Social Security Numident database access by Trump-Musk-DOGE: Proven vulnerability of identifying nearly all US likely transgender people with 99% confidence (2025-03-15)

(Contact: thezinniajones@protonmail.com)

Overview

- DOGE, using their recently obtained access to the Social Security Administration's (SSA) Numerical Identification System (Numident) database of every person in the country (Wired, 2025-03-13, https://archive.ph/vF0uL), **could use known published methodology from the Census Bureau from 10 years ago** (Harris, 2015, https://ideas.repec.org/p/osf/socarx/fj8y8_v1.html) **to generate a high confidence (90%, 95%, or 99%) list of all trans people in the United States.** This could likewise identify with 90-95-99% confidence whether a specific person is trans. The following operational definition has been used by the Census Bureau in the Harris study in 2015.
- Identification in Harris 2015 was **based on comparing a gender-typing of former deadname with current legal name, as well as optionally considering legal sex marker changes. This captures trans people who change their first and middle names in this way even without legal sex marker changes**.
    - "Any time an SSN is created or information associated with an existing SSN is changed, that event is registered as a claim [in Numident]"
    - "**Using information on name changes alone is a more inclusive strategy for identifying likely transgender individuals, since it does not require changes in sex-coding**, which typically require costly and not always preferred genital sexual reassignment surgery."

    - "This framework considers a name change to be consistent with a gender transition if (a) both the original and the new first name fall within the specified confidence threshold, (b) the gender of the original first name matches the original sex-coding on the account, (c) the gender of the new first name is different than the gender of the original first name, and (d) the name change is not reversed later."
    - "According to the most inclusive criteria, which require a name change consistent with gender transition using the 90 percent confidence threshold, 135,367 individuals in the data are likely to have transitioned gender between 1936 and 2010. As the name confidence thresholds become more demanding, moving rightward across the columns, the figures attenuate; **the 95 percent confidence threshold yields a count of 106,550 while 64,738 individuals have name changes that meet the 99 percent confidence threshold. These figures include any individuals who changed their first name from a traditionally female name to a traditionally male name (or vice versa)**, which is consistent with that person taking steps to be socially recognized as a gender distinct from the sex assigned at birth, regardless of whether that person has taken steps to alter his or her physical appearance through hormone treatment or surgery."
    - "The results … do not include those who have not legally changed their names, who have changed their names through common-law, whose name(s) do not meet my minimum confidence threshold, or whose gender transition did not involve a name change at all."
- **Linking with US Census Bureau census records** allowed for further detailed identification of trans people's "basic demographic characteristics and residential patterns".
    - "This is the first North American study to use administrative files to learn about transgender individuals. … to link the administrative records to **other data that has**

> **more information about likely transgender individuals' demographic and residential characteristics**."

Details and resources

- Makena Kelly, David Gilbert, Vittoria Elliott, Kate Knibbs, Dhruv Mehrotra, Dell Cameron, Tim Marchman, Leah Feiger, and Zoë Schiffer. (2025, March 13). Wired. https://www.wired.com/story/elon-musk-digital-coup-doge-data-ai/ ; https://archive.ph/vF0uL

> DOGE had installed a handpicked chief information officer at the SSA—the former CTO of a payments company headed by Jared Isaacman, a billionaire who once commanded two trips to space using SpaceX rockets and is Trump's current nominee to lead NASA. That new CIO, Michael Russo, had asked to bring on Akash Bobba, the former Palantir intern who had been working out of OPM, as an engineer.
>
> But there were "challenges" with Bobba's background check, Tiffany Flick, the acting chief of staff to the acting administrator, later stated in a sworn affidavit given in a suit against the SSA. Bobba wasn't brought on immediately. By February 10, seven days after he was requested to be onboarded, phone calls and emails started to come in—from Russo, Steve Davis, and others—making clear that Bobba was to be given access to SSA systems and data by the end of the day. As Russo and Davis "grew increasingly impatient" that evening, Flick recalled, Bobba was sworn in over the phone at 9 pm.
>
> Initially, Flick and officials from the CIO's office determined that Bobba would be given anonymized, read-only access to records in the Numerical Identification System, which contains information on everyone who has ever applied for a Social Security number. On

> February 15, Bobba reported that there were issues with the dataset he'd been provided. Russo demanded that Bobba be given full access to "everything, including source code," Flick recalled. This included the SSA's Enterprise Data Warehouse, which contains the "names of spouses and dependents, work history, financial and banking information, immigration or citizenship status, and marital status," according to Flick's affidavit.
>
> Later that day, the chief information officer for the whole federal government—a political appointee working out of the Office of Management and Budget—issued an opinion to Russo granting Bobba the access. Flick retired. In her affidavit, she expressed serious concerns about the potential for SSA records to be "inadvertently transferred to bad actors" and about "incredibly complex web of systems" being "broken by inadvertent user error."

