Improving COVID-19 case and immunization record linkage via non-probabilistic machine learning-based classification

Seth Rothbard, Chunyi Wu, Alex Cox, Meredith Cook, Sofia Husain, Terra Wiens, Isaiah Reed, Annie Khanani, Sean Coffinger

INTRODUCTION

- Needed to establish a linkage process between COVID-19 case data and vaccine records to better identify vaccine breakthrough cases.
- Initially used a deterministic linkage technique because it was quick and simple to set up.
- Deterministic linkage became increasingly inadequate as it was too inflexible to capture inexact record matches, which disproportionally failed to link people belonging to minority groups.
- This became especially problematic as the use of the linkage results expanded to include predictive modeling of COVID-19 and inform public policy.
- To reduce existing surveillance biases, an alternative linkage technique was necessary to establish a more robust surveillance system.

METHODS

Washington State's Center for Heath Statistics (CHS) machine learning-based classification method was proposed. Uses two machine learning models:

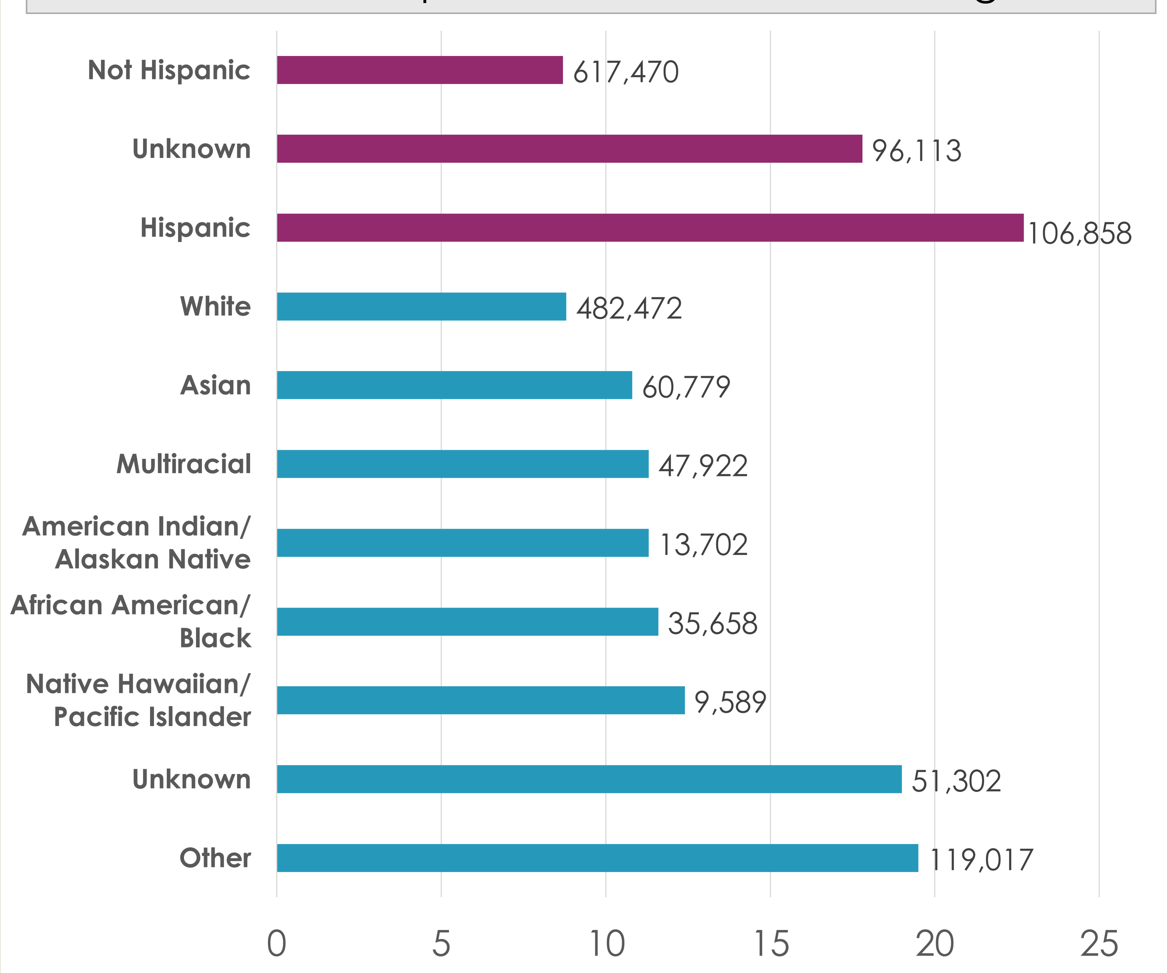
- **Radial Support Vector Machine (SVM)**
- **Random Forest (RF)**

Models were trained on vaccine and case data collected from Washington State residents. Testing was conducted on all historical COVID-19 case and vaccine records and underwent extensive QA prior to, during, and after it was transitioned into production. Post-transition QA was conducted at two time points: **Nov. 2021** and **Apr. 2022**.

