Biostatistics, 2026, **27**(1), kxag002
https://doi.org/10.1093/biostatistics/kxag002
Article

Dynamic case-control sampling for rapid estimation of vaccine effectiveness against an emerging infectious disease variant

Taylor M. Fortnam[1,*], Laura C. Chambers[2], Alyssa Bilinski[1,3], Ewa King[4], Richard C. Huard[4], Ellen Amore[4], Lisa M. Gargano[4], Jim McDonald[4], Philip A. Chan[5], and Joseph W. Hogan[1]

[1]Department of Biostatistics, Brown University School of Public Health, 121 S Main St, Providence, RI 02903, United States
[2]Department of Epidemiology, Brown University School of Public Health, 121 S Main St, Providence, RI 02903, United States
[3]Department of Health Services, Policy, and Practice, Brown University School of Public Health, 121 S Main St, Providence, RI 02903, United States
[4]Rhode Island Department of Health, 3 Capitol Hill, Providence, RI 02908, United States
[5]Department of Medicine, Brown University, 222 Richmond St, Providence, RI 02903, United States

*Corresponding author: Department of Biostatistics, Brown University School of Public Health, 121 S Main Street, Providence, RI 02903, United States. Email: taylor_fortnam@brown.edu.

Summary

New SARS-CoV-2 variants arise frequently with different viral properties that can impact the effectiveness of the vaccines. Updating estimates of vaccine effectiveness (VE) in public health surveillance can be limited by the necessity of conducting a distinct study that entails analysis of prospective cohort data or using a test-negative design. We introduce a method for dynamically updating estimates of VE using data that accumulate in real time. Our method uses dynamic case-control sampling to estimate VE against a newly emerging variant relative to a previous variant. Dynamic case-control sampling is a technique that continuously updates VE estimates by comparing individuals infected with a newly emerging variant (defined as "cases") to those infected with a previously circulating variant (defined as "controls"). We use this estimate in combination with information about VE from the previous variant (these estimates are typically available from larger, traditional studies) to infer VE against the emerging variant. We demonstrate the utility of this method on the BA.1 and BA.2 sub-lineages of the Omicron variant. The method produces estimates of VE comparable to those produced using traditional methods, although with increased SE. The increase in error, however, is reasonable given a much smaller sample size than

Received: September 25, 2024. **Revised**: February 19, 2026. **Accepted**: February 20, 2026

other studies, and error ranges of the estimates could be significantly improved by sequencing a larger proportion of identified cases. Our method, which assumes only a fraction of the new cases are being sequenced, can be applied by health departments using routinely collected data to produce timely, rigorous VE estimates to rapidly identify potential changes in VE.

Keywords COVID-19, genomic surveillance, infectious disease surveillance, Omicron variant, vaccine effectiveness

1. Introduction

Following the emergence of the Alpha SARS-CoV-2 variant in December 2020, the COVID-19 pandemic has been marked by regular evolution and rapid spread of new immune-evasive variants. In the United States, the first case of the Omicron variant of SARS-CoV-2 was identified on 1 December 2021, and it was circulating widely in every region of the country by mid-December (Centers for Disease Control and Prevention 2024), causing case and hospitalization spikes. Preliminary studies of the BA.1 sub-lineage of the Omicron variant demonstrated reduced vaccine effectiveness (VE) in laboratory settings (Lu et al. 2022; Wilhelm et al. 2022) and were available rapidly, but do not map directly to specific health outcomes. In vivo estimates produced from observational studies showing reduced VE against symptomatic disease and hospitalization (Andrews et al. 2022; Buchan et al. 2022; Collie et al. 2022; Eggink et al. 2022; Spensley et al. 2022) were not available until February to July 2022. By late March, 2022, the BA.2 sub-lineage, which contained additional mutations unique from those observed in the BA.1 sub-lineage of the Omicron variant, had already become dominant (Yu et al. 2022), and comparable VE against symptomatic disease and hospitalization for the BA.2 sub-lineage compared to the BA.1 sub-lineage was demonstrated (Kirsebom et al. 2022). Public health officials must be able to quickly assess the degree of vaccine effectiveness (VE) against new strains of the virus to anticipate the impact of new variants and need for mitigation measures. Obtaining reliable estimates of VE often involves conducting a prospective cohort or test-negative case control study (Abu-Raddad et al. 2021; Lin et al. 2022; Rosenberg et al. 2022), both of which can require large sample sizes (World Health Organization 2021) and substantial time for cases to accumulate. Although a test-negative design requires a smaller sample size to produce an estimate with similar SE to a cohort study, bias is often introduced in the sampling of test-negative controls. Specifically, controls cannot be sampled in the same way as those who have tested positive; hence likelihood of seeking care may differ by vaccination status, symptom status, or unmeasured characteristics, making it difficult to ascertain VE against any infection rather than just symptomatic infections.

In addition, genomic sequencing is costly and typically only available for a sub-sample of positive tests. Approaches to estimating variant-specific VE include (i) computing VE for cohorts in time periods during which each variant is dominant (Seppälä et al. 2021; Tartof et al. 2022), which does not make use of sequenced samples and relies on the assumption that all infections are due to a single variant at a given time, (ii) sequencing thousands of samples for a test-negative design (Bruxvoort et al. 2021; Puranik et al. 2021; Tseng et al. 2022), despite the cost and time-lag for sequencing, and (iii) by relying on S-gene target failure, which is useful for differentiating between the Omicron and Delta variants, but cannot be used to differentiate between sub-lineages of the Omicron variant (Rahimi and Talebi Bezmin Abadi 2022). With the rapid dominance of new variants and these limitations of existing study methodologies, estimates of VE may be biased and typically are not available until a strain is already dominant and response options are limited.

The screening method (Farrington 1993) based on case-only data can be used to estimate VE when information about population proportion of vaccine coverage is also available. This proportion is often derived using information from census data, together with information about number vaccinated. However, because the sample of sequences for cases identified as the BA.1 or BA.2 sub-lineage is not necessarily a random sample from population, it is difficult to estimate the population proportion vaccinated in the population represented by the sampled cases. Surveillance of VE against influenza are often focused on examining seasonal strains (Kelly et al. 2009;

Jester et al. 2018), rather than on dynamic updating of VE within a wave of infection. van der Laan and Gilbert (2025) propose a method that addresses the limitations of traditional methods for estimating VE by introducing a semi-parametric estimator for strain-specific VE and establishing that this estimator may yield more robust estimates of the odds ratio than logistic regression under certain conditions. They also apply their method to estimate VE in a setting in which viral load information was available and produce an estimate of VE against a specific strain *post hoc*, rather than dynamically updating estimates.

In this paper, we propose a novel surveillance method for obtaining and updating estimates of VE against infection with an emerging variant using dynamic case-control sampling. The method is based on cases for which genomic sequencing is available, minimizing mis-classification bias relative to methods implementing calendar-based classification and can be applied in settings where only a subset of cases are sequenced. Our method is motivated by the Rhode Island Department of Health's (RIDOH) practice of genotyping approximately 10% of cases at the time the work is initiated. We use genomic surveillance data as cases of a new variant or sub-lineage arise to produce and continuously update VE estimates. Our approach uses routinely collected data and can be applied without the need to sample additional data on test-negative controls.

Our method additionally accounts for differences in population-level transmission and certain measured characteristics of diagnosed infections. We show that by mid-to-late January 2021 our method would have obtained stable estimates of VE against infection with the BA.1 sub-lineage that are consistent with VE results published in much larger studies and without some of the limitations introduced by cohort and test-negative designs. We also show via simulation that under different ordering of case-accumulation, we can expect results similar to those observed in the application. Further, with a higher proportion of samples sequenced for both the Delta and Omicron variants, VE estimates could be produced with increased precision and estimates may stabilize more quickly. The method can still be used even with home-based testing replacing much of the need for testing performed in formal/lab settings, provided that additional assumptions about whether those who report for testing constitute a representative sample hold.

In the following sections, we introduce notation and derive our VE estimator, then describe it's application to producing updating estimates of VE using routine COVID-19 surveillance data in Rhode Island. We display the properties of this estimator in a simulation framework. Finally, we discuss the benefits and limitations of estimating VE in this manner.

2. Materials and methods

2.1. Study design and notation

We aim to estimate VE against an emerging variant. Let S be an integer-valued random variable denoting SARS-CoV-2 infection status, with $S = 0$ indicating uninfected and $S = s$ (for $s > 0$) indicating infection with variant (or viral subtype) s. Let V represent integer-valued vaccine status, with $V = 0$ corresponding to being unvaccinated and $V = v$ (for $v > 0$) representing level of vaccination received (eg none, partial, full, booster). Encoding S and V as integers sets a natural baseline at zero without implying order.

Following epidemiologic convention, we define the VE in terms of the risk of infection with a specific SARS-CoV-2 subtype $S = s$ if vaccinated at level $V = v$ relative to the risk of infection if not vaccinated ($V = 0$). We denote this quantity by $\mathrm{VE}_v(s)$, and write

$$\mathrm{VE}_v(s) = 1 - \frac{P(S = s \mid V = v)}{P(S = s \mid V = 0)},$$

or 1 minus the relative risk of infection for those vaccinated at level v versus those unvaccinated.

In defining VE against infection with variant $S = s$, we assume the denominator for "risk of infection" includes only those infected with variant s and those uninfected; ie excluding other variants. Formally, let $\mathscr{S} = \{0, 1, 2, \ldots, K\}$ represent the possible infection status values at a specific point in time, with $s = 0$ denoting "uninfected" and the integer values $s > 0$ indexing infection

with specific variants. We assume, for a variant $s \neq 0$,

$$\frac{P(S = s \mid V = v, S \in \mathscr{S})}{P(S = s \mid V = 0, S \in \mathscr{S})} \approx \frac{P(S = s \mid V = v, S \in \{0, s\})}{P(S = s \mid V = 0, S \in \{0, s\})}.$$

The approximation is implicit in most published reports of variant-specific VE (Lauring et al. 2022; Link-Gelles et al. 2023). It can be expected to hold when the set $\mathscr{S} \backslash \{0, s\}$ is small, a condition that is largely met in the setting our method addresses. We further note that it is only required for the approximation to hold for the relative risk; it is not necessary for the respective numerator and denominator probabilities to be approximately equal. When risk of infection is low, relative risk is well-approximated by the odds ratio, and VE against infection with variant s can be represented as

$$\begin{aligned} \mathrm{VE}_v(s) &= 1 - \frac{P(S = s \mid V = v)}{P(S = s \mid V = 0)} \\ &\approx 1 - \frac{P(S = s \mid V = v) \,/\, P(S = 0 \mid V = v)}{P(S = s \mid V = 0) \,/\, P(S = 0 \mid V = 0)}. \end{aligned} \tag{1}$$

The approximation (1) can be viewed as an odds ratio where $S \in \{0, s\}$. Now consider estimating VE against an *emerging* variant s' when reliable estimates of VE against a previously circulating *index* variant s are available. The VE against the index variant is therefore given by $\mathrm{VE}_v(s)$, and we assume there is a reliable estimate available from (for example) a randomized trial or a large-scale observational study. For example, during the emergence of the Omicron variant, VE for the Delta variant had been estimated from several large cohorts and reported in peer-reviewed literature and elsewhere (Seppälä et al. 2021; Gram et al. 2022; Tartof et al. 2022).

Owing to a diminishing number of individuals who remain free of *any* previous infection, it may be difficult to generate an estimate of VE that compares risk or odds of infection by the emerging variant relative to not being infected. However, it may be possible to generate reliable estimates of a relative measure that captures the effect of vaccination on infection with an emerging variant s' relative to an index variant s. Restricting attention to those who are infected by either s or s', let

$$\psi_v(s', s) = \frac{P(S = s' \mid V = v)/P(S = s \mid V = v)}{P(S = s' \mid V = 0)/P(S = s \mid V = 0)}.$$

denote the odds ratio quantifying the association of vaccination at level v against emerging variant s' relative to an index variant s; ie where $S \in \{s', s\}$. Using this formulation, the standard VE measure (1), quantifying effectiveness against infection with variant s relative to not being infected, can be written as $\mathrm{VE}_v(s) = 1 - \psi_v(s, 0)$.