- Makena Kelly and David Gilbert. (2025, March 13). These Are the 10 DOGE Operatives Inside the Social Security Administration. Wired. https://www.wired.com/story/doge-operatives-access-social-security-administration/ ; https://archive.ph/8Q7j9

> The operatives—whom the government did not name in its filing—are, according to internal documents, Akash Bobba, Scott Coulter, Marko Elez, Luke Farritor, Antonio Gracias, Gautier Cole Killian, Jon Koval, Nikhil Rajpal, Payton Rehling, and Ethan Shaotran. This team appears to be among the largest DOGE units deployed to any government agency.
>
> …
>
> Many of them have worked or interned at Musk companies such as Tesla and Space X, and the majority of them have also appeared at other government agencies in recent

weeks, as part of DOGE's incursion into the government. Musk has made wild claims about the social security system, calling it a "Ponzi scheme" and falsely claiming that millions of 150-year-olds were fraudulently collecting benefits.

According to the SSA court filing and accompanying sworn statements, seven of them have read-only access to several datasets including the Master Beneficiary Record, which contains detailed information about individuals and their benefits. Those same DOGE representatives also have read-only access to Numident, a database containing information about everyone who's ever applied for a Social Security number, as well as data referred to as "Treasury Payment Files Showing SSA Payments from SSOARS." (The Social Security Online Accounting and Reporting System is a set of systems containing "information on the SSA's financial position and operations," according to the SSA.)

These records contain a great deal of personally identifying and financial information; in filings the government says DOGE accessing them is necessary to "detect fraud."

…

Sources have told WIRED that one of the tasks the DOGE cohort will be assigned is how people identify themselves to access their benefit payments. Experts with decades of experience at the agency are now worried that DOGE operatives working across multiple agencies increases the risk of SSA data being shared outside of the agency or that their inexperience will lead to them breaking systems entirely.

In the SSA filing, lawyers for the agency claim that the DOGE operatives have "no access to SSA production automation, code, or configuration files." A previous sworn statement from Tiffany Flick, the agency's former acting chief of staff, claims that its CIO, Michael Russo, was DOGE-aligned and demanded that Bobba be given access to "everything, including source code."

…

According to an affidavit filed on Friday as part of a lawsuit designed to halt what the suit called DOGE's "unprecedented" seizure of SSA data, Flick outlined how she tried to educate Russo on how information at SSA is handled and the measures in place to prevent fraud.

"Mr Russo seemed completely focused on questions … based on the general myth of supposed widespread social security fraud, rather than facts," Flick said, adding that Russo was unwilling to understand SSA's complex systems and instead seemed fixated on conspiracy theories about fraud within the system, such as Musk's claim that millions of 150-year-olds were receiving benefit payments. Flick also wrote that she was "not confident" that DOGE operatives had "the requisite knowledge and training to prevent sensitive information from being inadvertently transferred to bad actors."

…

Many of the other DOGE names listed in SSA's internal records have previously been reported by WIRED as young, inexperienced technologists working at other government agencies.

Individuals having access to the records of multiple agencies is highly unusual and problematic, experts say. "Federal law places strict controls on personal data held by agencies, including limits on cross-agency transfers and rigorous training requirements for personnel who have a legitimate need for access," says John Davisson, the director of litigation at the Electronic Privacy Information Center. "Ignoring those safeguards and haphazardly putting systems at multiple agencies under the thumb of a single engineer obliterates those protections. They're hotwiring the federal government with a total disregard for privacy and data security."

