USG COVID-19 Response Team / CoVPN Vaccine Efficacy Trial Immune Correlates Statistical Analysis Plan

USG COVID-19 Response Team / Coronavirus Prevention Network (CoVPN) Biostatistics Team

Peter B. Gilbert[1,2*], Youyi Fong[1,2], David Benkeser[3], Jessica Andriesen[1], Bhavesh Borate[1], Marco Carone[2], Lindsay N. Carpp[1], Iván Díaz[4], Michael P. Fay[5], Andrew Fiore-Gartland[1], Nima S. Hejazi[1,4,6], Ying Huang[1,2], Yunda Huang[1], Ollivier Hyrien[1], Holly E. Janes[1,2], Michal Juraska[1], Avi Kenny[2], Kendrick Li[2], Alex Luedtke[7], Martha Nason[5], April K. Randhawa[1], Lars van der Laan[6], Brian D. Williamson[1], Wenbo Zhang[2], Dean Follmann[5]

[1]Vaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutchinson Cancer Research Center, Seattle, Washington

[2]Department of Biostatistics, University of Washington, Seattle, Washington

[3]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia

[4]Department of Population Health Sciences, Weill Cornell Medical College, New York, New York

[5]National Institute of Allergy and Infectious Diseases, Bethesda, Maryland

[6]Division of Biostatistics, School of Public Health, University of California, Berkeley, California

[7]Department of Statistics, University of Washington, Seattle, Washington

Correspondence: *pgilbert@fredhutch.org

April 18, 2022

Version 0.4

Contents

List of Tables

List of Figures

1 Prelude

The scientific community has responded to the COVID-19 pandemic with admirable global cooperation and solidarity, characterized by rapid sharing of results and data in an effort to urgently re-focus research toward the common goal of developing prevention and treatment modalities to help turn the tide of the COVID-19 pandemic. As spoken by WHO Director-General Dr. Tedros Adhanom Ghebreyesus at the 2020 Aspen Security Forum, "Our best way forward is to stick with science, solutions and solidarity and together we can overcome this pandemic." World Health Organization (2020)

In this spirit, we are making this Statistical Analysis Plan (SAP) publicly available at a relatively early and intermediate stage. The SAP is a work in progress that will continue to be developed and refined over the coming weeks. Our hope is that fellow statistical scientists and scientists of other disciplines will bring new insights and offer input to maximize the scientific knowledge pertaining to immune correlates of protection that can be learned from COVID-19 vaccine efficacy trials. We invite collaboration and are eager to explore opportunities for working with others on COVID-19 immune correlates analyses.

We envisage three applications of this SAP. First, as the Coronavirus Prevention Network Statistical Center and COVID-19 Response Biostatistics Team our group is responsible for statistical design and analysis of immune correlates for the United States Government (USG)/COVID-19 Response Team phase 3 trials, and this SAP serves as a master protocol-type SAP for harmonized immune correlates analyses across the trials. Second, researchers conducting additional clinical trials may work collaboratively with our group to co-conduct the immune correlates analysis. Third, researchers conducting clinical trials may use this SAP as a resource for immune correlates analyses conducted on their own, either by implementing this SAP or components therein, or by selecting methods and code from it to adopt for their own SAPs.

We are implementing the SAP in R. The R scripts are hosted on a Github code repository CoVPN/correlates_reporting and will be made publicly avail-

able as soon as they are ready. The first publicly available download will be focused on immunogenicity characterization and correlates of risk analysis and the second one will be focused on correlates of protection analysis. This collaborative Github repository will include reproducible reports implementing the SAP on a mock/practice COVID-19 vaccine efficacy trial data set.

Please direct communication related to this SAP to Peter B. Gilbert at the Fred Hutchinson Cancer Research Center (pgilbert@fredhutch.org).

