Prioritizing Equitable Representation, Sustainability, and Accuracy: The Deployment of Machine Learning Linkage Strategies During the COVID-19 Pandemic

Center for Health Statistics (CHS)[1], Center for Data Science (CDS)[2] & Office of Immunization (OI)[3]

3/31/2023

Authors:
Sean Coffinger[1], MA
Sean.Coffinger@doh.wa.gov

Seth Rothbard[2], MPH
Seth.Rothbard@doh.wa.gov

Chunyi Wu[2], PhD/MPH
Chunyi.Wu@doh.wa.gov

Alex Cox[2], SM
Alex.Cox@doh.wa.gov

Meredith Cook[3], PhD
Meredith.Cook@doh.wa.gov

Katie Hutchinson[1], PhD/MSPH
Katie.Hutchinson@doh.wa.gov

DOH 422-261 March 2023, to request this document in another format, call 1-800-525-0127. Deaf or hard of hearing customers, please call 711 (Washington Relay) or email doh.information@doh.wa.gov

Table of Contents

1 Executive Summary

On December 15, 2020, the first COVID-19 vaccination was recorded in Washington State [1]. As more of the population received vaccinations, it was important to assess the vaccination status of COVID-19 cases and identify vaccine breakthrough cases – where a positive lab test, either PCR or antigen, occurs at least 14 days after someone received their last recommended dose of an authorized, age-appropriate COVID-19 vaccine. To be successful, the Washington State Department of Health (DOH) needed to strengthen public health surveillance by establishing a robust linkage process between COVID-19 case and vaccine reports. This undertaking was complex. It involved linking COVID-19 case and vaccination records that are stored and managed in separate databases, with no shared unique identifying field like a person ID.

Due to the urgent need for linked data, a deterministic record linkage was originally implemented to pair the two datasets. However, as the numbers of cases and vaccine reports rapidly increased along with the volume of breakthrough cases, this strict and inflexible approach became inadequate. Inexact record matches were being missed, which disproportionally impacted BIPOC (Black, Indigenous, and People of Color) populations. Furthermore, as demand for and use of the linkage results expanded to high visibility efforts like predictive modeling of COVID-19 that informed public policy, a more robust system of record linkage was warranted.

A multi-center project team with representation from Washington State's Center for Health Statistics (CHS), Office of Immunizations (OI) and Center for Data Science (CDS) collaborated to adapt and implement a machine learning-based classification model originally developed by CHS. The statistical method calculates various distance metrics and flags candidate pairs, trains Radial Support Vector Machines (SVM) iteratively, and applies a consensus-based approach using Random Forest (RF) models for accurate and conservative record linkage. This n-gram based stacked classification model was selected because it was shown to be accurate, easy to automate and deploy, and specifically designed to address some of the health equity concerns stemming from known biases in traditional linkage methodologies.

The transition to a machine learning linkage resulted in approximately 11 percent more links between COVID-19 case and vaccine records compared to a deterministic linkage. Extensive quality assurance testing of the model demonstrated minimal increases in erroneous links, ensuring high linked data quality. Notably, the largest proportional increase was observed among BIPOC populations, specifically individuals identifying as Hispanic (22.7 percent), 'Other' race (19.5 percent), Native Hawaiian/Pacific Islander (12.4 percent), and Black/African American (11.6 percent). By identifying more links in these BIPOC groups and improving representation of BIPOC in the linked data, we aimed to promote equity in downstream data analysis and policy related decisions.

In conclusion, deterministic linkage strategies are insufficient for equitable surveillance when compared to a machine learning based classification. The implementation of the machine learning based linkage allowed DOH to assess the vaccination status of COVID-19 cases more accurately and equitably, while also strengthening other key surveillance efforts.

2 Background

On December 15, 2020, the first COVID-19 vaccination was recorded in Washington State, beginning a wide-reaching campaign to better protect the state's population from the most severe adverse health outcomes of COVID-19 infection [1]. To assess the degree of protection provided by COVID-19 vaccination, cases and vaccine records must be analyzed together. This involved linking COVID-19 case and vaccination records that are stored and managed in separate databases and do not share a unique identifying field like a person ID. Initially, the vaccination status of individuals was determined by self-reporting, through case investigation interviews and manual confirmation via matching vaccine administration records in Washington State's Immunization Information System (IIS). These preliminary strategies quickly became unsustainable due to the volume of cases and vaccinations. As a result, contacting every case was not possible, leaving linked data incomplete. It was necessary for public health surveillance to establish a robust linkage process between COVID-19 case and vaccine records to better identify vaccine breakthrough cases.

Due to the urgency to stand-up a record linkage process, a deterministic record linkage strategy was initially implemented to link COVID-19 case records to COVID-19 vaccine records. This methodology linked cases to vaccine records only if their first name, last name, and date of birth all matched exactly across records (here referred to as a 'deterministic' linkage). The Washington State Department of Health (DOH) routinely implements deterministic linkages because they are simple, quick to establish, and result in few false links due to the strictness of the model. This methodology was sufficient at the time, as the scope was initially focused on detecting possible vaccine breakthrough cases for case investigation purposes and not to determine population level protection. Due to a large degree of internal and external scrutiny and the sensitive nature of determining individual vaccination status for case investigations, it was necessary to maintain strict linkage criteria to identify linked records and keep Type I errors, falsely linking records belonging to different individuals, as low as possible.

While the strict linking criteria for the deterministic linkage methodology ensured very few false links, it became apparent that many links were being missed due to the variable data quality of records—in particular, the consistency of the spelling of names and formats of dates of birth across records. For example, there are instances where names are misspelled, transcribed incorrectly, parts of names are missing entirely, or a diminutive form of a name is used in one record and not the other. This inconsistency hampers the efficacy of a linkage as it can be difficult to determine if mismatching information across records truly refers to the same person.

Inexact but true matches need to be identified to accurately assess vaccination coverage, and overly strict rules enable inequities in data quality to manifest as inequity in linked data representation used for downstream analysis. It is important to acknowledge that structural racism has deep influences on the presence and quality of health data and on data science in general [2]. For example, it is known that BIPOC persons (Black, Indigenous, and People of Color) are more likely to have incomplete or incorrect information in their health records and are therefore less likely to be successfully linked in a record linkage [3,4]. One reason for this is the data collection systems in place are often designed with traditionally "Western" name and date standards in mind. Issues arise when failing to account for names with diacritical marks (e.g. letters with accent marks like á or the letter "n" with a tilde: ñ), different naming structures (such as multiple surnames or different ordering norms), names which transliterate into English with multiple correct spellings, and different date structures [3]. This illustrates how a linkage that requires an exact match across all records therefore disproportionally excludes non-White and/or Hispanic residents of Washington State.

The deterministic linkage was active during the late spring to the fall of 2021, and the amount of case and vaccine data requiring linkage was increasing. During this time period, Washington expanded its phased vaccine rollout to include a broader swath of its population, and the Delta variant caused a spike in COVID-19 cases. Concurrently, the results of the linkage were being used by an expanding number of different groups for a variety of purposes. This included monitoring vaccine effectiveness against new variants, updating the definition of a vaccine breakthrough case, and predictive modeling of COVID-19 infection trends, which were used to inform public policy in Washington State and a report on COVID-19 incidence and mortality published by the Center for Disease Control (CDC) [5]. Because the deterministic linkage methods resulted in systemic bias, such biases were inherited in any use of the data. It follows that continuing to use results from the deterministic linkage would disproportionately limit accurate surveillance of BIPOC communities and undermine the validity of DOH's evidence-based health recommendations [6]. A more robust system of record linkage was warranted.

When exploring alternatives, DOH first considered a traditional probabilistic record linkage strategy. Considering that this linkage process would be primarily carried out using the R statistical software, a probabilistic strategy was appealing, as there was a well-vetted "out-of-the-box" linking package available for use [7]. However, upon consideration, this option was found to be suboptimal.

The first problem was many proposed probabilistic linking strategies rely on the Fellegi-Sunter statistical framework for record linkage [8]. A key assumption underpinning the validity of the Fellegi-Sunter model is the conditional independence of the variables used to link records [8-10]. In practice, this assumption is often violated. For example, it would be dubious to claim that first name and sex assigned at birth are independent because first names are often associated with one sex more than the other. If first name and sex assigned at birth are present in a probabilistic linkage, the assumption of conditional independence would therefore not hold. The real-world consequence of this is usually written off as minor, and the statistical weights of the model are believed to be generally good enough to accurately link data [9]. However, relying on a linkage strategy which violated a statistical assumption which negatively impacts the accuracy of the linkage—even in a minor way—was nonetheless a detriment to this approach.

