nature portfolio

Peer Review File

Impact of unequal testing on vaccine effectiveness estimates across two study designs: a simulation study

Corresponding Author: Dr Korryn Bodner

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
During the height of the COVID-19 pandemic when mass testing was available, it was found that vaccinated individuals were more likely to test than unvaccinated individuals. This well written paper explores the impact of this on estimates of vaccine effectiveness, using the two most common designs for VE studies: the cohort and test negative designs.

Overall, the study is well thought out and explained. It offers some useful insights into the biases of each design that go beyond the original question of unequal testing. Although the content is technical, the authors have managed to write the paper in a way that should be accessible to epidemiologists working in the area of vaccine effectiveness. The methods are explained well and code is available for reproducibility. The figures are clear and informative. This is a good paper.

I have a couple of concerns about the test negative methods (lines 459 to 462) and corresponding results (lines 171 to 179). If I understand the methods correctly, a single individual can contribute multiple outcomes in the test negative design if they have an etiology of COVID-like symptoms before a SARS-CoV-2 infection. Whereas negative tests associated with an etiology of COVID-like symptoms are removed if they occur after SARS-CoV-2 infection (lines 461-2). I believe that it could be this imbalance (negative then positive is possible, but positive then negative is not) that induces the bias referred to as “misclassification”. In practice, there is no issue with including multiple tests per person in a test negative design study, so long as each symptomatic episode included is independent. The simulation study clearly involved a huge amount of work and I would not suggest a re-run, but I would like to see some discussion around the inclusion of negative controls in the study. Secondly, I object to labelling the bias “misclassification” because the negative outcomes are true negatives at the given point in time (in this sense they are not misclassified), and are independent of the unrecorded SARS-CoV-2 infection that occurred earlier. Since outcome misclassification in a test negative design would typically refer to false positive and false negative tests at the time they were given (your simulated tests had only true outcomes at the time given), this could be confusing for readers. I ask the authors to please find a new name for this bias.

I have a small number of additional very minor comments:
1) Line 359. Note the statement here is true when the reasons for unequal testing are unmeasured.
2) Line 430. To improve clarity, please explain what is meant by vaccine efficacy against infectiousness.
3) Lines 436 and 437. Brackets need closing.
4) Lines 445 and 456. Should VE be written as a function of time? i.e. VE(t)
5) Line 805. Check spelling: values.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors use an agent-based model to evaluate the bias introduced by unequal testing in a retrospective cohort and test-negative design. They found biases to be most prominent in the retrospective cohort design. The authors evaluated the extent of bias under different efficacies against susceptibility sampling periods, efficacy against infectiousness, and

probability of transmission. Many of my comments surround model assumptions and implementation.

The authors assume the sensitivity and specificity of these tests are perfect. The time the test is conducted during the infectious period is critical, as sensitivity will not be constant. Although the average time in testing is moderately different between vaccinated and unvaccinated individuals (~0.5 days), the temporal sensitivity begins to decline rapidly after symptom onset (https://doi.org/10.1038/s43856-022-00147-y). With unvaccinated individuals vaccinating later in the unequal testing scenario, the likelihood of a false negative increase relative to vaccinated individuals. Also, since the authors are exploring COVID-like symptoms, the specificity can influence vaccine efficacy with false positives. However, I expect this assumption around perfect specificity not to affect the results.

The efficacy is evaluated at the start of the epidemic, cumulatively throughout all stages. The stage/duration in which the effectiveness is evaluated would also impact these estimates. For example, during the growth phase of an epidemic, many individuals will be in the early stages of symptom onset. During the declining phase, more infected individuals would be in the later stages of infection. The test's sensitivity is much higher during the early stages of the symptomatic phase compared to testing in the late stages of infectiousness. Thus, the efficacy estimates could differ if only evaluated during the growth phase, peak phase, or only during the decline phase of the epidemic. There is a paragraph about the biases introduced and proposed solutions in the context of an emerging and evolving outbreak. However, it is unclear how estimates would differ when the VE study is initiated at different stages of the epidemic.

Efficacy against susceptibility was implemented as "all-or-nothing" protection, providing 100% protection in susceptibility for a subset of people. This assumption implies that protected individuals will never be infected throughout the simulation. Implementing a leaky vaccine would lead to eventual infection of these individuals, likely later in the simulation periods, after experiencing multiple exposures. Thus, this scenario could produce an efficacy against susceptibility that drastically differs from that in the paper (as I expect to decrease with the duration of the simulation). I would suggest exploring the effects of this assumption, as past research has highlighted the discrepancies between these two approaches (https://doi.org/10.1038/srep15468).

For clarity, the authors should highlight the differences in vaccine uptake between high and low healthcare access. I found it buried in the supplement when they could express it in Table 1.

Could the authors explain how the value of the proportion with COVID-like symptoms from alternate etiologies affect their results?

In the discussion, the authors discuss the potential differences in the likelihood of asymptomatic infection between vaccinated and unvaccinated individuals. Could the authors elaborate on the possible influence of a heterogeneous infectious period between vaccinated and unvaccinated individuals?

Table S1. The authors should clarify that this time is relative to the time of symptom onset.

In the discussion, it would be interesting to hear the author's views on the biases of unequal testing in evaluating natural immunity.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors compared the TND and cohort designs using a realistic scenario of unequal testing between the two groups (vaccinated and unvaccinated). The paper provides good insights about the properties of both designs. However, there is one major point that should be addressed.

