ADGN: An Algorithm for Record Linkage Using **A**ddress, **D**ate of Birth, **G**ender and **N**ame

Stephen Ansolabehere and Eitan D. Hersh *

November 16, 2016

Abstract

This paper presents an algorithm for record linkage that uses multiple indicators derived from combinations of fields commonly found in databases. Specifically, the quadruplet of Address (A), Date of Birth (D), Gender (G), and Name (N) and any triplet of A-D-G-N (i.e., ADG, ADN, AGN, and DGN) also link records with an extremely high likelihood. Matching on multiple identifiers avoids problems of missing data, inconsistent fields, and typographical errors. We show, using a very large database from the State of Texas, that exact matches using combinations A, D, G, and N produce a rate of matches comparable to 9-Digit Social Security Number. The rate of false negatives appears to be less than 1 percent. Further examination of the linkage rates show that reporting of the data at a higher level of aggregation, such as Birth Year instead of Date of Birth and omission of names, makes correct matches between databases highly unlikely, protecting an individual's records.

*Stephen Ansolabehere is the Frank G. Thompson Professor of Government, Harvard University, 1737 Cambridge Street, Cambridge MA, 02138. Eitan D. Hersh is Assistant Professor of Political Science, Yale University, 77 Prospect Street, New Haven CT, 06520.

1 Introduction

In 2011, the Texas State Legislature passed Senate Bill 14 (S. B. 14), which required voters to show one of six forms of government-issued photo identification: a state driver's license or ID card, a concealed handgun license, a U. S. passport, a military ID card, or a U. S. citizenship certificate with a photo. The law carved out exceptions for older people and people with disabilities, but it was one of the strictest voter identification laws in the country[1]. The Department of Justice refused to grant pre-clearance of the law under Section 5 of the Voting Rights Act. Texas sued, and a four-year legal battle began. Texas lost that original suit, *Texas v. Holder*, as well as a subsequent suit under Section 2 of the Voting Rights Act, *Veasey v. Abbott*, of intentional racial discrimination. To gauge how many voters lack the requisite ID and whether the law would have racially disparate effects, both the State of Texas and the Department of Justice conducted database-matching of the 13 million records on the Texas Election Administration Management (TEAM) voter file with the 25-million-record file of State of Texas IDs and various national lists maintained by the federal government.

The Texas registration-to-ID empirical investigation is one of the largest database matching efforts ever conducted in which information about the match rates and uniqueness of specific identifiers is publicly available. It offers a unique picture of the identifiability of individuals and the matchability of public databases using commonly recorded personal information, such as names and dates of birth. As a result, what we learn about the identifiability of individuals from this example not only sheds light on the policy questions about voter ID raised by the particular court case, but it also sheds lights on policies concerning data privacy and data management. The empirical task in the Texas case was first to estimate what percent of people would be affected by S. B. 14 (i.e. likely voters who lack a

[1]See, e.g. Manny Fernandez and Erik Eckholm, "Federal Court Rules Texas' ID Law Violates Voting Rights Act," *New York Times*, July 20, 2016.

qualifying ID) and then to estimate whether the implementation of the law was more likely to affect registered Black and Hispanic voters than registered White voters. What we learn about the ability to link records with a high degree of accuracy is informative for research and applications in a range of areas from voting rights to public health to criminology to marketing to government censuses.

The social science problem presented in the Texas case is, in fact, one of the most important database management issues today. How can researchers link records across multiple databases with personal information in the absence of a unique common identifier, typically 9-digit Social Security Numbers (SSN9)? Some states, such as Georgia, do maintain a common identifier to link records across databases maintained by the state government (Hood III and Bullock III, 2008). But identifiers like these are often unavailable or are sufficiently incomplete that they are not useful for matching. The challenge is how to use personal information, such as names, addresses, and dates of birth, that are available on all files to link the databases in a manner that minimizes the rates of false positives (matches between files that in fact correspond to different individuals) and false negatives (failure to match records on two files that correspond to the same individual) (Fellegi and Sunter, 1969).

Working on behalf of the United States in its litigation against the state of Texas, we developed a robust algorithm, building on prior work by Sweeney (2002), for matching databases using information in the Address (A), Date of Birth (D), Gender (G), and Name (N) fields. The algorithm, first, standardizes databases. Second, multiple identifiers are constructed using information from all four A, D, G, and N fields as well as triplets of those for fields (i.e., ADG, ADN, DGN). Finally, two records are considered a match if there is an exact match to any one of the multiple indicators. Using multiple indicators constructed from ADGN, the voter registration file from the Texas Election Administration Management system (TEAM) was matched to databases of acceptable forms of state and federal identifi-

cation, specifically, the Department of Public Safety (DPS) list of driver licenses, state IDs, and concealed handgun permits, the Department of State list of passports, the Department of Defense lists of military personnel (with active military ID), the Department of Veterans Affairs lists of veteran identification holders, the Social Security Administration disability list, and the Immigration and Naturalization lists of naturalized citizens with immigration papers. The number of records for which no matching record on one of the identification files could be found measured the number of people presumed to lack an acceptable ID under Texas law.

To validate the algorithm, we benefited from one unusual feature of the data: a subset of the voter registration records contained nine-digit Social Security Numbers and/or driver license numbers, identifiers that were also available in some of the databases to which we were linking. This allowed us to determine that 98 percent of records that matched to SSN also matched using some combination of three of A, D, G, and N. Similarly, 98 percent of those that matched using ADGN were subsequently matched using SSN9. Inspection of the data after matching reveal only a small percent of false positives. Of the entire database composed of 13.5 million records, only one-half of one percent were false negatives, cases that were not matched because of insufficient information, but should have been matched.