Secondly, linking records using a probabilistic linkage involves setting a probability threshold. This cutoff depends on the tolerance for false links. If a given record pair does not meet or exceed that threshold, they are not linked. The threshold can be manually adjusted and evaluated, but potential links outside the threshold require manual investigation—creating resource burden. Minimizing the volume of manual review is paramount for sustainable advanced linkages. Additionally, quantitative evaluation of the linkage results is made more difficult because the determination of a matching field is typically classified as a binary (i.e., either the variable comparison matches or not based on predefined logic). This binary result is then weighted and summed with other field comparisons. There often is no quantification of *how* close the fields actually are, resulting in flawed logic for complex and inexact comparisons.

Furthermore, traditional probabilistic methods do not enable continuous learning on decision criteria informed by manual quality assurance analysis. This makes it very difficult to customize algorithms to target subpopulation inequities. Even if decision boundaries and weighted scores are disaggregated by demographics initially, that boundary remains static unless manually reviewed and adjusted.

Alternatively, machine learning models re-ingest manually reviewed and corrected comparisons to learn from its mistakes. This leads to fewer comparisons requiring manual review over time and more accurate models. As a result, given the ability of machine learning classification models to be continually re-trained, machine learning linkage techniques are less susceptible to bias when compared to deterministic and probabilistic linkage strategies.

Lastly, probabilistic strategies, particularly out-of-the-box solutions, have demonstrated higher rates of error to identify the same number of links as targeted machine learning strategies. Again, extensive manual review can mitigate this, but the accuracy of the models is fundamentally different, with false detection rates being much higher in probabilistic strategies. This is a frequent finding, and one example is presented in Appendix G using the presented data.

With these limitations in mind, a record linkage methodology developed within Washington State's Center for Health Statistics (CHS) was proposed as an alternative: a non-probabilistic machine learning-based classification strategy. It addressed most of the concerns with the probabilistic linkage, was shown to be accurate, and was easy to automate and deploy to less experienced analysts performing quality assurance. This methodology performed remarkably well in previous linkage projects where vital records (birth and death records) and hospitalizations records were linked. Given these factors, the machine learning approach was chosen to replace the deterministic linkage.

3 Methods

For full methodological details of the implemented machine learning strategy, including procedural documentation beyond what is included in the methods section and appendices, please contact WA DOH CHS and refer to project MALAYAN CIVET (Machine Learning Auto-bot Yoking Advanced Networks of Covid Infection and Vaccine Event Tracking). Best practices were implemented to protect individual identities and protected health information.

3.1 Data Sources and Variables

Dose level vaccine data were pulled from the COVID-19 vaccine repository created from the Washington State Immunization Information System (WAIIS). COVID-19 case data were provided via the Washington Disease Reporting System (WDRS). Seven raw identifier variables existed in both datasets:

1. First Name
2. Middle Name Initial
3. Last Name
4. Sex
5. Birth Date (DOB)
6. Phone Number
7. Postal Code (ZIP)

The seven shared variables were then standardized, cleaned, and manipulated to enable comparison. Alternative fields were created to reflect the multiple ways the same information in each field could be interpreted. Alterative fields included splitting potentially compound names and including differing values of the same field from various sources. Other manipulated, or transformed, variables to account for common data input mistakes were also created as alternative variables. Table 1 displays a few commonly transformed alternative variables and their corresponding descriptions. A complete list of variables used for linkage is provided in Appendix A.

Table 1 – Description of transformed variables used in downstream machine learning classification algorithm.

Transformed Alternative Variable	Parent Variable	Description
DOB Switch	DOB	YYYYDDMM rather than YYYYMMDD
First Name Initial	First Name	The first letter initial of the primary source first name
First Name Frequency	First Name	The scaled proportion of first name incidence in that dataset
Last Name Frequency	Last Name	The scaled proportion of last name incidence in that dataset[i]

3.2 Candidate Link Identification

We merged all variables from each data source to form a pairwise comparison data frame where each row represented one individual possible link that contained both COVID-19 case and vaccine information. The “fuzzy” join contained custom criteria to avoid blocking on DOB, as many other strategies require (see appendix H for detailed definitions of fuzzy matching and blocking). Here, we “fuzzy” block on DOB if at least one other field matches deterministically. The fuzzy join criteria for inclusion into the candidate pair frame is listed in Table 2. Un-joined pairs were discarded as non-links.

Table 2- Inclusion criteria for blocking. Condition 1 and 2 must be met in one of the various ways.

Condition	Vaccine Field	Case Field	Inclusion Criteria	Or
1	DOB	DOB	Hamming Distance Below or Equal to 1	Or
	DOB	DOB Switch	Hamming Distance Below or Equal to 1	--
		AND		
2	First Name	First Name	Exact Match	Or
	Last Name	Last Name	Exact Match	Or
	Phone	Phone	Exact Match	Or
	ZIP	ZIP	Exact Match	--

3.3 Calculated Fields

Candidate pairs that were successfully merged then underwent field comparisons to create calculated distance metrics, relevant flags, agree/disagree comparisons, and other notable calculations. Name fields were compared via cosine and Jaro-Winkler distances. Dates of birth were compared using Hamming distance (see appendix F for detailed description of distance metrics employed). Name initials (first and middle), sex, and ZIP codes were all compared via a combination of two flags: one indicated if the values matched or not, and one indicated if one or more of the compared fields was missing. Lastly, phone numbers were compared via Damerau-Levenshtein distance. An exhaustive list of metrics used in downstream machine learning models and the description of each calculation is provided in Appendix B.

Each candidate pair received 15 calculated fields (or flags) which then were utilized in model training, development, and deployment.

3.4 Model Training

Supervised model training occurred iteratively and exclusively utilized Radial Support Vector Machines (SVM) for predictions and corrections. To assist in manual review for supervised training, a Platt Scaled SVM was trained with name matches and non-matches from a previous project conducted by CHS with a similar population distribution. These pseudo-probabilistic scores were used to inform the reviewer on how to initially sort sampled data.

Beginning with a random sample of 10,000 pairs, the calculated fields described in section 3.3 were calculated. The Platt Scaled SVM was then applied and grouped to the nearest tenth (resulting in 11 groups ranging from 0.0 – 1.0+). These groups were then resampled and 100 pairs from each group were manually reviewed. A novel, not scaled, radial SVM was then trained on this sample data, and manual scaling using logistic regression was used to select target groups for resampling. For example, if groups with scaled probabilities between .6 and .8 had samples with both links and non-links, this group was likely resampled to bolster the training set. Additionally, if groups 1.0+ and 0 only possessed links of one kind, they were likely to be omitted from further sampling. A detailed procedure of the sampling and resampling process for this project is found in Appendix C.

3.5 Model Testing

The final trained data were applied to the entirety of the historical dataset. Quality assurance (QA) and evaluation was initially conducted by the primary developer and included manually checking high cumulative scores, or summed distance calculations, and utilizing dimensionality reduction techniques (UMAP and TSNE) to visually look for outliers. Visually identified outliers were identified by checking clusters with erroneous links and manually observing hyperplane segmentations, a method that exemplifies one of the benefits of using SVMs. After the initial screening for false detections and missed links, additional reviewers were tasked with targeted quality assurance tasks. These tasks were reviewed by both the primary developer and the additional reviewer. In the case of a disagreement, a discussion would take place with the primary developer having the final say in the decision. No significant disagreements were identified in this project. The three tasks targeted the following: Type II errors, failing to link records belonging to the same individuals, caused by very common names, Type I errors caused by high cumulative scores, and Type I errors caused by sex disagreements with inexact name comparisons. In total, 2,423 comparisons that were deemed possible were manually reviewed and incorporated into the training set.