Bobba, a former Palantir intern and recent UC Berkeley graduate, was first appointed to the Office of Personnel Management (OPM) before moving to the General Services Administration (GSA). According to Flick's testimony, "There were challenges with Mr. Bobba's background check that took a few days to resolve." Flick did not expand on what those issues were but stated that Bobba was eventually granted access to sensitive SSA systems after Russo and Musk lieutenant Steve Davis directly pressured top SSA administrators. The acting commissioner was ultimately replaced by Leland Dudek, a mid-level staffer who was, Flick asserted, on leave after having communicated with DOGE outside normal channels—something he later bragged about in a LinkedIn post.

…

In a Tuesday meeting, United States DOGE Service administrator Amy Gleason told staff that Musk-affiliated engineers and some legacy USDS workers would be headed to SSA to improve "identity proofing," say sources who were in the meeting. The US DOGE Service is a permanent rebranding of the US Digital Service. Identity proofing is the

> process SSA uses to identify that users are who they say they are in order to access their benefits. The process includes registering with identity apps such as the government's own Login.gov or third-party services such as ID.me.
>
> …
>
> "It's a valid fear that personally identifiable information will be exfiltrated or source code messed with without necessary controls and rigor," John McGing, a former SSA employee who worked at the agency for almost four decades, tells WIRED.

- Benjamin Cerf Harris. (2015). Likely Transgender Individuals in U.S. Federal Administrative Records and the 2010 Census. https://ideas.repec.org/p/osf/socarx/fj8y8_v1.html

Harris (2015) relevant excerpts

> Abstract
>
> This paper utilizes changes to individuals' first names and sex-coding in files from the Social Security Administration (SSA) to identify people likely to be transgender. I first document trends in these transgender-consistent changes and compare them to trends in other types of changes to personal information. I find that transgender-consistent changes are present as early as 1936 and have grown with non-transgender consistent changes. Of the likely transgender individuals alive during 2010, the majority change their names but not their sex-coding. Of those who changed both their names and their sex-coding, most change both pieces of information concurrently, although over a quarter change their name first and their sex-coding 5-6 years later. Linking individuals to their 2010 Census responses shows my approach identifies more transgender members of racial and ethnic minority groups than other studies using, for example, anonymous online surveys. Finally, states with the highest proportion of likely transgender residents have

state-wide laws prohibiting discrimination on the basis of gender identity or expression. States with the lowest proportion do not.

[…]

1 Introduction

[...]

This paper identifies likely transgender individuals in a large, confidential, national data source and then links those individuals to their 2010 Census records. Specifically, I analyze the Social Security Administration (SSA) Numerical Identification System (the "Numident"), which contains information on first and middle names, sex-coding, and date of birth for every holder of a Social Security Number (SSN). I identify all adult SSN holders who permanently change their first names—or their middle names, in cases of gender-neutral first names—from traditionally male to traditionally female names (or vice versa). Because the Numident includes information on sexcoding, I also identify those who permanently change their sex-coding in the same direction as their name change. From October 2002 through the summer of 2013, the SSA required evidence of genital sexual reassignment surgery (SRS) for a person to change the sex-coding on his or her records. Therefore, while looking at name changes alone corresponds to a person's presentation of gender in society, looking at coincident changes in names and sex-coding during certain periods in the SSA's history corresponds to a definition of transgenderism that depends on surgical intervention.

Using these two main identification strategies, I explore the following research questions: First, how many adult SSN holders have changed the personal information on their accounts in ways that are consistent with a gender transition, and to what extent have the rates of transgender-consistent changes grown or decreased since the creation of the SSA in 1936? Second, among those who were alive on April 1 2010—the 2010 Census day—how many change their names only, and how many change both their names and their sex-coding? Of those who change both their name and their sex-coding, do they typically change their name first and their sex-coding later, or do they typically change both their name and their sex-coding simultaneously? Finally, by linking to the 2010 Census, I am able to answer questions about basic demographic characteristics and residential patterns.

I find that since the SSA's inception in 1936, as many as 135,367 individuals changed their name or sex-coding in ways that are consistent with a gender transition. While transgender-consistent claims registered with the SSA are evident in the agency's earliest years, the frequency of such claims increased dramatically during the SSA's history. This growth primarily mirrored growth in all other types of claims, reflecting both a growing population as well as an increased use of the SSN as a universal identifier. Of those 135.4 thousand likely transgender individuals, 89,667 were alive during the 2010 Census. Most had changed their name from one gender to another, but 21,833 had also changed their sex-coding.

