

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

A machine learning approach to linking medicaid and education records in New York city

Katharine Sadowski^{1,*}, Andy Martens², Kira Argenio², Sophia Day², Kevin Konty², Michael Gaile³, and Maria D. Fitzpatrick^{3,4}

Submission History

Submitted:	02/03/2026
Accepted:	21/08/2026
Published:	13/10/2026

¹Stanford University Graduate School of Education

²New York City Department of Health and Mental Hygiene Office of School Health

³Cornell University Jeb E. Brooks School of Public Policy

⁴NBER, 1050 Massachusetts Ave, Cambridge, MA 02138, USA

Abstract

Introduction

Linking education and health records can offer critical insights into how policies and interventions shape children's outcomes, yet these data linkage efforts are often constrained by inconsistent identifiers, incomplete population overlap, and variable data quality. We develop a supervised machine learning framework to link New York City Department of Education (DOE) and Medicaid records, one of the first large-scale linkages of its kind in the United States. Our two-step approach first deduplicates records within each dataset and then applies a random forest model to identify cross-dataset matches, integrating blocking, probabilistic scoring, and clustering to improve upon deterministic and off-the-shelf probabilistic linkage techniques while maintaining computational efficiency. Applying this framework, we achieve a high quality linkage, identifying matches for approximately 71% of records across the two systems. Match rates are consistent across birth cohorts and most demographic groups, though lower for some populations because the two systems capture partially overlapping subsets of children. Rather than optimising for a 100% match rate, we propose a framework for evaluating linkage quality under uncertainty, emphasising precision, transparency, and population representativeness. This approach provides scalable guidance for linking complex administrative systems and enables new opportunities to study how educational and health policies interact.

Keywords

administrative data linkage, administrative health care data, education records

Highlights

- Linking longitudinal administrative education and health records can generate critical evidence for researchers and policymakers studying the intersection of child health and academic outcomes, yet such linkage is challenging in the absence of a shared universal identifier and with only partially overlapping populations.
- We implemented a two-step supervised machine learning framework that first trains separate random forest models to de-duplicate records within each dataset, using within-dataset unique identifiers as training labels, and then applies both models to candidate cross-system record pairs incorporating blocking, probabilistic scoring, and one-to-one pruning to identify cross-system matches.
- In these two partially overlapping education and health data sources, we found that 71% had high quality matches, with match rates remaining stable across birth cohorts and increasing for individuals with longer observation histories.
- Match rates varied across racial groups and residential location, with lower rates for some subgroups reflecting differences in population coverage, name commonality, and administrative representation rather than deterioration in model performance.
- When linking partially overlapping administrative systems, researchers should prioritise linkage stability, representativeness, and precision over maximising match rates, and explicitly evaluate whether subgroup differences reflect true population non-overlap rather than data or model deficiencies.

*Corresponding Author:

Email Address: ksadow@stanford.edu (Katharine Sadowski)

Introduction

In the United States, children spend a substantial portion of their lives in schools, which serve not only as educational environments but also as key sites for the delivery of health services. Schools across the country provide health related programming such as asthma management initiatives, vision screenings, school-based health centers, and mental health interventions. While education records have long been used to evaluate how programs affect academic outcomes, understanding their effects on health outcomes remains challenging because health and education data systems operate independently, lack shared identifiers, and are governed by different regulatory frameworks.

Record linkage, the process of identifying records that refer to the same individual across datasets, offers a potential solution. The U.S. Census Bureau has been a leader in developing probabilistic linkage methods that estimate the likelihood that two records represent the same person when common unique identifiers are unavailable [1, 2]. These approaches, widely used in research linking administrative and survey data, have demonstrated that probabilistic and machine learning-based models can outperform deterministic methods, particularly when identifiers are incomplete or contain errors [3–5].

Beyond Census applications, other countries have made significant progress in linking administrative data across health and education systems. For example, countries with national health registries, such as the United Kingdom, Scotland, and Norway, have used population-level identifiers to study long-term associations between early education, health, and well-being [6–8]. However, these contexts differ substantially from the United States. In countries with national registries, every individual is represented, allowing researchers to assume near-complete population coverage and a shared “truth” against which linkage can be validated. In contrast, U.S. data systems are decentralised and fragmented across states and local agencies, creating unique methodological challenges for linkage and evaluation.

In this paper, we present a new framework for linking Medicaid claims with New York City Department of Education (DOE) student records, representing one of the first large-scale linkages of its kind in the United States. Using enrollment records from the DOE and Medicaid, we employ a supervised machine learning framework to improve on standard deterministic and probabilistic linkage methods. Our approach utilises carefully constructed training labels, enhanced feature engineering and calibrated thresholds to optimise the trade-off between precision and recall during de-duplication and cross-dataset matching. We achieve a high quality linkage, identifying matches for approximately 71% of records across the two systems, and we show how to interpret this match rate in the context of partially overlapping populations.

More broadly, our goal is to provide a transparent, generalisable framework for assessing linkage quality under uncertainty. Because many administrative datasets, especially those spanning sectors such as education, health, and social services, represent only subsets of the broader population, achieving a 100% match rate is neither possible nor desirable. Rather than optimising for complete linkage, we offer a set of principles and diagnostic tools that can be applied across

administrative contexts to evaluate linkage performance in real-world settings where full coverage is unattainable by design.

Methods

Data

Our study focuses on two populations in New York City: students and Medicaid enrollees. The New York City Department of Education (DOE), the largest school district in the United States, serves roughly one million students each year. New York City offers grades PK through 12. A substantial proportion of these students come from economically disadvantaged backgrounds; approximately 74 percent were classified as such in the 2023–24 school year [9]. We obtained enrollment records for all students attending school in New York City between the 2006–07 and 2024–25 school years with 21,078,606 total observations, representing 3,451,692 unique student IDs.

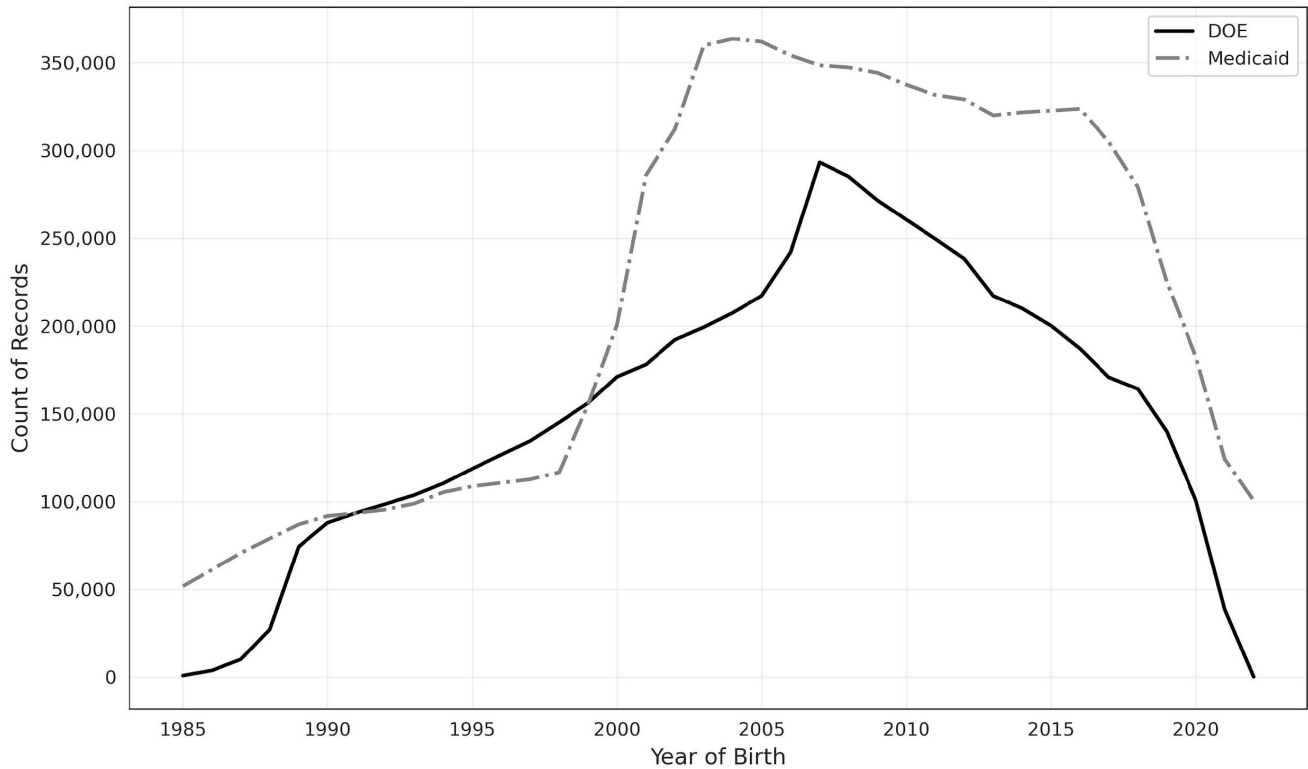
Medicaid is the largest public insurance program for low-income Americans, providing access to a wide range of medical services. Medicaid enrollment records were obtained from the Salient New York State Medicaid Enterprise System (NYSMES) and include all enrollments for individuals residing in New York City from January 2006 through June 2025. For consistency with the school-aged population, we restrict the sample to children between ages 3 and 20 during the observation window totalling 23,641,337 observations representing 3,541,508 unique enrollment IDs.¹

Although there is no shared identifier across sources, both follow a highly parallel structure. Each dataset includes its own internal unique identifier, along with individuals’ first and last names, date of birth, record year, gender, race, and residential ZIP code. This close alignment in available fields allows us to construct and apply a common matching model across datasets, as discussed below.

Figure 1 presents the distribution of unique person identifiers by year of birth from 1986 to 2022 in the underlying DOE enrollment and Medicaid administrative data, before applying de-duplication or matching procedures. While the figures highlight substantial overlap across cohorts, the overall distributions differ. The DOE data follow an approximately normal distribution, as expected for administrative education records concentrated around school-aged children. In contrast, the Medicaid data exhibit a bimodal distribution, with a pronounced increase in enrollment among individuals born between 2000 and 2004. This jump likely reflects several factors including major expansions in children’s Medicaid eligibility during the early 2000s (including indirect spillover from the Children’s Health Insurance Program expansion), improvements in administrative data capture, and a policy shift from strict re-certification toward continuous eligibility for Medicaid-eligible children [10–12]. Differences in the cohort

¹NYC DOE enrollment data covers all students in grades PK–12. While most students are between ages 4 and 18, children may enroll in PK in the calendar year they turn 4, meaning some begin at age 3 if they turn 4 by December of that school year. In addition, students with special needs or those held back may remain enrolled until age 21. To ensure complete matching, we therefore use Medicaid records for individuals ages 3 through 20.