2 Introduction

2.1 Antibody Assays and Day 57 Markers

This SAP describes the statistical analysis of antibody markers measured at a key time point post last vaccination as immune correlates of risk and as various types of immune correlates of protection against primary and secondary endpoints in COVID-19 Response Team / CoVPN COVID-19 vaccine efficacy (VE) trials. For definiteness, we assume this time point for antibody measurements is Day 57, a typical time point for a two-dose vaccine; for a one-dose vaccine the key time point would likely be around Day 29. The antibody markers of interest are measured using one of three kinds of humoral immunogenicity assays [more detail on assay types (2) and (3) can be found in Sholukh et al. (2020)]:

(1) **bAbs**: **Binding antibodies** to the vaccine insert SARS-CoV-2 proteins;

(2) **Pseudovirus-nAbs**: **Neutralizing antibodies** against viruses **pseudotyped** with the vaccine insert SARS-CoV-2 proteins; and

(3) **Wild Type Live virus-nAbs**: **Neutralizing antibodies** against **live** "vaccine insert-matched" wild type SARS-CoV-2 (or recombinant "vaccine insert-matched" SARS-CoV-2 harboring a reporter gene within the viral genome).

For example, the following assays are expected to be used:

(1) **bAb assay**: The MSD-ECL Multiplex Assay (MSD-ECL = meso scale discovery-electrochemiluminescence assay)

The MSD assay measures binding antibody to antigens corresponding to: Spike (an engineered version of the Spike protein harboring a double proline substitution (S-2P) that stabilizes it in the closed, prefusion conformation [McCallum et al. (2020)]); the Receptor Binding Domain (RBD) of the Spike protein; and Nucleocapsid protein (N), which is not contained in any of the COVID-19 vaccines.

This assay has a standard curve to interpolate arbitrary units/ml; an 8 point dilution curve on each sample with 5-fold dilutions starting at 1:20; an 8 point dilution curve on VRC control sera; and includes Positive, Negative and Intermediate controls. Binding antibody to N are not of interest as a potential immune correlate; these data are included only for immunogenicity evaluation. Based on the starting dilution of the standards and samples at 1:20, the lower limit of detection (LLOD) for the endpoint titer of the assay is 1:20. Binding antibody readouts below the LLOD are assigned the value LLOD/2 = 10. Values between the LLOD and the lower limit of quantitation (LLOQ) are taken as their actual numeric value.

The bAb assay readouts are in units AU/ml, where AU stands for arbitrary units from a standard curve. The process of validating the assay defined a lower limit of detection (LLOD), an upper limit of detection (ULOD), a lower limit of quantitation (LLOQ), an upper limit of quantitation (ULOQ), and a positivity cut-off for each antigen that defines positive vs. negative response. These values are as follows:

- bAb Spike:
  - Pos. Cutoff = 1204.71 AU/ml
  - LLOD = 34.18 AU/ml
  - ULOD = 19,136,250 AU/ml
  - LLOQ = 199.64 AU/ml
  - ULOQ = 1,128,438.87 AU/ml
- bAb RBD:
  - Pos. Cutoff = 517.86 AU/ml

  - LLOD = 58.59 AU/ml
  - ULOD = 8,201,250 AU/ml
  - LLOQ = 125.9678 AU/ml
  - ULOQ = 598,133.3615 AU/ml
- N:
  - Pos. cutoff = 9779.62 AU/ml
  - LLOD = 39.06 AU/ml
  - ULOD = 21,870,000
  - LLOQ = 1870.70 AU/ml
  - ULOQ = 239,449.31

The Vaccine Research Center established factors for converting the MSD assay readouts from AU/ml to WHO International Units/ml, which is the same thing as Binding Antibody Units/ml (BAU/ml). For the three binding antibody variables CoV-2 Spike IgG, CoV-2 RBD IgG, and CoV-2 N IgG, these conversion factors are 0.0090, 0.0272, and 0.0024, respectively. These conversion factors are applied, such that all binding Ab readouts are reported in WHO International Units/ml with notation BAU/ml, following the WHO recommendation, for all analyses. These conversion factors are also applied to yield the LLOD, ULOD, LLOQ, and ULOQ on the WHO BAU/ml scale. The following shows the assay limits on the BAU/ml scale:

- bAb Spike:
  - Pos. Cutoff = 10.8424 BAU/ml
  - LLOD = 0.3076 BAU/ml
  - ULOD = 172,226.2 BAU/ml
  - LLOQ = 1.7968 BAU/ml
  - ULOQ = 10,155.95 BAU/ml
- bAb RBD:

  - Pos. Cutoff = 14.0858 BAU/ml
  - LLOD = 1.593648 BAU/ml
  - ULOD = 223,074 BAU/ml
  - LLOQ = 3.4263 BAU/ml
  - ULOQ = 16,269.23 BAU/ml
- bAb N:
  - Pos. Cutoff = 23.4711 BAU/ml
  - LLOD = 0.093744 BAU/ml
  - ULOD = 52,488 BAU/ml
  - LLOQ = 4.4897 BAU/ml
  - ULOQ = 574.6783 BAU/ml

All values below the LLOD are assigned the value LLOD/2. For immunogenicity reporting, values greater than the ULOQ are not given a ceiling value of the ULOQ, the actual readouts are used. For the immune correlates analyses, values greater than the ULOQ are assigned the value of the ULOQ.