People who change both their names and their sex-coding most commonly change both pieces of information concurrently, although just over a quarter change their name first and their sex-coding 5–6 years later. Most are in their mid-thirties when they begin to register these changes with the SSA, although (male to female) transgender women frequently begin the process somewhat later in life.

Linking individuals to their responses in the 2010 Census provides even more detail. While the likely transgender individuals in the SSA data are more likely to report their race as White alone, the proportion who identify as Black alone is greater than in other studies that tend to underrepresent non-white populations. Similarly, I am able to identify a greater proportion of likely transgender individuals who report Hispanic origin than in previous work, suggesting my approach may be beneficial in identifying transgender individuals who are members of racial and ethnic minority groups.

Another important finding from the linked data is that individuals in the SSA files who are likely to be transgender are much less likely than non-transgender individuals to respond to the Census question on sex. Those who do respond to the question are much more likely than non-transgender individuals to check both options (that is, both "M" and "F") on the Census questionnaire. Nearly all surveys, censuses, and forms impose binary responses to questions about gender. These findings point both to the limitations of framing questions about gender in this way as well as to the need for more research on how transgender individuals interpret such questions.

Finally, the linked data allow me to investigate patterns of residential sorting across states. I find that likely transgender individuals in the SSA files are overrepresented in states with laws banning discrimination on the basis of gender identity or expression. In addition, the states where likely transgender individuals are most underrepresented have no anti-discrimination laws.

I am unaware of any other study that looks at the history of transgender individuals' interactions with the SSA, the ways likely transgender individuals respond to the question about sex on the Census, or residential sorting patterns of transgender individuals in the U.S. While previous research has documented in-sample demographic characteristics and transition pathways, none is based on so many observations. To my knowledge, each of the paper's findings—particularly in the U.S. context—is an empirical contribution.

This paper also represents a major methodological contribution. The use of administrative records in transgender-related research is rare. I know of only two other studies that use administrative data: Veale (2008) uses New Zealand passport data, and Weitze and Osburg (1996) uses German court records. This is the first North American study to use administrative files to learn about transgender individuals. It is also the first study, to the best of my knowledge, to link the administrative records to other data that has more information about likely transgender individuals' demographic and residential characteristics. This approach is a major innovation in its own right, but it also paves the way for future research using linked data to learn about earnings and employment, marriage and divorce rates, household composition, and incarceration rates of likely transgender individuals.

[…]

3 Analytic Framework

3.1 SSA administrative records

The data used in this paper are administrative records from the SSA, provided to the Census Bureau under Titles 5, 13, and 42 of the U.S. Code. 5

5 Specifically, 5 U.S.C. §552a (b) (4), 13 U.S.C. §6, and 42 U.S.C §902 and §1306. Ensuring confidentiality of the data are the primary concern at the Census Bureau; data stewardship training is required annually for all staff, and severe penalties exist for any misuse of data.

The SSA Numident is an administrative database containing the name, date of birth, and a sex-coding for every SSN holder. Furthermore, the Numident contains a record for every claim that changes the information associated with a given SSN. The Census Bureau acquires the Numident through a data sharing agreement with SSA, and uses the data to facilitate record linkage and statistical operations. In addition, the SSA sends quarterly updates to inform the Census of the creation of new SSNs as well as changes and corrections of information associated with existing SSNs. Thus, the Numident contains a record of every claim for the population of SSN holders as of the most recent update. This study evaluates all the Numident records from 1936 to the end of 2010. These administrative records, alone, are sufficient to answer the first three research questions about the historic trends in likely-gender transitions, the number of likely transgender SSN holders as of the 2010 Census, and the paths people take with regard to timing of name and sex-coding changes. Later, I will discuss how linking individuals to their 2010 Census responses will help answer the remaining research questions on the demographics and geographic characteristics of the likely transgender individuals in the SSA data. 6

6 The Census Bureau is analyzing the use of administrative records to improve data quality and reduce respondent burden in its surveys and censuses. Record linkage in Census Bureau processes or products requires administrative records with consistent and high-quality information on name, date of birth, and sex. Researching the transgender population provides

valuable insight to researchers and practitioners by exploring how name and sex-coding changes legitimately appear in the SSA file and other administrative data sources. This is an important step in understanding that changes in first name and sex data are not necessarily anomalies or errors.