Figure 1: Distribution of Person-Level Identifiers by Year of Birth in Raw DOE Enrollment and Medicaid Administrative Data



(birth year) distributions across the two datasets, together with improvements in administrative reporting over time, help explain why not all records have corresponding matches across sources.

Study Design

We design a two-step linkage algorithm with the goal of creating a unique person-level link between New York City Medicaid and Department of Education (DOE) records. We follow the standard pipeline for probabilistic matching: pre-processing, blocking, and matching [13], and extend it using supervised machine learning to improve accuracy and scalability. We calculate linkage rates across race and time to evaluate the model's performance and ensure equitable matching across demographic groups.

The first step involves de-duplication within datasets. Each dataset contains strong within-system unique identifiers, but a small share of individuals are assigned multiple identifiers due to administrative errors such as enrollment lapses, typographical mistakes, or re-enrollment under a new ID. We trained two supervised models, one for Medicaid and another for DOE, to identify and consolidate duplicate records within each dataset so that each individual is represented by a single record.

The second step focuses on cross-dataset linkage. We use the models trained in step one to evaluate candidate pairs across the Medicaid and DOE datasets. This step moves from searching for duplicates (i.e., records belonging to the same person) within a single dataset, to identifying potential matches between the two systems. Records were blocked on key identifiers to reduce computational complexity and then scored using the models to classify true matches. Creating

the models that are used to eventually link the two datasets is a large reason why we do not simply de-duplicate based entirely on unique ID numbers. This gives us an advantageous situation, where ID numbers are strong enough to be used as truth labels in model training, but can be improved upon as final duplicate labels.

This staged approach improves upon deterministic and purely probabilistic methods by first addressing data quality issues within each dataset and then applying machine learning to identify cross-dataset matches. Within these two steps, our methodology consists of two sub-steps: 1) pre-processing data and 2) de-duplication. The de-duplication tasks consist of blocking the datasets to generate candidate pairs, training a supervised machine learning model, scoring candidate pairs with match probabilities, and applying clustering to group matched records. We provide a detailed description of each process in the following sections.

Data Pre-Processing

We start by pre-processing both datasets individually. Pre-processing ensures the data are accurate, consistent, and comparable across sources by correcting errors, standardising formats, and handling missing or duplicate information. The goal is to improve data quality so that subsequent linkage steps, such as blocking and matching, are more reliable and efficient. Both Medicaid and DOE include structured administrative records containing personal identifiers, such as name, date of birth, gender, zip code, and race. For each dataset, we conduct the following cleaning steps: 1) Separate date of birth into day, month, and year; 2) Remove leading and trailing whitespace from strings; 3) Separate name suffixes into their own variable; 4) Standardise all ZIP codes to five

digits (e.g., leading zeros if necessary); 5) Remove extraneous characters from string fields, e.g. '3', '/', ','; 6) Standardise race and gender labels²; and 7) De-duplicate exact copies. After performing these steps, we move onto de-duplication.

Data De-Duplication

Once standardised through pre-processing, we proceed to de-duplication to ensure each individual is represented by a single record. This step reduces redundancy and computational complexity before cross-dataset linkage. Because de-duplication requires comparing millions of records, we use a blocking strategy to limit comparisons to records sharing key attributes such as name or date of birth, dramatically reducing runtime while retaining nearly all potential true matches.

De-duplication is necessary for several reasons. First, individuals often appear multiple times within each dataset, once per school year in the DOE data and once per month enrolled in the Medicaid data, making direct linkage computationally prohibitive and noisy. Second, unique identifiers cannot be relied upon exclusively: individuals may reenter these systems and occasionally receive new IDs following administrative updates. Finally, because our broader goal is to develop a linkage model that functions even when unique identifiers are unavailable allowing the trained model to be repurposed for linking the DOE and Medicaid data, we use probabilistic supervised learning methods to de-duplicate each dataset prior to cross-dataset linkage.

Following pre-processing, we implement the following steps for the de-duplication process: 1) Block each dataset against itself to generate candidate record pairs for comparison; 2) Train a random forest classifier to distinguish true duplicate pairs from non-matches; 3) Score each candidate pair using the trained model to assign a match probability; and 4) Apply clustering to group together records identified as duplicates.

Blocking Procedure

The goal of blocking is to create a manageable set of candidate pairs for de-duplication. Instead of comparing every record in the file to every other record, which is computationally infeasible, blocking groups records that are plausibly the same person and only compares records within the same block. A candidate pair is thus a pair of records that share a block key and will be evaluated by the model; it is not yet judged to be a match.

We start the process with name standardisation. Names in administrative data are highly noisy: they contain typographical errors, alternate spellings, nicknames, multi-part

structures, and occasional first/last name transpositions. To make blocking effective, we standardise names using complementary techniques: 1) Phonetic coding (Soundex): Soundex assigns the same code to similarly sounding names (e.g., *Michael/Michelle*). This captures phonetic similarity that exact string comparisons miss; 2) Split-name fields: Multi-part names (e.g., a first name with two parts *Mary Ann*) are split into separate fields so that alternative recording conventions, using or omitting middle names, do not prevent matches.

We next move to name clustering producing cluster-based identifiers that summarise the overall similarity between names that we then use for blocking. We recognise that a substantial literature on clustering algorithms for duplicate detection has compared a range of approaches, including partitioning, hierarchical methods, and centroid-based methods, finding that no single algorithm is universally optimal and that the choice involves tradeoffs across accuracy, scalability, and noise tolerance [14, 15]. However, we do not use clustering here for the de-duplication or entity resolution itself, but rather to improve our blocking and so the production of candidate record pairs. With blocking, we aim to generate a pool of candidate records pairs wide enough to contain virtually all true matches yet refined enough to enable computational scalability, a primary methodological constraint in production-scale administrative record linkage, including at the U.S. Census Bureau [16, 17] where tractable blocking methods are used precisely because they enable linkage at scale.

We construct name clusters by first computing character n-grams (2-grams): We compute 2-gram profiles (e.g., John → "jo", "oh", "hn") to detect substring-level similarity, such as John within Johnathan; then we create sparse vectors for each record. These vectors are the length of the entire set of 2-grams in the data, populated with the number of times each n-gram occurs in the first and last name. We also evaluated 3-gram profiles, which yielded sparser vectors and less even cluster size distributions; longer n-grams are poorly suited to short names, where a four-character given name produces only two 3-grams, and are less tolerant of single character error. We therefore use 2-grams throughout.

We then use MiniBatchKMeans—a form of k-means clustering appropriate for sparse vectors—to generate 1,000 clusters. Each cluster contains groups of records with somewhat similar n-grams, and as such, with names that bear a resemblance to one another. Each record is then assigned the identifier of its cluster. These cluster identifiers serve as robust, noise-tolerant features for blocking.

The number of name clusters ($k = 1,000$) reflects a trade-off between comparison cost, which falls with k , and coverage, since finer partitions risk assigning records of a true pair to different clusters. Comparing cluster size distributions across $k \in \{250, 500, 1,000, 2,000, 4,000\}$ (Supplementary Appendix 1), total comparisons decline smoothly with k while the largest cluster does not: for last names it is essentially unchanged between $k = 1,000$ and $k = 4,000$ (90,911 and 89,543 records). Because the largest block dominates comparison cost, further increases yield little benefit while fragmenting the distribution. Clusters with fewer than five records rise from zero at $k \leq 1,000$ to 2 at $k = 2,000$. We therefore set $k = 1,000$, which at least for this clustering algorithm and in this NYC population, eliminated the very large clusters present at lower values

²DOE and Medicaid report demographic fields using different coding conventions. In the Medicaid records, gender is reported in various formats (e.g., 'Male,' 'male,' 'M'), which we standardize to align with the DOE categories ('Male,' 'Female,' and 'X'). While Medicaid records biological sex, which remains stable throughout our sample, the DOE allows for self-reported gender, which may vary over time. For consistency across datasets, we refer to this variable as 'gender' throughout the paper. Race categories also differ slightly across systems: Medicaid reports "Asian or Pacific Islander," "American Indian," and "Multiple Races," while DOE uses "Asian," "Native American," and "Multi-racial." For consistency across datasets and to ensure comparability during linkage, we harmonize all records to the DOE categories: Hispanic, Black, Asian, White, Multi-racial, and Native American.

without producing sparsely populated ones. Because our multi-pass design requires a pair to be captured in only one pass, coverage does not depend on any single cluster assignment.

We then move to blocking. After constructing name clusters, we implement a multi-pass blocking strategy to maximise recall while keeping the candidate set tractable. We considered 14 candidate blocking variables (name groups, Soundex codes, split-name fields, DOB variants, geographic variables, name cluster, etc.). From these we curated 24 five-variable combinations such that each combination included at least one name variable (either first or last) and at least two DOB variants; each combination contained at most one first name variable and at most one last name variable. To capture potential issues with non-standardised names and dates of birth, we created combinations with swapped entries (e.g., first and last name swapped) for a total of 84 blocking passes (see Supplementary Appendix 2).

Each pass generates blocks by matching records on the chosen variables (exact matches on grouped codes or DOB variants; approximate matches where applicable), and all within-block record pairs are added to the union of candidate pairs. We take the union across all passes to prioritise recall: a pair need only appear in one pass to enter the candidate set. The candidate pairs produced by these 84 passes serve two purposes: a labelled subset if used to train the classifier models in the de-duplication pipeline, and the full set is scored by the trained classifier to identify within-dataset duplicates. In total, 17,729,937 candidate pairs were generated for the DOE data, and 13,929,606 for the Medicaid data. Multi-pass blocking increases recall (fewer true matches are missed) at the cost of more candidate pairs; the specific combination rules above were chosen to balance recall and computational cost. We report blocking diagnostics, including reduction ratio, pairs per record, and candidate recall against a labelled validation set, in Supplementary Appendix 3.

Modelling Training and Scoring

To enable cross-dataset linkage between Medicaid and DOE records, which do not share a common identifier, we trained random forest models to identify record pairs belonging to the same person. This serves a dual purpose: first, to de-duplicate records within each dataset so that each person is represented by a single record, and second, to generate a model that can later be applied to link the two datasets. Random forests were selected for their balance of interpretability, performance, and ability to flexibly model interactions between predictors. We used the unique IDs in each dataset as truth labels.