(2) **Pseudovirus-nAb assay**: A firefly luciferase (ffLuc) reporter neutralization assay for measuring neutralizing antibodies against SARS-CoV-2 Spike-pseudotyped viruses

Based on the Duke assay from the Montefiori lab, serum inhibitory dilution 50% titer (ID50) and serum inhibition dilution 80% titer (ID80) values are estimated based on a starting serum dilution of 1:10, with eight 5-fold dilutions. Thus 1:10 is the LLOD on the scale of the assay. The process of validating the assay defined the LLOD, LLOQ, and ULOQ for ID50 and ID80 as follows:

- ID50:
  - LLOD = 10
  - LLOQ = 18.5

  - ULOQ = 45118
- ID80:
  - LLOD = 10
  - LLOQ = 14.3
  - ULOQ = 10232

ID50 and ID80 values below the LLOD are assigned the value 10/2 = 5. Values between the LLOD and the LLOQ are taken as their actual numeric value. For immunogenicity reporting, values greater than the ULOQ are not given a ceiling value of the ULOQ, the actual readouts are used. For the immune correlates analyses, values greater than the ULOQ are assigned the value of the ULOQ. This is done so as to not unduly influence the correlates analyses by high outlying values, given the expectation that the most relevant marker dynamic range for correlates is much lower than the ULOQ.

ID50 and ID80 values are reported in international units based on the report from David Montefiori "Reagent Calibration Report: First WHO International Standard for SARS-CoV-2 Immunoglobulin in a Neutralization Assay" (May, 2021). This report derived calibration factors based on arithmetic means:

- Calibration factor ID50: 0.242
- Calibration factor ID80: 1.502

The original readouts are calibrated to the IU scale by multiplying each original ID50 value by 0.242, and multiplying each original ID80 value by 1.502, and units are reported as calibrated ID50 (cID50) and calibrated ID80 (cID80). Consequently, the LLOD, LLOQ and ULOQ for cID50 and cID80 are as follows in International Units:

- cID50:
  - LLOD = 2.42
  - LLOQ = 4.477
  - ULOQ = 10919

- cID80:
  - LLOD = 15.02
  - LLOQ = 21.4786
  - ULOQ = 15368

(3) **Wild Type Live virus-nAb assay**: An assay measuring antibody-mediated neutralization of live wild-type SARS-CoV-2 (WA isolate, passage 3, Vero-E6 cells).

The WT live virus-nAb marker is defined as MN50 calculated using the Spearman-Karber method. The Battelle assay has the following parameters, in original MN50 units and in International Units (IU) with WHO IU conversion factor 0.276:

- LLOD: 82.11 MN50 (22.66 IU)
- LLOQ: 159.79 MN50 (44.1 IU)
- ULOQ: 11,173.11 MN50 (3,083.74 IU)

Values below the LLOD are assigned the value LLOD/2. Values between the LLOD and the LLOQ are taken as their actual numeric value. Values greater than the ULOQ are assigned the value of the ULOQ.

Throughout this SAP we assume that all three types of assays have validated versions that are applied uniformly to samples collected in one or several late-stage SARS-CoV-2 vaccine efficacy trials. Samples from the same trial are expected to be assayed by the same lab that performs one of these immunoassays.

Based on each immunoassay applied to paired serum samples collected from participants on Day 1 (baseline, pre-vaccination) and Day 57 (post-vaccination visit), the following set of antibody markers is defined for immunogenicity and immune correlates analyses.

- For bAb: $\log_{10}$ IgG concentration (BAU/ml) at each time point, and the difference in $\log_{10}$ concentration (Day 57 minus Day 1) representing $\log_{10}$ fold-rise in IgG concentration from baseline to 28 days post dose

two. These markers are defined for each antigen Spike, RBD, and N.