Table 1 shows the layout of the SSA files, using fictitious data and variable names for illustrative purposes. The first column shows each record's Protected Identification Key (PIK), or unique person identifier used by the Census Bureau to link individuals' records across data sets while simultaneously protecting their confidential information. Section 3.2 discusses PIK assignment and record linkage in greater depth. Within a given PIK, each row represents a claim. Any time an SSN is created or information associated with an existing SSN is changed, that event is registered as a claim. Table 1 provides examples of claims associated with changes to given names, surnames, date of birth, and sex-coding. For instance, John Doe (PIK=01) corrected a transposed month and day in his date of birth record; Jane Smith (PIK=02) changed her surname to Doe, consistent with the popular convention of married women adopting their husbands' surnames; and John Miller's (PIK=08) apparently mis-entered sex-coding was corrected when he was under 2 years old.

In order to identify likely transgender people, the first step is to distinguish people whose first name or sex-coding changed from those whose first name and sex-coding were stable. In the example in Table 1, PIKs 01–06 are "stable", while PIKs 07–10 feature a change in the first name, sex-coding, or both.

Table 2 shows the total number of claims, the total number of unique PIKs, and the average number of claims per PIK within the data. These figures are also given for the set of PIKs with stable first name and sex-coding information, by sex, and the set of PIKs featuring a change in first name, sex-coding, or both. After dropping those who were not at least 16 years old as of April 1 2010 (Census day), my data include 828.0 million claims associated with 374.2 million unique PIKs. The average person had 2.2 claims associated with their Numident record. For the vast majority of PIKs—347.8 million, or 93.0 percent—we observe no first name or sex-coding changes. Nevertheless, these "stable" records have 2.2 claims associated with their Numident record on average. Changes to these records might include corrections to erroneous dates of birth (as in John Doe's (PIK=01)), changes of surnames (as in Jane Smith's record (PIK=02)), or records of death (as in Jane Johnson's record (PIK=09)). There is, however, substantial heterogeneity across males and females with regard to the average number of claims per person. While the average male with a stable record makes 1.8 claims, the average female makes 2.4 claims. This is likely due to the common custom of adopting the surname of one's husband. The remaining 26.3 million "non-stable" records exhibit at some point a change in their first names, sex-coding, or both. Of the non-stable records, there were 24.7 million changes to first names but not sex-coding (93.6 percent), 1.4 million changes to sex-coding but not name (5.2 percent), and only 328.4 thousand changes to both (1.2 percent).

3.1.1 Modeling gender transitions

Only some of the people who have changes to their first names or sex-coding (or both) are likely to be transgender. To identify those who are most likely to have undergone a gender transition, I look first for those who changed their name from a traditionally male name to a traditionally female name, or vice versa. Among those whose name changes are consistent with gender transition, I then identify people whose sex-coding changed in the same direction as their first name. To simplify my analysis and to reduce the likelihood of false-positive assignment of transgender status, I ignore those who changed their first name or sex-coding before turning 16 years old as well as those who later changed their first name or sex-coding back to the original version. Note that these rules need not apply in principle. People can and do transition before turning 16. Further, gender can be a fluid concept; people can (and do) transition from one gender to another and back again.[7] These rules are instead meant to limit the possibility of falsely identifying a person as transgender.

3.1.2 Determining the gender of a name

Using the 347.8 million stable records, I construct name-sex crosswalk tables showing, for every name, the proportion whose sex-coding is “M” and the proportion that is “F”. Since the gender of names changes over time, I generate these crosswalks by birth-decade (Barry III and Harper, 1993; Lieberson et al., 2000; Rossi, 1965).[8]

I then link the crosswalk, by name and birth decade, onto the 25.0 million records with non-stable first names, first by the original first name and then by the new first name. I can identify

[7] For example, Weitze and Osburg (1996) identified a small number of people who transition from one gender to another and back again in Germany.

[8] Some records list the first name as a single letter (e.g., J DOE). Other names are so rare that they appear fewer than 5 times in a decade. These names are not included in the crosswalks.

name changes that are consistent with gender transitions by comparing the likelihood that the original first name is male to the likelihood the new first name is male.