RESULTS

- The machine learning linkage **captured more links among every race and ethnicity group relative to the deterministic linkage** with the **largest proportional increase among non-White and/or Hispanic/Latino groups.**

Number of SVM + RF links and percent increase of links by ethnicity (maroon bars) and race (teal bars) using SVM+RF compared to deterministic linkage



Total number of links by SVM+RF presented at end of bars

Transitioning to a **machine learning linkage** yielded **11-38% more links** between COVID-19 case and vaccine records compared to a deterministic linkage. The **biggest increase** was among minority groups.

Summary of linkage method QA in Nov. 2021 vs. Apr. 2022

	A	B	C	D	E	F	G
QA Time Point	Number of records*	Linkage method	Number of links	Number of linkage disagreements (links not captured by the other method)**	Sample size of model disagreements for review	Number of false links among sample	Estimated false discovery rate assuming model agreements are correct: (D • (F/E)) / C
Nov. 2021 (pre-Omicron)	5,018,916 vaccine records 1,402,141 case records	Deterministic	736,564	90	90	40	0.005%
		SVM + RF	820,441	84,060	1,000	3	0.031%
Apr. 2022 (post-Omicron)	5,697,536 vaccine records 2,179,497 case records	Deterministic	876,778	962	962	139	0.016%
		SVM + RF	1,213,434	113,523	4,000	47	0.11%

* Records included in the linkage will be larger than reported values in Washington State as the inclusion criteria for the linkage differed from reporting criteria

** The difference between C & D columns for SVM + RF fields will not equal the C column for the deterministic linkage method as there was different inclusion criteria between the linkage methods

Stage	Methods
Model Strategy	Methodology adapted from Washington State CHS. • Radial Support Vector Machine & Random Forest models. **Both must agree.** This strategy demonstrated several advantages compared to other types of linkages such as: links non-exact matching records while **maintaining a low error rate**, can be **improved via QA**, runs relatively **quickly** and **manual burden is low**, and follows statistical assumptions.
Model Training	Used a **nestled sampling technique**. First a random sample of **10,000 COVID-19 case and vaccine record pairs** in Washington State was taken. **Five rounds of sampling without replacement** was carried out representing a total of about **5,500 record pairs**. Each round of sampling was followed by manual classification whether the record pair was a true link and models were trained based on those results.
Model Testing	The models were applied to **all historical COVID-19 case and vaccine data**. A field summarizing all distance metrics within a record pair was created to aid quality assurance and future manual review.
Quality Assurance	**Three** groups, totaling over **2,400 record pairs** were identified for QA: 1. Records containing common names which were not linked. **Type II error check** 2. Records linked despite overall high distance scores. **Type I error check** 3. Records linked despite name and sex disagreements. **Type I error check**

DISCUSSION

Transitioning to a machine learning linkage **increased the number of links, especially among non-White and Hispanic/Latino groups**. The increased number of links was associated with a *slightly higher false linkage rate*. While the rate of false links did increase, the real-world impact of this lower specificity resulted in a small amount of manual review. Model specificity could be improved by including more identifier linking variables. The higher yield of links was **consistent over time** based on QA analysis from Nov. 2021 and Apr. 2022.

CONCLUSION

Deterministic linkage strategies are insufficient for equitable surveillance when compared to a machine learning based-classification. This insufficiency was highlighted during the Omicron wave.

- The machine learning linkage enabled the WA DOH to **better assess the vaccination status of all COVID-19 cases** among other key surveillance efforts.

Washington State Department of HEALTH

DOH 422-265 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

CONTACT

Seth Rothbard, MPH
Seth.Rothbard@doh.wa.gov

***Poster was awarded the Outstanding Poster Presentation Award at the CSTE 2023 Conference**