1) It is not clear how time was took into account in the analysis of the data generated from the simulation (which does not appear to take time into account). In any study involving vaccine effectiveness, it is necessary to control for time (e.g. comparing positive and negative tests during the same week). This is due to the role of a variety of time-related factors, particularly depletion bias. It is necessary to control for time in simulation studies, just as in real world data (see: https://academic.oup.com/aje/article/191/5/800/6515669). Figure 3 appears to have the values for each time point. However, in the supplementary figures 7/8, these appear to show the cumulative number of individuals at the end of each follow-up period. This is not appropriate – even RCTs use incidence rate ratios (or hazard ratios) and not risk ratios to assess efficacy (https://www.nejm.org/doi/full/10.1056/NEJMoa2110345). Also, in the supplementary figures 4/5/6, it is not clear how the cumulative proportion of individuals can decrease over time.

This point is directly related to the statement in lines 266-269: “In contrast, biases produced by the test-negative design were more likely than biases produced by the cohort design to be influenced by the length of the sampling period, the efficacy against infectiousness, and the probability of transmission.” It is likely that a longer sampling period contributes to these biases. In this scenario, the cohort should be analysed using a Pox or Poisson model accounting for person-time at risk (VE_HR or VE_Rate ratio). The corresponding test negative design analysis will compare individuals tested (+ve or -ve) in the same week week (risk set sampling)

In lines 88 to 95, discussing the selection bias induced in the TND, this depends on the criterion used for assessing if the causal effect is identifiable, please check the work of Shrier et. Al. (https://academic.oup.com/ije/article/52/6/1968/7224569?searchresult=1). This type of problem would affect the estimation of risk ratio and not the odds ratio.

2) The paper of Suah et. al. also compares the test negative and cohort design using a variety of scenarios (including difference in testing). It would be good addendum for the discussion section to compare the findings of the current paper to that one.
https://www.medrxiv.org/content/10.1101/2022.08.25.22279235v1.full.pdf

Minor points:
1) The terms “susceptibility”, and “effectiveness against infectiousness” presumably refer to “effectiveness against progression to symptoms”. If so, it would be better to use terminology similar to Kahn et. al.

2) There are some pieces of discussion in the results section: lines150-155.
3) It is more likely that a vaccine will be “leaky”, rather than all or nothing (https://www.nature.com/articles/s41467-023-40750-8)
4) An additional limitation was not accounting for re-infections, which is a common scenario in any respiratory disease.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The revisions have helped to improve clarity of the manuscript. I have no further comments.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors replied satisfactorily to most of the previous points.

However, the issue regarding the calculation of OR from the TND remains unclear. In the manuscript, the cumulative proportion of individuals who tested positive for covid-19 is used in relation to vaccinated/unvaccinated individuals and other aetiologies (vaccinated/unvaccinated). However, as shown in the supplementary figures regarding the proportion of individuals testing positive for other aetiologies (which can be reduced once individuals who test positive for COVID-19 are removed from this pool), this condition would create a scenario of the numbers (tests in previous time) used for controls will be “removed' over time. For example:

OR time 100:

VAC+= 250/ VAC - = 400 / UNVAC+=300/UNVAC-=425

20 new positive tests in vac /unvac. Among them, 5 were individuals who tested positive for other aetiologies

OR time 120

VAC + = 270/ VAC- = 395 / UNVAC+ 320 / UNVAC- = 420

This approach would remove individuals who tested negative (positive for other diseases) previously and tested positive at the current time. The test-negative approach works by using the tests as a time-fixed point. However, this approach should retain the negative test of a person prior to the positive test.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
The changes to the manuscript and the response from the authors are both appropriate, and I have no further comments or suggestions.

(Remarks on code availability)

RESPONSE TO REVIEWERS

Reviewer #1 (Remarks to the Author)

R1: During the height of the COVID-19 pandemic when mass testing was available, it was found that vaccinated individuals were more likely to test than unvaccinated individuals. This well written paper explores the impact of this on estimates of vaccine effectiveness, using the two most common designs for VE studies: the cohort and test negative designs.

Overall, the study is well thought out and explained. It offers some useful insights into the biases of each design that go beyond the original question of unequal testing. Although the content is technical, the authors have managed to write the paper in a way that should be accessible to epidemiologists working in the area of vaccine effectiveness. The methods are explained well and code is available for reproducibility. The figures are clear and informative. This is a good paper.

RESPONSE: We thank Reviewer 1 for their positive feedback and the comments which we feel have helped strengthen the manuscript.

R1: I have a couple of concerns about the test negative methods (lines 459 to 462) and corresponding results (lines 171 to 179). If I understand the methods correctly, a single individual can contribute multiple outcomes in the test negative design if they have an etiology of COVID-like symptoms before a SARS-CoV-2 infection. Whereas negative tests associated with an etiology of COVID-like symptoms are removed if they occur after SARS-CoV-2 infection (lines 461-2). I believe that it could be this imbalance (negative then positive is possible, but positive then negative is not) that induces the bias referred to as “misclassification”. In practice, there is no issue with including multiple tests per person in a test negative design study, so long as each symptomatic episode included is independent. The simulation study clearly involved a huge amount of work and I would not suggest a re-run, but I would like to see some discussion around the inclusion of negative controls in the study.

RESPONSE: As noted, individuals could test multiple times over the sampling period and could have both positive and negative test results. However, for any given point in the sampling period, which represents the point when a VE study is conducted, individuals contribute only once (as a case or a control) to the VE estimate. In our study, for those who had tested positive during the sampling period, it is the first positive test, or if the individuals never tested positive, only the first negative test. This selection of tests is often adopted by healthcare settings, administrative records and public health surveillance systems (Suah et al. 2022). Upon re-reviewing our methods, we realized this information was missing. We have now added how we handled multiple tests as described above (Lines 529-531; Lines 542-545) and have included Suah et al. 's (2022) description for when this design is commonly used (Lines 545-547).