Moreover, VE against infection with an emerging variant can be expressed in terms of (i) VE for an index variant s, and (ii) the odds ratio capturing the effect of vaccination on infection with s' compared to s. Specifically, we can write $\psi_v(s', 0)$ as

$$\begin{aligned} \psi_v(s', 0) &= \frac{P(S = s' \mid V = v)/P(S = 0 \mid V = v)}{P(S = s' \mid V = 0)/P(S = 0 \mid V = 0)} \quad \text{(by definition)} \\ &= \frac{P(S = s' \mid V = v)/P(S = s \mid V = v)}{P(S = s' \mid V = 0)/P(S = s \mid V = 0)} \times \frac{P(S = s \mid V = v)/P(S = 0 \mid V = v)}{P(S = s \mid V = 0)/P(S = 0 \mid V = 0)} \\ &= \psi_v(s', s)\psi_v(s, 0), \end{aligned}$$

so that vaccine efficacy against the emerging variant s' can be represented as

$$\begin{aligned} \mathrm{VE}_v(s') &= 1 - \psi_v(s', 0) \\ &= 1 - \psi_v(s', s)\psi_v(s, 0) \\ &= 1 - \psi_v(s', s)\left\{1 - \mathrm{VE}_v(s)\right\}. \end{aligned} \tag{2}$$

This motivates our general approach to calculating $\mathrm{VE}_v(s')$, the VE against a new variant relative to being uninfected. First, using an accumulating database that contains information on infections from both the new variant s' and an index variant s, we use dynamic case-control sampling to

generate real-time updates for $\psi_v(s', s)$. The sampling is dynamic because the estimates of $\psi_v(s', s)$ can be updated in real time as new cases accumulate. Next, we combine this with available estimates of VE against the index variant, $\text{VE}_v(s)$ to derive updated estimates of $\text{VE}_v(s')$ using (2). As we show in Section 3, the updated estimates can be calculated daily to support ongoing public health surveillance.

For the estimates of VE against the emerging variant to be as precise and locally relevant as possible, we need to address several challenges. First, we need to generate reliable estimates of $\psi_v(s', s)$ from a sample of cases with sequenced virus, where selection into the sequenced sample is potentially nonrandom relative to the population of interest. Second, we need to incorporate uncertainty about existing estimates of VE against the index variant. Third, we note that comparative estimates of VE across different variants may not take into account differences in transmissibility, or attack rate, between the 2 variants.

The following sections detail our approaches to each of these challenges.

2.2. Estimation of relative vaccine effectiveness against an emerging variant

Our method uses individual-level data on all diagnosed infections during a defined analysis period, with associated demographic information and sequencing information for a subset of infections. For each individual, let G denote the indicator of whether a genomic sequence is available for the confirmed case (1 if yes, 0 if no). As above, let S represent infection status, with $S = 0$ denoting no current or previous infection and $S = 1, \ldots, J$ indexing viral subtype among those who are currently infected. Recall that $V \in \{0, 1, 2, 3\}$ represents vaccine status (not vaccinated, partial, full, and boosted, respectively). Finally, let X denote a vector of individual-specific covariates.

Availability of a genomic sequence is needed to classify the viral subtype for an individual infection. Because sequenced cases are a subset of all diagnosed infections, we introduce a model for the selection mechanism that leads to sequencing. For each new infection, let G denote the indicator of whether a genomic sequence is performed (1 = yes, 0 = no), and let $\pi(X_i) = P(G = 1 \mid X_i)$ represent the selection mechanism as a function of individual covariates X.

Using data from those with viral subtype available (those with $G_i = 1$), we obtain information on the variant or sub-lineage, and divide the sample of sequenced diagnosed infections into cases, defined as those infected by an emerging variant ($S = s'$), and controls, who are infected by an index variant ($S = s$). In this sample, vaccine status V indicates vaccination status of the individual as of the diagnosis date.

We estimate $\psi_v(s', s)$ via weighted logistic regression on a matched set of cases and controls. The matching is done using propensity scores estimated from a model $P(S = s' \mid X_i) = g(X_i; \beta)$, where g is a regression model such as logistic regression; in our application we use full matching (Hansen 2007; Stuart et al. 2011) to make maximum use of the available data. This creates Q matched sets such that within each matched set q, there are n_q controls and n'_q cases; each matched set contains at least 1 case and at least 1 control, and no infections are discarded.

The model used to estimate $\psi_v(s', s)$ is specified as

$$\begin{aligned} \text{logit}\left\{P(S = s' \mid X_i, V_i)\right\} &= \sum_{v=0}^{3} \theta_v \mathbb{I}(V_i = v) + h(X_i; \alpha) \\ &= \sum_{v=0}^{3} \log\left\{\psi_v(s', s)\right\} \mathbb{I}(V_i = v) + h(X_i; \alpha), \end{aligned} \tag{3}$$

where the vaccine status indicators $\mathbb{I}(V_i = v)$, for $v = 0, 1, 2, 3$, are covariates; $\theta_v = \log\left\{\psi_v(s', s)\right\}$ are their respective coefficients; and $h(X_i; \alpha)$ is a user-specified function of individual covariates X_i included to adjust for any remaining imbalances and increase efficiency. The exponentiated coefficients $\exp(\theta_v)$ from the fitted model are estimates of $\psi_v(s', s)$.

The model is fitted using weighted maximum likelihood with weights

$$w_{iq} = \frac{n'_q}{n_q}\left(\frac{\pi(X_i;\hat{\beta})}{\sum_i \pi(X_i;\hat{\beta})}\right)^{-1}. \tag{4}$$

The weights account for the varying numbers of cases and controls in the matched sets and for sample selection of infections that have a genomic sequence available for ascertaining viral subtype. We use weighted logistic regression to estimate an odds ratio interpretable as an estimate of relative VE, however the method is not dependent on the use of logistic regression; we could carry out inference using any valid model for a binary endpoint.

2.3. Inference about vaccine effectiveness against a new variant

To estimate and draw inferences about VE against a new variant, we combine estimates of relative VE derived using methods described in Section 2.2 with existing estimates of VE against the index variant. Estimates and uncertainty measures for VE against the index variant are derived from published randomized trials and observational studies.

Recall from (2) that VE against the emerging variant is given by $\text{VE}_v(s') = 1 - \psi_v(s',s)\{1 - \text{VE}_v(s)\}$. Noting that $\text{VE}_v(s')$ is a function of $\psi_v(s',s)$ and $\text{VE}_v(s)$, our approach is to use simulation to approximate its sampling distribution and generate inferences about $\text{VE}_v(s')$. Because $\log\{\psi_v(s',s)\} = \theta_v$ is estimated using (weighted) maximum likelihood, we approximate the sampling distribution of $\log\left\{\hat{\psi}_v(s',s)\right\}$ using $\mathcal{N}(\hat{\theta}_v, \hat{\sigma}^2_{\theta_v})$, where $\sigma^2_{\theta_v} = \text{var}(\hat{\theta}_v)$. Using SEs and CIs extracted from the literature on VE Using we make a similar assumption about estimates of $\text{VE}_v(s)$—namely that the distribution of $\widehat{\mu}_v = \log\left\{\widehat{\text{VE}}_v(s)\right\}$ can be approximated by $\mathcal{N}(\hat{\mu}_v, \hat{\sigma}^2_{\mu_v})$.

To simulate from the sampling distribution of $\text{VE}_v(s')$, we proceed in 3 steps: (i) simulate a pair $(\tilde{\theta}_v, \tilde{\mu}_v)$ from the respective sampling distributions given above, (ii) set $\tilde{\psi}_v(s',s) = \exp(\tilde{\theta}_v)$ and $\widetilde{\text{VE}}_v(s) = \exp(\tilde{\mu}_v)$, and (iii) use (2) to calculate $\widetilde{\text{VE}}_v(s')$ from $\tilde{\psi}_v(s',s)$ and $\widetilde{\text{VE}}_v(s)$. These steps can be repeated a large number of times (eg 10^5) to simulate replicates from the sampling distribution of $\widehat{\text{VE}}_v(s)$, and from there to calculate an estimate and associated CI.

2.4. Note on differences in infectivity of index and emerging variant

The infectivity rate can differ between strains or variants of a virus: when a new variant emerges, mutations can cause changes in the infectivity of the virus (Jalali et al. 2022) as well as impact the effectiveness of existing vaccines at neutralizing the virus in the body (Otto et al. 2021). If we define $\lambda_v(s) = P(S = s \mid V = v)$, we can then write

$$\begin{aligned}\psi_v(s',s) &= \frac{P(S=s' \mid V=v)/P(S=s \mid V=v)}{P(S=s' \mid V=0)/P(S=s \mid V=0)} \\ &= \frac{\lambda_v(s')/\lambda_v(s)}{\lambda_0(s')/\lambda_0(s)}.\end{aligned}$$

Now define $\kappa = \lambda_0(s')/\lambda_0(s)$, so that κ reflects the rate of infectivity in unvaccinated individuals among those with the emerging variant relative to the index variant. Thus, it is important to note that an observed estimate of the relative VE $\psi_v(s',s)$ may be partially attributable to actual differences in effectiveness (numerator) and differences in infectivity (denominator).

3. Estimating Vaccine Effectiveness Against BA.1 and BA.2 Sub-Lineages In Rhode Island

3.1. Overview

In this section, we separately estimate VE against each of the BA.1 and BA.2 sub-lineages of the Omicron variant, using Delta as the index variant. Specifically, we calculate relative VE, $\psi_v(s',s)$, where the BA.1 and BA.2 sub-lineages are labeled with s' and the Delta variant is labeled with s.

Hence we treat the Omicron variants as "cases" and the Delta variant as "control" in the estimation of relative VE.

As indicated in Section 1, only a subset of SARS-CoV-2 positive specimens collected in Rhode Island are sequenced. Using (4), sampling weights that are inversely proportional to $\pi(X_i;\beta)$ are used to ensure that sequenced specimens are a representative sample of all infections diagnosed after the conception of RIDOH's genomic sequencing program. Using weighted logistic regression (3), we calculate absolute VE for both the BA.1 and BA.2 sub-lineages by combining our estimate $\hat{\psi}_V(s',s)$ of relative VE with inferential information about VE against the Delta variant derived from published reports.

To illustrate the dynamic nature of the method for updating VE against the BA.1 and BA.2 sub-lineages as new sequences are obtained, we define $\psi_V(s',s;t)$ as the relative VE on day $t = 1,2,3,\ldots$, where $t = 1$ is the first day on which an observed sequence indicates that the viral strain is that of an emerging variant. We provide inferences for

$$\mathrm{VE}_V(s';t) = 1 - \psi_V(s',s;t)\{1 - \mathrm{VE}_V(s)\} \tag{5}$$

as a function of t, summarizing both point estimates and SE over time. Our estimates of VE for the BA.1 sub-lineage are compared to estimates that were later published using much larger datasets. For the BA.2 sub-lineage, we were unable to find directly comparable estimates of VE against any infection but do contextualize our results by providing literature estimates of VE against symptomatic infection with the BA.2 sub-lineage that were published later.

3.2. Formation of analysis dataset

As of the time this analysis was completed, RIDOH collected demographic information associated with all SARS-CoV-2 positive specimens, linked these individual records to a vaccination registry, and selected a sample of SARS-CoV-2 positive specimens for sequencing from various labs in the state by geographic area and hospitalization or "breakthrough" status ("breakthroughs" are COVID-19 infections that occur more than 14 d after vaccination). The demographic variables included in the covariate vector X_i are age, sex, race, congregate care status, and a ZIP-code-based 3-tier community-level COVID-19 risk classification. The covariate vector X can be elaborated to include information about timing of vaccination and prior infection status.

We start with infections diagnosed after 1 January 2021, when the RIDOH genomic surveillance program was initiated, and include data through 13 June 2021. During this period, specimens were collected from 283,385 individuals with at least 1 laboratory-confirmed SARS-CoV-2 infection reported to RIDOH. Of these, 14,862 individuals have at least 1 sequenced result. The file containing these data was finalized by RIDOH on 7 April 2023. We then restricted our sample to those aged 16 and over (owing to vaccine eligibility criteria at the time) and having no previous documented infection, resulting in 11,907 observations. We exclude those with any previous, documented infection because previous infection has been demonstrated to confer some immunity (Wilhelm et al. 2022). At the time, there was still a relatively large pool of individuals without any documented infection, although this subset will dwindle over time. Finally, we limited the analysis to sequenced diagnosed infections classified as variants relevant to our analysis, retaining infections classified as the Delta variant diagnosed on or after 1 November 2021, and those classified as the Omicron variant diagnosed on or after 24 November 2021 [when the Omicron variant was first reported to the WHO (Perrine and CDC COVID-19 Response Team 2021)]. Appendix Figs S3 and S6 depict timing of sequence acquisition and a flowchart depicting inclusion criteria.