3.6 Historical Linkage Calculation, Predictions, QA and Retraining

Upon the completion of review by multiple reviewers and retraining, the model was then applied to a second run on the full pairwise candidate dataset. Runtime for this second iteration was 2.2 days to fully analyze the 1,402,141 case records and the 5,018,906 vaccination records for links utilizing unoptimized parallelization. Unoptimized parallelization was implemented to minimize runtime while prioritizing stability in spite of computational and hardware constraints. After the pairwise calculations, the comparisons were fed into trained SVM and random forest (RF) models trained on the training set established above. Default parameters were used in both SVM and RF models apart from SVM using a radial, rather than a linear, model. Links were determined only if both the RF and SVM agreed that the comparison was a link. This stacked approach ensured the mitigation of Type 1 errors, false detections, and erred on the conservative side in linking two records.

3.7 Ongoing Automation

A significant benefit to utilizing machine learning in data linkage is the ability to establish an ongoing automated process that learns from minimal manual QA. After the historical information was linked, an automated process to identify new links from incoming vaccination and case data was established. A diagram explaining this process is presented in Appendix D.

4 Results

The results in this section reflect findings from the historical run and exclude links established via the automated process established thereafter. The date of the historical run was October 21, 2021. Results presented in the main text of this report will focus on comparing the machine learning strategy to deterministic methods (the previous methodology). A post-hoc evaluation that compares machine learning to out-of-the-box probabilistic methods can be found in Appendix G.

After standardization, cleaning and preprocessing, the total number of records in each data source are presented in Table 3.

Table 3 – Total number of records in each data source. Note the number of records were higher than reported, due to to our liberal inclusion criteria.

Dataset	Records (n)
Case Records	1,402,141
Vaccine Records	5,018,916

After blocking, calculations and classification identified positive matches and links were compiled. The number of links found by the SVM + RF model were compared to deterministically linked pairs (Exact match of DOB and name). Overall, the machine learning method found 11.39 percent more links than deterministic methods (Table 4 & Figure 1).

Table 4 – Total number of links identified by deterministic and machine learning approaches.

Method	Links (n)
Deterministic	736,564
Machine Learning (SVM + RF)	820,441

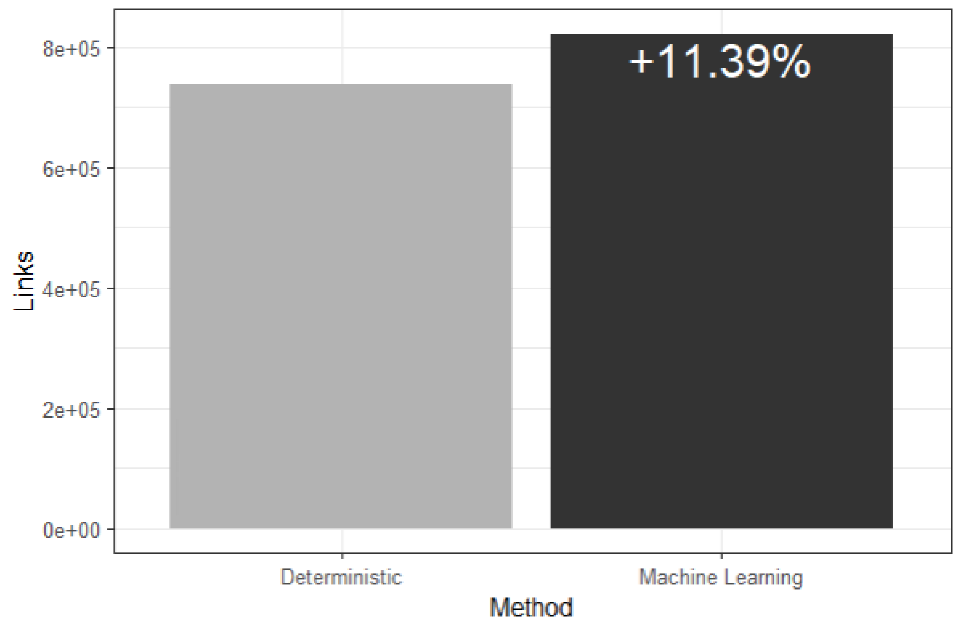


Figure 1 – Percent increase in total links by method.

To evaluate these results, a targeted quality assurance analysis was conducted, focusing on those links identified in one method but not the other. A total of 90 deterministically linked record pairs were determined to be non-links by the machine learning model, and the machine learning model found 84,060 linked pairs that failed to achieve deterministic criteria. From these record comparisons, we manually reviewed all 90 deterministic links the machine learning model missed and a sample of 1,000 records that failed deterministic criteria that the model classified as links. In total, 40 of the 90 links found by deterministic methods were true non-links, which in turn indicated the machine learning model failed to capture 50 true links. Additionally, three of the 1,000 sampled links made by the machine learning model were errors. Table 5 displays the sample and estimated 'full' false discovery rate (FDR) for each approach. The estimated full FDR assumes agreements between the two methods are true positives, then extrapolates FDRs at the sample FDR rate for method disagreements. FDRs for both models increased over time and were the highest during the Omicron COVID-19 wave, where each model observed a three-fold increase in FDR. We cover the impacts of Omicron on FDR in the discussion.

Table 5 – Quality Assurance sample description with calculated sample and estimated FDR along with the maximum observed FDR during the Omicron wave *(peak FDR between November 2021 – April 2022).

Method	Exclusively Identified Links	Sample For Review	Sample FDR	Estimated Full FDR	Max Omicron Observed FDR*
Deterministic	90	90	.44	0.005%	0.016%
SVM + RF	84,060	1000	.003	0.035%	0.11%

The machine learning approach found a total of 83,877 additional links between the historical datasets at a slightly higher Type I error rate (demonstrated above). A breakdown of the types of inexact links found is displayed in Table 6. Counts were recorded at the link level with multiple vaccines and cases allowed for each person. Furthermore, each inexact record link was allowed to be represented in multiple groups. Inexact name comparisons and DOBs were most prevalent.

Table 6 – Types of inexact links identified by the machine learning approach. Proportions were calculated with a total denominator of n = 820,441 (The number of links found using Machine Learning Method).

Group	Count (n)	Proportion (%)
Inexact DOB	13,413	1.6%
M/F Sex Disagreements	6,562	0.8%
Name Gender Disagreements	2,107	0.3%
Middle Name Disagreements	6,598	0.8%
Inexact First Name	42,067	5.1%
Inexact Last Name	39,716	4.8%
Inexact First & Last Name	3,648	0.4%
DOB Switch	2,195	0.3%

Linked records were then disaggregated by race and ethnicity to evaluate proportional increases for each subgroup. The overall increase of identified links provided by implementing the machine learning approach was 11.4 percent. Proportional increase for Black/African Americans, Native Hawaiian/Pacific Islanders, Individuals identifying as 'Other' race, and Hispanics were all above that baseline rate. Table 7 and 8 display the increases for each race and ethnicity subgroup.

Table 7 – Links identified by each method disaggregated by Race.

Race	Deterministic	SVM + RF	Percent Increase (Δ %)
American Indian/Alaskan Native	12,307	13,702	11.3%
Asian	54,867	60,779	10.8%
Black/African American	31,952	35,658	11.6%
Native Hawaiian/Pacific Islander	8,533	9,589	12.4%
White	443,138	482,472	8.8%
Other	99,598	119,017	19.5%
Multiracial	43,047	47,922	11.3%
NA	43,122	51,302	19.0%

Table 8 – Links identified by each method disaggregated by Ethnicity.

Ethnicity	Deterministic	SVM + RF	Percent Increase (Δ %)
Hispanic	87,060	106,858	22.7%
Not Hispanic	567,916	617,470	8.73%
NA	81,588	96,113	17.8%

Similar to the approach of Table 6, we investigated the inexact links captured by the machine learning model versus deterministic approaches and disaggregated results into two groups: White/Non-Hispanic links and BIPOC (Non-White and/or Hispanic links). This disaggregation is displayed in Figure 1. Appendix E tabulates this data and omits records with missing race/ethnicity data. The BIPOC group had proportionally more variation on truly linked records compared to their White/Non-Hispanic counterparts in every comparison. In other words, BIPOC records had more inexact matches that traditional methods would miss with deterministic or fuzzy methods.