Using multiple indicators offers a method for minimizing non-matches due to missing data, typographical errors, or variations in names and addresses. No single field determines whether a match occurs. The success of this approach, ultimately, depends on the uniqueness of the identifiers. Even in a large state, such as Texas, combinations of triplets of A, D, G or N proved to have a very high rate of uniqueness. This approach differs from probability-based methods, such as using fuzzy matching in constructing linkage indicators (Guth, 1976; De Brou and Olsen, 1986), or scoring criteria for identifying the likelihood of a match (Berlin and Rubin, 1995; Larsen and Rubin, 2001). Fuzzy matching and probabilistic criteria also are designed to address issues of inconsistently recorded information, missingness and typo-

graphical errors. These methods, however, can be computationally very intensive, and they can reduce the amount of control that researchers have over the linkage process.

Our contribution is threefold. First, we develop specific protocols for string standardization. Second, we posit and demonstrate that multiple linkage indicators improves considerably on methods using single linkage indicators. Third, the frequencies of various fields and the ability to match on them are immediately instructive about the identifiability of individuals using personal information commonly available on databases.

2 Nature of the Problem

2.1 Matchability and Identifiability

Record linkage is performed with two different objectives: matching of entire databases ("matchability") and identification of individuals ("identifiability"). Both of these applications seek to link databases with a high degree of accuracy. By accuracy, we mean that record linkage algorithms seek to minimize false positives (matches of two records that correspond to different individuals) and false negatives (non-matched records when the individual is, in fact, on both lists) (Fellegi and Sunter, 1969; Berlin and Rubin, 1995). The same technology is usually deployed for each purpose, but whether the goal is matchability or identifiability can affect certain trade-offs between minimizing false positives and minimizing false negatives.

The goal of the first objective – matchability – is to measure features of the population represented by that database. For example, the U. S. Census Bureau in 1980 and 1990 matched the Post Enumeration Survey to the Enumeration datafile in order to estimate the size of the population missed by the enumeration (Winkler, 1999, 2006; Larsen and Rubin, 2001). In such research, the effort is to develop an unbiased estimate, so researchers might tolerate some random error in the matching process — that is, some false positives — so long

as those are balanced by false negatives (Berlin and Rubin, 1995). In our effort on behalf of the Justice Department, our goal was analytics: we sought to determine the percent of people on the voter list who have an ID and to determine how that percent varied by racial group. We sought to minimize false positives and false negatives. Some were unavoidable but we sought to reduce them to be well below the observed rate of true non-matches and well below the racial differential in such non-matches.

The second objective – identifiability – is to link databases in order to find an individual and extract that person's information. In that context, the decision rule for what counts as a match may differ. For example, in linking health records across providers, the critical statistic might be the incidence of false positive matches, as any false positive might be life-threatening. Not surprisingly, many of the advances in record linkage and database matching have come out of the fields of health care, but they have spread to other areas (Pacheco et al., 2008; Mohanty et al., 2015). In the social sciences, when linking survey respondents to public records for the purpose of studying individual-level correlates (as opposed to studying population estimates), reducing false positives is typically the priority (e.g. Hersh and Goldenberg (2016); Ansolabehere and Hersh (2012). When the goal is identifiability, the minimization of false positives might be so important that the algorithm should tolerate a much higher rate of false negatives in order to avoid false positives.

In the Texas case, our goal was matchability - to determine what percent of people one one database (TEAM) appear in other databases, and whether the appearance on TEAM but not other database was correlated with race. The lessons of that research carry over to identifiability. In particular, the ability to identify individual records in any database database and to link that individual's records across databases can have both harms and benefits to the person whose information is at stake. There are benefits that come from empowering the individual or the people working on her or his behalf to better manage personal data. There is the threat or harm of identity theft, as might happen if one's

financial records are breached. There is also the harm that might result if private and sensitive information is released, say, revealing private communications to an employer or health records to an insurance company. The ability to identify an individual's information in multiple datasets and to link those datasets, then, has value to the individual to the extent that it allows her or him to manage and/or protect personal information.

2.2 How Unique is My Information?

The ability to link records with a high level of accuracy depends on the frequency with which pieces of information on lists, such as names and ZIP codes, occur in the population. The more unique a characteristic the more power it offers for record linkage. Combining information from different fields increases the power further. Some basic statistics from the Texas data files can help us understand which fields are particularly powerful for matching. The following exercise reveals that information extracted from the address, date of birth gender, and name fields can almost surely identify individuals uniquely.

Before proceeding with this exercise, a note about data. Most of the analysis in this article comes from the expert reports filed in court proceedings and are based on our analysis of the databases from the trial. In preparation for this article, however, we obtained (in October 2016, through a public records request) an additional copy of the Texas voter file. This file allows us to do some additional analysis of the Texas file beyond what we did while serving as experts. Note, however, that the file we obtained in 2016 is not the full statewide list of 13.5 million registrants. Instead, Texas transmitted the vote history file of 7.9 million individuals. This file does not list individuals who are registered but failed to cast ballots in recent elections. For the illustrative purposes below (i.e. to study the uniqueness of various fields), a database of 7.9 million records is sufficiently large (larger than the complete voter files of all but a handful of states) to make the point.

First, consider just one piece of information from addresses - five-digit ZIP Code (ZIP5).

In the State of Texas, there are 2,257 five-digit ZIP Codes listed in the 2016 voter list. Not every ZIP5 contains exactly the same number of voters. The most common ZIP Code is 75070, which has 32,724 persons. The five-digit zipcode alone reduces the dimensionality of the problem tremendously, from 1 in 8 million to 1 in 33,000. The uniqueness of address information can be refined further by using street number (that is the number of the building in which a person lives). The most common street number in Texas is #201. 16,199 Texas voters, or 0.2 percent, have that street number in the state of Texas. All other numbers are less common. The most common ZIP5-street number combination is street number 6 in ZIP code 77381, which is shared by 558 persons. All other combinations are less common.