- For PsV nAb: $\log_{10}$ serum inhibitory dilution 50% titer (ID50) and serum inhibition dilution 80% titer (cID80) at each time point, as well as the $\log_{10}$ fold-rise of these markers over Day 1 to Day 57.
- For WT live virus nAb: $\log_{10}$ serum MN50 at each time point, as well as the $\log_{10}$ fold-rise of this marker over Day 1 to Day 57.

For two-dose vaccines, the immunogenicity and correlates analyses may also include the same antibody markers measured at the second-dose sampling time point, which we refer to as the Day 29 time point. In this SAP we include contingency sub-sections marked '[With Day 29 Markers]' to describe how the SAP is augmented to include the Day 29 antibody markers in the analysis.

3 Study Cohorts and Endpoints

3.1 Study Cohort for Correlates Analyses

Finalization of the primary study cohort for correlates analysis will take place before unblinding case/non-case and randomization arm information for correlates analyses. The default is for the primary study cohort to be the same as the cohort used in the primary analysis of vaccine efficacy against the primary endpoint in the protocol, except that availability of a Day 1 and Day 57 blood sample for antibody testing is also required. It may also be required to have results from all tests for SARS-CoV-2 infection on Day 57 samples [serology and/or nucleic acid amplification test (NAAT)].

Typically the primary analysis cohort is baseline SARS-CoV-2 negative participants in the per-protocol cohort, with the per-protocol cohort defined as those who received all planned vaccinations without any specified protocol deviations, and who were SARS-CoV-2 RT-PCR negative at the terminal vaccination visit. We refer to this cohort representing the primary population for correlates analysis as the Per-Protocol Baseline Negative Cohort. We will wait to fully understand all of the antigen and serology testing data that are available in the data set to finalize the definition of the primary

analysis cohort. If a vaccine has high vaccine efficacy, it is possible that rare vaccine breakthrough cases will be individuals who were infected before the second vaccine dose, or soon after the first dose, but had unusually long time periods between SARS-CoV-2 acquisition and symptomatic infection (COVID) diagnosis. This situation could complicate the interpretation of correlates analyses. Therefore, we may conduct some correlates analyses that use a stringent criterion to include vaccine breakthrough cases in the analyses, such as requiring antigen negative tests at both the dose two visit and Day 57 and all serologies negative through Day 57.

As the primary analysis of vaccine efficacy is conducted in baseline negative individuals, correlates of risk (CoR) and correlates of protection (CoP) analyses are only done in baseline negative individuals, and the analysis of data from baseline positive individuals is for purposes of immunogenicity characterization, given too-few anticipated vaccine breakthrough study endpoints for CoR/CoP assessment (although if there are many baseline positive vaccine breakthrough endpoint cases that baseline positive subgroup analyses may be considered). In baseline negative individuals, antibody marker data in placebo recipients is relevant for verifying the expectation that almost all Day 57 marker responses will be negative, given the lack of SARS-CoV-2 antigen exposure.

3.1.1 [With Day 29 markers]

If Day 29 markers are included, analyses of Day 29 correlates are done in the same cohort as studied for the analyses of Day 57 markers, except the time origin in correlates analyses is set at the Day 29 visit and the set of participants included in the analysis is augmented with intercurrent cases. Intercurrent cases are defined as participants who were diagnosed with the COVID primary endpoint $\geq$ 7 days post Day 29 visit plus baseline negative cases with endpoint $\geq$ 6 days post Day 57 visit. These intercurrent cases are not included in the Day 57 marker correlates analyses for which cases are counted starting 7 days after the Day 57 visit. Thus, intercurrent cases are exactly the set of cases included in Day 29 correlates analyses but excluded from Day 57 correlates analyses. Analyses that include both Day 29 and 57

markers use the same cohort as for analyses of Day 57 markers only.

3.2 Study Endpoints

Endpoints for per-protocol correlates analyses of Day 57 markers are included if they occur at least 7 days after the Day 57 visit, to help ensure that the endpoint did not occur prior to Day 57 antibody measurement. Thus participants with a per-protocol endpoint diagnosed earlier are excluded from the Day 57 marker per-protocol correlates analyses.