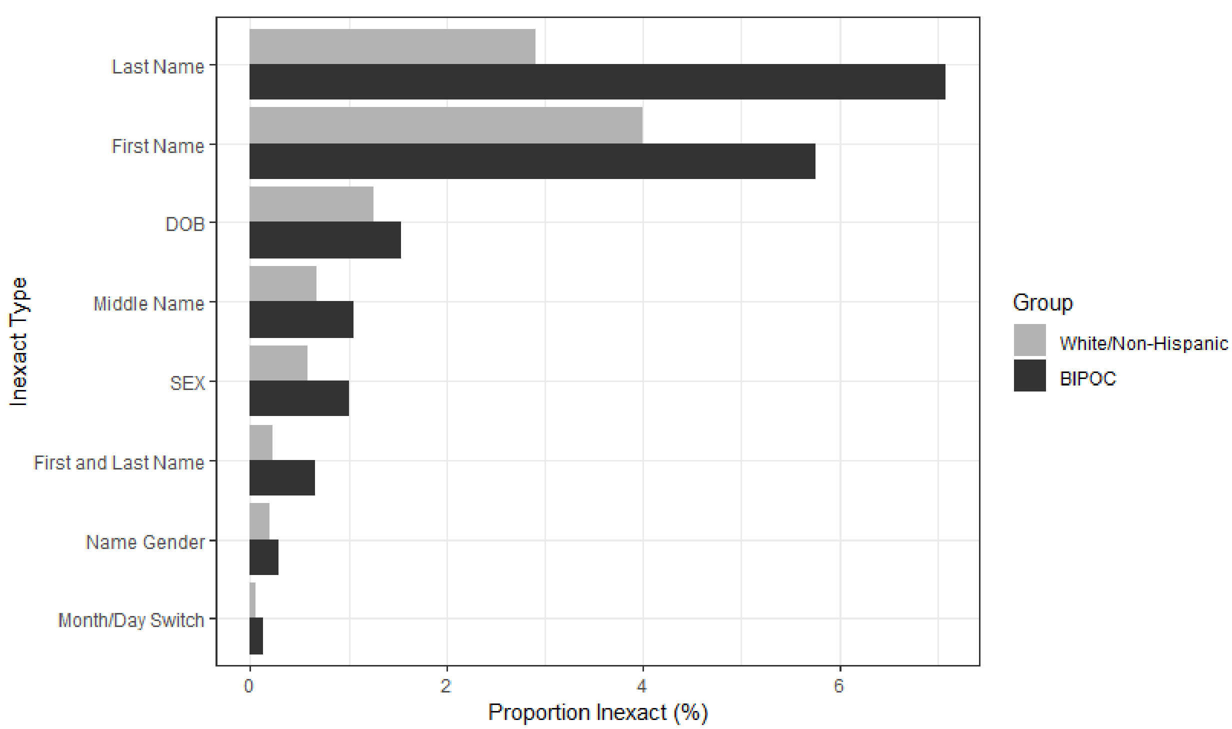


Figure 2- A comparison of inexact link rates between majority subpopulation (White/Non-Hispanic) and BIPOC subgroup (Non-White and/or Hispanic). Note the higher proportion of inexactness in all types for BIPOC subpopulations.

5 Discussion

The results of transitioning from a deterministic to a supervised machine learning linkage strategy led to several key improvements in COVID-19 case vaccination status surveillance. First, the non-probabilistic machine learning classification strategy established more links between case and vaccination records compared to the deterministic linkage. Next and most important, when analyzing the increase in links by race and ethnicity groups, the greatest proportional increase in the number of links was among BIPOC subgroups. Based on the total amount of links captured from the machine learning linkage, the largest groups of inexact matching links are from inexact first name, last name, and date of birth record pairs. And lastly, while the machine learning linkage consistently captured more links than the deterministic linkage, the magnitude of that increased linkage rate also increased over time.

The large increase in links among BIPOC subgroups was likely due to higher error rates in their vaccine and/or case records (Table 9). As established in previous research, BIPOC subgroups are more likely to have missing or incorrect information in their health records compared to White and non-Hispanic groups [3,4]. The machine learning classification method of establishing links allows for records containing errors or lacking information in certain fields to be linked. The increase in the number of links identified among Hispanic people was by far the largest. One reason for this finding was the machine learning linkage methodology was designed to robustly account for naming conventions more common among Hispanic people. For example, people who have double surnames can have one of their surnames mistakenly present in their middle name field or captured differently in data collection systems using different hyphenation or spacing standards or have a part of their name missing entirely. In each of these situations, the machine learning linkage can account for the differences and establish a link, indicating a key strength of the machine learning linkage over probabilistic methods. Using the SVM portion of the stacked classification method, clusters of links are created, enabling different decision thresholds for each cluster. For example, higher variability present in Hispanic/Latino last names can be accounted for and addressed by the model. Additionally, this enables the use of nonlinear dimensionality reduction techniques (TSNE and UMAP) to visually investigate clusters graphically. This kind of clustering is not a component of current probabilistic methods. Furthermore, the same probabilistic thresholds are applied uniformly to all potential links, even if there is an uneven distribution in the type of inexact matching information. This typically results in probabilistic methods requiring much more manual customization and review than machine learning methods. The stacked machine learning approach we implemented addresses the shortcomings of less specific manual review prevalent in probabilistic linkages through clustering, allowing for much more targeted checks and therefore less required manual review overall.

Another finding was a high relative improvement of linkage rates among people whose race was missing or entered as “Other”. The proportional increases in the number of links among these groups more closely mirrored the increases among people of color, suggesting that these populations are disproportionately BIPOC. This clearly demonstrates that despite the considerable gains in the number of links among non-White/Hispanic groups, they are still underrepresented in the final linkage products. Though still underrepresented, it is paramount to understand that the flexibility and robustness of the machine learning linkage approach improved linkage rates among BIPOC residents in a way that would not be feasible using other linkage techniques.

When comparing the estimated false linkage rates of the deterministic and machine learning linkage strategies, the deterministic linkage demonstrated consistently lower rates. This finding was expected, as the strict criteria necessary to establish a link using this methodology results in very few incorrect links. In this case, the only way a false link can be established using a deterministic strategy is through linking records from two different people who happen to have the exact same first name, last name, and date of birth. This is highly unlikely to occur in any significant frequency. The higher false linkage rate for the machine learning linkage was expected, as allowing inexactly matching records to still be linked intrinsically allows for a greater degree of potential error. This intentional increase in error was deliberate to enable greater increases in representation.

In terms of linking COVID-19 case and vaccine administration records, it was found that the consequence of this slightly higher false linkage rate was marginal when compared to the benefit of increasing the number of captured links, especially among BIPOC individuals. The negative effect of a higher false linkage rate was further mitigated by routine manual QA, which was established in the machine learning linkage process where questionable links are routinely reviewed and corrected if necessary. No analogous routine QA process could be established for the deterministic linkage, as there was no quantification of how similar record pairs were to each other.

Further analysis of false linkage rates over time showed both methodologies became less accurate when comparing rates from November 2021 to April 2022. During this time, both linkage strategies experienced an approximate three-fold increase in their false linkage rates. Crucially, the Omicron COVID-19 wave resulted in an unprecedented level of increases in COVID-19 cases over this time period. Additionally, the FDA expanded the emergency use authorization for COVID-19 boosters for all individuals age 18 and older in mid-November 2021 [11]. Consequently, there was a higher than usual number of new COVID-19 vaccine administrations during the winter months of 2021 and early 2022. Compounding the problem of the incredibly high volume of data to be collected over a short period of time was the heterogeneity in Washington State's health care providers and their data collection systems. As a result, the quality, completeness, and timeliness of the case and vaccination administration data submitted to DOH varied considerably over this period. This had a profound impact on the efficacy of the deterministic linkage. Over this time, the deterministic linkage captured about 140,000 new links compared to the machine learning linkage, which captured about 400,000. It was clear that the inflexibility of the deterministic linkage was a significant impediment during the Omicron wave.

Beyond capturing significantly more links, the machine learning linkage methodology improved upon the deterministic linkage in other ways. After the initial historical batch run, the machine learning linkage ran quicker compared to the deterministic linkage. This is due to the deterministic linkage completing a full historical linkage during every run. This is significantly slower than the machine learning linkage methodology shown in Appendix D, where only new vaccine doses and cases are eligible to be linked. Another improvement relative to the deterministic linkage was the iterative machine learning process. Not only were the linkage models able to be improved and more accurately link cases and vaccine records over time, but the amount of manual QA necessary to carry out this task decreased. Through manual review, the boundary line between links and non-links became clearer. Beyond the first few months of runs, the amount of manual review dropped from several pairs of links every day, to about two or three pairs of records per week.