Second, consider Date of Birth (DOB). The most common Birth Year on the Texas voter file is 1957, which contains approximately 1 in 50 people. The most common Date of Birth is August 30, 1960, which 652 persons share.

Third, consider Gender. Knowing an individual's gender cuts the matching problem in half.

Fourth, consider Name. The most common last name (ignoring first names) in the State of Texas voter file is Smith, which 73,971, or a little less than 1 out of 100, persons have. Matching on last name is quite powerful. In addition, first names and middle names refine the name field further. The most common First Name-Last Name combination in Texas is Maria Garcia, held by 1,914 voters. Adding the middle initial, the most common combination is Maria D. Garcia, a named shared by 296 of 7,880,488 voters.

Combining these four pieces of information, assuming they are independent, reveals that the ZIP5, Gender, DOB, and Last Name of an individual is unique to 1 is 2.7 billion individuals. The risk of a non-unique individual identified by these characteristics, then, is less than .003.[2] Hence, Address-DOB-Gender-Name (just last name) will almost surely identify every individual on a large data file, such as the State of Texas Voter File and Department of

[2]That is, (1/2.7 billion)/(1/7.9 million).

Public Safety Identification file, and can be used to link records across the databases. Even three pieces of information can almost surely identify people.

If ZIP Code, gender, date of birth, and last name were always available and always consistent across databases, our work would be done. Alas, in practice, a linkage strategy must deal with three common data recording issues: missing fields, inconsistent data fields, and typographical errors. Incomplete or missing fields will mean that the complete Identifier A-D-G-N cannot be constructed. If Date of Birth is missing, the D-portion is blank. If an exact match is the criterion, then a false negative may be generated because no match is possible on account of the missing data.

In the October 2016 voter list from Texas, out of 7,880,488 unique individuals, 19,172 were missing a residential ZIP code (though many of these had a mailing ZIP code), 16,208 were missing a street number, none were missing a listed birthdate, though, for instance, 5,196 individuals had a birthday listed as January 1, 1900. Some individuals were missing gender, but gender is easily imputable by first names. For the records containing gender, we calculate the percentage of each first name that is female, and assign female to names that are likely to be female but who are listed without a gender designation. We do the same for male names.

Inconsistent data fields arise for many reasons, such as alternative spellings of names, nicknames, name changes, and alternative addresses. A person may use a nickname on one file and a proper first name in another file. Name inconsistencies arise often with first names, and also when people marry and take the names of their spouses. Address fields also exhibit inconsistencies from one file to another. A person may use a home address in one file and a business address in another. Address fields, such as street suffixes and apartment numbers, are not treated similarly across databases. The data are correct; the individual may simply be identified different ways.

Typographical errors and erroneous fields arise because of keystroke errors in databases.

They may occur at a low frequency, but will reduce the accuracy of the matching process. They may be particularly common in databases like voter registration files, where citizens often submit hand-written applications for registration and clerks record the fields into a database. Errors and inconsistencies in fields will create false negatives, non-matches that should be matches, because the entry in one segment of A-D-G-N on one of the databases is incorrect and does not match the A-D-G-N in the target database with which a match is sought. Even when the incidence of key-punching errors is low, the probability of *at least one error* in a series of multiple indicators is not trivial. Erroneous fields, such as invalid numbers, also arise. For example, SSN9 fields sometimes contain 111111111, 123456789 or 999999999, which are invalid SSN9s.

3 Matching Process

The matching algorithm proceeds in four parts.

Database Preparation Databases are prepared and standardized.

Creation of Identifiers Identifier values used to link records in one database to records in another database are constructed by combining multiple individual fields.

Record Linkage and Matching One-to-many matches are conducted between the databases. That is, the algorithm matches each unique identifier on the TEAM database to all records on the identification database that have the corresponding value of the identifier. (One-to-one matching may alternatively be appropriate depending on the particular features of the matching databases.)

Data Gathering Appended to the voter file data are fields indicating every match found to a record on a state or federal database.

The result of this methodology is to produce a MATCH list and a NO MATCH list. This section describes the algorithm and results of the matching process in greater detail.

The first phase of the matching algorithm, Database Preparation, standardizes the coding of database fields to facilitate matching. Different databases store the fields in different ways. The database preparation in the algorithm standardizes the coding of dates of birth, genders, addresses and names, and identifies invalid or missing values (such as 111111111 for Social Security Numbers). Invalid identifiers are discovered by assessing frequencies of duplicates by identifier. As stated above, an identifier like month-day-year of birth ought to be distributed predictably across a population. Birthdays that occur at unusually high rates often indicate an issue like birthdays listed as 01/01/1901 or 11/11/1111. Analysis of duplicates also reveals completely duplicate cases. In the Texas voter file, we identified approximately two hundred records in which all information was identical.

Standardization of addresses involved extracting just the ZIP5 and street number fields. ZIP5 and street numbers are useful components of an address field because they are numeric and they are stored similarly across databases. Other components of addresses, like street names, street suffix, and apartment number pose difficulties because they are stored differently in different databases (e.g. First Street vs. 1st Street vs. First ST, etc), and because fields composed of lengthy strings will have higher incidence of typos.

Standardization of dates of birth involved converting all fields to Month-Day-Year combinations.

Standardization of Gender coded all indicators of Male to 0 and all indicators of Female to 1.