We engineered features to quantify similarity between record pairs, capturing both exact matches and approximate similarities. Our approach follows standard practices in the record linkage literature [18–21], which emphasise combining multiple similarity metrics to improve match accuracy. We include both traditional and computationally efficient measures of string similarity, such as Jaro-Winkler, Hamming distance, Soundex, to capture phonetic and typographical variation in names. While more complex hybrid similarity models can yield marginally higher accuracy, we opted for these standard measures due to their strong performance and substantially lower computational cost given the scale of our data.

In addition, we construct custom name-part representations that compare multi-part names by splitting them into substrings and boolean indicators for common name order switches (e.g., first and last name reversed), which are both standard in administrative linkage models [1, 2]. To illustrate how these features combine in the model, consider a candidate pair with “Martha Ann Doe” (DOE) and “Martha Doe” (Medicaid), with otherwise identical date of birth, race, gender, and zip code. The model receives this pair as a vector of similarity features: a low Jaro-Winkler similarity on the full first name string (“Martha” vs. “Martha Ann”), a high partial-match score from the split-name representation (recognising “Martha” as a component of “Martha Ann”), an exact match on last name, and exact matches on date of birth, gender, and zip code. The random forest combines these features and outputs a high match probability, classifying the pair as a match. A deterministic rule requiring exact string equality on first name would reject the same pair despite agreement on every other field. We also include exact match indicators for gender, date of birth, race, and zip code. We further incorporate name frequency metrics to capture the relative distinctiveness of names, which recent studies have shown to improve match accuracy in large datasets [4]. Together, these features capture both the structural and probabilistic dimensions of record similarity while maintaining computational efficiency. These features extend the blocking metrics and allow the model to account for typos, alternative spellings, and name order differences.

Random forest hyperparameters were selected via grid search on 66% of the record pairs, with the remaining 34% held out as a test set. On this 66%, we used 3-fold cross-validation for the grid search. To prevent data leakage from individuals appearing in multiple administrative records across years (for example, a student appearing as a first-grader in 2012 and as a twelfth-grader in 2023), both the train/test split and the cross-validation folds were grouped at the individual level using a stable cross-time identifier constructed during data pre-processing. This identifier resolves cases where the same individual appears under multiple administrative IDs across years, ensuring that records belonging to a given individual share a single grouping value. We note that the practical risk of within-individual leakage is somewhat attenuated in this setting because the unit of analysis is the record pair rather than the individual record, and the features capture pairwise similarity (e.g., string distances between names) rather than identifying attributes of any specific person; however, we adopted group-aware splitting as a methodological precaution consistent with best practice in supervised record linkage. A small fraction of individuals may not have been fully resolved by the cross-time identifier, leaving some residual potential for within-individual leakage. We searched over maximum tree depth values of 10, 15, 20, 25, 30, and 35, with 20 providing the best cross-validated performance and selected as the final hyperparameter. Each leaf node required at least 100 samples, and the forest consisted of 500 trees. Pruning was not applied, as tree complexity was controlled through the depth and leaf size parameters. These choices balance model flexibility and generalisation, ensuring stable predictions while capturing complex interactions in our large dataset.

The models performed strongly in identifying duplicate records. After training the model on the DOE dataset, we

assessed its validity by using it to predict matches in the Medicaid data, where unique Medicaid identifiers allow us to directly verify true matches. We similarly evaluated the Medicaid-trained model by applying it to the DOE dataset. Under this cross-dataset validation, the DOE trained model achieved an F1 score of 97.5%, with 97.0% precision and 97.9% recall, while the Medicaid-trained model achieved an F1 score of 95.1%, with 95.2% precision and 95.0% recall. High recall ensures that nearly all true duplicates are correctly identified, while high precision minimises false positives, both of which are critical for maintaining accurate individual level records. The strong performance under cross-application validation indicates that the models generalise well beyond the training data and provide reliable inputs for cross-dataset linkage.

Clustering

After scoring candidate pairs with the random forest model, we convert the resulting pairwise match probabilities into clusters of records representing unique individuals. Because the model assigns only a probability that two records belong to the same person rather than a unique identifier, clustering is used to infer individual level groups. Each record is treated as a node in a graph, with edges drawn between nodes that exceed the match probability threshold. Through the transitive property of matches, records indirectly connected via other high-probability links are merged into the same cluster. For example, if Record A matches Record B and Record B matches Record C, all three are assigned to a single cluster, even if the direct probability between A and C is slightly lower. Each cluster is then assigned a new unique identifier, ensuring that every individual is represented by a single record within the dataset.

Thresholds for drawing edges are calibrated piecewise depending on whether record pairs share an original identifier. Pairs that share an existing ID are expected to score highly and are therefore held to a stricter threshold to prevent false merges, while pairs lacking a shared ID use a slightly lower threshold to allow for genuine duplicates introduced by administrative errors. To minimise the chance of merging distinct individuals (for example, twins with similar identifiers), exact first name agreement is required whenever a lower threshold is applied. Across all settings, thresholds are conservatively high, biasing against false positives even at the expense of leaving some duplicates unmerged.

Using this approach, we identified 3,541,508 unique clusters (individuals) in the Medicaid data and 3,451,833 in the DOE data. These de-duplicated, cluster assigned datasets serve as the foundation for cross-dataset linkage, ensuring that each person is represented by a single, unified record in each system.

This framework mitigates key sources of linkage uncertainty. De-duplication errors typically stem from either removing records that are not true duplicates or failing to remove true duplicates. Our conservative thresholding strategy prioritises the first type of accuracy, reducing the likelihood of false merges while allowing remaining duplicates to be reconciled during the clustering process. Finally, the structured nature of the data, consistent pre-processing across systems, and the choice of a random forest classifier

further enhance robustness. While alternative models such as XG Boost could yield incremental performance gains, our validation demonstrates that the current framework achieves high precision and recall with minimal risk of overfitting.

Linking DOE and Medicaid Records

After de-duplicating both datasets, much of the groundwork for record linkage has already been completed. Linking two datasets is conceptually similar to de-duplication: instead of linking a dataset to itself, we link the DOE and Medicaid datasets. Because of this similarity, much of the code and process developed for de-duplication, cleaning, name handling, blocking, and feature engineering, is reused with minimal modification. After blocking the DOE and Medicaid records, we generated 31,898,140 candidate pairs to consider.

Some differences arise at the model scoring stage. For each candidate pair, we use both random forest models previously trained during de-duplication (one on DOE data, one on Medicaid data) and take the average of the two match probabilities. This approach leverages the patterns learned from each dataset while maintaining consistency in feature representation.

We validated the model in two ways. First, we compared the random forest match probabilities to an unsupervised Fellegi-Sunter model, finding a correlation of $r=0.87$, indicating strong agreement [22]. Second, we manually reviewed a sample of candidate pairs within decile threshold ranges signifying high-, medium-, or low-probability matches. High-probability matches generally had exact or near-exact agreement on all fields, medium probability matches often shared similar names, and low-probability matches shared very few characteristics.

Selecting a Match Probability Threshold

We selected a baseline match threshold of 0.32 using the bimodal-valley approach to threshold selection in classical record linkage [23]. The distribution of predicted match probabilities is bimodal (Figure 2), with a high-density mode of clear non-matches near zero, a high-density mode of clear matches near one, and a low-density region between them. We selected 0.32 within this low-density region, prioritising balance between false-positive and false-negative errors under the assumption that their costs are approximately equivalent for our de-duplication application.

We further validated this threshold against an independent benchmark: the Fellegi-Sunter (1969) match weight, the canonical metric in classical record linkage. For each candidate pair we computed the Fellegi-Sunter weight and compared the resulting distribution to our model's classifications (Figure 3). The Fellegi-Sunter distribution is itself bimodal, with a non-match mode at lower weights and a match mode at substantially higher weights, separated by a low-density region. Our probability threshold of 0.32 partitions candidate pairs at this same natural separation: pairs classified as matches by our model (grey bars) correspond almost entirely to the high-weight Fellegi-Sunter mode, while pairs classified as non-matches (black bars) correspond to the low-weight mode. The concordance between our calibrated probability threshold

Figure 2: Medicaid and DOE Model Match Probability Comparison

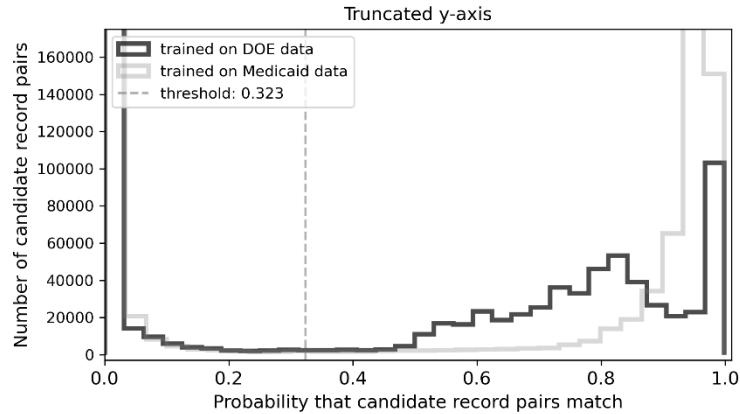
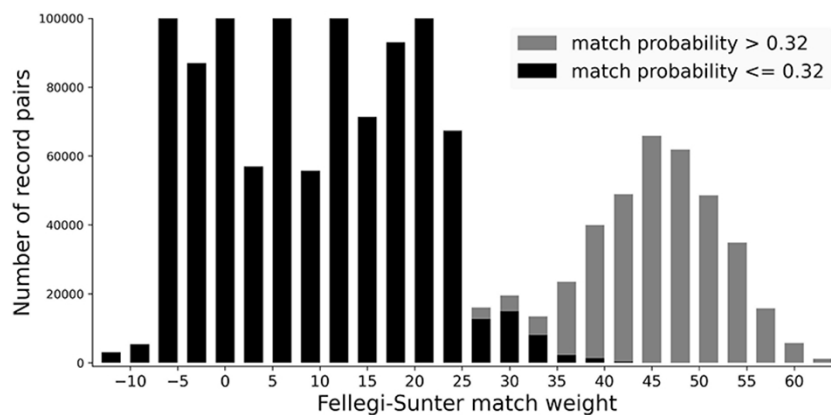


Figure 3: Fellegi-Sunter Match Weight



and the Fellegi-Sunter weight-based separation supports the threshold choice.