The final analysis sample comprises data from 5,751 individuals, of which 2,220 (39%) had infections with the Omicron BA.1 sub-lineage (cases), 1,462 (25%) with the Omicron BA.2 sub-lineage (cases), and 2,069 (36%) with the Delta variant (controls). Details on all sub-lineages observed in the available data, and the breakdown of included sub-lineages of the Omicron variant over time, are shown in the Appendix (Section SA.1 and Appendix Fig. S3). Characteristics of cases and controls are summarized in Table 1.

Table 1 Characteristics of cases (BA.1 or BA.2) and controls (Delta), as well as all diagnosed infections and sequenced diagnosed infections.[a]

Variable	Measure	BA.1 ($N = 2{,}220$)	BA.2 ($N = 1{,}462$)	Delta ($N = 2{,}069$)	Sequenced diagnosed infections ($N = 14{,}862$)	Diagnosed infections ($N = 283{,}385$)
Age:	Mean (SD)	42.5 (18.7)	44.0 (19.8)	43.8 (19.3)	35.9 (21.4)	35.0 (21.4)
	# Missing	0	0	0	0	5
Sex[b]:	Female	1,010 (45%)	809 (55%)	903 (44%)	7,135 (48%)	113,676 (40%)
	Male	895 (40%)	642 (44%)	800 (39%)	6,357 (43%)	97,635 (34%)
	Other	7 (0.3%)	8 (0.5%)	8 (0.3%)	35 (0%)	658 (0%)
	Declined to State	< 5	0	< 5	11 (0%)	280 (0%)
	# Missing	305 (14%)	8 (0.5%)	357 (17%)	1,324 (9%)	71,136 (25%)
Community risk level[c]:	High	360 (16%)	163 (11%)	361 (17%)	2,832 (19%)	60,154 (21%)
	Moderate	503 (23%)	269 (18%)	502 (24%)	3,419 (23%)	62,900 (22%)
	Low	1,276 (57%)	998 (68%)	1,130 (55%)	8,198 (55%)	137,499 (49%)
	# Missing	81 (4%)	32 (2%)	76 (4%)	413 (3%)	22,832 (8%)
Race/Ethnicity[d]:	White	1,095 (49%)	813 (56%)	1,097 (53%)	8,054 (54%)	133,168 (47%)
	Black	106 (5%)	40 (3%)	87 (4%)	775 (5%)	12,314 (4%)
	Hisp./Latino (any race)	243 (11%)	99 (7%)	199 (10%)	2,256 (15%)	43,504 (15%)
	Asian	63 (3%)	60 (4%)	41 (2%)	339 (2%)	5,394 (2%)
	American Indian	6 (0%)	7 (0%)	9 (0%)	66 (0%)	850 (0%)
	Pacific Islander	< 5	< 5	< 5	12 (0%)	194 (0%)
	Other	23 (1%)	17 (1%)	36 (2%)	230 (2%)	3,942 (1%)
	Multiple Races	25 (1%)	19 (1%)	27 (1%)	289 (2%)	4,162 (1%)
	Declined to State	76 (3%)	31 (2%)	64 (3%)	420 (3%)	8,081 (3%)
	# Missing	580 (26%)	374 (26%)	506 (24%)	2,421 (16%)	71,776 (25%)
Congregate care status:	Resident	93 (4%)	17 (1%)	102 (5%)	435 (3%)	6,759 (2%)
	Employee	75 (3%)	36 (2%)	39 (2%)	301 (2%)	6,156 (2%)
	Not Congregate	2,003 (90%)	1,396 (95%)	1,922 (93%)	140,09 (94%)	268,785 (95%)
	# Missing	49 (2%)	13 (1%)	6 (0%)	117 (1%)	1,685 (1%)
Vaccination status:	Unvaccinated	549 (25%)	269 (18%)	1,017 (49%)	8,727 (59%)	154,675 (55%)
	Partial Primary Series	118 (5%)	72 (5%)	59 (3%)	502 (3%)	9,637 (3%)
	Primary Series	843 (38%)	306 (21%)	938 (45%)	3,927 (26%)	80,873 (29%)
	Booster	710 (32%)	815 (56%)	55 (3%)	1,706 (11%)	38,200 (13%)
Time between completion of primary series and diagnosis[e]:	14 to 90 d	57 (3%)	10 (.7%)	47 (2%)	418 (3%)	8,766 (3%)
	91 to 180 d	143 (6%)	24 (2%)	126 (6%)	1,253 (8%)	17,918 (6%)
	181 to 270 d	355 (16%)	72 (5%)	618 (30%)	1,565 (11%)	36,072 (13%)
	> 270 d	284 (13%)	200 (14%)	147 (7%)	691 (5%)	18,117 (6%)
Time frame:	First diagnosis	26 November 2022	19 December 2021	01 November 2022	1 January 2021	1 January 2021
	Last diagnosis	1 May 2022	31 May 2022	25 January 2022	6 March 2022	12 June 2022

[a]For the population of all sequenced first diagnosed infections, we only include infections with diagnosis date occurring on or after 1 January 2021. Vaccination status is defined as of the infection diagnosis date.

[b]The sex variable was re-leveled to 4 categories: female; male; other; and unknown, which included "Declined to State" and "Missing". The results using the re-leveled variables and the original levels of these variables are shown in the Appendix.

[c]ZIP-code-based 3-tier community risk classification created by RIDOH to help guide COVID-19 surveillance and response efforts.

[d]The race/ethnicity variable was re-leveled to 5 categories: Hispanic/Latino (any race); Black; White; Other, which included Asian, Amer. Indian/Alaska Native, Native Hawaiian/Pac. Islander, Other, and Multiple Races; and Unknown, which included Declined to State and Missing.

[e]Among those who have completed the primary vaccination series and received no booster doses.

3.3. Inference about vaccine effectiveness against Omicron BA.1 and BA.2

On each day of the surveillance period, cases (BA.1 and BA.2 infections) and controls (Delta infections) are fully matched into sets via propensity score matching on age, sex, race/ethnicity, congregate care status, and a ZIP-code-based 3-tier community risk classification based on community characteristics such as population density, sociodemographics, and COVID-19 burden (quantified as low-, moderate-, and high-risk). The matching is done separately for the BA.1 vs Delta comparison and the BA.2 vs Delta comparison.

We compute relative VE in terms of $\psi_v(s', s; t)$ via weighted logistic regression as described in Section 2.2. Vaccine levels v are defined as

$$v = \begin{cases} 0, \text{ if unvaccinated} \\ 1, \text{ if partially completed primary vaccination series} \\ 2, \text{ if completed primary vaccination series} \\ 3, \text{ if completed primary vaccination series and at least one booster dose} \end{cases}$$

but encoded in the model as nominal values.

Using the approximation in (1), and following the method described in Section 2.3, we combine published estimates of VE against infection with the Delta variant with our estimates of $\psi_v(s', s; t)$ to generate inferences for VE against infection with the BA.1 or BA.2 sub-lineages

Table 2 Estimated odds ratios $\psi_\nu(s', s)$ and 95% CIs from the weighted logistic regression for each sub-lineage of the Omicron variant relative to the Delta variant by vaccination status when adjusting for variables used in the matching.[a]

	Estimate (CI) for $\psi(s', s)$	
Vaccination status	$s' =$ BA.1	$s' =$ BA.2
Unvaccinated	...	...
One dose of 2-dose series	3.74 (2.70, 5.23)	6.12 (4.10, 9.19)
Completed primary series	1.77 (1.52, 2.05)	1.25 (1.03, 1.52)

[a]Values $\psi(s', s) > 1$ indicate that VE against infection with the Delta variant is greater than that for each of these sub-lineages of the Omicron variant.

of the Omicron variant. Specifically, recall that VE against infection with the Delta variant is $1 - \psi_\nu(s, 0)$ so that, following (2), VE against the Omicron variant (relative to being unvaccinated) is $1 - \psi_\nu(s', s; t)\psi_\nu(s, 0)$. Although we do have complete vaccination information, owing to limited number of sequences for individuals with only 1 dose of the 2-dose primary vaccine series ($\nu = 1$) or with a vaccine booster ($\nu = 3$) during the analysis period, we restrict our inferences about VE for sub-lineages of the Omicron variant to those having completed the 2-dose primary vaccination series.

4. Results

4.1. Inferences about vaccine effectiveness using full dataset

We first analyze the full dataset, comprising 2,220 BA.1, 1,462 BA.2, and 2,069 Delta infections. Table 2 displays the estimates and 95% CIs for $\psi(s', s)$ for the BA.1 and BA.2 sub-lineages (s') relative to the Delta variant (s), yielding $\hat{\psi}(\text{BA.1}, \text{Delta}) = 1.77$ and $\hat{\psi}(\text{BA.2}, \text{Delta}) = 1.25$, indicating that full primary vaccine series is more effective against Delta than against either BA.1 or BA.2.

Table 3 shows the corresponding estimates and 95% CI for VE against BA.1 and BA.2 for those having a complete primary vaccination series, calculated using the methods described in Section 2.3. The entries in the table use estimates of VE against the Delta variant reported by a variety of larger-scale studies (Bruxvoort et al. 2021; Puranik et al. 2021; Seppälä et al. 2021; Gram et al. 2022; Tartof et al. 2022; Tseng et al. 2022). The derived VE against the BA.1 and BA.2 sub-lineage are lower than estimates of VE against the Delta variant from the literature. Estimated VE for the BA.1 sub-lineage ranges from 10% to 77%, and for the BA.2 sub-lineage ranges from 36% to 84% (Table 3). The range in point estimates is driven by variability in estimates of VE against infection with the Delta variant from the literature, which may result from differences in time since vaccination or differences among those presenting for testing by study. However, most of our estimates of VE against the Delta variant fall in the range of 60% to 65%.

For several reasons, we use the estimates of VE against infection with the Delta variant from Tseng et al. (2022) to conduct further analyses and as a basis for simulations reported in Section 4.1. We rely on this study when computing estimates because (i) they provide estimates of 2-dose VE at different ranges of time since vaccination, (ii) utilize S-gene target failure to determine variants rather than just a Delta-dominant time period, (iii) conducted the study in the United States and therefore likely had social-distancing and vaccination protocols that were more similar to those in Rhode Island than those in other countries, and (iv) and their resulting estimates were similar to what was reported by other studies.

4.2. Dynamic updating of vaccine effectiveness against an emerging variant

The key motivation for our method is to generate real-time updates of VE against an emerging variant using accumulating information on sequenced infections. Once VE for the index variant has been established, inferences about VE for an emerging variant are updated on a daily basis. To illustrate how accumulating information informs inference for VE against an emerging variant over time, we ordered the emerging Omicron variant infections by diagnosis date and used our

Table 3 Inferences for VE given complete primary series against infection with the BA.1 and BA.2 sub-lineages of the Omicron variant in Rhode Island based on VE estimates against the Delta variant from 6 previously published studies and estimates of relative VE from Table 2.