Standardization of names involved the most attention to subtle details. We constructed two different name indicators, one for Last Names only and the other for Last Name, First Name and Middle Initial. The latter offered a higher level of uniqueness but also more false negatives owing to variations in first names, especially due to nicknames. Rather than use

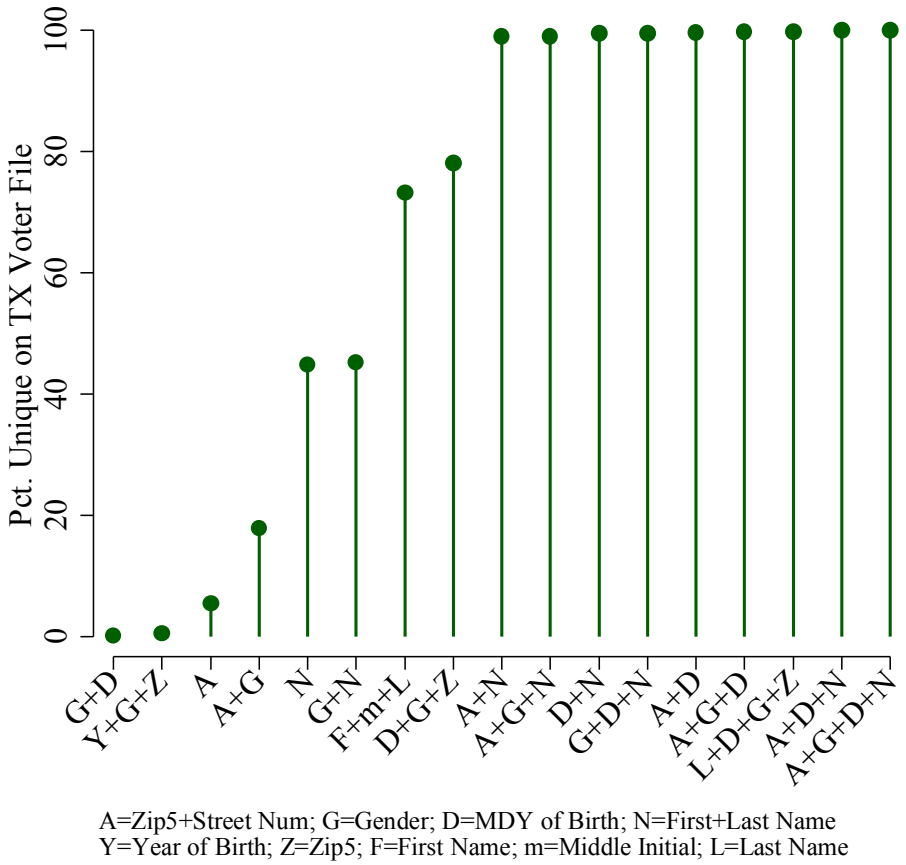
a name dictionary for first names and nicknames, we simply altered the identification fields. Standardization of last names required several procedures. First, all names were converted to uppercase letters. Second, the algorithm removed all apostrophes, hyphens, and other markings. For examples, the last name O'Conner occurs as O'CONNER in some databases and OCONNER in others; elimination of extraneous marks changes all such cases to OCONNER. A variant on the name matches accounted for variations in usage of hyphenated or compound last names, such as maiden names used as middle names. In addition to attempting matches on last names, we broke each hyphenated name into two parts and attempted to find matching records using each name. Third, all spaces in last names were removed. VAN HALEN or De La Rosa became VANHALEN and DELAROSA. Fourth, signifiers such as II, III, IV, and JR were removed from the ends of names, as these are often not standard.

The second part of the algorithm develops multiple identifiers for purposes of record linkage. The algorithm builds identifiers by combining fields related to Address, Date of Birth, Gender, and Name. In total, 6 different primary identifiers were constructed in the registration database (TEAM) and in the license database (DPS). Each identifier corresponds to a particular combination of fields. For example, Combination A consists of First Name, Last Name, Date of Birth, Gender, Street Number, and 5-digit ZIP Code. A sample version of Combination A for a man named John Smith, born on January 1, 1960, and living at 100 Main Street in the ZIP Code 78610 would be JOHNSMITH00101196010078610.

Importantly, each of these indicators is nearly always unique identifying, so long as data is not missing in the component parts. To illustrate this, consider Figure 1. The calculations shown here are similar to ones we performed based on the original Texas file, but these come from the 2016 voter list obtained from Texas. The figure shows 17 combinations of components of A-G-D-N, some of which we did use in matching and some of which we did not. There are a number of potential combinations that are not unique to individuals. For example, first name + last name or first name + last name + gender are only unique to

about 50% of registered voters. Street Number + Zipcode + gender is only unique to about 20% of voters. However, all of the combinations that we used in our algorithm were unique for over 99% of registered voters. This is important because it means that if a person is identified in another database with the same combination, it is very unlikely that this is a different person from the one we are trying to link.

Figure 1: Percent Unique of A-G-D-N Combinations in the Texas voter file



Source: October 2016 list of Texas voters.

The third stage of the process, the Record Linkage and Matching phase, conducts one-to-many matches and performs multiple sweeps for each identifier (ADGN, ADG, ADN, DGN). That is, a match was said to exist for each record in TEAM for which one or more records could be found in a corresponding database. Multiple sweeps provide a guard against false

negatives arising due to typographical errors, missing fields, or inconsistencies. For example, a person may have one last name in one database but another last name in another database, say because of a typo or a name change. He or she would be matched on Address, Data of Birth, and Gender. The algorithm will match the record on the identifiers that do not contain each of these categories of fields, thus avoiding non-matches due to typographical errors, nicknames, missing fields, and other inconsistencies between databases. A record is determined to have found a match if a given identifier in TEAM is identical to at least one corresponding identifier in an identification database.