Figure 1 defines five study endpoints that are assessed in COVID-19 vaccine efficacy trials, where all trials use COVID (symptomatic infection) as the primary endpoint. While the severe COVID endpoint is of paramount clinical importance, likely the number of events at the time of the first correlates analysis will be too small to assess correlates against this endpoint, such that correlates analyses will be done once more endpoints have accrued through longer-term follow-up.

In contrast, depending on the estimate of vaccine efficacy, there may be enough data to assess correlates against the endpoints non-severe COVID, SARS-CoV-2 infection (COV-INF), asymptomatic infection (ASYMP-COV-INF) at or shortly after the time of the first correlates analysis, and viral load at COVID diagnosis. Similar statistical methods can be used for each endpoint, with some distinctions that we discuss below in "General Statistical Issues in Correlates Assessment."

When a correlates analysis is done, all available follow-up for participants is included through to the time of the data base lock for the correlates analysis, for every CoR and CoP analysis that is conducted. This means that the time of right censoring for a given failure time endpoint will be the first event of loss to follow-up or the date of administrative censoring defined as the last date of available follow-up. For CoP analyses, which use both vaccine and placebo recipient data and leverage the randomization, follow-up is censored at the time of unblinding. In general all blinded follow-up is included and no post-unblinding follow-up is included.

3.2.1 [With Day 29 markers]

If Day 29 markers are included, analyses that study Day 29 markers count study endpoints starting 7 days post Day 29 visit, instead of starting 7 days post Day 57 visit. Analyses that include both Day 29 and Day 57 markers (for correlates analyses this only includes the multivariable correlates of risk superlearning objective) use the same cohort and endpoints as analyses of Day 57 markers only.

4 Objectives of Immune Correlates Analyses of a Phase 3 Trial Data Set

4.1 Characterize Vaccine Immunogenicity

There are two objectives to characterize the binding and neutralizing antibody immunogenicity of the vaccine:

Stage 1 To characterize vaccine immunogenicity (bAb, nAb) at Day 1, 29, 57

Stage 2 To characterize vaccine immunogenicity/durability (bAb, nAb) over time (Day 1, 29, 57, 209, 394, 759)

4.2 Correlates of Risk and Correlates of Protection

We broadly classify the proposed analyses into two related categories: correlates of risk (CoR) and correlates of protection (CoP) analyses. CoR analyses seek to characterize correlations/associations of markers with future risk of the outcome amongst vaccinated individuals in the study cohort. CoP analyses seek to formally characterize causal relationships among vaccination, antibody markers and the study endpoint, and use data from both vaccine and placebo recipients. Table 1 summarizes these objectives and statistical frameworks that are commonly used to these ends.

The advantage of CoR analyses it that it is possible to obtain definitive answers from the phase 3 data sets, that is one can credibly characterize associations between markers and outcome. The advantage of CoP analyses is that the effects being estimated have interpretation directly in terms of how

an antibody marker can be used to reliably predict vaccine efficacy (the criterion for use of a non-validated surrogate endpoint for accelerated approval). The disadvantage of CoR analyses are that a CoR may fail to be a CoP, for example due to unmeasured confounding, lack of transitivity where a vaccine effect on an antibody marker occurs in different individuals than clinical vaccine efficacy, or off-target effects (VanderWeele, 2013). The disadvantage of CoP analyses is that statistical inferences rely on causal assumptions that cannot be completely verified from the phase 3 data, such that compelling evidence may require multiple phase 3 trials and external evidence on mechanism of protection (e.g., from adoptive transfer or vaccine challenge trials). Our approach presents results for both CoR and CoP analyses, seeking clear exposition of how to interpret results, the assumptions undergirding the validity of the results, and diagnostics of these assumptions and assessment of robustness of findings to violation of assumptions.

We conjecture that an antibody marker could qualify as a non-validated surrogate endpoint (meeting accelerated approval criteria) based on meeting all three conditions: (1) demonstration of a strong and robust CoR with confounding control; (2) external data supporting functionality and connection to a mechanism of protection; and (3) CoP analyses supporting that the biomarker is likely to be a CoP and not only a CoR. Mechanisms of protection as in (2) may be learned through passive antibody transfer studies and vaccine challenge studies in animals and/or humans.