This case study has several notable limitations and can be summarized thusly:

- Vaccination data used for the linkage was stored in a way that did not allow for easy correction or editing of that source data.
- A name frequency parameter was used as a variable in link classification that was based on the frequency of names within Washington State as a whole, but it did not factor name frequency within race groups.
- The standardization of name fields impacts names containing characters and punctuation not found in the standard 26 letter English alphabet more than other names. This forced standardization creates some error as characters are transformed to fit the English alphabet.
- The process of blocking record pairs, which reduces the amount of record comparisons that flow into the linkage models, can erroneously lead to missed links.

Vaccine records stored in the WAIIS, which is a live database, can be updated or corrected if necessary. However, due to data sharing limitations, the linkages were unable to use the live data. Instead, daily static snapshots of new vaccine doses were provided to the team conducting the linkage. The result was that the vaccine information available to link was limited to only what was contained in these snapshots. Any subsequent edits or corrections were either not available to be shared or entered as new, sometimes duplicative vaccine doses. One consequence of this was a higher-than-expected amount of inexact matching information. For the most part, the vaccine information was later corrected in the live database. But these corrections were not reflected in the source data that was run through the linkage. Another consequence was vaccine doses administered to the same person can be submitted multiple times, and each instance can have a different unique person ID number. These records are often merged and reconciled in the live database, but these corrections are not reflected in the static data used for the linkage. This can result in an artificially high number of links due to linking duplicate records. Deduplication procedures may remedy this, but all linkages without exception are susceptible to the deficits in data quality that they inherit.

Another limitation specific to the machine learning methods implemented here is the name frequency parameter used in link classification. While this metric did help separate classification clusters, a consequence of using this parameter was that records with very common names were less likely to be classified as a link. Furthermore, this parameter is not built for each race or ethnicity group. This means that common names in racial groups which are small compared to the overall size of the population are not accurately represented. An example of this is the name “Mohammed” or “Muhammad,” which is often cited as the most popular name in the world, but not in Washington State [12]. This name has a relatively low overall incidence in Washington State but has a relatively high incidence among smaller regional subgroups such as North African, Middle Eastern, or South Asian residents. In this case, it could lead to a higher false linkage rate among people with common names within smaller race groups.

The final set limitations discussed here are the preprocessing and blocking of data prior to any linkage algorithm implementation. Preprocessing the data standardizes alphanumeric values to enable algorithmic comparison. This forces all names and identifiers into standard Latin characters, removing any special characters and punctuation. This impacts non-western identifiers more, as the incidence of these characters and manipulations are more common in BIPOC subgroups. Furthermore, blocking strategies to remove insignificant comparisons has an error rate. This error rate is hypothesized to disproportionately impact BIPOC comparisons due to increased variability in data consistency. Newer blocking methods are warranted and currently being implemented to move away from logic-based filtering, which is necessary to maximize equitable representation downstream.

Moving forward, we have outlined some upstream data issues which need to be considered when evaluating the efficacy of this machine learning linkage. The issue of common names in smaller subgroups presents a challenge in accurately linking these populations. To take the example presented above, it was found that a disproportionate number of people with common North African, Middle Eastern, and South Asian names were mislabeled as links compared to other populations. It was also found that many residents belonging to these race groups have a birth date listed as January 1. Looking into this phenomenon further, it was revealed that US residents without accurate birth records are assigned January 1 as a birth date. Overall, this manifests itself as a subpopulation with higher-than-expected matching birthdates and a small number of very common names resulting in more false links. This also highlights a strength of the machine learning linkage process established here, as targeted QA can identify these clusters of cases for more rigorous review. This would be a far more effective solution compared to probabilistic or deterministic linkages that would require a significantly more burdensome and less specific manual review to correct.

6 Conclusions

When linking data from separate sources, establishing a linkage process that directly mitigates the inequities stemming from poor data quality is paramount for adequate surveillance. As demonstrated above, deterministic linkage techniques inherit biases present in source data and cannot easily overcome this limitation without correcting this data. While probabilistic linkage strategies can overcome some biases inherent in the data, they are limited in their capacity to improve accuracy for several reasons. Probabilistic linkage strategies contain arbitrary confidence thresholds and have demonstrated higher error rates without exhaustive and intensive manual review. Subsequent analyses and recommendations based on these linked data will be incomplete and often inadequate. More robust linkage strategies such as the machine learning methodology presented here have highlighted the insufficiencies of deterministic and probabilistic linkages, especially when issues regarding source data quality cannot be addressed.

Moving forward, we recommend surveillance systems and linkage projects that rely on deterministic linkages be heavily scrutinized. Another conclusion, which is critical when thinking about this machine linkage approach in public health practice, is its impact on health equity. Errors and missing information in health records are inevitable, and these errors disproportionately impact BIPOC populations. Solutions to remedy these upstream data quality problems are often difficult and involve substantial changes to data collection systems themselves. This is further complicated by the myriad of different submitters of health information and the flexibility required by the DOH to accept different data formats. While further efforts to standardize health information submission and methods of collecting health data will aid linkage efforts greatly, this is a lengthy process and there is an immediate need for high quality linked health information. Deterministic and probabilistic linkage strategies either fail to account for variable data quality among different subpopulations or are limited in their ability to target specific populations for manual review. Establishing a machine learning linkage enabled better vaccine breakthrough identification and allowed for more complete linked data to assess vaccination status of COVID-19 cases among all Washingtonians, regardless of race and/or ethnicity. The machine learning linkage strategy serves as a direct response to the inequities present in health data quality. The robustness and ease of maintenance of this linkage methodology makes it an invaluable tool for workers involved in public health surveillance and policy making.

7 Acknowledgements

This work was possible due to the contribution and review of numerous staff across multiple teams at Washington DOH, who worked tirelessly to stand up this project in the middle of the COVID-19 pandemic. Sofia Husain, Isaiah Reed and Annie Khanani from the Office of Communicable Disease Epidemiology (OCDE)'s Vaccine Preventable Diseases (VPD) team spearheaded COVID-19 breakthrough surveillance. They were instrumental in establishing the original deterministic linkage that was fundamental to this work and have provided critical input to the development and implementation of this data linkage as subject matter experts (SMEs) and primary users of these data. Terra Wiens, Jessica Marcinkevage, Odane Dunbar, Mariana Rosenthal, and Peter Dieringer from the Office of Immunization (OI) have been critical partners in the implementation of this work, including ensuring access to IIS data, supporting QA, analysis and dissemination of the results, and maintaining the linkage on a routine basis. Maya Bhat and Julian Kapoor support ongoing linkage efforts in CHS and have informed the current work as well as developed more advanced and novel machine learning approaches for record linkage. Dianna Hergott from the Center for Data Science (CDS) provided thoughtful peer-review and feedback on drafts of this report.

We extend a special thanks to Anthony Rivers, DOH Equity and Social Justice Manager, for composing a peer-review and agreeing to the inclusion of it in the report. Additionally, we extend a special thanks to Nikki Lanka, Office of Public Affairs & Equity (OPAE), for leading official DOH review processes and managing post-composition dissemination efforts.

Finally, this work would not have been possible without the tireless work of all those who make availability of robust vaccination and case data a reality, including local health jurisdiction investigators, health care providers, diagnostic laboratories and other entities who collect and report data to the Immunization Information System (IIS) and the Washington Disease Reporting System (WDRS), as well as the numerous DOH staff who manage these public health data systems, process incoming data and ensure data quality.