As shown in the Table 1, the algorithm implemented in the Texas case conducts two sorts of sweeps through the data to find matching records. The Primary Sweeps match on Combinations A-F and M, and are run on all TEAM records. A-F correspond to matches obtained through the ADGN link indicators. M and SSN are matches between Texas State IDs and SSN9 for those cases for which such fields are available on the TEAM database. The analysis presented here focuses on Sweeps A through F, M and SSN.

In addition, the implementation of the matching algorithm sought to find additional matches using a set of Secondary Sweeps performed on the TEAM records not matched in the Primary Sweeps. The secondary sweeps are shown in Table 1 as Combinations G - L. The secondary matches include additional variants, such as using middle initials, and matching separately on components of compound surnames. For example, we would attempt to match a voter named Ruth Bader Ginsberg to records of Ruth Bader Ginsberg, Ruth Ginsberg and Ruth Bader. For Federal databases, the Primary Sweeps are run against all qualifying Federal records with Texas addresses, while the Secondary Sweeps are run both against Texas-only records, as well as against the nationwide universe of the relevant Federal dataset.

By using multiple identifiers, the algorithm is developed to be sensitive to variations in names, such as nicknames and compound names, to typographical errors, and to missing information. By matching on identifiers constructed from a larger number of categories of

fields (three or four), the algorithm exhausts all possible linkages among the identifiers that have a high likelihood of finding unique matches.

Table 1: Combinations of Fields Used as Matching Identifiers, Details

		Matching Combinations
Texas DPS Databases	**Primary Sweeps** (All TEAM records)	A: FirstName + LastName + Gender + DOB + ZIP + Street Num. B: LastName + Gender + DOB + ZIP + Street Num. C: Gender + DOB + ZIP + Street Num. D: FirstName + LastName + ZIP + Street Num. E: FirstName + LastName + Gender + ZIP + Street Num. F: FirstName + LastName + DOB + Gender M: Texas Driver License Number
	Secondary Sweeps (TEAM Records with no Primary Match)	G: FirstName + LastName + MiddleInitial + DOB H: DOB + ZIP + SSN4 I: FirstName + LastName + DOB + SSN4 K: FirstName + LastName#1 + MiddleInitial + DOB L: FirstName + LastName#2 + Middle Initial + DOB SSN: 9-digit Social Security Number
Federal Databases	**Primary Sweeps** (All TEAM records against Fed. records with TX address)	Repeat Combos A-F above
	Secondary Sweeps (TEAM records with no primary match against Fed. records with TX address)	Repeat Combos. G-SSN
	Nationwide Sweeps (TEAM records with no primary or secondary match against nationwide Fed. records)	All sweeps without address criteria

The fourth phase of the matching process is the Data Gathering phase. The results of all matching sweeps are recorded for each individual TEAM record. Most records matched on all or almost all indicators, but some only matched on one or two indicators. We record

each indicator for which a match occurred. This stage also appends indicators of deceased records or expired licenses from the Texas DPS data to TEAM.

Regarding computing, it is worth noting that the TEAM database is 13 gigabytes; the DPS database is 25 gigabytes. On account of security, processing had to be done on a local machine. We, as well as the federal agencies responsible for some of the matching, had a variety of experiences with the time it took to process the matches. The first implementation of the algorithm was performed in 2012 as part of the case *Texas v. Holder*. One loop through the data to match on the string First Name-Last Name/Date of Birth/Address using STATA could take several hours. Some federal agencies using SQL also reported hours-long computing time. When conducting the match in the context of the 2015 case, *Veasey v. Perry*, we upgraded software to 8 core STATA (from 2 core) on a computer with processors to accommodate. The computer performed one iteration of the matching algorithm in less than 30 minutes, compared with several hours. That improvement in speed was critical to be able to validate each step of the algorithm and to train the algorithm to catch any errors, trap special cases, and measure performance of the matching routine.

4 Results

4.1 Rates of Matches and No-Matches.

The implementation of the algorithm developed for the United States in this case matched the entire TEAM database to 10 different state and federal databases. Table 2 below lists the number of records in TEAM that matched to each state or Federal database using that algorithm.

The most commonly held form of identification is a State of Texas Driver License, followed by a United States Passport. Just over 78 percent of records in TEAM matched to the DPS Driver License list, while 42 percent of records in TEAM matched to the DOS passport

database. The next most common form of ID is a DPS Personal (or State) ID, held by 10 percent of those in TEAM.

Table 2: Percent of TEAM Records that Match to a Given ID or Disability Database

Database	Total Matched	Pct. Matched
Texas IDs		
Driver License	10,663,738	78.60%
Personal ID	1,284,658	9.50
Concealed Handgun License	733,008	5.40
Election ID Card	163	<0.01
Federal IDs		
State (Passport)	5,731,163	42.30
Defense	638,354	4.70
Immingration	735,086	5.40
Veterans Health	296,005	2.20
Federal Disability IDs		
Social Security	804,338	5.90
Veterans Disability	188,516	1.40

Of the 13,564,416 records in the TEAM database, 12,593,640 matched to at least one record corresponding to acceptable photo ID issued by the State of Texas, and 6,305,182 records matched to at least one record corresponding to acceptable photo ID issued by the Federal government. By acceptable, we mean acceptable according to the particulars of the Texas law. Most of the records matched to the Federal databases also matched to a State of Texas identification database. For 608,470 records on the TEAM database (approximately 4.5 percent of all records in TEAM), no matching record was found on any of the state or Federal identification databases or to an identification database for which a disability exemption was granted.

Hence, 4.5 percent of records on the list of registered voters in Texas could not be found in any database of state or federal IDs deemed acceptable under S.B. 14.