Table 1: Correlates of Risk (CoRs) and Correlates of Protection (CoPs) Objectives for Day 57 Markers

Objective Type	Objective
CoRs (Risk Prediction Modeling)	**To assess Day 57 markers as CoRs in vaccine recipients** a. Relative risks of outcome across marker levels b. Absolute risk of outcome across marker levels c. Machine learning risk prediction for multivariable markers
CoP: Correlates of VE	**To assess Day 57 markers as correlates of VE in vaccine recipients** a. Principal stratification effect modification analysis b. Assesses VE across subgroups of vaccine recipients defined by Day 57 marker level in vaccine recipients
CoP: Controlled Effects on Risk and VE	**To assess Day 57 markers for how assignment to vaccine and a fixed marker value would alter risk compared to assignment to placebo**
CoP: Stochastic Interventional Effects on Risk and VE	**To assess Day 57 markers for how stochastic shifts in their distribution would alter mean risk and VE** (Hejazi et al., 2020)
CoP: Mediators of VE	**To assess Day 57 markers as mediators of VE** a. Mechanisms of protection via natural direct and indirect effects a. Estimate the proportion of VE mediated by a marker or markers

4.2.1 [With Day 29 markers]

If Day 29 markers are included, then each of the objectives for Day 57 markers is repeated for Day 29 markers. In addition, the multivariable CoR machine learning objective includes models that include both Day 29 and Day 57 markers.

4.3 Synthesis of the Phase 3 Correlates Analyses for Decisions

Establishment of an immunologic biomarker for approval/bridging applications is generally not based on pre-fabricated criteria nor a single type of correlates analysis. Therefore, the goal of the correlates analysis is to generate evidence about correlates from many perspectives, and to synthesize the

evidence to support certain decisions. Consequently, we believe there is value in assessing all of the types of correlates presented in Table 1 in each trial, given that the analyses address distinct questions. Obtaining a set of results from multiple distinct approaches that provide complementary and coherent support may increase the rigor and robustness of an evidence package supporting potential use of an antibody marker as a validated surrogate (for traditional approval) or as a non-validated surrogate (for accelerated approval) (Fleming and Powers, 2012); these uses of a biomarker are summarized below. However, the assumptions needed for valid inferences are somewhat different across the methods, and some of these assumptions have testable implications; therefore examination of the assumptions may lead to favoring some methods over others, and affect the synthesis and interpretation of results, and moreover if diagnostics support that some necessary assumptions are infeasible then certain analyses will be canceled, as described below.

Section 16 summarizes the approach to use and interpretation of the set of multiple correlates of protection methods. Furthermore, depending on the number of study endpoints in the vaccine and placebo arms at the time a trial delivers primary results, some of the Day 57 marker correlates types defined in Table 1 will be evaluable at the first correlates analyses, whereas others will not be evaluable until additional evaluable vaccine breakthrough endpoints have been observed.

As detailed in Table 4, some CoR analyses are done after there are at least 25 evaluable vaccine breakthrough cases, which is considered to be a minimal number to achieve worthwhile precision. On the other hand, the most nonparametric/flexible CoR analyses require more cases, as do the CoP analyses in general, given the need to adjust for all potential confounders in order to fully identify the causal effects parameters of interest and the greater challenge in estimation (compared to CoR analysis) posed by the need to deal with missing potential outcomes.

Finally, we note that meta-analysis of multiple VE trials will provide important empirical support for potentially establishing an immunologic surrogate endpoint, which underscores the necessity of standardizing the VE

trials (common study endpoints, common labs and immunoassays, common statistical methods and data analysis).

4.4 Additional Objectives Not Covered in this SAP

An additional objective that will be assessed, but will be described in a separate SAP, is to assess antibody markers over time beyond Day 57 through two years post vaccination, to assess "outcome-proximal correlates." While we do not formally develop methods to address this aim in this SAP, here we very briefly outline approaches that may be pursued to address such aims. For example, a CoR analysis might assess longitudinal markers in vaccine recipients through estimation of the hazard ratio of outcome across levels of the current value of the marker modeled via a linear mixed effects model (e.g., Fu and Gilbert, 2017). A CoP analysis might assess longitudinal markers as mediators of VE (mechanisms of protection), for example by assessing the proportion of VE mediated by the longitudinal marker(s) profile (e.g., Zheng and van der Laan, 2017).