8 Appendix A – Description of Data Variables

Cases	Description
CASE_ID	Unique identifier for WDRS COVID Cases
DOB #1	Primary DOB on record
DOB #2	Secondary DOB from alternative demographic table
PHONE #1	Primary phone on record
PHONE #2	Secondary phone on record
PHONE #3	Tertiary phone on record
SEX	Sex on record
ZIP	Zip/Postal code provided
MIDDLE INITIAL #1	Primary middle name first letter
MIDDLE INITIAL #2	Secondary middle name first letter
FIRST NAME #1	Primary first name on record
FIRST NAME #2-#3	Alternative names from other tables
FIRST NAME #1 INITIAL	First letter of primary first name
LAST NAME #1	Primary last name on record
LAST NAME #2-#3	Alternative names from other tables

Vaccines	Description
RecipientID	Unique identifier from IIS
DOB #1	Primary DOB on record
DOB #2	Secondary DOB from alternative demographic table
DOB Switch #1	Primary DOB with month and day switched (if valid)
DOB Switch #2	Secondary DOB with month and day switched (if valid)
PHONE #1	Primary phone on record
PHONE #2	Secondary phone on record
SEX #1	Sex on record
SEX #2	Sex from demographic table (if different from SEX #1)
ZIP #1	Zip/Postal code provided
ZIP #2	Alternate Zip/Postal code provided
MIDDLE INITIAL #1	Primary middle name first letter
MIDDLE INITIAL #2	Secondary middle name first letter
FIRST NAME #1	Primary first name on record
FIRST NAME #2 - #8	Alternative names from other tables and splitting
FIRST NAME #1 INITIAL	First letter of primary first name
LAST NAME #1	Primary last name on record
LAST NAME #2-#8	Alternative names from other tables and splitting
FIRST NAME FREQUENCY	The scaled frequency of the first name used in linkage comparison
LAST NAME FREQUENCY	The scaled frequency of the last name used in linkage comparison

9 Appendix B – Description of Calculations

Calculation	Description
DOB_HAM	Hamming distance between numeric date of birth. Either 0 or 1 due to our inclusion criteria for the pairwise data frame.
SEX_Disagree	Was there an explicit M/F disagreement between records? 0 was no, 1 was yes. If a U or blank was present, the value was 0 (no explicit disagreement by our criteria).
NAME_GENDER_Disagree	Probable gender prediction of first name disagree – This value comes from an index file produced from all-time WHALES birth record data. For each name in the birth record table, a sex at birth breakdown was calculated. Each name was designated as primarily male, primarily female, or in the case that there were equivalent number of each sex at births, neither primarily male/female. This table was joined in and if there was a M/F designation disagreement between the first names, a flag of 1 was produced. If there was no explicit disagreement in primary sex, the flag is set at 0.
FIRSTNAME_COSn2	Minimum bigram cosine distance of all first name field comparisons. This field breaks up each first name field into n-gram chunks of 2 letters then computes the cosine distance. We use cosine distance to capture compound name similarities because it cares less about vector length.
LASTNAME_COSn2	Minimum bigram cosine distance of all last name field comparisons. This field breaks up each last name field into n-gram chunks of 2 letters then computes the cosine distance. We use cosine distance to capture double surname similarities because it cares less about vector length.
FIRSTNAME_JW	Minimum Jaro-Winkler distance of all first name field comparisons. We use JW to capture misspelled names of similar lengths in addition to the cosine distance calculation.
LASTNAME_JW	Minimum Jaro-Winkler distance of all first name filed comparisons. We use JW to capture misspelled names of similar lengths in addition to the cosine distance calculation.
MINAME_Disagree	Are both middle initials present, and if so, do they disagree? If they disagree a value of 1 is flagged, if they agree or if at least one value is missing, a value of 0 is coded.
MINAME_Missing	Is there a missing middle initial in one or both sides of the comparison? This flag differentiates a match vs. a non-disagreement due to missing variables.
COMBNAME_COS	Trigram cosine distance of all first and last name fields (deduped) in one string. By including a compound first-last name field we are attempting to capture first-last name switches, and we include a trigram (n-gram size 3) cosine distance calculation to parse out cases that bigram similarity is higher than optimal because of the n-gram length. Trigram should correct these cases.
PHONE_DL	Damerau-Levenshtein distance of phone number. We use this rather than hamming to capture clerical errors with less position dependencies.
PHONE_MISSING	Is there a missing phone in one or both sides of the comparison? This flag differentiates a match vs. a non-disagreement due to missing variables.

ZIP_GEO	Estimated distance between two zipcode values. Zipcodes are translated to a centroid based lat/long set of variables provided in an index file ripped from Github. Then we calculate the megameter distance between the two lat/long pairs. Megameter was chosen to standardize between 0-1 without any distortion from distribution effects. Distances were maxed at 1 megameter, suitable for the range of WA.
ZIP_MISSING	Is there a missing zipcode value in one or both sides of the comparison? This flag differentiates a match vs. a non-disagreement due to missing variables.
FM_SWITCH_ALERT	Evidence of a first/middle name switch? If the first name initial of the case data matches the middle name initial of the vaccination data AND the first name initial of the vaccination data matches the middle name initial of the case data, then the flag is set at 1. Otherwise, it is set at 0.
FIRST_NAME_FREQ	Frequency of first name in dataset from the vaccination side.
LAST_NAME_FREQ	Frequency of last name in dataset from the vaccination side.

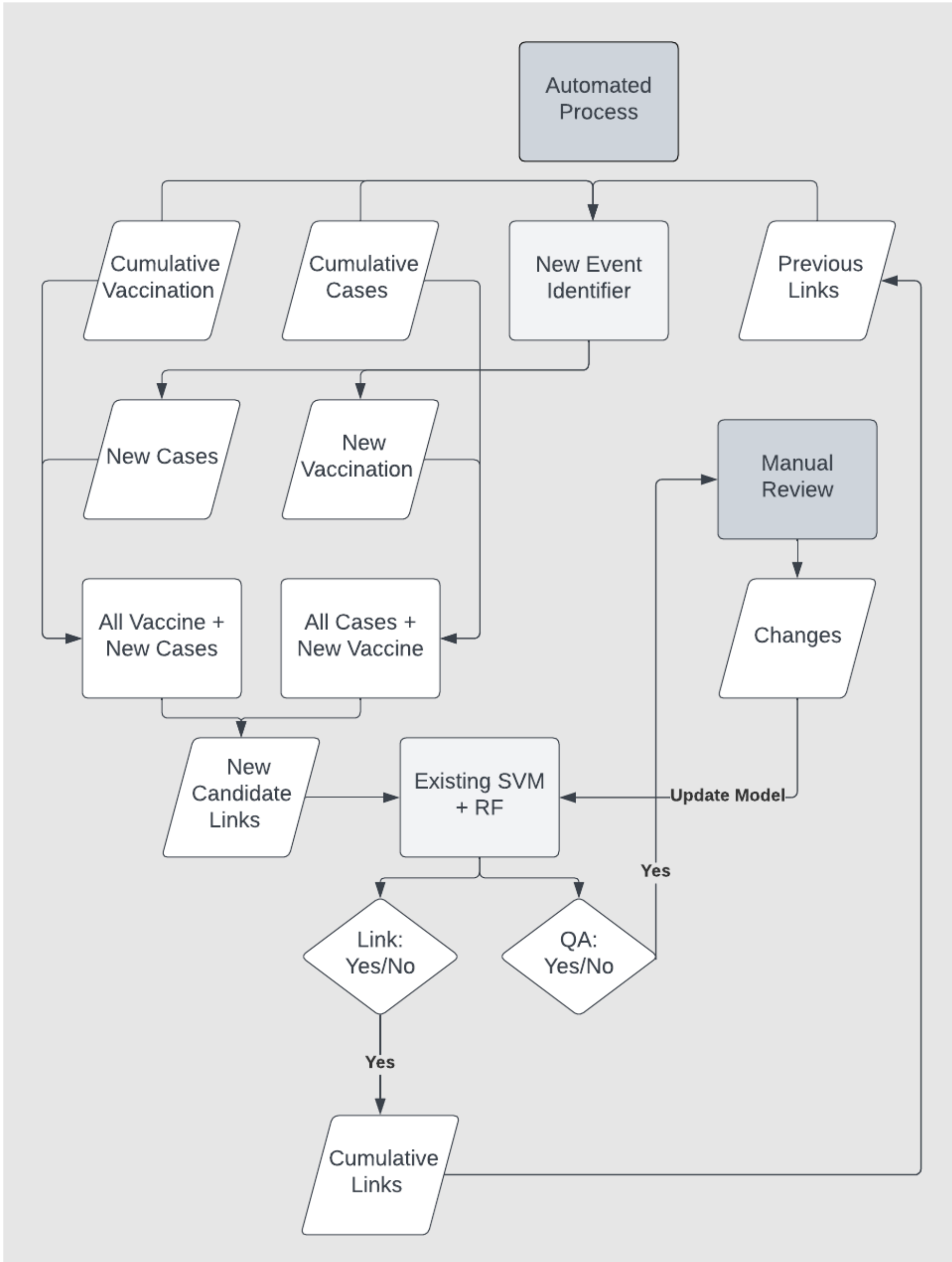
10 Appendix C – Sequential Training Methodology Notes