Table 3: Records Matched and Not Matched to State and Federal Databases, Excluding Deceased

Any State Record	Any Federal Record		
	No Match	Match	All
No Match	608,470	285,466	893,936
Match	6,573,924	6,019,716	12,593,640
All	7,182,394	6,305,182	13,487,576

4.2 Accuracy

4.2.1 Comparison with SSN9

We used matches to SSN9 as a "unique identifier" against which to test the accuracy of the primary matches using combinations of Address, Date of Birth, Gender, and Name. For the subset of records with SSN9 on TEAM (approximately half of the records), we examined cases that had SSN9 and for which there was NO MATCH between TEAM and a DPS record using combinations of Address, Date of Birth, Gender, and Name. That is, to test the validity of the Primary Matching algorithm, we conducted those matches for cases with SSN9. We then rematched the cases to DPS using SSN9, and calculated the percent of cases for which no Primary Match could be found but for which there was an SSN9 match.

Table 4: Comparison of SSN9 Match to ADGN Match

ADGN Match	SSN9 Match		
	No SSN Match	SSN Match	Total
No ADGN Match	1,207,739	135,686	1,343,425
ADGN Match	119,601	5,249,230	5,368,831
Total	1,327,340	5,384,916	6,712,256

Of the 5,384,916 records that match on SSN9 between TEAM and DPS, 2.5 percent were not matched using Address, Date of Birth, Gender and Name between TEAM and DPS. We further examined the set of cases for which there was a primary match using Address, Date of Birth, Gender, and Name. Of the 5,368,831 records on TEAM that match to one of the

primary indicators using Address, Date of Birth, Gender, and Name between TEAM and DPS, 2.2% were not matched using SSN9 to link records between TEAM and DPS. Hence, the Address, Date of Birth, Gender, and Name combinations could accurately match 97.5 percent of records (using SSN9 as the benchmark for validation). By comparison, SSN9, which is often relied on as a unique identifier, could match 97.8 percent of records (using Address, Date of Birth, Gender, and Name primary matches as a benchmark for validation). In other words, the primary matches on combinations of Address, Date of Birth, Gender, and Name are almost the functional equivalent to matching on SSN9.

The rate at which the primary sweeps using Address, Date of Birth, Gender, and Name combinations yield NO MATCH but for which a matching SSN9 exists is extremely low. It is, for example, comparable to the rates reported in the article of Professors Hood and Bullock on the Georgia ID law for 2008 (98.2%), and much lower than the figure they report for 2004 (78.5%).

The rate at which the primary sweeps using Address, Date of Birth, Gender, and Name combinations yield NO MATCH but for which a matching SSN9 exists is lowered further upon using DPS ID in the primary matches, upon conducting secondary matches, including SSN9, and upon using the federal data.

It is also worth noting that not all of the cases in this small subset of NO MATCHES are erroneous. Some of these NO MATCHES may reflect discrepant information on TEAM and DPS that would make it difficult to authenticate the person at the voting place, such as a name change and missing Date of Birth, or a change in name and address. For example, suppose there is a registered voter listed on TEAM by the name Jon Jones living at 100 Main Street. Suppose this same person is listed with a drivers license as Jonathan Jonas at 200 First Street. This record might have discrepancies on multiple indicators because this person moved residences and had a typo and/or nickname used in one but not all databases. This person would be a NO MATCH using our ADGN algorithm but a MATCH using SSN.

However, because of the discrepancies on the indicators, an election official might disqualify him from voting. Thus, it is not clear whether or not this mismatch ought to be considered a false negative or a true NO MATCH.

4.2.2 False Negatives

We conducted forensic analysis of the NO MATCH list to determine what was the likely incidence of false negatives. Again, a false negative is a TEAM record that should have matched to another database but failed to do so.

To investigate false negatives, we looked at records on TEAM that a.) did not match, and b.) had a DPS identification number (i.e. drivers' license number) listed in the voter registration record. Recall that a subset of voter registration records listed a DPS ID number. Most of the records with DPS IDs did not match because the IDs were for expired, revoked, or otherwise invalid licenses that would have not have been useable as a voter ID. Of the total NO MATCH list, 56.7% had a listed DPS ID, and of those 18.4% represented a valid ID under S.B. 14.[3]

Of 82,045 records that appeared to have valid DPS IDs but were not matching on ADGN, SSN, or DPS ID, we first confirmed that they were not matching and then we performed a visual inspection of the databases of 100 randomly selected cases. There are several reasons why these people would have a DPS ID and not be on the DPS. A non-exhaustive list includes the following: (1) The underlying DPS record may have been purged, perhaps because they were no longer valid and were removed as deadwood from DPS. (2) The DPS records provided to us may not be complete. (3) There might be typographical errors in the DL number (though, if that alone were the cause, the other matching sweeps likely would

[3]Our analysis here was based on a NO MATCH list of 786,726 records, different from the 608,470 records indicated above as the NO MATCH list. In the course of our work on behalf of the United States, we first estimated a NO MATCH list of 786,727. In the midst of the trial, Texas identified an additional list of valid ID holders, which brought the NO MATCH list down to 608,470. It was at the earlier stage of the trial that we performed the analysis in this section, so that is what we can report here. We have no reason to expect that the analysis would be different if performed on the revised NO MATCH list.

have caught some of these records).

While we could not determine definitively why 82,045 records contained DPS IDs but failed to match, we take these records to be potential candidates for false negatives - records for which a match might have been found but there was insufficient information on the voter records to perform a match. If the records without DPS IDs listed are similar to those with DPS IDs listed, we estimate the false negative rate at eight-tenths of one percent of all records on the TEAM voter file. That is, if we cannot determine why 18.4% of the 56.7% of NO MATCHES with DPS IDs failed to match, then plausibly we would not be able to determine about 18.4% of the 43.4% of NO MATCHES without listed DPS IDs.