			Vaccine effectiveness		
Location	Study type	No. of infections; No. of without infection	Delta ($VE_v(s)$)	BA.1 ($VE_v(s')$)	BA.2 ($VE_v(s')$)
California[a] (Tartof et al. 2022)	Cohort, Delta-dominant	197,535; 2,919,754	49 (46, 51)	10 (–5, 23)	36 (22, 48)
California[b] (Bruxvoort et al. 2021)	Case control, sequenced	2,027; 10,135	87 (84, 89)	77 (71, 82)	84 (79, 88)
California (Tseng et al. 2022)	Case-control, S-Gene Target Failure	26,683; 109,662	64 (60, 67)	36 (24, 47)	55 (44, 64)
Minnesota[c] (Puranik et al. 2021)	Case control Delta-dominant	25,869; 25,869;	59 (36, 75)	25 (–18, 56)	47 (15, 69)
Norway (Seppälä et al. 2021)	Cohort, sequenced	5,430; 4,199,429	65 (61, 68)	38 (26, 48)	56 (46, 65)
Denmark[d] (Gram et al. 2022)	Cohort; Delta-dominant and sequenced	34,636; 842,397	65 (64, 66)	38 (28, 47)	56 (47, 64)

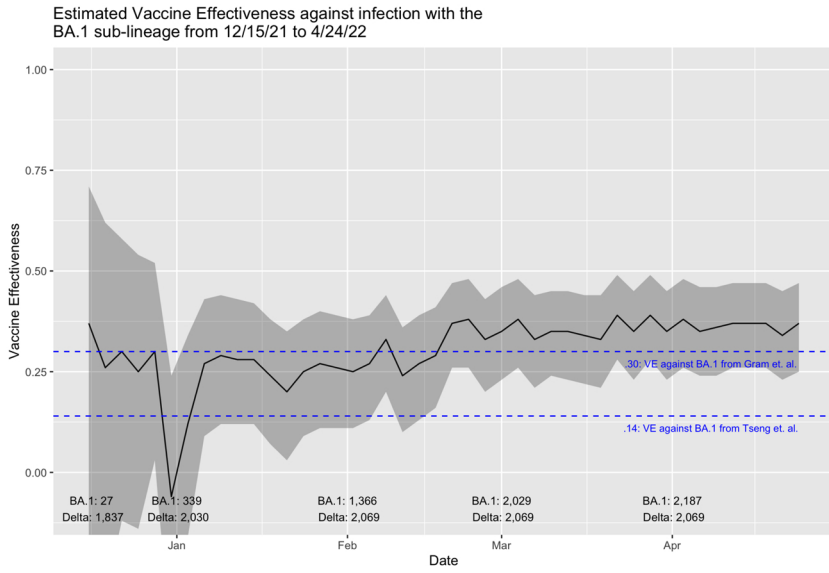
[a]Estimated effectiveness $\geq$ 7 mo after completion of primary vaccination series, Pfizer vaccine only.
[b]Moderna vaccine only.
[c]Investigated Moderna and Pfizer vaccines separately, estimates shown are averages.
[d]$>$ 120 d after completion of primary vaccination series, among individuals ages 12 to 59 yrs. Approximately 75% of samples were sequenced during the Delta-dominant portion of the study period.

methods to re-analyze the available data on each day. This generates successive inferences for $VE_v(s'; t)$ defined in (5).

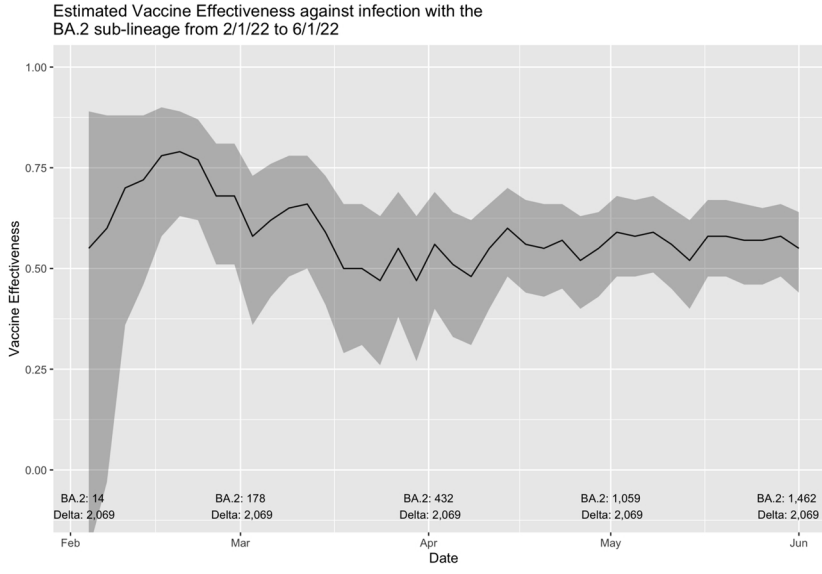
Figure 1a and b show the progression of the estimates of VE against infection with the BA.1 and BA.2 sub-lineages using previously published estimates of VE against infection with the Delta variant from Tseng et al. (2022). To produce each figure, we subset the data sequentially to include the available cases and controls by diagnosis date up to and including the date shown on the horizontal axes. In Fig. 1a, estimates of VE for a complete primary vaccination series against infection with the BA.1 sub-lineage began to stabilize after the accumulation of only a few hundred cases; CIs narrowed as cases accrued. For each sub-lineage, it took about 6 to 8 wks from the time of the first appearance of the new sub-variant for the estimate to stabilize.

Estimates produced for the BA.1 sub-lineage are comparable with those found in the literature (and shown as horizontal lines in Fig. 1a). Tseng et al. (2022) find that VE against infection with the BA.1 sub-lineage declines from 44% (35, 52) in the first 3 mo following completion of the primary vaccination series to 6% (0, 11) more than 270 d after completion. Similarly Gram et al. (2022) found that VE against infection with what they call "Omicron variant" (likely the BA.1 sub-lineage) for those 12 to 59 yrs declines from 40% (38.6, 41.3) in the first month after completion of the primary vaccination series to 12.6% (12.0, 13.3) more than 120 d after, with most of their estimates around 31 to 32%. An additional study by Eggink et al. (2022) also established that VE against any infection was reduced for the Omicron variant compared to the Delta variant, but did not provide an estimate for absolute VE.

We were unable to locate estimates of VE specifically against any infection with the BA.2 sub-lineage. While many studies emphasize decreasing effectiveness as time since vaccination increases, 2 estimate VE against symptomatic infection: a UK case-control study found that VE against symptomatic infection with the BA.2 sub-lineage (27.8% (25.9, 29.7)) was higher than that against the BA.1 sub-lineage (14.8% (12.9, 16.7)) (Kirsebom et al. 2022), similar to our result in that we



(a) Updating estimates of VE as cases accumulate, BA.1 sub-lineage



(b) Updating estimates of VE as cases accumulate, BA.2 sub-lineage

Figure 1 Progression of VE estimate over time against infection with the BA.1 or BA.2 sub-lineages of the Omicron variant. The estimates are computed using VE estimates against the Delta variant from Tseng et al. (2022) and our data on all available sequenced first infections in those aged 16 yrs and older prior to and including the date shown on the horizontal axis. Shaded regions indicate the 95% CIs for the estimate. Numbers of cases and controls for the estimate produced on the first date shown on the graph and the first of each following month are along the horizontal axis. Dashed lines (included for the BA.1 sub-lineage only) indicate VE estimates against infection with the Omicron variant from other studies: Tseng et al. (2022) and Gram et al. (2022).

also saw higher VE against the BA.2 sub-lineage, and a Qatar study estimated VE of the Pfizer vaccine of 51.7% (43.2, 58.2), also against symptomatic infection (Chemaitelly et al. 2022), similar to our result for VE against the BA.2 sub-lineage. However, it is not entirely appropriate to compare VE against any infection with VE against symptomatic infection. Our estimates are reasonable when compared with literature estimates of 2-dose VE against any infection with unspecified sub-lineages of the Omicron variant: 43% (42, 44) from Šmíd et al. (2022) and 55.2% (23.5, 73.7) from Hansen et al. (2021). Although our estimates are comparable to literature estimates, our credible intervals are wide due to the relatively small overall sample size and the fixed number of observations of the previously circulating variant once the emerging variant becomes dominant.

4.3. Simulation study of dynamic VE estimation properties

4.3.1. Objectives and specifications of the simulation study

We conduct a simulation study to investigate the properties of our dynamically updated estimates of VE against the emerging variant, including rate of convergence to the true VE and limiting width of the CIs for the estimate. The goal is to understand, for a fixed number of sequences for the index variant, how many sequences from the emerging variant are needed to arrive at an estimate and CI that would not appreciably improve with the addition of more sequences from the emerging variant. The simulation also assesses sensitivity of the evolving inference about relative VE to different ordering of the cases over time.

Recall that VE for the emerging variant is given by $\mathrm{VE}(s') = 1 - \psi(s', s)\{1 - \mathrm{VE}(s)\}$. Let N and N' denote the number of sequences from the index and emerging variants, respectively. For simplicity, we assume for the simulation that the N is fixed when accumulation of sequences on the index variant begins; hence, for fixed N, the precision of estimates of $\mathrm{VE}(s')$ will depend solely on N', and in particular $\mathrm{var}\left\{\widehat{\mathrm{VE}}(s')\right\}$ will depend, as a function of N', only on $\mathrm{var}\left\{\hat{\psi}(s', s)\right\}$. Moreover, because we are using case-control sampling to estimate $\psi(s', s)$, and because the number N of index cases (cases) remains fixed following emergence of a new variant, the degree of precision and the width of the CIs will eventually reach a limiting value that will not materially change even as N' increases. In short, the precision of our estimates of $\mathrm{VE}(s')$ is ultimately dictated by the number of sequences available for the index variant. From a practical point of view, health departments or public health authorities who implement this dynamic surveillance method can use this information to plan or budget the number of sequences to be collected as each new variant emerges.

The data were generated for the simulation by re-sampling the observed data under random case-accumulation scenarios. Covariate values associated with each case record were maintained while only altering the case date in order to maintain realistic observations while examining different case-accumulation patterns. Specifically, the data-generating mechanism was as follows: holding the number of sequences of the Delta variant (controls) constant, to mimic the sampling distribution of case sequences, we shuffled the observed diagnosis date for all records identified as the BA.1 sub-lineage to produce 100 permutations of the original data. Then, for each dataset, we mimicked the process of computing dynamic estimates of VE for the BA.1 sub-lineage by re-estimating $\psi_V(s', s; t)$ and the corresponding value of corresponding estimates of $\mathrm{VE}(s; t)$ using data available at 3-d increments. This analysis was carried out for each of the 100 permuted datasets, so that estimates, SEs, and CIs could be represented as a function of both number of accumulated sequences of the BA.1 sub-lineage N' and as a function of days t.

4.3.2. Simulation results

The results of each iteration of the simulation are shown in Fig. 2. The top panels show trace plots of $\hat{\psi}(s', s; N')$ and its SE as a function of N' and the bottom show trace plots of $\hat{\psi}(s', s; t)$ and its SE as a function of t. The plots in the right panel depict trace plots of estimated SE of $\log\left\{\hat{\psi}(s', s; N')\right\}$ and $\log\left\{\hat{\psi}(s', s; t)\right\}$, indicating that the estimated SE has very little variability after accumulation of just 250 cases or so, and that accumulating more than about 1,200 cases does not lead to appreciable improvement in precision. We focus on SE of $\log\left\{\hat{\psi}(s', s; N')\right\}$ because

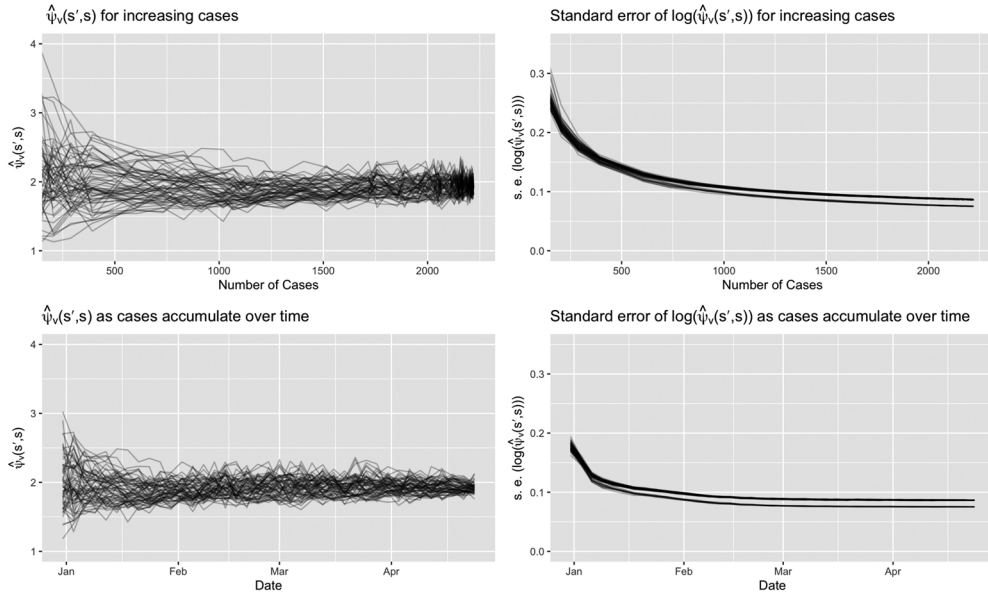


Figure 2 The left panels show the progression of $\psi_V(s', s)$ and the right panels show estimates of the SE of $\psi_V(s', s)$ over increasing cases and as cases accumulate over time. Each line shows the progression of an estimate throughout one iteration of the simulation. In the bottom 2 panels, the horizontal axes span the time period 30 December 2021 to 1 June 2022.

for a fixed number N of index (control) sequences and for a fixed value of VE and SE for the VE of the index variant, the remaining uncertainty in the estimate of VE for the emerging variant derives solely from uncertainty associated with the estimate $\hat{\psi}(s', s; N')$.

