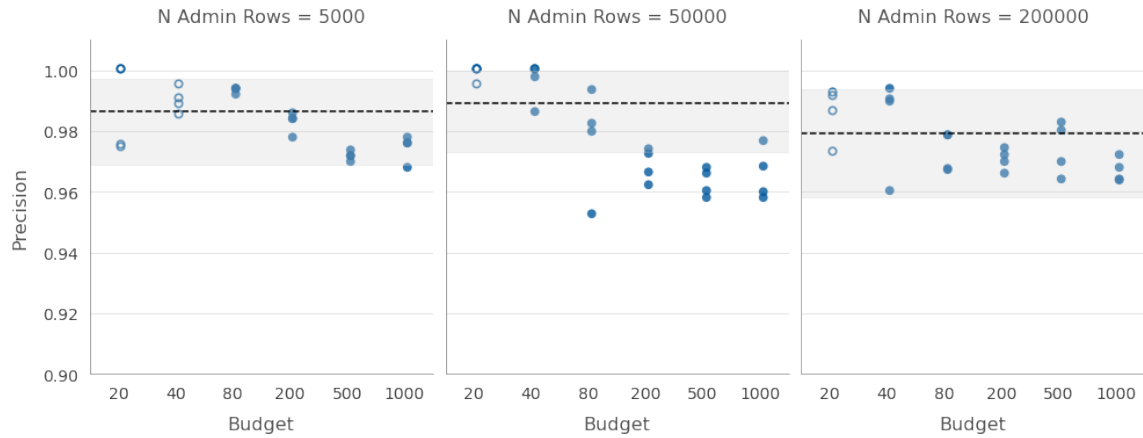
**(c)** Recall for different sample sizes



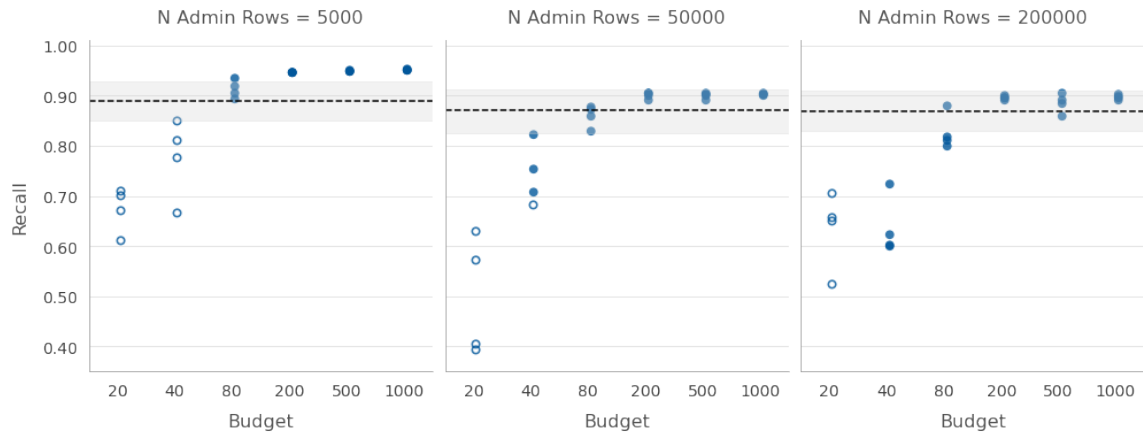
**Figure 3:** Performance of dedupe, not default



(a) Total errors using dedupe, not default



(b) Precision using dedupe, not default



(c) Recall using dedupe, not default

*Note:* Dashed line represents the performance of Name Match. Each dot represents a dedupe run. Each filled dot represents a run where 10 positive and 10 negative links were obtained.

---

## Supporting Information Captions

**S1 Appendix. Traditional Record Linkage Framework Performance.**

**S2 Figure. Performance under the deduplication framework, sample size 150,000, with bootstrapped 95% confidence interval for NameMatch performance.**

**S3 Figure. Performance under the traditional record linkage framework, sample size 150,000, with bootstrapped 95% confidence interval for NameMatch performance**