5 Implications

5.1 Racial Effects of S. B. 14

The purpose of this study was to estimate how many registered voters might be excluded by S. B. 14 and whether the law had disparate effects across racial groups. The final piece of the analysis was to examine racial differences in the likelihood of having an acceptable ID or not, that is, the likelihood of being on the NO MATCH list. Using racial data on individuals provided by the firm Catalist, we estimated that Blacks and Hispanics are significantly less likely to have the forms of ID required under S.B. 14. That is, the NO MATCH rate served as the dependent variable in an analysis estimating whether there were likely racial differences or disparate racial effects of S. B. 14. Table 5 provides the rates with registered voters in the TEAM database in each racial group were found to have NO MATCH to any of the corresponding ID databases.

Analysis of individual level data using the Catalist classification of race reveals significant differences in the NO MATCH rates of the different racial groups. The baseline universe of registered voters, which consists of all currently-registered voters in TEAM – after removing

Table 5: Number and Percent of instances of NO-MATCH and MATCH by Racial Group, Using Catalist Racial Classification

Race	NO-MATCH	MATCH	Total
Anglo	296,156 (3.6%)	7,949,860 (96.4%)	8,246,016
Black	127,908 (7.5%)	1,569,861 (92.5%)	1,707,769
Hispanic	174,715 (5.7%)	2,867,782 (94.2%)	3,042,497
Other	9,691 (2.0%)	481,621 (98.0%)	491,312
All	608,470 (4.5%)	12,879,124 (95.5%)	13,487,594

those who matched a record marked as deceased in a DPS ID file has 13,487,594 records. Of these, 8,246,016 are classified as Anglo according to Catalists estimates; 1,707,769 are Black; 3,042,497 are Hispanic; and 491,312 are Other Races.

The rate of non-matches between TEAM and identification databases varies by race. Of records identified as Anglo in the Baseline Universe of Registered Voters, 3.6 percent had no matching record in state or Federal identification databases. By comparison, no matching records were found for 7.5 percent of people identified as Black and 5.7 percent of people identified as Hispanic. The differences in rates of matching and non-matching across racial groups are statistically significantly different from one another. The difference between Blacks and Anglos in the rate of non-matching is 3.9 percentage points, and the difference between Anglos and Hispanics in the rate of non-matching is 2.1 percentage points. Both differences are highly significantly different from 0, using conventional hypothesis tests.

5.2 Implications for Administration of Election Laws

The task we sought to perform for *Veasey v. Abbott*, the Texas Voter ID case under Section 2 of the Voting Rights Act, was to identify all individuals on the TEAM database

that had a matching record on at least one of the relevant ID databases. We sought to answer a slightly different question in *State of Texas v. Holder*, the Section 5 case. In that analysis, we matched records using information that would be on IDs and could be used to authenticate people at the polling places. Specifically, we linked records using Name, Address, and Date of Birth. We did not use multiple identifiers, but instead linked using a single identifier that concatenated all three fields. The reasoning behind this approach was that a voter could potentially be disqualified, at the discretion of a poll worker, if the name or address or date of birth varied between the voter file and the DPS file, or was missing on the voter file.

In answering this question we discovered that three times as many people could not be matched between the TEAM and DPS files. Specifically, no exact match using Address, Date of Birth and Name was found for 1.9 million records of the 13 million records on TEAM. The racial differences were approximately the same magnitude as with the method using multiple indicators (see Ansolabehere (September 16, 2014.)).

The higher NO MATCH rate using this method is due to three factors. First, missing fields, especially missing birth dates, and typographical errors create no matches. Second, variations in address fields caused when people move account for additional instances of NO MATCH. Third, names, especially first names, vary across files. When matching individuals, such issues create obstacles to record linkage. However, in determining who may be affected by the ID law, these cases are instances where the vote may potentially be denied to a registered voter owing to an irregularity or variation in the records.

Take for example U. S. Senator W. Phillip Gramm. On the DPS file, he is recorded using his full name. On the TEAM file, he is recorded as Phil Gramm. If Senator Gramm attempted to vote, a poll worker could interpret the election code to mean that he lacked relevant ID. Comparing the two different approaches to matching provides bounds on the possible effects of the law. A more lenient interpretation of the law asks for which individuals

on TEAM is there no evidence of any record on any ID database, even if some of the information varies. This approach suggests a more exhaustive set of matching criteria. A more restrictive interpretation of the law would hold that any individual for whom the ID and TEAM records do not match on the key fields in the databases (Name, Address, and Date of Birth) might be denied the vote.

Comparing the two approaches is also instructive about the power of using multiple indicators or a single indicator. Using a single indicator generates a much larger set of non-matches than using multiple indicators – three times as large.

5.3 Implications for Identifiability and Field Redaction

The results of this study have important lessons for identifiability of records on publicly available databases, and how to redact data in order to protect individuals' identities. El Eman (2013) provides a thorough discussion of the risks of data breaches, as well as practical methods for de-identification. When and how de-identification occurs depends, ultimately, on what information is on a database and the uniqueness and frequency of specific pieces of information and combinations of pieces of information (see also Sweeney (September 29, 2015)). Many databases are protected by redacting a single field, such as name. Others are protected by redacting information up to a very high level of aggregation such as state. The cost of masking more detailed data is that researchers lose information that is relevant to their research, such as the effects of local context on sociological or political behavior. How unique are individual pieces of information and combinations of pieces of information in large databases?

The frequency of fields and combinations of fields in the Texas database reveals which fields are most informative about individuals' identities. Consult Figure 1 above.

For single fields, such as Name alone or Address alone, there is a low rate of uniqueness. However, combining fields increases uniqueness considerably. Strikingly each triplet formed

with A, D, G, or N has a uniqueness of over 99%. Even just two pieces – either Name and Date of Birth or Name and Address – identifies nearly all individuals in a database uniquely. 99.6 percent are unique with DoB and Name, 99 percent are unique with Name and Address.