Within the context of Table 1, PIKs 1–6 are used to construct the name-sex crosswalks. The name John is associated with an "M" sex-coding 100 percent of the time, and the name Jane is never associated with an "M." However, the name Val is associated with an "M" 50 percent of the time.

Figure 1 shows the distribution of likelihoods that a name is male. The sharply bi-modal shape of the distribution makes it immediately clear that most names are either strongly male or strongly female. Approximately 40 percent of all names in the 347.8 million stable records always belonged to people whose sex-coding was "M", while about 52 percent of all names belonged to people whose sex-coding was "F". Only about 3 percent of all names were perfectly gender-neutral. While Figure 1 characterizes the distribution of names, Table 3 presents the number of individuals with stable records whose names were male or female 90 percent, 95 percent, and 99 percent of the time. Panel A shows that most males had strongly male names. Only 6.3 percent of males had names that were associated with males in fewer than 90 percent of the cases, and 81.5 percent of males had names that were associated with males 99 percent of the time or more. Panel B shows similar results for females, although females are slightly more likely to have names that are female in fewer than 90 percent of the cases. The take-away message is that there is very little ambiguity in determining the gender of a person's name: very few given names are gender-neutral, and very few people have gender-neutral names.

3.1.3 Determining which name changes are likely transgender

I have no prior notion of what likelihood threshold is sufficient to confidently classify a name as likely male or likely female, so I present results for three progressively demanding sets of thresholds: 10 and 90 percent, 5 and 95 percent, and 1 and 99 percent. That is, for the first set, a name is categorized as likely male (female) if 90 percent or greater (10 percent or lower) of all stable records 9 with that name had a sex-coding that was "M" ("F"). I call this the 90 percent confidence threshold. I similarly define the 95 percent confidence threshold and the 99 percent confidence threshold. Note that name combinations that meet the 99 percent threshold also meet the 90 and 95 percent thresholds, so these are not mutually exclusive categories. If a person's first name is gender-neutral (i.e., if it does not meet the 90 percent confidence threshold or higher), then I use the gender of that person's middle name.

This framework considers a name change to be consistent with a gender transition if (a) both the original and the new first name fall within the specified confidence threshold, (b) the gender of the original first name matches the original sex-coding on the account, (c) the gender of the new first name is different than the gender of the original first name, and (d) the name change is not reversed later. 9

9 The requirement that the gender of the original name matches the sex-coding on the account ensures we do not falsely identify people as likely transgender if they change their gender non-conforming names to gender-conforming names later in life. That is, if a boy named Sue changes his name to Stu, requirement (b) in the list above prevents us from accidentally counting this person as likely transgender.

To illustrate the application of these rules, we return to Table 1. Jane Johnson's (PIK=09) name change from John is considered consistent with a gender transition, since both John and Jane surpass the 90 percent confidence threshold. The same logic applies to John Brown's (PIK=10) name change from Jane Brown. However, Jane Thompson's (PIK=07) name change from Val is not classified as consistent with a gender transition, since the name Val does not meet the minimum confidence threshold. If this person's original middle name was identifiably male, then we would classify Jane Thompson as transgender.

Of course, the actual cases in the Numident are not as straightforward as those described in Table 1. In particular, both accidental perturbations in how an individual's first name is spelled over the lifecycle (e.g., transposing letters), as well as intential differences (e.g., using nicknames), will lead that person to be identified as a name-changer and potentially transgender. Falsely identifying someone as a name-changer is problematic in its own right, but it can lead to serious measurement error when we also use the gender of a person's middle name. For example, it is common for married women to list their maiden surname in the middle name field. Since many surnames are also used as male's names (e.g., Harrison, Jefferson, etc.), false-positive identification of a woman as a likely transgender man can be an issue. To illustrate using fake names, suppose "MARY JANE RICHARD" gets married to "WILLIAM WILLIAMS." Mary files an SS-5 to adopt her married name and change her maiden name to her middle name. Either when filling out the form or when transcribing it, two letters in her first name are accidentally transposed, so her new name is encoded as "MAYR RICHARD WILLIAMS". Because her first name changes from "MARY" to "MAYR", she falls into the set of potential transgender individuals. "MARY" is a female name, but "MAYR" is neither male nor female. If we use the gender of her new middle name, “RICHARD”, she will be falsely classified as likely transgender, since her first name changed from the traditionally female “MARY” to the traditionally male “RICHARD”. To avoid these scenarios, I use the SAS SPEDIS function to eliminate people with small spelling distances between the new first name and the original first name. I also exclude using the gender of the (new) middle name if that middle name ever appeared as that person’s last name. The benefit of these rules is that they reduce the chances of false-positive classification of individuals as likely transgender. The cost, however, is many transgender people in the data will be identified as non-transgender. Using information on sex-coding, however, can reduce even further the chances of false-positive measurement error.