In applications where the cost of false positives (incorrectly matching records belonging to different individuals) exceeds the cost of false negatives (failing to match records belonging to the same individual), a more conservative threshold may be appropriate. We provide such an alternative at 0.75, identified through visual inspection of the predicted probability distribution in Figure 2 as the point beyond which the right mode (clear matches) becomes dominant and the contribution from the valley region effectively vanishes. We note that threshold selection in this region of the distribution is inherently a judgement call: a more risk-averse application could reasonably select an even higher threshold, such as 0.85 or 0.90, at the cost of further reductions in recall. While less precisely grounded than the 0.32 threshold, the 0.75 cutoff shifts the precision-recall trade-off toward higher precision at modest cost to recall: applying it reduces the average yearly match rate in the DOE data from 77.6% to 74.1% and in the Medicaid data from 83.5% to 79.4%. We discuss both thresholds because different applications may warrant different choices depending on the relative costs of false-positive and false-negative errors.

Applying the Threshold and Pruning to a Final Crosswalk

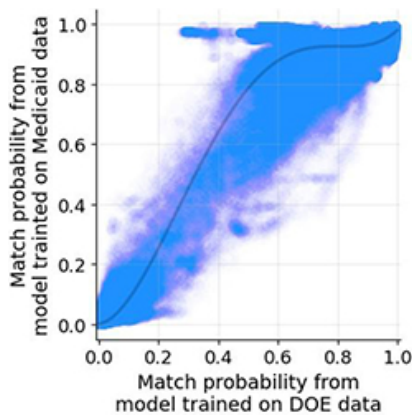
We filtered out matches below a 0.32 probability threshold. High confidence matches (exact agreement on name, DOB,

gender, zip code, and race) were linked immediately. We then applied a pruning algorithm: 1) Sort by student ID and select the highest-probability Medicaid match; 2) Repeat by Medicaid ID to select the highest-probability student match; 3) For reciprocal matches, accept the link; 4) For conflicts (e.g., Student A $\rightarrow$ Medicaid B but Medicaid B $\rightarrow$ Student C), we applied tie-breaking bonuses: higher priority for exact name matches and a smaller bonus for DOB matches. Student IDs can have more than one match in cases where the records are exactly duplicates. These are often cases where a Medicaid recipient experienced an enrollment lapse and erroneously received a new Medicaid ID. These are generally the same person and to keep the full set of Medicaid records they are linked to the same student. The final output is a “crosswalk” dataset, pairing DOE Student IDs with Medicaid IDs and assigning each link a match probability, ready for downstream analysis.

Concordance Between Medicaid- and DOE Trained Models

We next assess the agreement between the Medicaid-trained and DOE trained match models. Ideally, matches with low predicted probabilities in the Medicaid model would also be low in the DOE model, and similarly for high-probability matches. The distribution of match probabilities shows that most matches cluster near 0 or 1, indicating strong concordance between the two models (Figure 4).

Figure 4: Correlation Between Medicaid and DOE Trained Models



Quantitatively, 90.8% of matches fall within ± 0.1 of each other, highlighting the robustness of the predictions across data sources. Nevertheless, a non-trivial subset of matches receives higher probabilities from the Medicaid model than from the DOE model (and vice versa). This divergence may reflect differences in the underlying data quality suggesting that each model contributes complementary information to the overall linkage process.

Results

Following the de-duplication of both administrative datasets, we identified 3,451,692 unique individuals in the DOE records and 3,541,508 unique individuals in the Medicaid records. The cross-dataset linkage process generated 31,898,140 candidate record pairs, yielding 2,479,742 unique linked individuals between the two systems. This corresponds to universal match rates of approximately 71.8% for DOE records and 70.0% for Medicaid records. That is, for every 100 children in the NYC DOE records, 71.8 are successfully linked to a Medicaid record; for every 100 children in Medicaid records, 70.0 are linked to a DOE record. These figures indicate substantial overlap between the two populations, consistent with the expected share of public school students enrolled in Medicaid. The remaining unmatched records largely reflect differences in population coverage, such as privately insured students or Medicaid-enrolled children not attending NYC public schools during the study period.

While these overall linkage rates provide a high-level summary of performance, they may obscure meaningful variation across time or population subgroups. In particular, earlier birth cohorts exhibit lower Medicaid enrollment, as shown in Figure 1, which could naturally reduce observed match rates. Similarly, variation in racial and ethnic composition across cohorts may influence name-based matching if not properly accounted for. To better understand these dynamics and assess the stability of the linkage process, we evaluate match performance along two dimensions: 1) temporal variation in match rates across birth cohorts, and 2) match rates by individual's demographic characteristics and variable missingness.

Match Rates by Birth Cohort

We first examine match rates by birth cohort to assess whether linkage quality varies over time. Stable match rates across cohorts would indicate consistent model performance, while systematic deviations could reflect changes in administrative coverage or data quality. The average yearly match rate is 77.6% for DOE records and 83.7% for Medicaid records.³ Across both samples, match rates remain high and broadly stable, following an arc that peaks in the early 2000s before experiencing a modest decline (Figure 5). This trajectory likely reflects two intersecting factors: the expansion of Medicaid eligibility and administrative record-keeping in the early 2000s [11, 12], and differences in the length of longitudinal observation across cohorts. Midperiod cohorts benefit from extended follow-up, increasing the number of observed records and thus the cumulative opportunities for successful linkage. In contrast, later cohorts have shorter observation windows, which mechanically reduce the probability of capturing sufficient identifying metadata for a match. Consequently, these patterns appear to be driven primarily by variation in data coverage and duration of follow-up rather than deterioration in model performance.

This interpretation is reinforced when examining match rates by the number of years an individual is observed in the data (Figure 6). Match rates increase substantially with longer observation histories. In Medicaid, match rates rise steeply from approximately 25% among individuals observed for a single year to roughly 80% by five years of enrollment, after which they plateau. DOE records begin at a higher baseline (approximately 55% with one year observed) and increase steadily with additional years. These patterns underscore that linkage success improves with repeated administrative observations, consistent with the cohort-level trends described above.

Finally, the low match rates for the earliest birth cohorts (1985–1986) are mechanical. DOE enrollment records begin in the 2006–07 school year, meaning individuals born in these years would have been beyond typical school age when the data start. As a result, only those who were retained or receiving special education services would appear in the DOE records, leading to limited coverage and mechanically lower Medicaid match rates for these cohorts.

Match Rates by Demographics and Missingness

Next, we examine whether linkage performance varies across the attributes used in the matching process—race, gender, residential borough, and attribute missingness—and assess whether differences in match rates can be explained by variation in sample composition.⁴ Table 1 highlights that the

³It is important to note that the match rates presented in Figure 5 represent the proportion of records matched within each cohort, rather than the proportion of unique individuals matched. Because children may appear in multiple years of administrative records, cohorts with higher record persistence (i.e., more school enrollments or Medicaid claims per child) will mechanically exhibit higher record-level match rates even if the underlying unique individual match rates remain constant. The universal unique individual match rates reported above therefore provide a more conservative and comparable measure of linkage success across the full study population, while the cohort-specific rates illuminate how match quality varies across time periods.

⁴Residential zip code is used directly in the matching model. However, because New York City contains nearly 180 distinct zip codes, we