To potentially help addressing these aims, antibody markers should be measured at the time of COV-INF and COVID diagnosis. An open question is whether and how these measurements may be used in outcome-proximal correlates analyses, for example by assuming that the observed marker values on or near the day of diagnosis were present on the date of SARS-CoV-2 acquisition. Justification for this assumption for the COVID endpoint would derive from information that COVID tends to occur within only several days of SARS-CoV-2 acquisition, implying insufficient time for the infection to make new antibodies that would complicate the interpretation of the vaccine-elicited antibodies. It is possible that this condition could only be verified in a subset of cases, in which case validation-set missing data statistical methods may be fruitful. More validation work will be required before methods would be used treating marker values at diagnosis as present at endpoint diagnosis and caused solely by the vaccine (i.e., not also caused by natural infection). For the COV-INF endpoint it is less feasible / possible to use the COV-INF sampling marker value to infer the marker value at acquisition, given the unknown period of weeks or months that may have elapsed between acquisition

and seroconversion.

Outcome-proximal correlates analyses may be especially relevant if Day 57 antibody markers tend to be generally high in vaccine recipients and this fact leads to a failure of the correlates analyses to identify a Day 57 correlate: if antibodies wane over time then the outcome-proximal correlates analyses could be more sensitive to detect a correlate.

Additional objectives that may be addressed with the at-diagnosis samples include: (1) to characterize abnormal responses, possibly relevant for safety signals; and (2) to assess the effect of pre-existing antibodies on the active immune response to infection and disease.

5 Applications of Immune Correlates Analyses: Vaccine Approval Pathways and Standards of Evidence

Suppose that one or more phase 3 trials demonstrates beneficial vaccine efficacy against the primary clinical endpoint (e.g., symptomatic infection, i.e. COVID) meeting pre-specified success criteria, and correlates analyses of Day 57 antibody marker data are conducted based on the clinical data and antibody data from the phase 3 trial(s). These correlates analyses, combined with additional data supporting the role of antibody markers as mechanisms of protection or as surrogates of mechanisms of protection, can buttress two potential applications of an antibody marker (Table 2).

Table 2: Two Potential Vaccine Approval Pathways Based on a Day 57 Antibody Marker Endpoint

Traditional Approval	If the marker is scientifically well-established to reliably predict vaccine vaccine efficacy, then subsequent efficacy trials may use the marker as the primary endpoint a. Same vaccine for different populations b. Possibly new vaccines in the same class for the same or different populations
Accelerated Approval	If the marker is judged reasonably likely to predict vaccine efficacy but not yet scientifically well established, then accelerated approval based on the marker endpoint may be possible (requires verification of beneficial clinical VE in post marketing studies) a. Same vaccine for different populations b. Possibly new vaccines in the same class for the same or different populations

Fleming and Powers (2012) defined a *validated surrogate* as a marker that is appropriate for use as an outcome measure for traditional approval of a specific class of interventions against a specific disease, when such interventions are deemed safe and have demonstrated strong evidence that risks from off-target effects are acceptable. They also defined a *non-validated surrogate* as a marker appropriate for use as an outcome measure for accelerated approval as one established to be "reasonably likely to predict clinical benefit" for a specific disease setting and class of interventions. These definitions provide two goalposts for immune correlates analyses of COVID-19 VE trials.

Table 3 summarizes one possible set of requirements for a Day 57 antibody marker to be accepted as a *validated surrogate* for a COVID-19 disease endpoint for use in approving COVID-19 vaccines for specific populations (e.g., SARS-CoV-2 seronegative adults) using Fleming and Power's definition. These potential requirements are conjectures provided for conceptualization purposes, and are not based on COVID-19 regulatory guidance documents.

Table 3: Potential Traditional Approval Requirements for a Day 57 Antibody Marker