The imbalance (i.e. individuals moving from a negative outcome to a positive outcome but not the reverse) exists irrespective of testing differences. When calculating VE for symptomatic SARS-CoV-2 infection, we compare cases (all those who have been infected in

a given sampling period) with controls (those who have never been infected). Over the course of an epidemic, the pool of individuals who have never been infected (or tested positive) naturally shrinks, reducing the number of potential controls. When a vaccine is efficacious and provides all-or-nothing protection, a subset of vaccinated individuals remains fully protected throughout the epidemic and thus cannot move from the negative outcome to a positive outcome. If the epidemic is large, this process can result in controls that no longer accurately reflect the population's vaccination distribution later in the epidemic. This is why, when testing is equal, we can observe an overestimate of VE at later epidemic stages (although note this effect is very small when efficacy against susceptibility is low as there are very few vaccinated individuals who are fully protected). However, the existence of the imbalance and its differences by vaccination alone, do not explain why VE becomes underestimated when testing is unequal or why the magnitude of bias due to unequal testing grows significantly larger over time when efficacy against susceptibility is lower.

Recognizing the consequences of this different imbalance by vaccination status is important to understanding why the VE estimates from the test-negative design can change over time - particularly in the equal testing scenario. To address this, we have now added to the Discussion a description of this process and its effect on VE estimates in the equal testing scenario (Lines 333-339):

> "Even with equal testing, some continued attention may still be warranted as VE estimates can become overestimated when the epidemic is large and the sampling period includes later epidemic stages. This overestimation occurs because, as the epidemic progresses, the pool of individuals who have never been infected with symptomatic SARS-CoV-2 shrinks - especially among the unvaccinated - leading to a control group that no longer accurately reflects the population's vaccination distribution."

R1: Secondly, I object to labelling the bias "misclassification" because the negative outcomes are true negatives at the given point in time (in this sense they are not misclassified), and are independent of the unrecorded SARS-CoV-2 infection that occurred earlier. Since outcome misclassification in a test negative design would typically refer to false positive and false negative tests at the time they were given (your simulated tests had only true outcomes at the time given), this could be confusing for readers. I ask the authors to please find a new name for this bias.

RESPONSE: We understand that this misclassification terminology could cause confusion with misclassification due to testing sensitivity. However, we propose that it is important to make clear that individuals do become misclassified in the equation as individuals who should be classified as "having had a SARS-CoV-2 infection" are instead classified as "never having had a SARS-CoV-2 infection". To address the potential confusion while still acknowledging that misclassification is occurring, we have done the following: we 1) changed the "Misclassification" heading in the Supplementary Material to be "Missed Detection Misclassification Example"; 2) removed the term "false negative" from Supplementary Figure 3 as the term is closely linked to testing sensitivity; 3) updated the text in the Results section and Supplementary Material so that we reference missed detections when referring to misclassification; and 4) added a more thorough explanation of how this misclassification due to missed detection works in the Results and Supplementary Material.

For example, in Supplementary 1: Missed Detection Misclassification Example, we have now expanded the explanation as follows (Lines 58-62):

> "individuals can be misclassified in the test-negative design when they have or have had a symptomatic SARS-CoV-2 infection but have been classified as "negative". This occurs when individuals have had both a symptomatic SARS-CoV-2 infection and COVID-like symptoms from other etiologies during the same sampling period but were only tested when they had COVID-like symptoms".

R1: I have a small number of additional very minor comments:

R1: 1) Line 359. Note the statement here is true when the reasons for unequal testing are unmeasured.

RESPONSE: We agree and have now added "when the mechanisms driving unequal testing are unmeasured" to the end of that sentence (Line 429)

R1: 2) Line 430. To improve clarity, please explain what is meant by vaccine efficacy against infectiousness.

RESPONSE: To improve clarity, we have now explicitly outlined the mechanism of vaccine efficacy against infectiousness in the Methods section as requested. Specifically, we added: "Efficacy against infectiousness was implemented as a reduced probability of SARS-CoV-2 transmission for vaccinated individuals infected with the virus, given contact with a susceptible individual" (Lines 509-511)

R1: 3) Lines 436 and 437. Brackets need closing.

RESPONSE: Done

R1: 4) Lines 445 and 456. Should VE be written as a function of time? i.e. VE(t)

RESPONSE: Yes, this is correct and the formulas have been revised accordingly.

R1: 5) Line 805. Check spelling: values.

RESPONSE: Fixed, thank you.

Reviewer #2 (Remarks to the Author)

R2: The authors use an agent-based model to evaluate the bias introduced by unequal testing in a retrospective cohort and test-negative design. They found biases to be most prominent in the retrospective cohort design. The authors evaluated the extent of bias under different efficacies against susceptibility sampling periods, efficacy against infectiousness, and probability of transmission. Many of my comments surround model assumptions and implementation.

RESPONSE: We thank Reviewer 2 for their feedback. In response to their comments, we have added clarifications to the text as well as run additional analyses.

R2: The authors assume the sensitivity and specificity of these tests are perfect. The time the test is conducted during the infectious period is critical, as sensitivity will not be constant. Although the average time in testing is moderately different between vaccinated and unvaccinated individuals (~0.5 days), the temporal sensitivity begins to decline rapidly after symptom onset (https://doi.org/10.1038/s43856-022-00147-y). With unvaccinated individuals vaccinating later in the unequal testing scenario, the likelihood of a false negative increase relative to vaccinated individuals. Also, since the authors are exploring COVID-like symptoms, the specificity can influence vaccine efficacy with false positives. However, I expect this assumption around perfect specificity not to affect the results.