We computed estimates and CIs for $\text{VE}_V(s')$ using the same procedure described in Section 2 using VE estimate 64% and CI $(60, 67)$ for the Delta variant reported in Tseng et al. (2022). In this simulation, it took accumulation of 583 cases on average for the credible interval to fall within 35 percentage points, but around 1,700 cases on average for the credible interval to fall within 30 percentage points, indicating that increasing cases only decreases the interval width marginally, partially due to the dependence of the estimate on a fixed-size interval for the estimate of $\text{VE}_V(s)$. Moreover, the estimate of $\text{VE}_V(s')$ changes little after accumulation of about 1,200 cases. Referring to Appendix Fig. S5 and based on the rate at which RIDOH was generating sequences for new cases of the BA.1 sub-lineage, a reliable estimate of VE using this method could have been produced by early January 2022, illustrating the potential of this method for rapid assessment of VE. If capacity for sequencing were expanded, a reliable estimate could have been generated even earlier. For comparison, other estimates of VE against the BA.1 sub-lineage from observational studies started appearing in the published literature by September 2022 (Gram et al. 2022) and January 2023 (Tseng et al. 2022) The values and intervals of our estimates also encompass values comparable to those reported in other studies (Table 3) and no longer fluctuate widely after about 6 to 8 wks, suggesting that our estimates quickly become stable (Fig. 1a and b).

Notably, our resulting intervals are large, especially when compared to other reported VE estimates from larger studies. The precision of the estimate of $\text{VE}_V(s_*)$ that we produce depends both on the number of available cases and controls and the error associated with the estimate of VE against the previous variant ($\text{VE}_V(s_0)$). Due to the small size of Rhode Island, and the practice of sequencing only about 10% of documented infections at the time, we have effectively utilized all available sequenced information on these variants, as evidenced by the observation that cohort and case-control studies of comparably small sizes would have produced similarly wide intervals (World Health Organization 2021).

Although some sacrifice is made in the efficiency of our method for estimating VE in comparison to other study types because of the reliance on external estimates, our method does not require collection of extensive data as in a cohort study or sampling of test-negative controls for a case-control studies, which can introduce bias when using this type of design to estimate VE against any infection. Estimation of VE against emerging variants becomes more challenging as new variants arise with increasing speed: for instance, estimates of VE against infection with the BA.1 sub-lineage were more readily available in the literature than estimates of VE against the BA.2 sub-lineage. With our method, the credible interval associated with the estimate of ψ stops getting smaller despite additional cases because the number of controls stops increasing once the emerging variant takes hold, limiting the possibility for further reduction in the margin of error. In a larger population or with sequencing of a higher proportion of cases, it would be possible to produce a precise estimate more quickly, although based on the results of the simulation, additional cases beyond around 1,700 may not increase the precision by a large amount. Timing of case and control accumulation induces variability among the specific trace plots but does not appear to impact SE as a function of time. Factors such as waning VE over time and differential uptake of vaccines are not addressed in the simulation and may be sources of bias. These are discussed further in the Discussion (Section 6) and the Appendix (Section SA.8).

5. Discussion

We formalize and demonstrate a novel method for VE estimation that can be used for continued surveillance of emerging variants by relying on data that is readily available in most state health departments. We apply this method to estimate VE against any infection with the BA.1 and BA.2 sub-lineages of the Omicron variant. We use a simulation to show, for a scenario consistent with our data and varying case-accumulation patterns, the expected rate of convergence to a VE estimate, along with the expected degree of uncertainty, as a function of time. Our method yields estimates of VE around 35% to 38% against infection with the BA.1 sub-lineage and 47% against infection with the BA.2 sub-lineage for individuals with a complete primary vaccination series. Our estimates are comparable to estimates of VE against BA.1 from larger published studies (Eggink et al. 2022; Gram et al. 2022; Tseng et al. 2022). We could not use our data to produce reliable estimates of VE for a complete primary series plus 1 or more booster doses because there were only a small number of boosted individuals infected with the Delta variant.

A key feature of our method is the use of viral sequence data that allows clear identification of different strains. Sequencing is expensive and those selected for sequencing may not be a random sample of the population; we use reweighting to address this. Some studies attempt to estimate strain-specific VE without sequencing large numbers of records. Tseng et al. (2022) and Eggink et al. (2022) use S-gene target failure to distinguish between the Omicron (BA.1) variant and the Delta variant. Gram et al. (2022) uses infections from the Omicron-dominant time period, rather than sequencing thousands of samples. Identification of S-gene target failure is useful for differentiating between the BA.1 Omicron sub-lineage and the Delta variant, but cannot be used to distinguish between BA.2 and Delta (Rahimi and Talebi Bezmin Abadi 2022), and therefore may not be a feasible technique for future variants and sub-lineages. Further, cohort studies that classify viral subtypes based on a time period during which a specific variant is dominant (Tartof et al. 2022) rely on the assumption that all infections at a given point in time are with a single variant. Appendix Fig. S5 indicates that is not the case for our data.

If variants were defined using a distance metric as in an article by Magaret et al. (2024), we could consider estimating a coefficient associated with the continuous distance metric, so that it could be interpreted as an estimate of the increase in risk of infection relative to a base variant. Under a multiple discretely defined variant scenario, we could also define 1 variant as the reference variant and estimate relative VE for each discretely defined variant relative to the reference, however this may introduce additional limitations under small or moderate sample sizes.

In comparison to other studies estimating VE against the Omicron variant (Eggink et al. 2022; Gram et al. 2022; Tseng et al. 2022) estimating VE, our method produced comparable estimates using only a few hundred cases. This is roughly the size needed for a test-negative case-control

study (World Health Organization 2021), but avoids the introduction of bias from sampling test-negative controls. Our estimate is derived dynamically using routinely collected surveillance data, potentially generating VE information prior to the completion of larger-scale retrospective studies. Our approach is not designed to replace larger nationwide cohort studies such as Gram et al. (2022), which ultimately yield precise estimates and important epidemiologic information such as subgroup effects. Rather, it enables public health officials at a federal, state, or local level to monitor VE using the sequencing information already available to them on a proportion of cases, and to rapidly update VE estimates as cases accrue, making it ideal for surveillance purposes. The precision of our estimates depends on the number of sequenced samples; for the period covered by our analysis, about 10% of new infections were sequenced by Rhode Island Department of Health. Higher proportions could be chosen to target specific levels of precision in the VE estimates.

Vaccine-related recommendations and advice by public health authorities typically rely on multiple indicators and sources of information such as case and hospitalization rates, other surveillance data, evidence from scientific literature, and public health and policy context. Updated VE from our method would be 1 such indicator, but the method is not designed for stand-alone surveillance to trigger decisions based on specific thresholds. Our method also can easily apply to other new variants and can be computed quickly as data accumulates, provided that the method for selecting which samples to sequence remains consistent. From a surveillance standpoint, continued monitoring for potential changes to VE in the context of emerging variants can provide a rapid alert for public health officials, and the use of estimation of VE for surveillance purposes relies on the ability to obtain a quick estimate so that health departments can act in a timely manner. In principle, the dynamic estimate of relative VE could be used to identify waning immunity, especially with additional samples.

Our method relies on several key assumptions, including these: (i) the population is equally susceptible to infection during the time period that each variant or sub-lineage is dominant, (ii) the vaccine effect is durable enough to compare cases and controls at differing lengths of time since vaccination, within the study time-frame, and (iii) that VE, conditional on the covariates, is independent of propensity to get tested. In addition to these assumptions, our estimates of VE against the emerging variant depend on estimates of VE against the previously circulating variant, which can vary substantially by study and are estimated in different populations. This introduces potential issues with transportability and precision of our VE estimates for the emerging variant. Finally, sequencing results may not be available until a period of time after the initial specimen collection; thus, the timeliness of VE estimates from this method in practice will depend on the timeliness of the sequencing process.

Regarding assumption (1), Section 2.4 describes a parameterization that allows differences in susceptibility to infection by subtype due to differential attack rates. Assumption (2) would be violated if vaccine efficacy wanes over time. In Appendix Section SA.8.1, we show that if individuals received their vaccination in a fixed time period, and if VE declines at an equal rate over time regardless of variant, the later sampling of cases from the emerging variant (BA.1 and BA.2 in this case) will lead to downward bias of its VE; ie an under-estimate of the true VE. We also show that if the emergence of a new variant leads to higher rate of vaccination, the VE for the emerging variant will have downward bias. Violations of assumption (2) can also be mitigated by appropriate sampling of the index variant. In our application, we sampled Delta infections in a way that maximized overlap of the infection dates corresponding to Delta and Omicron variants; see Appendix Section SA.2 for more detail. Finally, our simulation study shows the impact of randomly reordering the case ascertainment dates, which is an indirect way to understand the potential impact of waning VE. While the reordered case dates lead to variation in the path of the time-dependent relative VE for the Omicron BA.1 sub-lineage relative to the Delta variant, there does not appear to be a systematic impact of reordering on relative VE inferences at the end of the follow-up period.

Because our sample is drawn from those who had a positive test for SARS-CoV-2 infection, assumption (3) corresponds to a missing-at-random assumption. This assumption is implicit in many study designs used to estimate VE, including the test-negative design. Formulating a sensitivity

analysis is possible in principle, but the complete absence of information about those who are infected but not tested presents practical challenges.

The assumption that VE estimates of the index variant from other published studies is transportable to the target population is difficult to verify without individual-level data. If the published studies report subgroup effects, normalization to the target population is possible by estimating subgroup-specific effects and then averaging over the target population distribution of subgroup variable; see Appendix Table S5 (Appendix Section SA.4). For our analysis, we used the index-variant VE estimate from a single study from the United States where the population mix was similar to that of our sample.

An alternative approach is to use meta-analytic methods to generate an estimate (or, using Bayesian methods, a distribution) of VE against the index Delta variant. This is discussed further in the Appendix (Section SA.7). It would be possible to use the calculation in Section 2.4 and estimates of differential attack rates between variants (Jalali et al. 2022) to produce an adjusted estimate of relative VE dependent on a particular estimate of relative infectivity; we leave this as a possible avenue for future work. In terms of quantifying uncertainty about the VE estimate, particularly related to sampling weights, in practice we would recommend using bootstrap resampling for the daily updates. Owing to the computational demands of carrying out full matching on each bootstrap sample, bootstrap implementation using 100 draws for selected single-day updates each took just under an hour. We did find that for selected days, the bootstrap SEs are slightly higher; however, due to the computational burden of doing this over more than 100 d, our analysis reports robust SEs that treat the weights as known.

Despite the limitations, our dynamic case-control method delivered estimates of VE against the emerging Omicron BA.1 variant that turned out to be consistent with estimates produced by larger studies in similar populations. The VE estimate stabilized after about 2 mo of daily updates, using data from 2,030 index Delta cases (available near the beginning of the surveillance period) and about 1,500 cases of the emerging BA.1 sub-lineage.

Our analysis points to reduced VE against infection with the BA.1 or BA.2 sub-lineages, relative to Delta, and provides methodology that other health departments can apply rapidly to monitor VE against emerging strains of the virus. Our method would provide estimates with smaller error bounds with access to more sequenced samples, such as in a larger state or with additional resources for sequencing of a large number of samples. Nonetheless, our method for producing dynamically updating estimates can be used in practice by departments of health or large health systems to identify potential changes in VE. Although developed using features of the data available specific to the COVID-19 pandemic, it may also be generalized to other infectious diseases that are monitored in a similar manner.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions.