3.1.4 Sex-coding changes

Using information on name changes alone is a more inclusive strategy for identifying likely transgender individuals, since it does not require changes in sex-coding, which typically require costly and not always preferred genital sexual reassignment surgery. Nevertheless, identifying those who changed the sex-coding associated with their SSNs is interesting in its own right and can provide additional reductions of false-positive measurement error. Returning to the example in Table 1, we see that four years after Jane Johnson (PIK=09) changed her first name from John, she changed her sex-coding from “M” to “F”. At the same time that John Brown (PIK=10) changed his name from Jane, he also changed his sex-coding from “F” to “M”. John Miller’s (PIK=08) sex-coding change, however, is not consistent with a gender transition for two reasons. First, his first name did not change. Second, his sex-coding changed before he turned 16. Such scenarios are relatively common among sex-coding changes, and appear to reflect corrections to mis-entered sex-coding during the process of applying for an SSN, which is why the age-rule is important for our estimates.

Once likely transgender individuals are identified in the SSA data (allowing the identification to vary by confidence threshold of names and whether or not their sex-coding changes on the account) it is not difficult to calculate the number transgender-consistent claims per year and the number of SSN holders identified as likely transgender who were alive during the 2010 Census. Similarly, it is straightforward to explore how many people change their names and sex-coding at the same time, how many change their names first, and (for the latter group) what the average number of years is between the name and the sex-coding change. In order to learn about race, ethnicity, and geographic characteristics, however, we must link the SSA records to the most recent decennial census.

3.2 Record linkage

To ensure non-disclosure of SSNs other Personally Identifiable Information (PII), the Census Bureau assigns each individual a PIK, which is a unique identifier used internally by the Census to link individuals across data sets for statistical and research purposes. The Census Bureau's Center for Administrative Records Research and Applications (CARRA) uses probability record linkage techniques and personal information such as name, date of birth, and residential location to assign PIKs to individuals' census records, where possible, through the Person Identification Validation System (PVS) (see Wagner and Layne, 2014, for more details). The Census Bureau developed the PVS as part of its ongoing research to improve data quality and reduce costs of data collection.

The 2010 Census collected information on 308,745,538 individuals residing in the U.S. during April 2010.10

10 This includes residents of group quarters, Puerto Rico, and those under 16 years of age.

Of these, 280,989,153 records (91.0 percent) were assigned PIKs. These 281.0 million records included 10,486,988 duplicate PIKs. This occurs when two census records have the same probability of being a match to the person represented by a given PIK. I am able to resolve 53 percent of the duplicates by dropping records of 5,552,757 individuals who were identical to their duplicates in terms of state of residence, reported sex, race, Hispanic origin, and age (the main census variables used in my analysis). Because the information is identical for these records, dropping all but one can be done without introducing new measurement error.

I next reduce the number of duplicates that need to be resolved by linking the remaining census records (including the remaining 4.9 million duplicates) to the 89,667 individuals in the SSA Numident who were determined to be likely transgender. The match yielded 90,686 potential linkages, of which 88,663 are unique matches and 1,019 are duplicates. I begin by keeping only duplicates with the closest agreement in age between the SSA records and the census records, which eliminates 346 duplicate records. I then randomly select one of the within-state duplicates, which eliminates 397 additional duplicates. This will not introduce bias to my estimates of the geographic distribution of likely transgender individuals, but it will introduce attenuation bias to estimates of race, Hispanic origin, sex-reporting, and age. Finally, I randomly drop the remaining 276 duplicate PIKs. I am left with 89,667 valid links, a 100 percent match rate.