RESPONSE: We appreciate Reviewer 2's insight on the importance of testing sensitivity, its variability across the infection period, and its potential implications for our VE estimates. Regarding the difference in the average time of testing between vaccinated and unvaccinated individuals, as noted, this difference by vaccination status is only about half a day. In Figure 1 of Wells et al. (2022), a half day difference translates to about a 5% or less difference in testing sensitivity by vaccination status, which we do not expect would have a substantial impact on our results.

Regarding testing specificity, SARS-CoV-2 diagnostic PCR tests have consistently demonstrated very high specificity (> 97%; Kostoulas, Eusebi and Hartnack 2021), which also reduces our concerns as well that false positives are likely to meaningfully bias VE estimates for SARS-CoV-2.

R2: The efficacy is evaluated at the start of the epidemic, cumulatively throughout all stages. The stage/duration in which the effectiveness is evaluated would also impact these estimates. For example, during the growth phase of an epidemic, many individuals will be in the early stages of symptom onset. During the declining phase, more infected individuals would be in the later stages of infection. The test's sensitivity is much higher during the early stages of the symptomatic phase compared to testing in the late stages of infectiousness. Thus, the efficacy estimates could differ if only evaluated during the growth phase, peak phase, or only during the decline phase of the epidemic. There is a paragraph about the biases introduced and proposed solutions in the context of an emerging and evolving outbreak. However, it is unclear how estimates would differ when the VE study is initiated at different stages of the epidemic.

RESPONSE: Regarding the potential changing impact of testing sensitivity on the VE estimates over the course of the epidemic, we distinguish between an individual's phase of infection and the epidemic phase. In the simulations, we assumed that individual testing probabilities remained the same over the course of an epidemic. This means that, regardless of the epidemic stage, the timing of testing relative to symptom onset for each individual was the same. Consequently, the average testing sensitivity is not changing over the course of the epidemic.

However, although the average testing sensitivity remains constant over time, we wondered if the effects of testing sensitivity could change over the course of an epidemic. This arose

as a study by Endo et al. (2020) found that the degree of bias from testing sensitivity can be influenced by the case ratio (i.e. the ratio between the true disease cases and controls), which can change over time. To conduct this analysis, we varied testing sensitivity (from 0.5 - 1) and measured its effect on the magnitude of bias at two key stages of the epidemic: 1) the inflection point, when the epidemic experienced its highest epidemic growth point; and 2) the epidemic peak. Our main findings were: 1) varying testing sensitivity had no effect on the magnitude of bias due to unequal testing in the cohort design, 2) it had a minor effect in the test-negative design with the size of this effect increasing slightly as the epidemic progressed; and 3) that the direction of the effect of testing sensitivity depended on vaccine efficacy. Despite these findings of some effect for the test-negative design, the overall influence of testing sensitivity on the magnitude of bias due to unequal testing remained small (less than 0.05).

We have now added the methodology, the findings, and an explanation for why the effects of testing sensitivity changed across levels of efficacy to the manuscript. These additions resulted in a new paragraph in the Results section (Lines 189-202), three supplementary figures (Supplementary Fig. 2, 5-6), new sentences to our Methods section (Lines 591-597), and a new supplementary section (Supplementary 1: Effects of Testing Sensitivity)

R2: Efficacy against susceptibility was implemented as "all-or-nothing" protection, providing 100% protection in susceptibility for a subset of people. This assumption implies that protected individuals will never be infected throughout the simulation. Implementing a leaky vaccine would lead to eventual infection of these individuals, likely later in the simulation periods, after experiencing multiple exposures. Thus, this scenario could produce an efficacy against susceptibility that drastically differs from that in the paper (as I expect to decrease with the duration of the simulation). I would suggest exploring the effects of this assumption, as past research has highlighted the discrepancies between these two approaches (https://doi.org/10.1038/srep15468).

RESPONSE: We acknowledge that the assumption of “all-or-nothing” protection in the model simplifies the dynamics of vaccine-induced protection. As you noted, implementing “leaky” protection, where all vaccinated individuals receive partial protection, would likely lead to eventual infection of these individuals, particularly over longer periods of time. The assumption of “all-or-nothing” protection was chosen to provide a baseline for understanding how unequal testing influences bias in VE estimates without introducing additional complexities of the differential depletion by vaccination status associated with the leaky vaccine.

We recognize this is an important limitation of our study and so we have explicitly noted it in the Discussion (Lines 416-424). Although we did not run additional analyses assuming a leaky vaccine type, existing studies (Lewnard et al. 2018; Kahn et al. 2022) provide insights that allow us to infer how assuming leaky-type protection may affect the magnitude of bias due to unequal testing over time. On Lines 419-424, we have now added these expectations along with our reasoning behind them:

> “Had we assumed “leaky” protection and not excluded those with past infection[57], the increasing magnitude of bias in our symptomatic VE estimates with unequal testing would likely have increased further. This is because even

> with equal testing, “leaky” protection introduces downward bias in VE estimates over the course of an epidemic, as more uninfected and unvaccinated individuals who are not at risk of infection are included in the analysis[25,57].

R2: For clarity, the authors should highlight the differences in vaccine uptake between high and low healthcare access. I found it buried in the supplement when they could express it in Table 1.

RESPONSE: We have added the differences in vaccine uptake between high and low healthcare access for each testing scenario in Table 1. Additionally, we also included the testing probabilities for those with low and high healthcare access so that all important values needed to recreate the testing scenarios are now available in Table 1.

R2: Could the authors explain how the value of the proportion with COVID-like symptoms from alternate etiologies affect their results?