Author contributions

All authors contributed to the development of the evaluation concept. T.M.F., L.C.C., A.B., and J.W.H. contributed to the writing and statistical analysis. All authors contributed to interpretation of the findings and critical revision of the article.

Supplementary material

Supplementary material is available at *Biostatistics Journal* online.

Funding

This work was funded by the Rhode Island Department of Health.

Conflicts of interest

None declared.

Data availability

All analyses were completed using R v4.0.2. Optimal full matching was performed using the MatchIt (Stuart et al. 2011) and optmatch (Hansen 2007) packages. The study was classified as exempt by the RIDOH Institutional Review Board and counts of 1 to 4 were suppressed in output tables and reports in accordance with RIDOH's Small Numbers Policy. Two code files are available on Figshare at https://doi.org/10.6084/m9.figshare.c.7952267.v1. The first of these produces data in a similar format to that used in the analyses, with a working example of how similar estimates can be produced on the synthetic data, and the second provides code for the analyses in the*46pt manuscript.

References

Abu-Raddad LJ, Chemaitelly H, Butt AA, National Study Group for COVID-19 Vaccination 2021. Effectiveness of the BNT162b2 COVID-19 vaccine against the B.1.1.7 and B.1.351 variants. N Engl J Med. 385:187–189. https://doi.org/10.1056/NEJMc2104974

Andrews N et al. 2022. COVID-19 vaccine effectiveness against the Omicron (B.1.1.529) variant. N Engl J Med. 386:1532–1546. https://doi.org/10.1056/NEJMoa2119451

Bruxvoort KJ et al. 2021. Effectiveness of mRNA-1273 against Delta, Mu, and other emerging variants of SARS-CoV-2: test negative case-control study. BMJ. 375:e068848.https://doi.org/10.1136/bmj-2021-068848

Buchan SA, et al. 2022. Estimated Effectiveness of COVID-19 vaccines against Omicron or Delta Symptomatic infection and severe Outcomes. JAMA Netw Open. 5:e2232760. https://doi.org/10.1001/jamanetworkopen.2022.32760.

Centers for Disease Control and Prevention. 2024. COVID data tracker. Department of Health and Human Services, CDC. https://covid.cdc.gov/covid-data-tracker.

Chemaitelly H et al. 2022. Duration of mRNA vaccine protection against SARS-CoV-2 Omicron BA.1 and BA. 2 subvariants in Qatar. Nat Commun. 13:3082.

Collie S, Champion J, Moultrie H, Bekker L-G, Gray G. 2022. Effectiveness of BNT162b2 vaccine against Omicron variant in South Africa. N Engl J Med. 386:494–496. https://doi.org/10.1056/NEJMc2119270

Eggink D et al. 2022. Increased risk of infection with SARS-CoV-2 Omicron BA.1 compared with Delta in vaccinated and previously infected individuals, The Netherlands, 22 November 2021 to 19 January 2022. Euro Surveill. 27:2101196. https://doi.org/10.2807/1560-7917.ES.2022.27.4.2101196

Farrington C. 1993. Estimation of vaccine effectiveness using the screening method. Int J Epidemiol. 22:742–746.

Gram MA et al. 2022. Vaccine effectiveness against SARS-CoV-2 infection or COVID-19 hospitalization with the Alpha, Delta, or Omicron SARS-CoV-2 variant: a nationwide Danish cohort study. PLoS Med. 19:e1003992. https://doi.org/10.1371/journal.pmed.1003992

Hansen BB. 2007. Optmatch: flexible, optimal matching for observational studies. New Funct Multivariate Anal. 7:18–24.

Hansen CH et al. 2021. Vaccine effectiveness against SARS-CoV-2 infection with the Omicron or Delta variants following a two-dose or booster BNT162b2 or mRNA-1273 vaccination series: A Danish cohort study (Version 3). medRxiv. https://doi.org/10.1101/2021.12.20.21267966

Jalali N et al. 2022. Increased household transmission and immune escape of the SARS-CoV-2 Omicron compared to Delta variants. Nat Commun. 13:5706.https://doi.org/10.1038/s41467-022-33233-9

Jester B et al. 2018. Mapping of the us domestic influenza virologic surveillance landscape. Emerg Infect Dis. 24:1300–1306.

Kelly H et al. 2009. Estimation of influenza vaccine effectiveness from routine surveillance data. PLoS One. 4:e5079.
Kirsebom FCM et al. 2022. COVID-19 vaccine effectiveness against the omicron (BA. 2) variant in England. Lancet Infect Dis. 22:931–933.
Lauring AS et al.; Influenza and Other Viruses in the Acutely Ill (IVY) Network. 2022. Clinical severity of, and effectiveness of mRNA vaccines against, COVID-19 from Omicron, Delta, and Alpha SARS-CoV-2 variants in the United States: prospective observational study. BMJ. 376:e069761.
Lin D-Y et al. 2022. Effects of vaccination and previous infection on Omicron infections in children. N Engl J Med. 387:1141–1143. https://doi.org/10.1056/NEJMc2209371
Link-Gelles R et al. 2023. Estimation of COVID-19 mRNA vaccine effectiveness and COVID-19 illness and severity by vaccination status during Omicron BA.4 and BA.5 sublineage periods. JAMA Netw Open. 6:e232598.
Lu L et al. 2022. Neutralization of severe acute respiratory syndrome coronavirus 2 omicron variant by sera from BNT162b2 or CoronaVac vaccine recipients. Clin Infect Dis. 75:e822–e826. https://doi.org/10.1093/cid/ciab1041
Magaret CA et al. 2024. Quantifying how single dose ad26. cov2. s vaccine efficacy depends on spike sequence features. Nat Commun. 15:2175.
Otto SP et al. 2021. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. Curr Biol. 31:R918–R929. https://doi.org/10.1016/j.cub.2021.06.049
Perrine CG, CDC COVID-19 Response Team. 2021. SARS-CoV-2 B.1.1.529 (Omicron) variant – United States, December 1–8, 2021. MMWR Morb Mortal Wkly Rep. 70:1731–1734. https://www.cdc.gov/mmwr/volumes/70/wr/mm7050e1.htm
Puranik A et al. 2021. Comparison of two highly-effective mRNA vaccines for COVID-19 during periods of Alpha and Delta variant prevalence. https://www.medrxiv.org/content/early/2021/08/21/2021.08.06.21261707
Rahimi F, Talebi Bezmin Abadi A. 2022. The Omicron subvariant BA.2: birth of a new challenge during the COVID-19 pandemic. Int J Surg. 99:106261. https://doi.org/10.1016/j.ijsu.2022.106261
Rosenberg ES et al. 2022. COVID-19 vaccine effectiveness in New York State. N Engl J Med. 386:116–127. https://doi.org/10.1056/NEJMoa2116063
Seppälä E et al. 2021. Vaccine effectiveness against infection with the Delta (B.1.617.2) variant, Norway, April to August 2021. Euro Surveill. 26:2100793.https://doi.org/10.2807/1560-7917.ES.2021.26.35.2100793
Šmíd M et al. 2022. Protection by vaccines and previous infection against the Omicron variant of severe acute respiratory syndrome coronavirus 2. J Infect Dis. 226:1385–1390.
Spensley KJ et al. 2022. Comparison of vaccine effectiveness against the Omicron (B. 1.1. 529) variant in hemodialysis patients. Kidney Int Rep. 7:1406–1409.
Stuart EA, King G, Imai K, Ho D. 2011. Matchit: Nonparametric preprocessing for parametric causal inference. J Stat Softw. 42(8):1–28.
Tartof SY et al. 2022. Effectiveness of a third dose of BNT162b2 mRNA COVID-19 vaccine in a large US health system: a retrospective cohort study. Lancet Regional Health Am. 9:100198. https://doi.org/10.1016/j.lana.2022.100198
Tseng HF et al. 2022. Effectiveness of mRNA-1273 against SARS-CoV-2 Omicron and Delta variants. Nat Med. 28:1063–1071. https://doi.org/10.1038/s41591-022-01753-y
van der Laan L, Gilbert PB. 2025. Semiparametric inference for relative heterogeneous vaccine efficacy between strains in observational case-only studies [preprint], arXiv, arXiv:2303.11462.
Wilhelm A et al. 2022. Limited neutralisation of the SARS-CoV-2 Omicron subvariants BA. 1 and BA. 2 by convalescent and vaccine serum and monoclonal antibodies. EBioMedicine. 82:104158.
World Health Organization. 2021. Sample size calculator for evaluation of COVID-19 vaccine effectiveness. https://apps.who.int/iris/rest/bitstreams/1337428/retrieve
Yu J et al. 2022. Neutralization of the SARS-CoV-2 Omicron BA.1 and BA.2 variants. N Engl J Med. 386:1579–1580.

APPENDIX

Included and excluded sub-lineages

The BA.1 sub-lineage, consisting of 2,220 individuals, includes samples classified as BA.1 (1,266 or 57%), BA.1.1 (938 or 42%), B.1.1.529 (11 or 0.5%), BA.1.1.15 (< 5), BA.1.1.18 (< 5), BA.1.14 (< 5), BA.1.15 (< 5), and BA.3 (< 5). The BA.2 sub-lineage, consisting of 1,462 individuals and containing all BA.2 sub-lineages except BA.2.12.1 and BA.2.75, includes samples classified as BA.2 (1,364 or 93%), BA.2.1 (< 5), BA.2.10 (5 or 0.3%), BA.2.12 (14 or 1%), BA.2.18 (< 5), BA.2.26 (< 5), BA.2.3 (33 or 2%), BA.2.3.4 (< 5), BA.2.37 (< 5), BA.2.6 (< 5), BA.2.7 (10 or 0.7%), and BA.2.9 (27 or 2%). Sub-lineage classifications were made by RIDOH.

Appendix Figure 3 displays the distribution of records classified as the Delta variant and sub-lineages of the Omicron variant over time, showing the number of sequenced diagnosed infections classified in the three categories Omicron, Delta, or non-Variant of Concern/Variant of Interest/Variant Being Monitored (Non-VOC/VOI/VBM) by date of the positive PCR test result. As diagnosed infections of some sub-lineages of the Omicron variant increased, these quickly became dominant variants, as was seen in many other geographic areas.

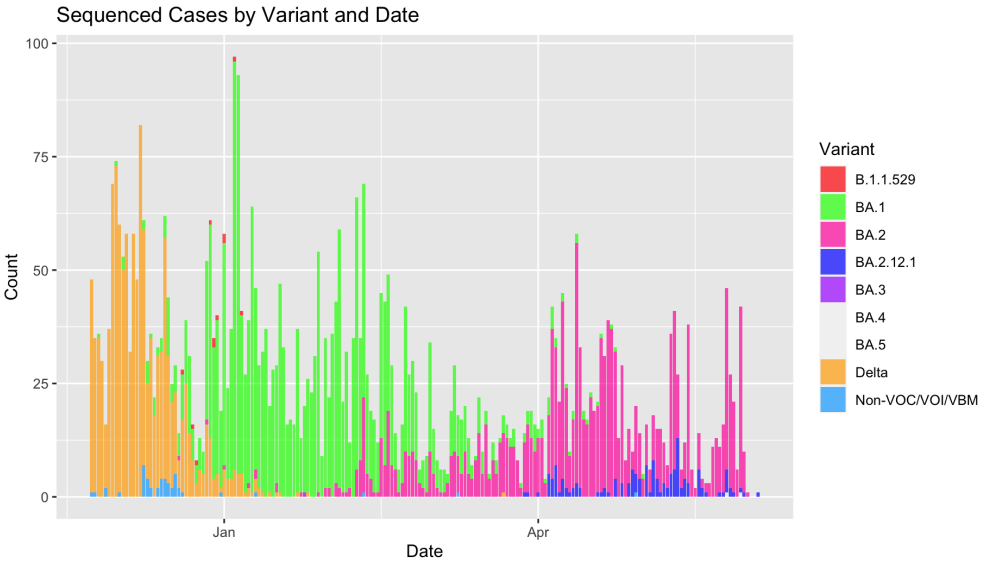


Fig. 3. The horizontal axis spans the time period from November 24, 2021, to June 13, 2022. The vertical axis indicates the count of sequenced diagnosed infections. Non-VOC/VOI/VBM stands for non-Variant of Concern/Variant of Interest/Variant Being Monitored.