The data also reveal that there are some simple protections against identifiability that preserve information for researchers. Specifically, two pieces of information that make individuals highly identifiable, especially in combination, are Date of Birth and Name. Clearly Name should be masked, but what about age, one of the most common social science indicators? Fortunately, Year of Birth, even in combination with other pieces of information adds little leverage over the identification of individuals. The combination of Year of Birth, gender, and ZIP5 (a combination sometimes provided in de-identified surveys) is only unique to 0.42% of individuals. Also, ZIP5, a level of aggregation far lower than many databases present, has very little effect on identification, in the absence of Date of Birth and Name. This suggests that more geographic information could be made available on surveys and other databases, and that would likely improve the integration of individual level data we geographic or contextual data.

6 Conclusion

Record linkage and database matching have moved out of the fields of computer science and statistics and out of the closeted world of secure databases, such as databases maintained by the Census Bureau. We have examined one case, the case of S. B. 14, of the application of record linkage methods to a significant public policy challenge, the protection of voting rights. This study was conducted in the context of a particular case and a particular set of databases. All of the information in this article is available in public records produced for the courts in *State of Texas v. Holder* and *Veasey v. Abbott* or is based on our analysis of a publicly obtained list of Texas voters.

The value of this information goes far beyond these cases. Society has now entered an era where extensive amounts of information about individuals are available not only on single databases, but by linking databases together. Theories of data and database security have guided development of many tools for linking records and protecting data, but decisions about how to match databases and how to protect information require information about how unique data actually are. The goal of this study is to inform decisions both about how to match databases efficiently and accurately and how to protect sensitive information about individuals.

References

Ansolabehere, Stephen. September 16, 2014. "Corrected Supplemental Report." Marc Veasey, et al. v. Rick Perry, et al. In the United States District Court for the Southern District of Texas, Corpus Christi Division. Civil Action No. 2:13-cv-193. Document 600-1.

URL: *http://moritzlaw.osu.edu/electionlaw/litigation/documents/Veasey6552.pdf*

Ansolabehere, Stephen and Eitan Hersh. 2012. "Validation: What Survey Misreporting Reveal about Survey Misreporting and the Real Electorate." *Political Analysis* 20(4):437–459.

Berlin, Thomas and Donald B. Rubin. 1995. "A Method for Calibrating False-Match Rates in Record Linkage." *Journal of the American Statistical Association* 90(430):694–707.

De Brou, David and Mark Olsen. 1986. "The Guth Algorithm and the Nominal Record Linkage of Multi-Ethnic Populations." *Historical Methods* 19(1):20–24.

El Eman, Khaled. 2013. *Guide to the De-Identification of Personal Health Information*. Boca Raton: CRC Press.

Fellegi, Ivan P. and Alan B. Sunter. 1969. "A Theory of Record Linkage." *Journal of the American Statistical Association* 64(328):1183–1210.

Guth, Gloria J. A. 1976. "Surname Spellings and Computerized Record Linkage." *Historical Letters Newsletter* 10(1):10–19.

Hersh, Eitan and Matthew Goldenberg. 2016. "Democratic and Republican Physicians Provide Different Care on Politicized Health Issues." *Proceedings of the National Academy of Science* 113(42):11811–11816.

Hood III, M.V. and Charles S. Bullock III. 2008. "Worth a Thousand Words? An Analysis of Georgia's Voter Identification Statute." *American Politics Research* 36(4):555–579.

Jaro, Matthew A. 1989. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 84(406):414–420.

Larsen, Micahel D. and Donald B. Rubin. 2001. "Iterative Automoated Record Linkage Using Mixture Models." *Journal of the American Statistical Association* 96(453):32–41.

Li, Xiaochun and Changyu Shen. 2013. "Linkage of Patient Records from Disparate Sources." *Statistical Methods in Medical Research* 22(1):31–38.

Mohanty, April F., Jacob Crook, Christina Porucznik, Erin M. Johnson, Robert T. Rolfs and Brian C. Sauer. 2015. "Development and evaluation of a record linkage protocol for Utahs Controlled Substance Database." *Health Infomatics Journal* Forthcoming:1–9.

Pacheco, Antonio G., Valeria Saraceni, Suely H. Tuboi, Lawrence H. Moulton, Richard E. Chaisson, Solange C. Cavalcante, Betina Durovni, Jos C. Faulhaber, Jonathan E. Golub, Bonnie King, Schechter Mauro and Lee H. Harrison. 2008. "Validation of a Hierarchical Deterministic Record-Linkage Algorithm Using Data From 2 Different Cohorts of Human Immunodeficiency Virus-Infected Persons and Mortality Databases in Brazil." *American Journal of Epidemiology* 168(11):1326–1332.

Sweeney, Latanya. 2002. "k-Anonymity: A Model for Protecting Privacy." *International Journal of Unicertaity, Fuzziness, and Knowledge-Based Systems* 10(5):557–570.

Sweeney, Latanya. 2005. "Privacy-Enhanced Linking." *ACM SIGKDD Explorations Newsletter* 7(2):72–75.

Sweeney, Latanya. September 29, 2015. "Only You, Your Doctor, and Others May Know." Technology Science.
URL: *http://techscience.org/a/2015092903/*

Winkler, William E. 1999. "The State of Record Linkage and Current Research Problems." Census Bureau Research Report Series (#93-8), Statistical Research Division, U.S. Census Bureau.

Winkler, William E. 2006. "Overview of Record Linkage and Current Research Directions." Census Bureau Research Report Series (Statistics #2006-2). Statistical Research Division, U.S. Census Bureau.

Winkler, William E. 2014. "Matching and Record Linkage." *WIREs Comput Stat* 6:313–325.