RESPONSE: We have now added an explanation about how changes in the prevalence of COVID-like symptoms from alternate etiologies can affect the magnitude of bias. Specifically, in the Results section, we have added starting on Line 206:

> This increase in the magnitude of bias occurred because, as the prevalence of COVID-like symptoms increased, unvaccinated individuals experienced a higher proportional increase in the number of individuals being tested and classified as “negative” than vaccinated individuals. The reason for this higher proportional increase is that, as the prevalence of COVID-like symptoms from alternate etiologies increased, those unvaccinated individuals who were unlikely to be tested gained more opportunities to test and be classified as “negative”.

The details and calculations supporting this statement can be found in a newly created Supplementary material section (Supplementary 1:Proportional Increase Calculations: Vaccination Status and COVID-like Symptoms due to Alternate Etiologies).

R2: In the discussion, the authors discuss the potential differences in the likelihood of asymptomatic infection between vaccinated and unvaccinated individuals. Could the authors elaborate on the possible influence of a heterogeneous infectious period between vaccinated and unvaccinated individuals?

RESPONSE: As we assumed that vaccinated and unvaccinated individuals are randomly mixing, we do not believe a heterogeneous infectious period by vaccination status alone should substantially affect our results as any increase in the infectious period for one group would increase transmission to both vaccinated and unvaccinated individuals. However, if a longer infectious period also coincided with a longer duration of symptoms, then we can easily see how this heterogeneity could have a significant effect on the size of the bias. This is because heterogeneity in symptom duration by vaccination status could alter the size of the testing window by vaccination status. As studies have shown that vaccination can reduce symptom duration (e.g. Mohr et al. 2023, BMJ Open), we recognize this as an important limitation we had previously missed. We have now included this limitation in the Discussion

in the paragraph about potential consequences of heterogeneity in symptom severity by vaccination status (Lines 403-408):

> "Vaccination for SARS-CoV-2 has been shown to reduce both severity of symptoms and symptom duration [52–54], which could lead to vaccinated individuals testing less frequently when symptomatic or having fewer days to test. This could, in turn, reduce the magnitude of bias due to unequal testing. However, the magnitude of bias could remain unchanged if vaccinated individuals were also more likely than unvaccinated individuals to get tested with milder symptoms and to test earlier."

R2: Table S1. The authors should clarify that this time is relative to the time of symptom onset.

RESPONSE: We have updated Table S1 heading to be "Average Day of First Test Relative to Symptom Onset". We have also clarified this in Table S1's caption and in the paragraph on Timing of Testing in the supplementary material.

R2: In the discussion, it would be interesting to hear the author's views on the biases of unequal testing in evaluating natural immunity.

RESPONSE: We have now added a paragraph to the Discussion that expands on how unequal testing could similarly influence estimates of other protective effects, such as natural immunity (Lines 374-383).Specifically, we note (Lines 376-380):

> "Had we replaced vaccination with prior infection in our study and instead estimated the protection of natural immunity, we would expect that unequal testing by prior infection would lead to biased estimates. This expectation aligns with other studies that have noted the potential for such unequal testing to compromise protective effect estimates[49,50].

In this same paragraph, we also acknowledged that "Testing differences by prior infection status have been observed[48]" (Line 376), referencing Freedman et al 2024, further underscoring the broader utility of our findings for studies assessing the protective effects of natural immunity.

Reviewer #3 (Remarks to the Author)

R3: The authors compared the TND and cohort designs using a realistic scenario of unequal testing between the two groups (vaccinated and unvaccinated). The paper provides good insights about the properties of both designs. However, there is one major point that should be addressed.

RESPONSE: We thank Reviewer 3 for their in-depth review and their recognition of the value of this manuscript. We hope that our revisions adequately address both the major and minor points they raised.

R3: 1) It is not clear how time was took into account in the analysis of the data generated from the simulation (which does not appear to take time into account). In any study involving vaccine effectiveness, it is necessary to control for time (e.g. comparing positive and negative tests during the same week). This is due to the role of a variety of time-related factors, particularly depletion bias. It is necessary to control for time in simulation studies, just as in real world data (see: https://academic.oup.com/aje/article/191/5/800/6515669). Figure 3 appears to have the values for each time point. However, in the supplementary figures 7/8, these appear to show the cumulative number of individuals at the end of each follow-up period. This is not appropriate – even RCTs use incidence rate ratios (or hazard ratios) and not risk ratios to assess efficacy (https://www.nejm.org/doi/full/10.1056/NEJMoa2110345). Also, in the supplementary figures 4/5/6, it is not clear how the cumulative proportion of individuals can decrease over time.

This point is directly related to the statement in lines 266-269: “In contrast, biases produced by the test-negative design were more likely than biases produced by the cohort design to be influenced by the length of the sampling period, the efficacy against infectiousness, and the probability of transmission.” It is likely that a longer sampling period contributes to these biases. In this scenario, the cohort should be analysed using a Cox or Poisson model accounting for person-time at risk (VE_HR or VE_Rate ratio). The corresponding test negative design analysis will compare individuals tested (+ve or -ve) in the same week week (risk set sampling)

RESPONSE: In real world data, factors such as vaccination coverage, immune escape, behaviours, and vaccine effectiveness (due to waning) often vary over time, necessitating adjustments to control for their influence on VE estimates. These time-related factors can influence VE estimates from both the cohort design and the test-negative design. In our simulations, however, all such factors were deliberately held constant over time as part of the model design. This is why we do not observe any bias in the VE estimates from the cohort design in the equal testing scenario. We recognize these assumptions may not have been made sufficiently explicit in the original manuscript. To address this, we have ensured that each relevant section of the methods now specifies that these factors were assumed to remain constant over time (Lines 512-513; Line 564) Additionally, we have reiterated this in the simulations and analysis section of the Methods: “As previously noted, vaccination coverage, immune escape, and differences in testing were all assumed to be constant over time, and no additional time-related confounders were included in the model” (Lines 572-574)