4 Results

4.1 Results from administrative data

4.1.1 Transgender-consistent claims, 1936–2010

Table 4 shows the number of individuals appearing in the SSA Numident between 1936 and 2010 whose records meet the criteria, described above, for being likely transgender. The top Panel A gives results for the entire universe of SSN holders, while Panels B and C give estimates for Male to Female (MTF) and Female to Male (FTM) likely transgender individuals, respectively. The first row within each panel gives the number of people who changed their first name in a way consistent with gender transition, and the second row gives the number who changed their sex-coding in the same direction as their name change. Thus, the second row within a panel imposes more restrictive requirements for being identified as likely transgender than the first row within a panel. Moving from left to right across columns gives the results using the 90 percent, 95 percent, and 99 percent confidence thresholds for the gender of the original and new first names. Since these confidence thresholds are increasingly demanding, the most conservative results appear in the bottom-right cell of each panel, and the most inclusive estimate appears in the top-left cell.

According to the most inclusive criteria, which require a name change consistent with gender transition using the 90 percent confidence threshold, 135,367 individuals in the data are likely to have transitioned gender between 1936 and 2010. As the name confidence thresholds become more demanding, moving rightward across the columns, the figures attenuate; the 95 percent confidence threshold yields a count of 106,550 while 64,738 individuals have name changes that meet the 99 percent confidence threshold. These figures include any individuals who changed their first name from a traditionally female name to a traditionally male name (or vice versa), which is consistent with that person taking steps to be socially recognized as a gender distinct from the sex assigned at birth, regardless of whether that person has taken steps to alter his or her physical appearance through hormone treatment or surgery. The second row within each panel shows results requiring individuals to alter their sex-coding in the same direction as their name change. Between 1936 and 2010, the number of adult SSN holders who changed their first name and sex-coding in the same direction ranges from 21,981 (using the 99 percent confidence threshold) to 30,006 (using the 90 percent confidence threshold). It is difficult to know how the SSA responded to requests to change the sex-coding on a person's account during the first several decades of its existence, however from 1980 through 2013, evidence of scheduled or completed SRS was necessary to change the sexcoding on a person's account. The bottom row of each panel, therefore, may be include individuals whose gender transition included surgical intervention; uncertainty about how individual cases were handled prior to 1980 as well as how closely later policies were followed should be taken into account when interpreting these figures.

We now turn to the bottom two panels, which break Panel A out by gender. Looking at the bottom rows of Panels B and C, which require corresponding name and sex-coding changes, likely transgender women (MTF) make up a larger share of the total number of likely transgender individuals than do likely transgender men (FTM). This is consistent with much of the previous literature, 12 which finds that the rate of MTF transitions involving SRS is higher than FTM transitions involving SRS (see, for example, APA, 2000; Bakker et al., 1993; Tsoi, 1988; Pauly, 1968; W˚alinder, 1968). However, this relationship is reversed when using name changes but not sex-coding changes. Two possible mechanisms may underlie this reversal. On the one hand, since for over 30 years the SSA required evidence of scheduled or completed genital SRS to change the sex-coding on one's account, differences in the cost, complexity, and effectiveness of surgical interventions for FTM transitions versus MTF transitions could curtail the number of people who have changed their name and sex-coding from female to male, but not the number who have changed their names only. On the other hand, the results using name-changes alone in Panel C may suffer from more measurement error than the corresponding results in Panel B. Table 2 showed that females file more claims than males. Since females names are longer and more complex, variant spellings and errors in writing the first name (such as transposing two letters) can shift that person into the group being checked for likely transgender status (Lieberson and Bell, 1992). If the "new" first name is unique, then the algorithm will look to their middle name. Since females are more likely to be given male names or to use maiden surnames (which are often male first names) as middle names, they could be falsely identified as likely transgender (Goldin and Shim, 2004). I have taken several steps to minimize this type of measurement error, but I cannot eliminate it entirely. The results from the linked Census data in Section 4.3.3, support the second story, although I cannot rule out the possibility that both mechanisms are at play.

The results in Table 4 do not include those who have not legally changed their names, who have changed their names through common-law, whose name(s) do no meet my minimum confidence threshold, or whose gender transition did not involve a name change at all. Furthermore, my results do not include people who do not possess SSNs, who transitioned before attaining an SSN, or who transitioned before turning 16. For these reasons, these results are likely to undercount the number of transgender people in the data and should not be construed as an estimate of the total population of people who have transitioned gender over the time frame. Nevertheless, the results can inform us of general features of the data and trends in certain types of claims over time.