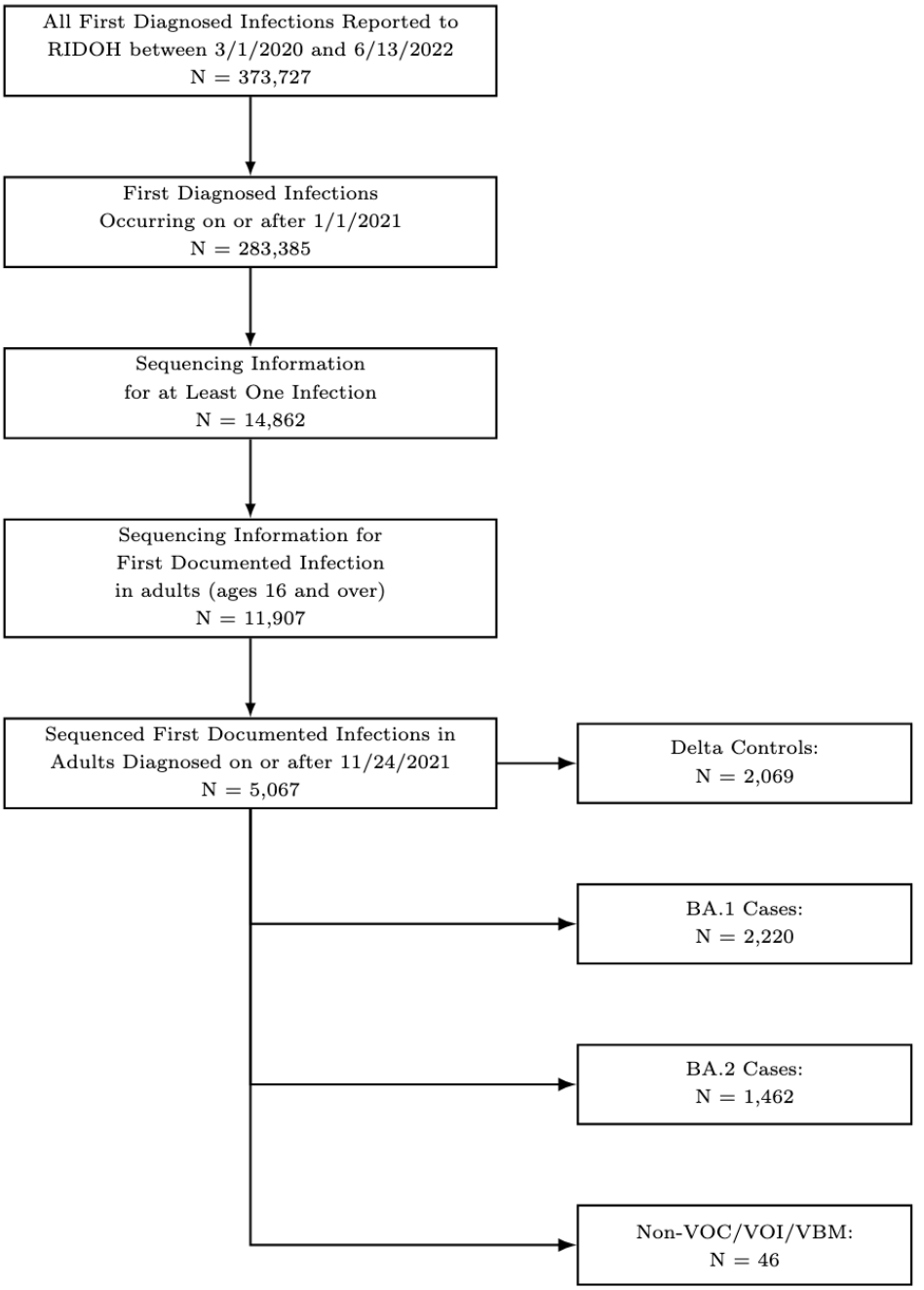


Fig. 4. Flowchart detailing sample sizes following application of inclusion and exclusion criteria.

Use of contemporaneous controls

As mentioned previously, we only used infections with the Delta variant that occurred on or after November 1, 2021, to define a sample with cases and controls occurring within the same range of months. We investigated the effect of using all available infections with the Delta variant on our estimates of relative VE and VE against infection with the Omicron variant. As seen in Appendix Table 4, there is evidence that features of infections with the Delta variant that occurred earlier in 2021 were more dissimilar to features of infections with these sub-lineages of the Omicron variant than later infections with the Delta variant. This effect may be due to waning immunity over time, changes in social distancing and other public health recommendations, or circulation of different sub-lineages of the Delta variant. While it is most-pertinent to use all available information on the emerging variant up to the time of the analysis, using all available information on the previously-circulating variant may increase the confoundedness with time between vaccination and infection and may not improve decision-making using this estimator due to differences in the groups experiencing infections during each phase. These considerations may become less relevant for endemic diseases that fluctuate seasonally as cases and controls may be more similar. The estimates of VE are impacted by (1) the number of index variant cases with sequence data, (2) the number of emerging variant cases with sequence data, and (3) the precision of VE estimates from the literature. The first two items are impacted by the absolute number of cases, capacity and funding for sequencing, and the extent of overlap between when the variants were circulating. However, as seen in the above analysis, considering a wider time window can lead to unbalanced case and control counts and inclusion of more heterogeneity in cases, such as increased differences in time between vaccination and infection, which may increase the impact of waning vaccine effectiveness. That is, there are only so many sequenced cases for each variant, and there are limitations to including cases from very different time frames due to other temporal trends. For the VE estimates to stabilize more quickly, the jurisdiction could either sequence a larger number of cases (which may or may not be possible given capacity and funding constraints) or, if possible based on when the index variant was circulating, consider a wider time window for the index variant (which may introduce more bias temporal trends).

Table 4. The table shows how the estimates of $\psi_v(s', s)$ and $\mathbf{VE}_v(s')$ change depending on how many infections with the Delta variant are included, when including sequentially fewer infections by date. The date column indicates the beginning of the range of included infections with the Delta variant. The OR column shows the odds ratio $(\psi_v(s', s))$ estimated using all infections with the Delta variant that occurred during the relevant time frame and all infections with the BA.1 sub-lineage. The values indicating estimated VE against infection with the BA.1 sub-lineage of the Omicron variant, shown in the $\mathbf{VE}_v(s')$ column, are computed using $\mathbf{VE}_v(s)$ estimates against infection with the Delta variant (64 (60, 67)) from Tseng et al. (2022)

Date	# Delta Infections	# Omicron Infections	OR (95% CI)	$\text{VE}_v(s')$
5/1/2021	5,068	2,220	2.11 (1.85, 2.41)	24 (11, 36)
9/1/2021	3,926	2,220	2.07 (1.82, 2.37)	25 (12, 37)
10/1/2021	2,931	2,220	2.04 (1.77, 2.35)	26 (13, 38)
11/1/2021	2,070	2,220	1.77 (1.52 2.05)	36 (24, 47)
11/24/2021	1,193	2,220	1.73 (1.46, 2.05)	38 (24, 49)

Effect of time since vaccination

As mentioned in section 6, by comparing individuals experiencing infections at differing lengths of time since vaccination, we are making the assumption that the vaccine effect does not wane meaningfully over the study duration. Vaccine durability is a limitation of other studies that investigate VE. Other studies have identified decreasing effectiveness as time since vaccination increases Tseng et al. (2022); Gram et al. (2022); Bruxvoort et al. (2021), so we selected contemporaneous controls (i.e. later Delta cases than those from the beginning of the Delta wave) in an effort to detect decreasing VE under a co-occuring effect of waning VE. In theory, we would be able to produce estimates of VE adjusting for time since vaccination. We tested approaches to adjusting for this by defining a variable for time between vaccination and infection and using this to match cases and controls. However, the utility of each adjustment was affected by the difference in the timing distributions between variants: the timing of vaccination relative to the timing of infections with the Delta variant was significantly longer in general than the timing of vaccination relative to infection with the Omicron variant, so it was difficult to produce a sufficient matching using this approach. We also tested producing stratified odds ratios, stratifying by groups defined by groupings of days since vaccination, but this produced subgroups that were too small and of incomparable size between cases and controls, further exacerbated by the already small state population and sample size.

However, in practice, it is difficult to stratify by length of time since vaccination when trying to produce an estimate quickly. When we partition cases of the BA.1 or BA.2 sub-lineage into categories based on length of time since vaccination, the number of samples becomes too small in some categories, especially in comparison to counts of infections with the Delta variant in the same categories, producing very wide credible intervals. For the BA.2 sub-lineage in particular, those infected with the BA.2 sub-lineage tended to get infected much longer after completing the primary series than those infected with the BA.1 sub-lineage, so we do not have a sufficient number of observations for those who got vaccinated within some time frames to make conclusions about VE.

Vaccine effectiveness differing by age group (transportability)

While we only have summary statistics on selected characteristics for sampled populations from other studies, we further explore our method by computing estimates of relative VE using Rhode Island data for specific population subsets and produce estimates of VE for these groups, motivated by observed differences in VE for those over 60 or 65 years in other studies (Gram et al., 2022; Tartof et al., 2022). We also observed decreased VE for older age groups (Appendix Table 5).

Table 5. Estimates of VE for age groups, relying on estimates of Delta VE for a primary series from Tartof et al. (2022). Unlike the estimates from the same study shown in Table 2, these estimates from Tartof et al. (2022) include adults within anywhere from 7 to 239 days following vaccination.

Age group	Primary Series VE, Delta	Primary Series $\psi_v(s', s)$, BA.1	Primary Series VE, BA.1
16-44 years	73 (71, 74)	1.77 (1.48, 2.13)	52 (42, 60)
45-64 years	73 (71, 74)	2.21 (1.65, 2.96)	40 (20, 56)
$\geq$ 65 years	61 (57, 65)	1.29 (0.78, 2.18)	48 (15, 70)
All $\geq$ 16 years	72 (71, 73)	1.77 (1.52, 2.05)	50 (42, 57)

Odds ratios for re-leveled factor variables

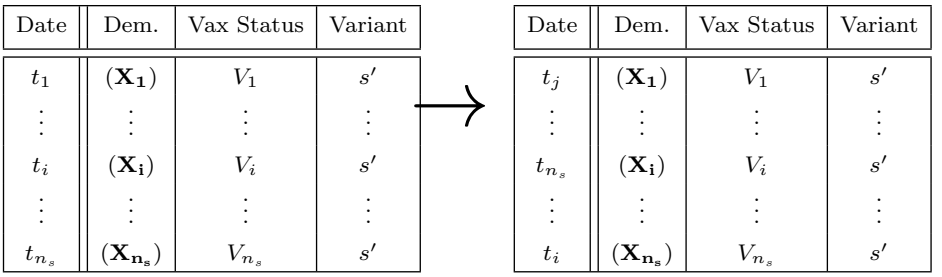
Appendix Table 6 shows the adjusted ORs when using the sex and race categorical variables at their original factor levels in the matching and in the logistic regression. The original levels are shown in Table 1 and the re-leveled categories are described in the footnote to Table 1.

Table 6. Odds Ratios and 95 % confidence intervals for the BA.1-Delta and BA.2-Delta comparisons by vaccination status from weighted logistic regression when adjusting for variables used in the matching, with factor variables consisting of original levels. Estimates of relative VE are displayed in the format of the odds of differing levels of vaccination status given having the Omicron variant (BA.1 or BA.2) compared to the Delta variant, using unvaccinated as the reference category.

Vaccination Status	BA.1 Adj. OR (95% CI)	BA.2 Adj. OR (95% CI)
Unvaccinated	–	–
One Dose of Two-Dose Series	3.87 (2.78, 5.44)	5.24 (3.54, 7.78)
Completed Primary Series	1.77 (1.52, 2.06)	1.16 (.95, 1.40)

Specifications for Simulation Study

The following diagram indicates how each simulated dataset was produced by randomly shuffling case accumulation dates. The set of controls was fixed for the purposes of the simulation.