In both study designs, the depletion bias as outlined by Khan et al. (2022) can give the impression of a waning vaccine. However, Khan et al. (2022) have shown that this problem only arises when a vaccine is leaky. When the vaccine instead provides all-or-nothing protection, as was assumed in our study, the depletion bias they describe does not affect VE estimates (Figure 3; Khan et al. 2022). Additionally, Khan et al. (2022) discovered that if a vaccine provided all-or-nothing protection, excluding individuals with past infection - their proposed fix for the depletion bias with the leaky vaccine - overestimated VE (Figure 3; Khan et al. 2022). This overestimation occurs because “excluding people with past infection with an all-or-nothing vaccine removes people for whom the vaccine did not work at all and focuses the analysis on those for whom the vaccine may be effective” (Kahn et al. 2022). In our analysis, limiting comparisons to individuals only tested in the same week would similarly

exclude those previously infected in prior weeks, potentially introducing an additional bias in the VE estimates.

One aspect of all-or-nothing protection that can affect VE estimates from the test-negative design in later stages of the epidemic is the retention of fully protected vaccinated individuals as controls (i.e., individuals who have never been infected with SARS-CoV-2). At the beginning of an epidemic, the composition of the controls mirrors the general population's vaccination distribution, resulting in unbiased VE estimates. However, as the epidemic progressed, more unvaccinated individuals than vaccinated individuals became infected and were removed from this group, leading to an overrepresentation of vaccinated individuals later in the epidemic and the potential for VE estimates to be overestimated when the epidemic is large. Note the overrepresentation does not stem from differential depletion, as both unvaccinated and unprotected vaccinated individuals are equally removed under the all-or-nothing protection model. Instead, it occurs because the fully protected vaccinated individuals remain in the control group throughout the epidemic, allowing them to make up a larger proportion of controls once many individuals have become infected. As a result of this retention, we found that when testing was equal, this source of bias could lead to biased VE estimates later in the epidemic. However, for VE estimates taken prior to the epidemic peak, the effect was non-existent or small. Additionally, in scenarios where testing differences had the greatest impact (i.e., when vaccine efficacy was low), the effect on VE estimates remained minor even at later epidemic stages (bias < 0.1). The existence and explanation for this bias has now been detailed in the Discussion (Lines 333-339):

> “Even with equal testing, some continued attention may still be warranted as VE estimates [from the test-negative design] can become overestimated when the epidemic is large, and the sampling period includes later epidemic stages. This overestimation occurs because, as the epidemic progresses, the pool of individuals who have never been infected with symptomatic SARS-CoV-2 shrinks - especially among the unvaccinated - leading to a control group that no longer accurately reflects the population’s vaccination distribution.”

Finally, regarding the supplementary figures 4,5, and 6 (now supplementary figures 8,9,10), only the cumulative proportions of individuals with COVID-like symptoms who have never had symptomatic SARS-CoV-2 infection can decrease over time. As described above, this decrease occurs because more of these individuals are becoming infected with SARS-CoV-2 over the course of the epidemic. This explanation has now been added to the captions of Supplementary Fig 8 ,9,10 and Supplementary Fig. 15-16 (which also show the cumulative proportions of individuals with COVID-like symptoms).

R3: In lines 88 to 95, discussing the selection bias induced in the TND, this depends on the criterion used for assessing if the causal effect is identifiable, please check the work of Shrier et. Al. (https://academic.oup.com/ije/article/52/6/1968/7224569?searchresult=1). This type of problem would affect the estimation of risk ratio and not the odds ratio.

RESPONSE: We will thank Reviewer 3 for recommending this paper as it is relevant to our work. We agree with Reviewer 3 that when vaccination is only associated with testing via SARS-CoV-2 infection (Figure 5A in Shrier et al. [2023]), the causal odds ratio is identifiable. Our Equal Testing scenario reflects this situation, explaining why we do not observe biased

VE estimates in this scenario, except for a slight overestimation late in the epidemic (Figure 3, Lines 156-164) - the cause of which is described in our response above.

In contrast, when there is a relationship between testing and vaccination, beyond the one that exists via SARS-CoV-2 infection, the causal odds ratio becomes not identifiable (Figure 5c in Shrier et al. [2023]). The situation depicted in Figure 5c in Shrier et al. (2023) resembles our Unequal Testing scenario, where an additional association exists between testing and vaccination (Figure 1) creates an "opening biasing pathway". In the Introduction, we had aimed to explain how this biasing pathway could arise when selecting on testing (Lines 88-98). In response to Reviewer 3's feedback, we have now refined the text from Lines 94-98 to more clearly convey that the relationship between vaccination and SARS-CoV-2 infection only becomes distorted when both independent associations are present, and testing is selected upon:

> "restricting the study sample to tested individuals may introduce a selection bias via selecting on a collider (also referred to as collider bias), in a situation when the exposure (vaccination) and outcome (infection) both independently affect the likelihood of being tested. Selecting on testing in this scenario distorts the association between vaccination and infection, thereby creating biased VE estimates[6,7]"

In our study design, in addition to selecting on testing, we also condition on symptomatic - aiming to block the selection bias described when selecting on testing. However, by additionally selecting on symptomatic, we also induced other artificial associations that bias the relationship between vaccination and SARS-CoV-2 infection. This is why, in our Unequal Testing scenario, we find underestimated VE (Figure 3, Lines 174-176). We have updated the Methods section to better explain how conditioning on symptomatic (in addition to testing) can create a biasing pathway, starting on Line 459:

> "Selecting additionally on symptomatic aims to block the selection bias [due to selecting on testing] but it can also lead to other bias. Specifically, conditioning on testing introduces a relationship between healthcare engagement and symptomatic ("d" in Fig. 1B), while additionally conditioning on symptomatic induces an association between healthcare engagement and SARS-CoV-2 infection ("e" in Fig. 1B). This induced association ("e"), together with the existing causal relationship between healthcare engagement and vaccination ("a" in Fig1B), results in confounding that biases the relationship between vaccination and SARS-CoV-2 infection (see Supplementary 2: Selection Bias for details)"

R3: 2) The paper of Suah et. al. also compares the test negative and cohort design using a variety of scenarios (including difference in testing). It would be a good addendum for the discussion section to compare the findings of the current paper to that one. https://www.medrxiv.org/content/10.1101/2022.08.25.22279235v1.full.pdf

RESPONSE: We appreciate Reviewer 3 highlighting this relevant work. As Reviewer 3 noted, Suah et al. (2022) also explored the impact of differences in testing by vaccination status on VE estimates, finding that lower testing among unvaccinated individuals compared to vaccinated individuals resulted in underestimates of symptomatic VE. They observed

these underestimates grew with testing differences and found that they were present in both the cohort and test-negative design, which is consistent with our findings. We have now added this comparison to the Discussion (Lines 290 - 294):

> "Similar to Suah et al.[28], we found that the presence of unequal testing by vaccination - where unvaccinated individuals tested less than vaccinated individuals - resulted in underestimation of the true symptomatic VE in both the cohort and test-negative designs with greater testing differences resulting in larger underestimates."

Additionally, we have also now cited Suah et al. (2022) in our introduction to direct readers to another study that directly compares these two study designs (Lines 107:110):

> "Various studies have qualitatively described the mechanisms by which unequal testing by vaccination could lead to residual confounding and/or selection bias. However, fewer studies have measured the direction and magnitude of bias in VE estimates due to unequal testing by vaccination in cohort vs. test-negative designs (but see Suah et al. [28])"

R3: Minor points

R3: 1) The terms "susceptibility", and "effectiveness against infectiousness" presumably refer to "effectiveness against progression to symptoms". If so, it would be better to use terminology similar to Kahn et. al.

RESPONSE: The terms efficacy against susceptibility and efficacy against infectiousness refer to a reduction in the acquisition of SARS-CoV-2 and the reduction in transmitting SARS-CoV-2 when infected and vaccinated, respectively. These terms were taken from Shim and Galvani (2012), which is referenced on Line 506. As neither of these terms refer explicitly to progression to symptoms, we did not use Kahn et al.'s terminology.

R3: 2) There are some pieces of discussion in the results section: lines150-155.

RESPONSE: Our explanation for our findings in the indicated results paragraph has now been been moved to the Discussion section (Lines 333-339)

R3: 3) It is more likely that a vaccine will be "leaky", rather than all or nothing (https://www.nature.com/articles/s41467-023-40750-8)

RESPONSE: We recognize that all-or-nothing protection may not align with recent evidence for SARS-CoV-2. However, we opted to assume "all-or-nothing" protection to provide a baseline for understanding how unequal testing influences bias in VE estimates without introducing additional complexities of the differential depletion by vaccination status associated with the leaky vaccine. Furthermore, this assumption allows our findings to extend beyond SARS-CoV-2, offering insights into scenarios where this form of protection is more applicable.

We recognize that in the context of SARS-CoV-2 and many other infectious diseases, not assuming leaky protection is an important limitation of our study. Therefore, we have now noted it as a limitation in the Discussion (Lines 416-417). However, we also expanded this paragraph to include expectations on how having a leaky vaccine may influence the magnitude of bias due to testing differences. We drew on insights from Lewnard et al. 2018 and Kahn et al. 2022 to provide reasoning behind these expectations, taking into account the potential effects of the depletion bias. Specifically, on Lines 419-424, we have noted:

> "Had we assumed "leaky" protection and not excluded those with past infection[57], the increasing magnitude of bias in our symptomatic VE estimates with unequal testing would likely have increased further. This is because, even with equal testing, "leaky" protection introduces downward bias in VE estimates over the course of an epidemic, as more uninfected and unvaccinated individuals who are not at risk of infection are included in the analysis[25,57].

R3: 4) An additional limitation was not accounting for re-infections, which is a common scenario in any respiratory disease.

RESPONSE: We agree that this is an important limitation. We have now addressed this by noting that our study assumes individuals can only be infected with SARS-CoV-2 once. In addition to noting this limitation, we also outlined how undetected prior infections may have affected our symptomatic VE estimates and additionally, estimates of hybrid immunity, when testing differences by vaccination status are present. Specifically, we have added (Lines 394-399):

> "The presence of prior infections can contribute to underestimating symptomatic VE[51]. Therefore, including prior undetected infections before the start of our sampling period may have further exacerbated the underestimation of VE due to unequal testing by vaccination status. Allowing for multiple infections throughout the sampling period and estimating hybrid immunity could also have underestimated this protective effect if testing was unequal and unvaccinated individuals - with or without natural immunity - remained a reference group."

In cases where we are assessing hybrid immunity, we also point out that the magnitude of the potential bias would "depend on how testing varied by vaccination status, prior infection status, and by the interaction between the two" (Lines 400 - 401)

RESPONSE TO REVIEWERS

Reviewer #3 (Remarks to the Author)

R3: The authors replied satisfactorily to most of the previous points.