1. Random sample of 10,000 pairs
2. Calculate distance metrics for all pairs
3. Group pairs into 11 stratified pseudo-probabilistic groups based off previously trained SVM on just name and DOBs
    a. Group 1 = 0 - .1
    b. Group 2 = .1-.2
    …

    k. Group 11 = 1+

4. Sample 1 - sampled 100 from each group above for manual review
5. 1,100 pairs reviewed, 17.4 percent of these records were found to be pairs
6. All pairs were plotted in a logistic function with the x-axis being SVM score and the Y being the binary (1-link, 0-not a pair)
7. A SVM was trained using this labeled data – here SVM_1
8. Sample 2 repeated the process of sample 1 while avoiding resampling those included in the first sample and only sampled pairs with SVM scores higher that .4 (There were no links in sample 1 at this value that were links)
9. 600 pairs reviewed, 32 percent of these records were found to be pairs
10. SVM_1 then predicted linkage status for pairs in sample 2, agreement was already at 99 percent
11. Sample 2 and Sample 1 were then added together to create a cumulative sample, which was then plotted in a logistic function to determine the next region of focus.
12. Cumulative sample #1 trained a new SVM
13. Sample 3 focused on pairs with name SVM scores between .72 and .93 determined by the logistic function in step 11. This 'zooming in' on areas where probabilities are non-zero and non-one is critical for our strategy and suitable samples from this region is crucial to capture the variance of this region.
14. 1,000 pairs were manually reviewed, 16.1 percent were found to be pairs. Agreement was above 99 percent with predicted values from the SVM trained on cumulative sample 1
15. Sample 3 was added to the cumulative sample creating Cumulative sample #2
16. And SVM was trained on cumulative sample #2
17. Another logistic function was fit to this cumulative sample
18. Sample 4 and 5 were utilized to fix a few bugs in the calculation fields in which the SVM was retrained using corrected values and changed predictions were reviewed manually. 4 was used to review new links and 5 was used to review newly rejected pairs.
19. TSNE and UMAP were used to visually observe clusters and identify and outlier clusters that were misclassified

11 Appendix D – Automated Process Schematic



12 Appendix E – Race/Ethnicity Disaggregation of Inexact Match Types

-Group	White/Non-Hispanic		BIPOC (Non-White and/or Hispanic)	
	Count (n)	Proportion (%)	Count (n)	Proportion (%)
Inexact DOB	5,512	1.26%	4,167	1.53%
M/F Sex Disagreements	2,541	0.58%	2,746	1.01%
Name Gender Disagreements	840	0.19%	805	0.30%
Middle Name Disagreements	2,933	0.67%	2,874	1.05%
Inexact First Name	17,435	3.99%	15,698	5.76%
Inexact Last Name	12,664	2.90%	19,309	7.08%
Inexact First & Last Name	966	0.22%	1,820	0.67%
DOB Switch	243	0.06%	369	0.14%
TOTAL	**436,688**	--	**272,761**	--

13 Appendix F – Description of Distance Metrics Employed

Distance Metric	Description	Formula
Hamming	Comparing two equal length strings via binary agreements and disagreements.	$\Sigma(Bit\ Disagreements)$
Cosine	Measure of how dissimilar/similar two vectors are based on their angle. Vector length does not influence score and uses q-gram components to calculate score.	$cosine\ distance = D_C(A,B) \coloneqq 1 - S_C(A,B)$ Where: $S_C(A,B) \coloneqq \cos(\theta) = \frac{A \cdot B}{\lvert\lvert A \rvert\rvert\ \lvert\lvert B \rvert\rvert} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2 \cdot \sum_{i=1}^{n} {B_i}^2}}$
Jaro-Winkler	Accounts for matching characters, length of strings, and number of transpositions.	$Jaro\ distance = sim_j = {}^1/_3\left(\frac{m}{\lvert s_1 \rvert} + \frac{m}{\lvert s_2 \rvert} + \frac{m-t}{m}\right)$ Where: 1. m is the number of matching characters and is not 0 2. $\lvert s_1 \rvert$ is the length of string s_1 3. t is the number of transpositions $Jaro - Winkler\ similarity = sim_j + \ell p(1 - sim_j)$ Where: 1. ℓ is the length of common prefix at the start of the string up to 4 characters 2. p is a constant scaling factor between .1 and .25
Damerau-Levenshtein	Minimal number of insertions, deletions and replacements needed for transforming string a into string b allowing transposition of adjacent symbols.	$d_{a,b}(i,j) = min \begin{cases} 0 & if\ i = j = 0 \\ d_{a,b}(i-1,j) + 1 & if\ i > 0 \\ d_{a,b}(i,j-1) + 1 & if\ j > 0 \\ d_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} & if\ i,j > 0 \\ d_{a,b}(i-2,j-2) + 1_{(a_i \neq b_j)} & if\ i,j > 1 \end{cases}$ Where: 1. $d_{a,b}(i-1,j) + 1$ is a deletion 2. $d_{a,b}(i,j-1) + 1$ is an insertion 3. $d_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)}$ is a match or mismatch 4. $d_{a,b}(i-2,j-2) + 1_{(a_i \neq b_j)}$ is a transposition
Agreement/ Missing	A combination of do the strings match exactly. If not was one or more string missing data?	1. $\begin{cases} a = b \\ a \neq b \end{cases}$ 2. $\begin{cases} a = NA \\ b = NA \end{cases}$

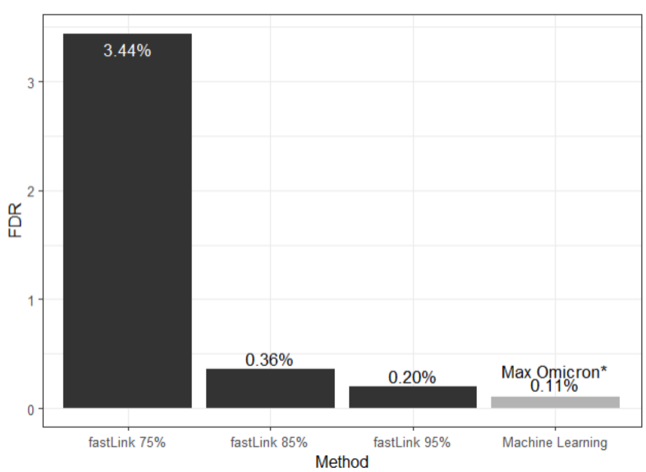
14 Appendix G – Probabilistic Results Compared to Machine Learning Strategy

To compare machine learning results with a probabilistic solution, a post-hoc analysis on machine learning results was conducted using the R package fastLink [13]. fastLink (FL) will serve as our "out-of-the-box" probabilistic comparison and is based on the Fellegi-Sunter methodology. The probabilistic model used all variables and did not block off anything deterministically. All parameters in the fastLink package utilized defaults except threshold values (accepted posterior probabilities).

Since the ML links have been reviewed and vetted, a known pool of inexact true links can be evaluated using these probabilistic methods. Machine learning (ML) inexact links were fed into the probabilistic matching function, and links were derived at multiple threshold levels. Inexact links here are defined as true links with one or more comparative disagreement in name or DOB. The gross number of inexact links identified by each method and threshold are presented in the table below.

Method	Threshold (pm%)	Inexact Links (n)
Machine Learning	--	42,123
fastLink	99	34,300
fastLink	98	37,743
fastLink	97	37,877
fastLink	95	38,112
fastLink	90	38,660
fastLink	85	38,809
fastLink	75	42,298

At higher confidence thresholds (probabilities above 85 percent) probabilistic methods do not capture as many inexact matches as ML. In addition to capturing less links, error rates are higher, as shown in the figure below.



The performance of Machine Learning methods on inexact matches, which we have shown to impact BIPOC persons more, indicates that ML methods outperform commonly implemented probabilistic solutions and can provide a greater degree of equity in linkage rates.