Figure 5: Match Rates Between DOE and Medicaid Records by Birth Year

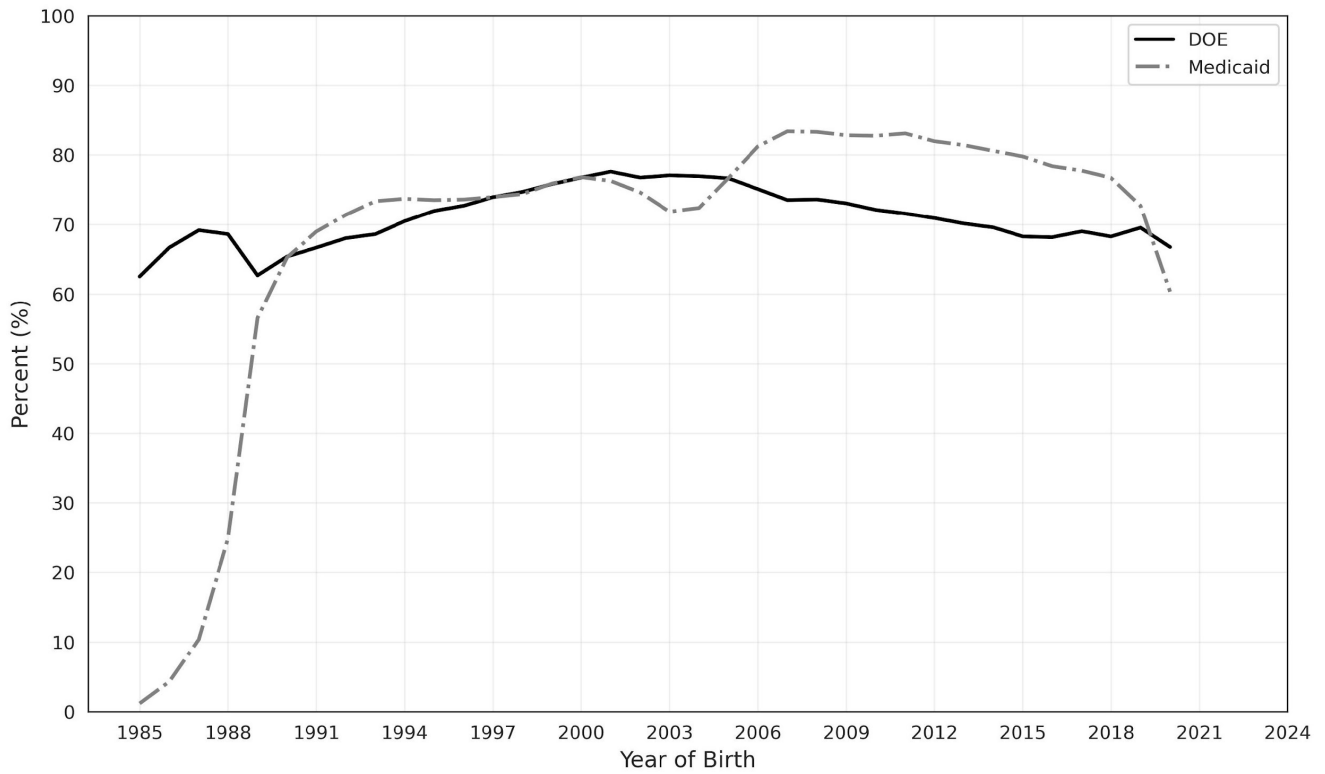
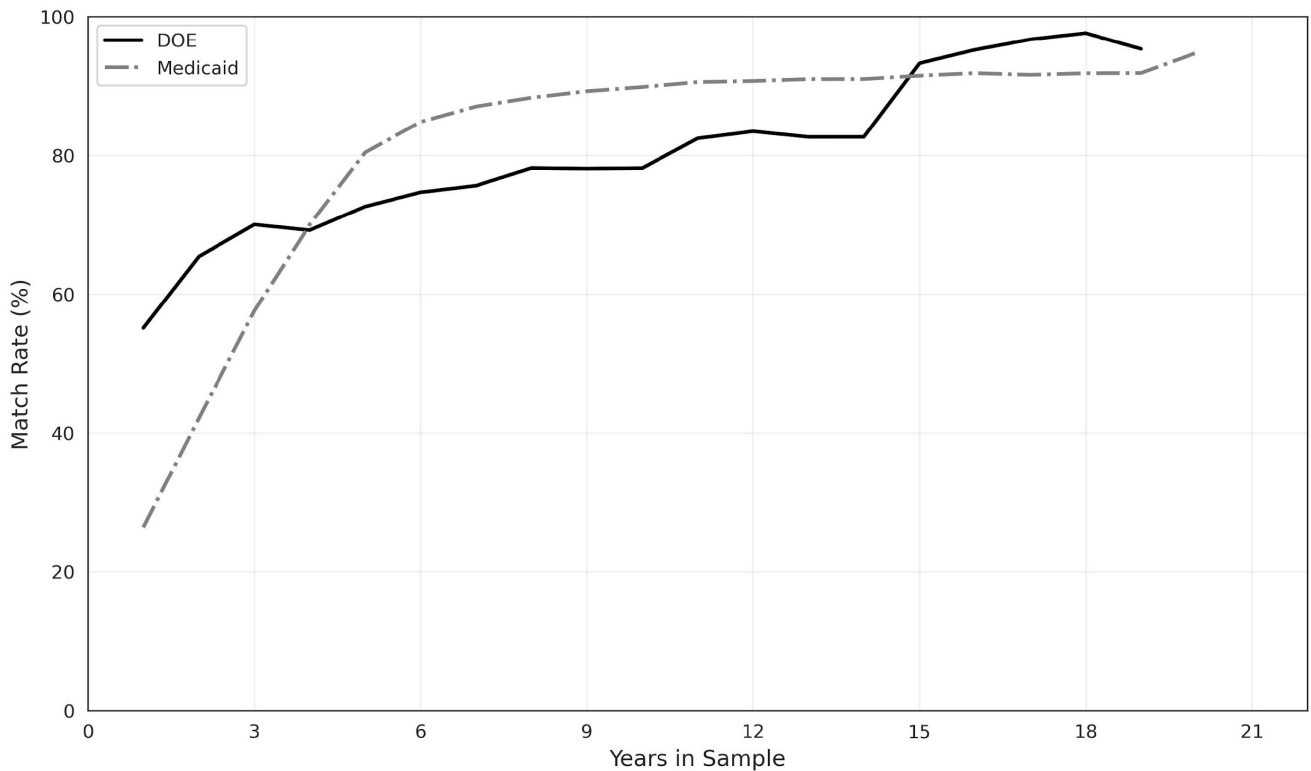


Figure 6: Match Rates by Years Observed in DOE Enrollment and Medicaid Administrative Data



DOE and Medicaid samples represent meaningfully different populations. While both datasets are roughly evenly split by aggregate zip codes to boroughs to facilitate clearer interpretation of spatial variation in match rates. Specifically, we classify Manhattan (100xx–102xx), the Bronx (104xx), Staten Island (103xx), Brooklyn (112xx), and Queens (110xx–111xx, 113xx–116xx).

gender, the DOE sample includes a higher share of Asian, Black, and White children, whereas the Medicaid sample includes a larger proportion of multi-racial children and records with missing race. Children in the Medicaid sample are also more likely to reside in Brooklyn and the Bronx, boroughs

that are, on average, lower-income than the rest of New York City.

Overall, missingness in the variables used for linkage is low. In both datasets, first name, last name, date of birth, and gender are fully observed. Race is missing for 0.4% of DOE records and 4.5% of Medicaid records, while residential borough is missing for approximately 4% of DOE records and 2.5% of Medicaid records. Consistent with expectations, records missing information used in the linkage process are less likely to match, and match rates among records with missing race or borough exhibit greater variation over time.

Match rates vary substantially across racial groups and are strongly related to representation in the underlying data. In both the DOE and Medicaid samples, White children and children with missing race information have noticeably lower match rates than other groups. We fit separate logistic regression models to estimate the association between each demographic characteristic (race, gender, and residential borough) and the probability of matching, controlling for the remaining two characteristics in each model. The results are presented as adjusted odds ratios with 99% confidence intervals. Our results indicate that, relative to White children, Asian, Black, Hispanic, and Native American children are between four and eight times more likely to successfully match (Table 2). These differences persist across datasets, suggesting that lower match rates among White children are not driven by a single data source but instead may reflect systematic differences in name commonality or population representation. In contrast, match performance does not differ meaningfully by gender: males and females are equally likely to be matched in both datasets (Table 2). Differences by residential borough are more pronounced. The Bronx consistently exhibits the highest match rates in both the DOE and Medicaid samples. In the DOE data, however, match rates vary more widely across boroughs, with Staten Island exhibiting the lowest match rate, approximately 61%. Staten Island is one of the wealthier boroughs in New York City and is less represented in the Medicaid sample, which may limit the availability of high quality candidate matches.

Taken together, these results suggest that differences in linkage performance are driven less by demographic characteristics themselves and more by differences in population coverage, name commonality, and sample overlap across data sources. Because New York City is highly diverse and naming conventions vary across linguistic and cultural contexts, name-based similarity measures—including phonetic encodings, string similarity, and embedding-based features—may perform differently across groups. Importantly, match rates are broadly stable over time within racial groups, indicating that the combined feature set is temporally robust, even if performance is weaker for groups with more common names or lower representation across linked datasets (Figures 7, 8).

Several mechanisms are consistent with the lower match rates observed for certain subgroups. First, records with missing information are mechanically more difficult to link because the algorithm has fewer attributes on which to establish similarity. Even small amounts of missingness can reduce the probability of identifying a high confidence match, particularly in large administrative datasets where many individuals share similar names or dates of birth.

Second, differences in match rates may reflect differential representation across the two data sources. Children from higher-income households are less likely to be enrolled in Medicaid, which mechanically limits the pool of potential matches regardless of linkage quality. As a result, lower match rates for groups that are disproportionately higher-income may reflect true population non-overlap between the DOE and Medicaid data rather than algorithmic failure.

As a concrete example, consider the pattern observed for White children where match rates for White children are approximately 50% in both the DOE and Medicaid samples. If lower match rates were driven solely by under-representation in Medicaid, one would expect lower match rates in the DOE data but comparatively higher match rates within the Medicaid sample, where Medicaid-eligible White children should still be linkable to DOE records. The similarity in match rates across data sources suggests two non-mutually exclusive interpretations. One possibility is differential institutional coverage. Medicaid-eligible White children may be less likely to appear in DOE administrative records, for example, due to private schooling or other alternative educational arrangements, while White children enrolled in DOE records may be less likely to qualify for Medicaid. A second, non-mutually exclusive possibility is reduced model performance for more common names, where higher name frequency increases ambiguity in name-based similarity measures. While the current analysis cannot distinguish between these explanations, the symmetry in match rates across data sources points to differential population overlap as an important factor shaping observed linkage performance.

Match Quality and Validation

Matches were evaluated based on their predicted match probabilities from the random forest model. Quantitative assessment of linkage accuracy relies on the cross-dataset validation described in Section 2dii, in which dataset-specific identifiers provide ground truth for record pairs. The manual inspection described here serves a complementary and distinct purpose. Rather than estimating error rates, it characterises the types of discrepancy present at each level of the probability distribution, so that model scores can be interpreted substantively and an appropriate acceptance threshold can be identified for our setting. The inspected samples are necessarily small relative to the scale of the linkage; they are intended to make the score interpretable, not to quantify accuracy.

Matched pairs were first stratified into equal-count deciles and a random sample of 50 pairs was inspected within each. Because the distribution of match probabilities is heavily skewed toward 1 (Figure 2), all ten deciles fell within a narrow high-probability range, and even the lowest contained only minor discrepancies such as an unmatched middle initial or a hyphenated surname truncated in one system. We therefore repeated the exercise using ten fixed-width bands of width 0.1, which span the full probability range and allow the transition from confident to uncertain matches to be observed directly.

Across bands, discrepancy patterns vary systematically. Pairs above 0.9 agree almost exactly, differing at most by a single character, a compound surname, or an additional middle

Table 1: Distribution of Records and Match Rates by Attribute

Variable	DOE		Medicaid	
	Percent (%)	Match rate (%)	Percent (%)	Match rate (%)
<i>Race</i>				
Asian	15.61	78.09	11.47	85.43
Hispanic	40.29	86.18	42.85	90.27
Black	27.16	81.08	20.40	88.49
White	14.60	49.72	11.79	51.52
Native American	0.86	81.10	0.16	83.48
Multi-racial	1.11	50.10	8.83	87.32
Missing	0.43	60.59	4.51	70.61
<i>Gender</i>				
Female	48.72	78.13	49.31	83.51
Male	51.28	77.14	50.69	83.52
X	0.02	55.63		
<i>Residential Borough</i>				
Queens	27.56	75.25	24.86	86.89
Bronx	22.00	87.52	24.96	89.00
Manhattan	10.79	71.02	10.16	85.92
Brooklyn	29.52	79.07	33.21	76.69
Staten Island	6.08	60.68	4.29	88.56
Missing	4.07	71.97	2.52	66.26

Note. All figures are based on the authors' calculations. The table includes the average yearly percent of children by race, gender, and residential borough in the DOE and Medicaid records in the first column, and the average yearly match rate in the second. The match rate is calculated as the average number of unique individuals who matched between the DOE and Medicaid files divided by the total number of unique de-duplicated individuals in the DOE or Medicaid records, respectively. While DOE records include three gender categories: Female, Male, and X, Medicaid records include only two categories: Female and Male.