For each iteration of the simulation, we apply the following process:

- Starting with the set of observed case records from RI with genomic sequencing information, separate into two sets, those with $S = s'$ and those with $S = s$.
- The set of case records with $S = s$ remains constant throughout.
- Within each set with $S = s'$, holding $(\mathbf{X_i}, \mathbf{V_i}, \mathbf{S_i})$ combinations constant, shuffle the case identification date.
- Bind the rows with the $S = s$ records to create a full set of cases ($S = s'$) and controls ($S = s$). For a fixed $N = n_{s'} + n_s$, this constitutes one simulated dataset.
- For each simulated dataset and for every day (date d) (arbitrary time increment within date range of observations):

  1. Gather all cases and controls up to and including those recorded on the date d.
  2. Compute matched sets using propensity scores/logistic regression.
  3. Compute $\psi_v(s', s)$ using weighted logistic regression on the matched dataset.

Figure 2 displays the progression of the estimate for each simulated dataset under randomly-shuffled case-accumulation patterns. The control records remain fixed throughout. Whether an individual was vaccinated by the time of their diagnosed infection is not being shuffled, only the date on which an observed individual observation is recorded is shuffled. Appendix Figure 5 displays the observed accumulation of cases and controls over time and new cases and controls by day.

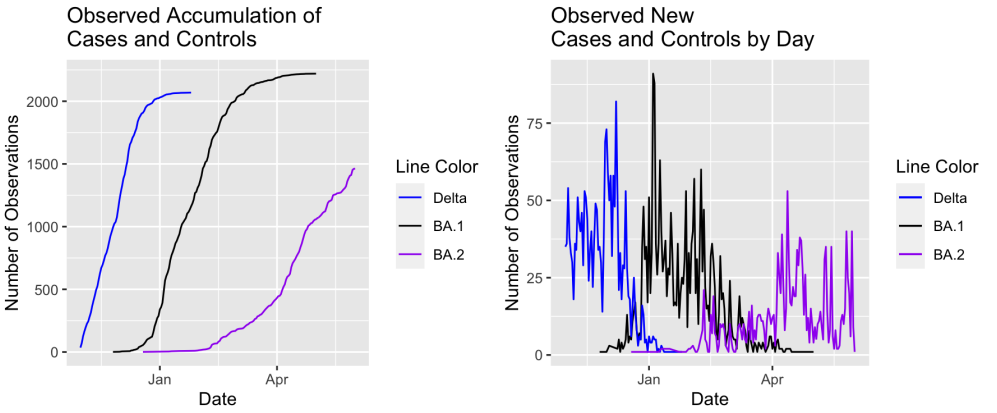


Fig. 5. The left panel shows the observed accumulation of diagnosed infections identified as each variant/sub-lineage. The right shows the number of new diagnosed infections observed each day. In both, line color indicates the sequenced variant or sub-lineage. The horizontal axis spans the time period 12/15/2021 - 6/1/2022.

Possible Extensions and Variations

Given the high variability in the estimates of VE against the index variant, we could consider using a meta-analytic estimate of VE. We can envision two different approaches. First, one could derive a meta-analytic summary of VE against the Delta variant based on a collection of published studies, weighting by the standard error from each study and use this to produce one estimate of VE against the BA.1 and BA.2 sub-lineages. Second, one could produce study-specific estimates of VE against the BA.1 or BA.2 sub-lineages using each Delta VE estimate and the BA.1 or BA.2 relative VE estimate, then combine them into a meta-analytic estimate of VE against BA.1 or BA.2, again weighting by the standard error of the resulting BA.1 or BA.2 VE estimates. Instead, we opted to select the Delta VE estimate where the design was most closely aligned with the setting generating our data from Rhode Island (e.g. the study was conducted in a large health system in the US, thereby using patients subject to similar vaccination eligibility timing, and in a population with more similar average demographics than in other countries).

In practice, however, a proper meta-analysis requires a thorough review of relevant literature and proper method for combining estimates across studies. It may also be necessary to assign weights to different studies, owing to the wide variation in sampling strategies and varying inclusion and exclusion criteria by study (e.g. time-since vaccination, patient age, vaccination eligibility). Additionally, at the time of writing, we did not find meta-analytic estimates of VE against the Delta variant, and given the motivating use of our method, it would be unlikely to have access to such estimates against the previous variant while the emerging variant is active. Due to these concerns, we ultimately would recommend reporting VE estimates as we have done in Table 3, accompanied by a full documentation of the source of the reference VE estimates. We believe that the estimate of VE against the previous variant would arise from literature estimates in comparable municipalities conducting VE studies, although a meta-analytic estimate could be applied for retrospective evaluation of the method.

In another possible extension that would allow us to validate our BA.2 VE estimates, we could ascertain the ratio of VE against symptomatic disease to VE against any infection, then introduce the assumption that the ratio is the same for BA.2 as for BA.1 (or any variant for which the information is available). Then, we could take a literature estimate of BA.2 VE against symptomatic disease and multiply it by the ratio to get a BA.2 VE estimate against any infection for comparison.

We could also consider a Bayesian approach using VE estimates from the previously-circulating variant (Delta) to estimate a distribution describing VE based on existing data and population-level characteristics. These estimates can be combined through a meta-analysis or adjusted for covariates to account for differences across studies, such as population age structure or vaccine type and recency. Then we could estimate a measure of relative VE for an emerging variant similar to that produced above, and combine it with the prior to update the new VE estimate using the prior knowledge about the existing distribution (or use it to assess how different it is from the prior). If demographic information is available, the prior could be tailored to match the study population before combining estimates.

Instead of sampling from the relevant normal distribution, we could use a nonparametric bootstrap approach to estimate the variability of the estimate. In this approach, we would repeatedly sample with replacement from all cases accumulated up to the relevant date, then compute weights reflecting the probability of being sequenced, then use the sequenced cases to compute a full matching for the relevant cases and controls, re-compute the weights, then produce the estimate. In this approach, re-doing the matching for each re-sampled dataset is computationally intensive: when estimating the BA.1 complete primary series relative VE by bootstrapping 100 iterations, we calculated an average estimate of 0.638, with standard error of 0.10 (on the log scale). For comparison, our estimate on the log scale is 0.57 with a standard error of 0.07. It took 51 minutes to compute 100 iterations. We computed the same estimates at two different time points as well: bootstrapping 100 iterations for all data accumulated up to March 1st yielded an estimate of 0.629 with a standard error of 0.108, with an actual estimate of 0.668 and standard error of 0.08 for the same time frame, and bootstrapping 100 iterations for all data accumulated up to February 1st yielded an estimate of 0.734 with a standard error of 0.107, with an actual estimate of 0.752 and standard error of 0.09 for the same time frame. We could have computed bootstrap estimates in this manner each day to produce error estimates for our dynamically-updating estimator, however, this process would have been time-consuming and computationally intensive, especially considering the approximately 130 days that constitutes the case-accumulation period for the BA.1 sub-lineage.

Potential Sources of Bias

We consider two scenarios that could contribute to bias in In the following subsections, we derive expressions for bias in vaccine effectiveness (VE) estimates obtained from a case-control design under two distinct time-related confounding mechanisms. First, we explore how bias arises due to waning vaccine-induced immunity when two variants circulate at different times and infections with the later-circulating variant occur at longer average times since vaccination. Then, we examine bias that results from a population-level increase in vaccination probability at a specific calendar time, which differentially affects the vaccination rates among cases (infected with the later-circulating variant) and controls (infected earlier). In both scenarios, the temporal mismatch between vaccination and infection windows leads to systematic bias in the estimated VE from the dynamic case-control sampling method that we apply. Both scenarios also indicate ways that we could set up simulations to further justify how these sources of these confounding may lead to bias.

Confounding Introduced by Waning Vaccine Efficacy Over Time

We could consider a simulated scenario to evaluate VE against an emerging variant under a confounding mechanism driven by calendar time (e.g 180 days). In this setting, individuals are vaccinated uniformly early in the analysis period (e.g., uniformly in days 0–60), and two variants arise at different times: we refer to the previously-circulating variant as Variant D and assume that infections occur uniformly from day 0 to 180, while infections with the emerging variant, Variant B, occur uniformly from day 120 to 300. Although the true vaccine efficacy (VE) is the same for both variants and governed by a waning function, the case-control comparison can introduce bias because of differences in time since vaccination at the time of infection.

In this scenario, vaccine efficacy is assumed to wane as a function of time since vaccination and is assumed to wane at the same rate for each variant. Specifically, we assume that the probability of infection for a vaccinated individual decreases exponentially with time since vaccination, parameterized by a waning rate $\lambda > 0$.

We model waning protection through a multiplicative hazard model: among vaccinated individuals, the probability of infection is reduced by a factor of $\exp(-\lambda\Delta)$, where Δ denotes time since vaccination and $\lambda > 0$ controls the strength of waning.

The vaccine effectiveness for each variant, conditional on time since vaccination, is given by $\mathrm{VE}_v = 1 - \exp(-\lambda\Delta_v)$, where Δ_v is the average time since vaccination for infections with variant $v \in \{D, B\}$. Due to the staggered circulation times of the two variants and the fixed vaccination schedule, the average time since vaccination is systematically greater for B infections ($\Delta_B > \Delta_D$). Thus, even though the waning model is the same across both variants, the distribution of Δ differs for cases (variant B) and controls (variant D), leading to differential attenuation of protection.

The estimated odds ratio from the case-control design is approximately $\widehat{\mathrm{OR}} = \exp\{-\lambda(\Delta_B - \Delta_D)\}$, implying an estimated VE of $\widehat{\mathrm{VE}} = 1 - \exp\{-\lambda(\Delta_B - \Delta_D)\}$. The bias in the estimated VE, relative to a baseline where $\Delta_B = \Delta_D$, is therefore $\mathrm{Bias} = \widehat{\mathrm{VE}} - \mathrm{VE}_{\mathrm{true}} = 1 - \exp\{-\lambda(\Delta_B - \Delta_D)\} - (1 - \exp\{-\lambda\Delta\})$. Since $\Delta_B > \Delta_D$, this results in underestimation of the true VE: the longer average time since vaccination among cases artificially lowers the observed VE, introducing a negative bias.

Bias from Time-Varying Vaccination Probability

Next, we consider a scenario where vaccine efficacy is constant over time, but the probability of vaccination increases at a known point in calendar time. One instance under which this scenario is plausible is if individuals become more likely to get vaccinated upon the emergence of a more-infectious variant. We assume that individuals are at risk of infection with variant D during days 0–180 and variant B during days 120–300. A population-level increase in vaccination occurs at day 120 and can be modeled as a step function: the probability of vaccination is π_{early} before day 120 and $\pi_{\mathrm{late}} > \pi_{\mathrm{early}}$ on or after day 120. This change induces time-varying confounding: because infection with variant B occurs later in time, individuals infected with B are more likely to be vaccinated than those infected with D.

Even though the true VE is fixed (i.e., $\mathrm{VE}_{\mathrm{true}} = 1 - \exp(-\beta)$), the case-control estimate is distorted by temporal trends in vaccine uptake. Specifically, cases are individuals infected with variant B (uniformly in days 120–300), and controls are individuals infected with variant D (uniformly in days 0–180). The expected probability of vaccination among cases is π_{late}, while the expected vaccination probability among controls is a weighted average:

$$P(Z = 1 \mid \text{control}) = \frac{1}{180}\left(120 \cdot \pi_{\mathrm{early}} + 60 \cdot \pi_{\mathrm{late}}\right).$$

Thus, the odds of vaccination among cases is inflated relative to controls, even though vaccination is equally protective across all individuals.

The estimated odds ratio becomes:

$$\widehat{\mathrm{OR}} = \frac{\pi_{\mathrm{late}}(1-B)}{B(1-\pi_{\mathrm{late}})},$$

where B is the average probability of vaccination among controls. Consequently, the estimated vaccine effectiveness is:

$$\widehat{\mathrm{VE}} = 1 - \frac{\pi_{\mathrm{late}}(1-B)}{B(1-\pi_{\mathrm{late}})}.$$

This expression is generally smaller than the true VE, and the difference increases as the gap between π_{late} and π_{early} widens. Thus, the case-control design yields a negatively biased VE estimate due to the correlation between vaccination probability and calendar time, which in turn aligns with differential susceptibility to variant B versus variant D.