15 Appendix H – Glossary of Terms and Definitions

Term	Definition
Deterministic Matching	In deterministic matching, a precise linkage between two datasets is established when every matching variable exhibits exact agreement.
Fuzzy Matching	Fuzzy matching is a data linkage approach used to match records from diverse datasets by considering approximate matches and accommodating variations, errors, and inconsistencies within the data. It enables the identification of potential matches based on the degree of similarity or dissimilarity between data values, allowing for more flexible and inclusive data linkage processes.
Blocking	Blocking is a data preprocessing technique designed to mitigate the inherent computational complexity in inexact matching processes. This process involves dividing datasets into smaller, more manageable subsets or blocks based on a defined set of variables. Subsequently, these distinct blocks are independently processed to identify potential matches, leading to a reduction in the total number of candidate comparisons and an overall enhancement in the efficiency of the inexact matching algorithm.
Hamming Distance	Hamming Distance quantifies dissimilarity between two equal-length strings by counting the positions at which corresponding elements in the two strings differ.
Damerau-Levenshtein Distance	Damerau-Levenshtein Distance measures the dissimilarity between two strings by calculating the minimum number of operations (insertions, deletions, substitutions, and transpositions of adjacent characters) required to transform one string into another.
Jaro-Winkler Distance	Jaro-Winkler Distance calculates string similarity by considering the number of matching characters, the frequency of transpositions, and the length of a shared prefix at the beginning of the strings, up to a maximum of four characters.
Cosine Distance	Cosine Distance, also known as Cosine Similarity, quantifies vector similarity based on the angle between them. It remains uninfluenced by vector length. In our application, a value approaching 0 signifies strong similarity, while 1 indicates dissimilarity.

Radial Support Vector Machines	Radial Support Vector Machines are a supervised machine learning algorithm within the Support Vector Machines (SVM) family, tailored for classification and regression tasks. They excel when dealing with complex, nonlinear data patterns. The term 'radial' is derived from their use of radial basis functions to transform data into higher-dimensional space, enhancing their ability to define decision boundaries. Radial SVM is known for its robust performance in identifying optimal decision boundaries or regression curves.
Random Forest Models	Random Forest is a machine learning technique used for classification and regression. It combines multiple decision trees, each trained on a different data subset through bootstrapping. RF models aggregate predictions from these trees, enhancing accuracy and robustness. They are known for their ability to handle complex data and avoid overfitting.

16 Equity and Social Justice Manager Peer-Review

In the early stages of the COVID-19 pandemic, cases involving COVID-19 were linked using vaccination records. Vaccination status was solely determined by self-reported information. However, this method became unsustainable, and data were incomplete due to the impossibility of contacting every new case. Therefore, there was an increasing need to establish a better linkage process to better identify vaccine breakthrough cases, which are instances where an individual tests positive for COVID-19 after being fully vaccinated.

A deterministic record linkage, which produces links based on common identifiers or variables, was initially used. But it was only able to link vaccination records if first name, last name, and date of birth matched across all records. Although deemed necessary, it quickly became apparent the quality of the data was flawed because of name and date of birth data entry errors during data collection. This is in addition to biases in data collection systems towards “traditional” American names. Diacritic marks, naming structures, name translation, and varying date structures create date entry discrepancies that disproportionately impact BIPOC communities. Therefore, BIPOC are less likely to have records linked due to the overly strict rules of the deterministic linkage strategy.

By mid-2021, the COVID-19 vaccine rollout was well underway with cases and vaccination data increasing. There was also an increased demand for the data to devise a strategy for the State of Washington’s response and efforts to control the spread of the COVID-19. Because of the deterministic methodology, there was a downstream, or reactive analysis, so it was determined the continued use of incomplete or inaccurate data would negatively impact the surveillance of BIPOC and increase health disparities.

The Department of Health also considered a traditional probabilistic linkage strategy, a strategy that expands the potential identifiers and can produce more reliable results than the deterministic linkage strategy. This also proved to be inadequate, with a higher rate of error due to the statistical assumptions, the need to minimize manual intervention, and the inability to customize algorithms to detect inequities.

Ultimately, the Washington State’s Center for Health Statistics developed a non-probabilistic machine learning-based linkage strategy, proven in previous linkage projects to be the most effective and reliable strategy, requiring less manual intervention.

In my review of the background for the implementation of the machine learning linkage strategy, I recommend this paper be accepted. The effort to be more inclusive by addressing disparities in data systems will undoubtedly have a positive impact on the health outcomes of groups experiencing disproportionate risk of disease.

Although I find no major issues in the background, I do find minor issues in certain limitations. Discrepancies created by the inability to use live data versus the snapshots of records and the duplication of records with different “person ID numbers” may impact the quality of the data as mentioned in the discussion. There could be some concern over the validity of the data because of the potentially misleading inflated number of linkages due to the duplications of records. However, notwithstanding the limitations, the ability to address inequitable data linkage by creating strategies able to provide more complete data is imperative to accurately assess health disparities and the impact of public health interventions.

Anthony Rivers

Equity and Social Justice Manager

Disease Control and Health Statistics

Washington State Department of Health

17 References

1. Gray L, Ellison J. In pandemic milestone, UW brings COVID-19 vaccines to frontline health care workers. UW News. Accessed August 31, 2023. https://www.washington.edu/news/2020/12/28/in-pandemic-milestone-uw-brings-covid19-vaccines-to-frontline-healthcare-workers/
2. Knight HE, Deeny SR, Dreyer K, et al. Challenging racism in the use of health data. *The Lancet Digital Health*. 2021;3(3):e144-e146. doi:10.1016/S2589-7500(21)00019-4
3. Grath-Lone LM, Libuy N, Etoori D, Blackburn R, Gilbert R, Harron K. Ethnic bias in data linkage. *The Lancet Digital Health*. 2021;3(6):e339. doi:10.1016/S2589-7500(21)00081-9
4. Bohensky MA, Jolley D, Sundararajan V, et al. Data Linkage: A powerful research tool with potential problems. *BMC Health Serv Res*. 2010;10(1):346. doi:10.1186/1472-6963-10-346
5. Johnson AG, Linde L, Ali AR, et al. COVID-19 incidence and mortality among unvaccinated and vaccinated persons aged ≥12 years by receipt of bivalent booster doses and time since vaccination - 24 U. S. Jurisdictions, October 3, 2021-December 24, 2022. *MMWR Morb Mortal Wkly Rep*. 2023;72(6):145-152. doi:10.15585/mmwr.mm7206a3
6. Gilbert R, Lafferty R, Hagger-Johnson G, et al. Guild: guidance for information about linking data sets. *Journal of Public Health*. 2018;40(1):191-198. doi:10.1093/pubmed/fdx037
7. Sariyar M, Borg A. The RecordLinkage package: detecting errors in data. *The R Journal*. 2010;2(2):61-67. Accessed August 31, 2023. https://journal.r-project.org/archive/2010/RJ-2010-017/index.html
8. Fellegi IP, Sunter AB. A theory for record linkage. *Journal of the American Statistical Association*. 1969;64(328):1183-1210. doi:10.1080/01621459.1969.10501049
9. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. *Int J Epidemiol*. 2016;45(3):954-964. doi:10.1093/ije/dyv322
10. Elfeky M, Verykios V, Elmagarmid A, Ghanem T, Huwait A. Record linkage: a machine learning approach, a toolbox, and a digital government web service. *Department of Computer Science Technical Reports*. Published online July 1, 2003. https://docs.lib.purdue.edu/cstech/1573
11. Food & Drug Administration. Coronavirus (COVID-19) update: FDA expands eligibility for covid-19 vaccine boosters. Published November 19, 2021. Accessed August 31, 2023. https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-expands-eligibility-covid-19-vaccine-boosters
12. Sabir RI, Nawaz F, Zakir U, Naeem M, Anjum T. Names and behavior: case of the name “muhammad.” *IJPT*. 2014;2(4). doi:10.15640/ijpt.v2n4a7
13. Enamorado T, Fifield B, Imai K. fastlink: fast probabilistic record linkage with missing data. Published online April 29, 2020. Accessed August 31, 2023. https://cran.r-project.org/web/packages/fastLink/index.html