Table 2: Adjusted odds Ratios for Match Success by Attribute

Variable	DOE		Medicaid	
	aOR	99% CI	aOR	99% CI
<i>Race (Reference: White)</i>				
Asian	3.603	(3.587, 3.620)	5.451	(5.421, 5.482)
Black	3.688	(3.673, 3.704)	7.049	(7.014, 7.084)
Hispanic	5.650	(5.627, 5.673)	7.977	(7.941, 8.012)
Multi-racial	0.967	(0.956, 0.978)	6.064	(6.026, 6.102)
Native American	4.293	(4.224, 4.362)	4.239	(4.088, 4.396)
Missing	1.478	(1.452, 1.505)	1.923	(1.911, 1.936)
<i>Gender (Reference: Female)</i>				
Male	0.948	(0.946, 0.951)	1.014	(1.011, 1.018)
X	0.465	(0.368, 0.588)		
<i>Residential Borough (Reference: Bronx)</i>				
Brooklyn	0.772	(0.768, 0.776)	0.690	(0.687, 0.693)
Manhattan	0.460	(0.458, 0.463)	0.826	(0.821, 0.831)
Queens	0.551	(0.549, 0.554)	1.090	(1.085, 1.096)
Staten Island	0.450	(0.447, 0.453)	1.949	(1.931, 1.968)
Missing	0.422	(0.419, 0.425)	0.331	(0.328, 0.334)

Note. All estimates are based on the authors' calculations. We report adjusted odds ratios (aORs) from separate models estimated for race, gender, and residential borough. Column (1) reports the point estimate of the adjusted odds ratio, and column (2) reports the corresponding 99% confidence interval. Odds ratios reflect the likelihood of a successful match relative to the omitted reference category in each specification. DOE records include three gender categories (Female, Male, and X), while Medicaid records include only two categories (Female and Male).

initial; no sampled pair in these bands was judged incorrect. Between 0.6 and 0.8, pairs typically share at least part of the

surname, a nickname or spelling variant of the given name, and one auxiliary field that is missing or discrepant. This

Figure 7: Match Rates by Race Over Time for DOE Records

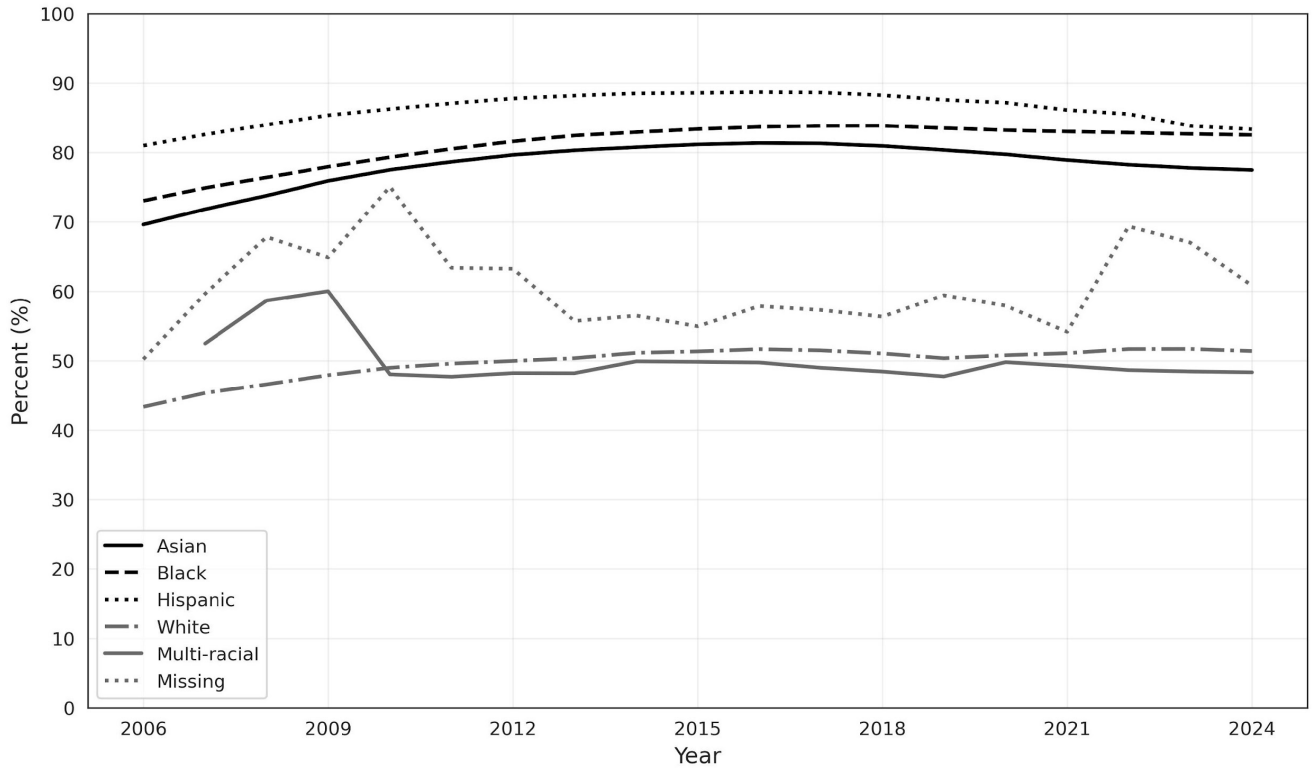
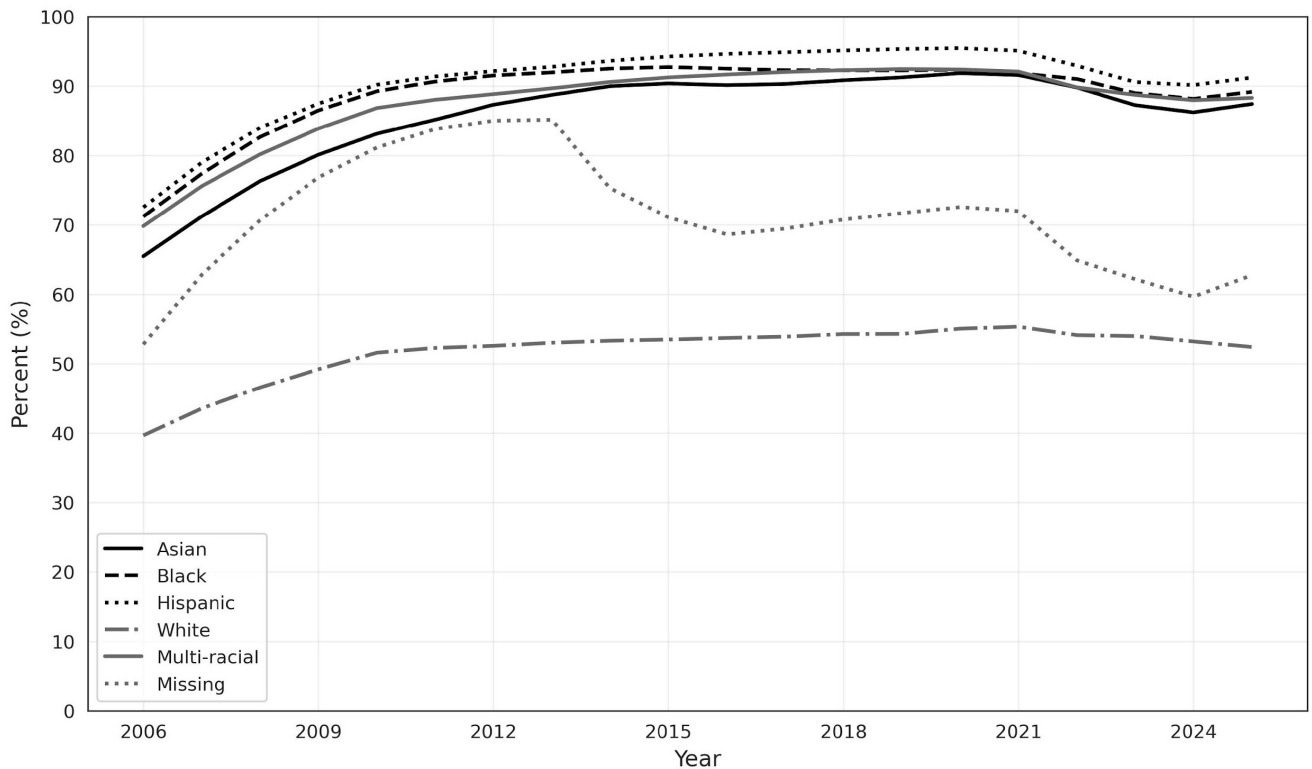


Figure 8: Match Rates by Race Over Time for Medicaid Records



variation is consistent with informal name recording and with fields that change over time. Near 0.4, agreement remains substantial but uncertainty increases: given names may be unrelated nicknames, and some pairs agree on given name, date of birth, race, and ZIP code while differing entirely on surname. Where the given name is common, such pairs are plausibly distinct individuals. The distinguishing feature

of lower probability pairs is not the accumulation of small differences but disagreement on the individual's name as a whole, which is the basis for our acceptance threshold. While missed duplicates cannot be quantified precisely, the one-to-one pruning step ensures that an unmatched duplicate leaves its original record correctly represented, minimising false merges.

Feature importance analysis further supports the interpretability of the model: name, date of birth, and gender are the strongest predictors of true matches, with ZIP code and race contributing meaningfully, though to a lesser extent (Supplementary Appendix 4). This aligns with expectations that names, gender, and birth dates provide the most distinctive identifying information in administrative records.

Together, these analyses confirm that the linkage process performs robustly across time and population subgroups. Variations in match rates primarily reflect genuine changes in the underlying populations served by DOE and Medicaid, rather than deficiencies in the model or data quality. These results underscore the validity of the linkage and highlight the potential of integrated education–health administrative data to support future policy and research applications.

Discussion

Building on our validation results, which demonstrate high linkage precision even under incomplete population overlap, this paper contributes to a growing literature on linking education and health administrative records. Unlike countries with national health registries, the United States lacks a unified system capturing all health encounters. Insurance coverage is fragmented, and while Medicaid offers near-universal coverage for low-income children, it does not include all students. New York City thus represents an ideal but complex case: it is home to a large and diverse population of Medicaid-enrolled children, but also a substantial share of students who are privately insured or uninsured. In this setting, perfect overlap between datasets is not achievable, and traditional linkage benchmarks that emphasise near-complete match rates are inappropriate. Instead, we introduce a framework for evaluating match quality under uncertainty, one that accounts for partial population overlap, variable data quality, and the realities of large-scale administrative systems.