However, the issue regarding the calculation of OR from the TND remains unclear. In the manuscript, the cumulative proportion of individuals who tested positive for covid-19 is used in relation to vaccinated/unvaccinated individuals and other aetiologies (vaccinated/unvaccinated). However, as shown in the supplementary figures regarding the proportion of individuals testing positive for other aetiologies (which can be reduced once individuals who test positive for COVID-19 are removed from this pool), this condition would create a scenario of the numbers (tests in previous time) used for controls will be “removed' over time. For example:

OR time 100:

VAC+= 250/ VAC - = 400 / UNVAC+=300/UNVAC-=425

20 new positive tests in vac /unvac. Among them, 5 were individuals who tested positive for other aetiologies

OR time 120

VAC + = 270/ VAC- = 395 / UNVAC+ 320 / UNVAC- = 420

This approach would remove individuals who tested negative (positive for other diseases) previously and tested positive at the current time. The test-negative approach works by using the tests as a time-fixed point. However, this approach should retain the negative test of a person prior to the positive test.

RESPONSE: We thank Reviewer 3 for their thoughtful comments, which we think have helped us further improve the paper

The reviewer raises an important point with the test-negative design - that our approach may imply the removal of controls over time when individuals who previously tested negative later test positive for COVID-19. They also note that in a standard test-negative design, negative tests should be treated as a time-fixed reference point.

First, the comment led us to better clarify in our manuscript what we mean when we discuss VE estimates “over time”. Second, we conducted additional sensitivity analyses to explore how accounting for time (versus not explicitly accounting for time) could influence the biases due to unequal testing.

The test-negative design is a type of retrospective case-control study. Individuals are classified as a “case” or a “control” after looking back at the sample at a given time-point. Thus, controls are not removed over time for a given study. When we conduct a test-

negative design with data up to day 100, we classify individuals as either positive or negative within a study's sampling period of 100 days. If we repeat the test-negative design on day 120, and an individual who had initially tested negative later tested positive, they are selected as a case and are never included as part of the control group in the test-negative study conducted using the 120 day sample To ensure this methodology is clear, we have added the sentence, "each individual included in a study receives only one classification" (Line 559).

In the OR time 100 and 120 examples provided by Reviewer 3, the cumulative proportions of those testing negative (and never positive) represent different (independent) VE studies with different sampling periods – not repeated VE measurements in a single study. This distinction is important because, if we track the timing of negative tests used in a symptomatic VE calculation within a single study, the cumulative proportion of negative tests would increase over time, as Reviewer 3 correctly noted. We conducted analyses that tracked sampled negative tests over time to confirm this pattern: for a study ending on day 70 – with equal testing and low vaccine efficacy against susceptibility (0.1)- the median cumulative proportion of those testing negative increased over time from 0.07 at t = 20, to 0.14 at t = 40, to 0.19 at t = 60 and to 0.22 at t=70.

To clarify that VE estimates shown over time represent comparisons across studies, we have revised the manuscript to explicitly state that these estimates come from different studies with distinct sampling periods. For example, we modified the Abstract to state, "we also found the magnitude of bias was moderated by the study's selected sampling period" (Line 58) and the Introduction to, "and then examined how biases varied across studies with different sampling periods within each study design (Objective 1)" (Lines 134-135). We also specified that RR(t) and OR(t) were each calculated "for a study conducted at time t;" (Lines 537, 555). Additional revisions were made throughout the manuscript to clarify this point, including in the Results (Lines 149, 158, 165, 168, 185-186), the Discussion (Lines 309, 345), the Figure 3 caption (Lines 901-902) and other sections of the Methods (Lines 607-610).

In response to Reviewer 3's questions about the potential influence of time and the benefits of treating tests from the test-negative design as time-fixed, we have now conducted additional sensitivity analyses that account for time when estimating symptomatic VE for the test-negative design. To conduct these analyses, we recorded the time of each sampled test and then included time from the start of the study (i.e., calendar time) as a covariate in a logistic regression model. A description of this methodology is provided in the Methods (Lines 560-563) and in Supplementary 2: Symptomatic VE Estimation. Additionally, we clarified the sampling criteria for positive and negative tests used in the symptomatic VE calculations to avoid any confusion about inclusion and exclusion criteria:

> "Individuals contributed only a single positive or negative test to a study: if an individual had at least one positive test, only the first positive test was selected; if an individual had multiple negative tests and no positive tests, one negative test was randomly selected" (Lines 563-566).

We did not retain an individual's negative test prior to their positive test because selecting a single test per individual within a study period is standard practice in the test-negative design.

We found that accounting for time led to very small differences in symptomatic VE estimates compared to our main results (before explicitly accounting for time). Specifically, “adjusting for time using a logistic model reduced the magnitude of bias for studies with shorter sampling periods, particularly with lower vaccine efficacy; however, the overall pattern of increasing bias with longer sampling periods remained unchanged” (Lines 195-198). These results have now been added as Supplementary Figure 6 and Supplementary Tables 4-5.

To align analyses across study designs, we also conducted sensitivity analyses that accounted for time for the cohort design. For these analyses, we used a Cox model—previously recommended by Reviewer 3. A description of this methodology is provided in the Methods (Lines 542-544) and in Supplementary 2: Symptomatic VE Estimation. Altogether, we found that although “the temporal pattern of the bias due to unequal testing was substantially reduced, especially with lower vaccine efficacy (0.1)” (Lines 174-175), large biases due to testing differences remained. These results have now been added as Supplementary Figure 2 and Supplementary Tables 2-3.