Rethinking Match Quality

Much of the existing linkage literature treats higher match rates as evidence of better linkage quality. However, when datasets represent related but nonidentical populations, maximising coverage may increase the risk of false positives, and the appropriate benchmark is not 100% but the expected size of the population intersection. The Census Bureau's Person Identification Validation System illustrates the upper bound: linking surveys to a near-universal administrative reference file yields Protected Identification Key assignment rates of roughly 90% for major surveys conducted since 2000 [2, 24]. When the two linked samples instead represent partially overlapping populations, the achievable rate is necessarily lower. In our context, the absence of perfect overlap between DOE and Medicaid records reflects real-world differences in eligibility and participation, not algorithmic failure. About 74% of NYC public school students are classified as economically disadvantaged each year [9]. The threshold for that classification (185% FPL) is higher than the Medicaid eligibility threshold for children in New York (154% FPL), so the share of students Medicaid-eligible each year is somewhat lower than 74%. However, kids move in and out of poverty over

time, so the share ever Medicaid-eligible at some point during their childhood is higher than the single year share. With these two effects pushing in opposite directions and no NYC-specific lifetime estimate available, we surmise ~75% is a reasonable approximation on the achievable match rate between DOE and Medicaid records. Our observed rates of 71.8% (DOE) and 70.0% (Medicaid) sit close to this ceiling and substantially exceed the 27% recovered by a deterministic match on cleaned first name, last name, date of birth, and gender.

We therefore emphasise quality over quantity focusing on whether matches are accurate, stable, and representative rather than pursuing an unattainable 100% match rate. A key feature of our approach is the use of dataset-specific unique identifiers for cross-validation: by applying a model trained on one dataset to the other—where independent identifiers provide a ground truth—we can rigorously evaluate model performance even when the underlying populations differ. This strategy allows us to maximise the information contained in each dataset while maintaining high confidence in the accuracy of identified matches.

Data Quality and Model Robustness

Administrative data often contain inconsistencies such as misspellings, missing fields, or reassigned identifiers. Our feature engineering approach mitigates these challenges by incorporating multiple name-similarity measures (e.g., Jaro-Winkler, n-grams, Soundex), name-part representations, and exact demographic matches. Feature importance analysis confirms that name and date of birth variables drive model performance, while race, ZIP code, and gender play secondary roles. This structure ensures that the model remains resilient even when auxiliary demographic fields are incomplete or inconsistently recorded, an especially important property in the context of Medicaid and education data, which are collected by distinct agencies for different administrative purposes.

Linking Under Uncertainty

We explicitly frame record linkage as an exercise in uncertainty management rather than a deterministic classification task. Our high thresholds for de-duplication bias the model toward false negatives (i.e., conservatively leaving some duplicates in) while avoiding false positives that would erroneously merge distinct individuals. The clustering and pruning steps then reconcile remaining uncertainty by identifying transitive relationships among records while enforcing one-to-one pairings across systems. This multi-stage design prioritises data integrity and interpretability over computational efficiency.

Evaluating and Validating Matches

To evaluate match quality, we implement stratified validation by match probability, inspecting random samples across deciles. This approach confirms that high-probability pairs reflect near-perfect administrative agreement, while lower probability pairs differ only in predictable, realistic ways such as hyphenation or middle initials. These qualitative diagnostics, combined with conventional precision, recall, and F1 metrics, provide a transparent framework for evaluating linkage quality when complete ground truth is unavailable. We separately

examine coverage showing that match rates differ across racial and ethnic groups but are stable within groups over time, indicating that linkage coverage is not sensitive to administrative change.

An important caveat applies to these subgroup diagnostics. Observed differences in match rates across racial and ethnic groups are not directly interpretable as differences in linkage performance. A lower match rate for a given group may reflect genuine differences in Medicaid eligibility and enrollment, such as a smaller true population intersection, or differential model performance (for example, if name-similarity features behave less consistently for names subject to greater variation in transliteration, ordering, or hyphenation across administrative systems). Our design cannot separate these mechanisms: doing so would require a subgroup-representative sample of individuals with known enrollment status in both systems, which no available data source provides. The explanations we offer for subgroup differences should therefore be read as hypotheses for future work rather than established findings.

Broader Implications

Our framework establishes a model for linking large-scale administrative datasets in contexts where populations partially overlap, identifiers are imperfect, and uncertainty is inherent. Rather than treating unmatched records as failures, we view them as an expected and interpretable outcome of data system boundaries. This perspective reframes linkage success in terms of accuracy, transparency, and reproducibility, offering guidance for other health, education, and policy professionals seeking to link heterogeneous administrative data sources.

Conclusion

We present a comprehensive framework for linking Medicaid and Department of Education records in New York City, addressing challenges posed by multiple records per individual, data inconsistencies, and the absence of a shared unique identifier. By combining systematic data pre-processing, supervised machine learning, and clustering, our approach balances precision and recall while minimising the risk of false matches. The resulting linkage achieves high match quality under realistic conditions of incomplete population overlap and administrative uncertainty.

Beyond this specific application, our framework offers generalisable guidance for diverse stakeholders linking heterogeneous administrative datasets. It provides practical recommendations for feature engineering, threshold calibration, and validation in contexts where perfect overlap between data sources is not expected. The resulting linked data enable rigorous analyses of the connections between education and health, supporting evidence-based policymaking and advancing public health surveillance at the intersection of social and health administrative systems.

Acknowledgements

The conclusions in this publication are those of the authors and do not represent the views of New York City Department of Health and Mental Hygiene. All errors are our own.

Statement on Conflicts of Interest

Maria Fitzpatrick received financial support from the New York City Department of Health and Mental Hygiene in support of this research. All other authors report no financial support or conflicts of interest related to this research.

Ethics Statement

The NYC DOHMH Institutional Review Board determined this study constitutes public health surveillance rather than work (45 CFR 46.102.I.2) and is therefore exempt from informed consent requirements.

Abbreviations

N/A

Data Availability Statement

Data cannot be shared publicly because it is protected by the Family Educational Rights and Privacy Act (FERPA) and the Health Insurance Portability and Accountability Act (HIPAA) and there is no mechanism available to make the data public. Data belong to the New York State Department of Health (NYSDOH) and New York City Department of Education and Department of Health and Mental Hygiene and consist of private and protected information pertaining to minors. The views expressed in this study do not represent those of the New York State Department of Health (NYSDOH) unless NYSDOH agrees to the conclusions.

AI Disclosure Statement

Portions of the manuscript were edited for grammar and clarity using assistive AI with ChatGPT from OpenAI. The AI tool was used solely for language refinement and did not contribute to the study design, analysis, interpretation, or conclusions. The authors reviewed and verified the output. The authors take full responsibility for the content of the manuscript.

References

1. Layne M, Wagner D, Rothhaas C. Estimating Record Linkage False Match Rate for the Person Identification Validation System. US Census Bureau [Internet]. 2014;CARRA Working Paper Series(Working Paper #2014-02). Available from: <https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-02.pdf>.
2. Wagner D, Layne M. The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software. US Census Bureau [Internet]. 2014;CARRA Working Paper Series (2014-01). Available from: <https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-01.html>.

3. Abramitzky R, Boustan L, Eriksson K, Feigenbaum J, Pérez S. Automated Linking of Historical Data. *Journal of Economic Literature*. 2021 Sep 1;59(3):865–918. <https://doi.org/10.1257/jel.20201599>
4. Enamorado T, Fifield B, Imai K. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. *Am Polit Sci Rev*. 2019 May;113(2):353–71. <https://doi.org/10.1017/S0003055418000783>
5. Ruggles S, Fitch CA, Roberts E. Historical Census Record Linkage. *Annu Rev Sociol*. 2018 Jul 30;44(1):19–37. <https://doi.org/10.1146/annurev-soc-073117-041447>
6. Libuy N, Harron K, Gilbert R, Caulton R, Cameron E, Blackburn R. Linking education and hospital data in England: linkage process and quality. *IJPDS*. 2021 Sep 16;6(1). <https://doi.org/10.23889/ijpds.v6i1.1671>
7. Maret-Ouda J, Tao W, Wahlin K, Lagergren J. Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data. *Scand J Public Health*. 2017 Jul;45(17 suppl):14–9. <https://doi.org/10.1177/1403494817702336>
8. Wood R, Clark D, King A, Mackay D, Pell J. Novel cross-sectoral linkage of routine health and education data at an all-Scotland level: a feasibility study. *The Lancet*. 2013 Nov;382:S10. [https://doi.org/10.1016/S0140-6736\(13\)62435-6](https://doi.org/10.1016/S0140-6736(13)62435-6)
9. NYCPS. New York City Public Schools Data at a Glance [Internet]. 2025 [cited 2025 Nov 19]. Available from: <https://www.schools.nyc.gov/about-us/reports/nycps-data-at-a-glance>.
10. Arsen E, Scaffidi S, Myers N. Lessons from the Great Recession: New York Medicaid Enrollment During the COVID-19 Crisis [Internet]. United Hospital Fund; 2020. Available from: <https://uhfnyc.org/publications/publication/ny-medicaid-enrollment-during-great-recession-and-covid-19/>.
11. Fiorentini C. Growth in New York's Medicaid Enrollment and Costs [Fiscal Brief] [Internet]. New York City Independent Budget Office; 2013. Available from: <https://www.ibo.nyc.ny.us/iboreports/medicaid2013.pdf>.
12. McMorro S, Kenney GM, Long SK, Goin DE. Medicaid Expansions from 1997 to 2009 Increased Coverage and Improved Access and Mental Health Outcomes for Low-Income Parents. *Health Services Research*. 2016 Aug;51(4):1347–67. <https://doi.org/10.1111/1475-6773.12432>
13. Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 [cited 2025 Oct 22]. (AHRQ Methods for Effective Health Care). Available from: <https://www.ncbi.nlm.nih.gov/books/NBK253313/>. PubMed PMID: 25392892.
14. Hassanzadeh O, Chiang F, Lee HC, Miller RJ. Framework for evaluating clustering algorithms in duplicate detection. *Proc VLDB Endow*. 2009 Aug;2(1):1282–93. <https://doi.org/10.14778/1687627.1687771>
15. Reas R, Ash S, Barton RA, Borthwick A. SuperPart: Supervised graph partitioning for record linkage [Internet]. Amazon Science; 2018. Available from: <https://www.amazon.science/publications/superpart-supervised-graph-partitioning-for-record-linkage>.
16. US Census Bureau. Four Cooperative Agreements: Census Bureau Research on Record Linkage and Entity Resolution. 2021.
17. Steorts RC, Ventura SL, Sadinle M, Fienberg SE. A Comparison of Blocking Methods for Record Linkage [Internet]. arXiv; 2014 [cited 2026 May 6]. Available from: <https://arxiv.org/abs/1407.3191>. <https://doi.org/10.48550/ARXIV.1407.3191>
18. Grzebala P. Private Record linkage: Comparison of selected techniques for name matching. *Browse all Theses and Dissertations* [Internet]. 2016;(1474). Available from: https://corescholar.libraries.wright.edu/etd_all/1474.
19. Jaro M. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*. 1989;84:414–20. <https://doi.org/10.2307/2289924>
20. Winkler W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods* [Internet]. 1990. Available from: <https://eric.ed.gov/?id=ED325505>.
21. Yancey WE. Evaluating String Comparator Performance for Record Linkage [Internet]. U.S. Census Bureau; 2005. (Research Report Series). Report nos.: 2005–05. Available from: <https://www.census.gov/content/dam/Census/library/working-papers/2005/adrm/rrs2005-05.pdf#:~:text=The%20purpose%20of%20the%20string%20comparator%20in,and%20pairs%20of%20strings%20that%20do%20not>.
22. Fellegi IP, Sunter AB. A Theory for Record Linkage. *Journal of the American Statistical Association*. 1969 Dec;64(328):1183–210. <https://doi.org/10.1080/01621459.1969.10501049>
23. Belin TR, Rubin DB. A Method for Calibrating False-Match Rates in Record Linkage. *Journal of the American Statistical Association*. 1995 Jun;90(430):694–707. <https://doi.org/10.1080/01621459.1995.10476563>
24. Alexander J, Genadek K. Using administrative records to support the linkage of census data: protocol for building a longitudinal infrastructure of U.S. census records. *IJPDS*. 2023 Jan 11;7(4). <https://doi.org/10.23889/ijpds.v7i4.1764>

Supplementary Appendices

Appendix A1: Cluster Size Distributions Across Candidate Values of k for First and Last Name

k (cluster size)	Populated clusters	Largest cluster	Median cluster	Clusters < 5	Within-cluster comparisons (billions)
<i>First Name</i>					
250	250	414,935	28,413.50	0	327.19
500	500	170,414	15,381.50	0	136.27
1,000	1,000	65,006	7,568.50	0	69.16
2,000	2,000	38,865	3,536.00	0	39.08
4,000	3,999	38,670	1,603.00	4	26.16
<i>Last Name</i>					
250	250	378,311	27,359.50	0	327.3
500	500	156,251	13,704.50	0	155.14
1,000	1,000	90,911	6,867.00	0	82.84
2,000	2,000	89,553	3,185.50	2	54.5
4,000	3,986	89,543	1,390.50	29	43.59

Note. Clusters were generated by applying MiniBatchKMeans to character 2-gram count vectors, separately for first and last names. Populated clusters may be fewer than k because k-means can return empty clusters. Clusters < 5 counts sparsely populated clusters, which risk assigning records for the same individual to different clusters. Within-cluster comparisons is the number of record pairs requiring evaluation if the cluster identifier were the sole blocking key ($\sum n(n-1)/2$), and indexes computational cost; because cost scales quadratically with cluster size, it is dominated by the largest clusters.



Appendix A2: Blocking Passes Used for Record Linkage Between DOE and Medicaid Data

Standard Pass Blocking Variables

1. name first 1, gender, dob year, dob month, dob day
2. cluster first, cluster last, gender, dob year, dob month
3. name first 2, name last 2, dob year, dob month, dob day
4. cluster last, gender, dob year, dob month, dob day
5. name last 1, gender, dob year, dob month, dob day
6. cluster first, cluster last, dob year, dob month, dob day
7. name first 1, name last 1, gender, dob year, dob day
8. name first 1, name last 1, gender, dob month, dob day
9. name last soundex 3, name first soundex 3, gender, dob year, dob day
10. name last soundex 3, name first soundex 3, gender, dob month, dob day
11. name last soundex 3, name first soundex 3, dob year, dob month, dob day
12. cluster first, gender, dob year, dob month, dob day
13. name first 2, gender, dob year, dob month, dob day
14. name first 1, name last 1, gender, dob year, dob month
15. name last soundex 3, gender, dob year, dob month, dob day
16. name first 2, name last 2, gender, dob year, dob day
17. name first 2, name last 2, gender, dob month, dob day
18. cluster first, cluster last, gender, dob year, dob day
19. name last soundex 3, name first soundex 3, gender, dob year, dob month
20. cluster first, cluster last, gender, dob month, dob day
21. name first soundex 3, gender, dob year, dob month, dob day
22. name first 2, name last 2, gender, dob year, dob month
23. name first 1, name last 1, dob year, dob month, dob day
24. name last 2, gender, dob year, dob month, dob day

Switched First and Last Name

25. first 1, gender, dob year, dob month, dob day
26. cluster first, cluster last, gender, dob year, dob month
27. name first 2, name last 2, dob year, dob month, dob day
28. cluster last, gender, dob year, dob month, dob day
29. name last 1, gender, dob year, dob month, dob day
30. cluster first, cluster last, dob year, dob month, dob day
31. name first 1, name last 1, gender, dob year, dob day
32. name first 1, name last 1, gender, dob month, dob day



Continued

Appendix A2: Continued

-
- 33. name last soundex 3, name first soundex 3, gender, dob year, dob day
 - 34. name last soundex 3, name first soundex 3, gender, dob month, dob day
 - 35. name last soundex 3, name first soundex 3, dob year, dob month, dob day
 - 36. cluster first, gender, dob year, dob month, dob day
 - 37. name first 2, gender, dob year, dob month, dob day
 - 38. name first 1, name last 1, gender, dob year, dob month
 - 39. name last soundex 3, gender, dob year, dob month, dob day
 - 40. name first 2, name last 2, gender, dob year, dob day
 - 41. name first 2, name last 2, gender, dob month, dob day
 - 42. cluster first, cluster last, gender, dob year, dob day
 - 43. name last soundex 3, name first soundex 3, gender, dob year, dob month
 - 44. cluster first, cluster last, gender, dob month, dob day
 - 45. name first soundex 3, gender, dob year, dob month, dob day
 - 46. name first 2, name last 2, gender, dob year, dob month
 - 47. name first 1, name last 1, dob year, dob month, dob day
 - 48. name last 2, gender, dob year, dob month, dob day
-

Switched Part of Name (e.g., name first 1/name first 2)

-
- 49. name first 1, gender, dob year, dob month, dob day
 - 50. name first 2, gender, dob year, dob month, dob day
 - 51. name last 1, gender, dob year, dob month, dob day
 - 52. name first 1, name last 1, gender, dob year, dob month
 - 53. name first 1, name last 1, gender, dob year, dob day
 - 54. name first 2, name last 2, gender, dob year, dob day
 - 55. name first 2, name last 2, dob year, dob month, dob day
 - 56. name first 1, name last 1, gender, dob month, dob day
 - 57. name first 2, name last 2, gender, dob year, dob month
 - 58. name first 2, name last 2, gender, dob month, dob day
 - 59. name first 1, name last 1, dob year, dob month, dob day
 - 60. name last 2, gender, dob year, dob month, dob day
-

Switched Month and Day

-
- 61. name first 1, gender, dob year, dob month, dob day
 - 62. cluster first, cluster last, gender, dob year, dob month
 - 63. name first 2, name last 2, dob year, dob month, dob day
-



Appendix A2: Continued

-
- 64. cluster last, gender, dob year, dob month, dob day
 - 65. name last 1, gender, dob year, dob month, dob day
 - 66. cluster first, cluster last, dob year, dob month, dob day
 - 67. name first 1, name last 1, gender, dob year, dob day
 - 68. name first 1, name last 1, gender, dob month, dob day
 - 69. name last soundex 3, name first soundex 3, gender, dob year, dob day
 - 70. name last soundex 3, name first soundex 3, gender, dob month, dob day
 - 71. name last soundex 3, name first soundex 3, dob year, dob month, dob day
 - 72. cluster first, gender, dob year, dob month, dob day
 - 73. name first 2, gender, dob year, dob month, dob day
 - 74. name first 1, name last 1, gender, dob year, dob month
 - 75. name last soundex 3, gender, dob year, dob month, dob day
 - 76. name first 2, name last 2, gender, dob year, dob day
 - 77. name first 2, name last 2, gender, dob month, dob day
 - 78. cluster first, cluster last, gender, dob year, dob day
 - 79. name last soundex 3, name first soundex 3, gender, dob year, dob month
 - 80. cluster first, cluster last, gender, dob month, dob day
 - 81. name first soundex 3, gender, dob year, dob month, dob day
 - 82. name first 2, name last 2, gender, dob year, dob month
 - 83. name first 1, name last 1, dob year, dob month, dob day
 - 84. name last 2, gender, dob year, dob month, dob day
-

Appendix A3: Blocking Diagnostics for DOE and Medicaid Record Linkage

Measure	Computation	Interpretation
DOE De-duplication	<u>17,729,937</u>	Reduces candidate pairs by
Reduction Ratio (RR)	$(21,078,606)^2$	~399 million times
Medicaid De-duplication	<u>13,929,606</u>	Reduces candidate pairs by
Reduction Ratio (RR)	$(19,256,218)^2$	~376 million times
Cross-dataset Link	<u>31,898,140</u>	Reduces candidate pairs by
Reduction Ratio (RR)	$(13,929,606 \times 17,729,937)^2$	~12.9 million times
Candidate Recall	<u>2,479,742</u>	Matches ~7.8% of candidate
	31,898,140	pairs; sufficient coverage

Note: Reduction ratios (RR) quantify the extent to which blocking reduces the number of record pairs that must be compared. A smaller RR indicates a more efficient blocking strategy. Candidate recall measures the proportion of true matches retained after blocking; maintaining sufficient recall ensures that computational gains from blocking do not come at the cost of missing valid links.

Appendix A4: Feature importance for matching