Requirements (1.-6. Required)	Endpoints and Evidence Bar
1. Strong evidence for CoR and CoP in vaccine recipients in animal and/or human challenge models	COVID and VL endpoints: Highly significant and predictive and Severe COVID : Point estimates in the right direction and COV-INF, ASYMP-COV-INF: No countervailing evidence[1]
2. Strong evidence that the marker is a mechanistic CoP or tightly correlated with a mechanistic CoP (likely deriving from animal challenge studies of vaccines or passively transferred antibodies)	Study endpoints used in challenge models such as subgenomic SARS-CoV-2 RNA
3. Supportive evidence from natural history studies of CoRs of re-infection in SARS-CoV-2 infected individuals	Same endpoints as in Phase 3 trial (COVID , severe COVID , ASYMP-COV-INF, COV-INF, VL Dx)
4. Phase 3 trial strong evidence as a CoR in vaccine recipients	COVID and ≥ 1 other endpoint: Highly significant and predictive and Point estimates in the right direction for the other endpoints Consistent results Day 29, 57 markers Require consistent results from multiple trials
5. Phase 3 trial strong supportive evidence as CoP, for at least one CoP type, plus point estimates in the right direction for the other CoP types (consistency of evidence)	COVID Point estimates of association/causal parameters in the right direction for the other endpoints[2] Require consistent results from multiple trials
6. Temporal ordering support for several of the above results, e.g., CoRs and CoPs are stronger for COVID occurrence proximal to vaccination than distal, synchronized with pattern of biomarker waning	COVID , severe COVID , ASYMP-COV-INF, COV-INF, VL Dx
7. Additional support from non-vaccine interventions, e.g., demonstration of a neutralization CoP for a monoclonal Ab	COVID , severe COVID , ASYMP-COV-INF, COV-INF, VL Dx

[1]Countervailing evidence could be any observations that provide evidence against a CoP, e.g., relative to Bradford-Hill criteria (see Section 16).

[2]Because CoPs can differ by study endpoint Plotkin (2010) and vaccine efficacy can differ by study endpoint, this criterion will not necessarily be important.

A potential goalpost for a *non-validated surrogate* for accelerated approval can be conceptualized as the same as that for traditional approval, with modifications:

- The package of evidence for the seven sources listed in Table 3 may be less stringent quantitatively, and not requiring success on all of the first six categories.
- Source 4 (Phase 3 CoR in vaccine recipients) would need to have strong evidence (highly statistically significant and highly predictive).
- The support for an immune correlate may be more restricted to a given study endpoint.
- It may no longer be required to have replication of results across two or more Phase 3 trials.

It is hypothesized that a single validated assay will yield a validated or non-validated surrogate endpoint, e.g., based on binding antibody IgG concentration or serum ID50 or cID80 titer to viruses pseudotyped with the Spike vaccine insert protein (or live SARS-CoV-2). However, the goalposts could potentially also be met by a synthesis biomarker aggregated from measurements from multiple validated assays if this aggregation substantially improves the correlate (e.g., a co-correlate Plotkin (2010); Plotkin and Gilbert (2018)). However, the preferred approach, for parsimony and practical utility, would be to define a correlates of protection as a single biomarker derived from a single assay.

5.0.1 [With Day 29 markers]

If Day 29 markers are included, then a validated surrogate endpoint or non-validated surrogate endpoint could be defined based on either a Day 57 time point or a Day 29 time point, and possibly also requiring both time points integrated into the same biomarker.

6 Timeline/Sequencing of Correlates Analyses

The correlates analyses are initiated by the availability of (a) a data set defined at or after the primary analysis data set triggered by the accrual of a certain number of primary endpoints (typically approximately 150 in U.S. phase 3 studies); and (b) Day 1, 57 antibody marker data from correlates-eligible COVID primary endpoint cases from at least 25 baseline negative vaccine recipients. The latter requirement ensures that there are enough endpoint cases to achieve worthwhile precision for CoR analyses. The HVTN 505 trial serves as a precedent where 25 evaluable vaccine recipient cases provided enough data to reasonably characterize correlates of risk for a preventive candidate HIV vaccine (Janes et al., 2017; Fong et al., 2018; Neidich et al., 2019; Gilbert et al., 2020b). In addition, simulation studies show that correlates analyses at 20 endpoints have notably lower precision.

Table 4 shows the minimum number of baseline negative vaccine recipient endpoints evaluable for correlates analyses that are required before conducting the various planned correlates analyses.

Table 4: Minimum Numbers of Evaluable Endpoints in Baseline Negative Vaccine Recipients to Initiate Correlates Analyses

Correlates Analysis Type	**Number**
CoRs (Risk Prediction Modeling)	
a. (Semi)parametric models with strongly parametrized associations:	
Cox, hinge/threshold logistic regression	25
b. Flexible parametric models: Generalized additive model	35
c. Nonparametric thresholds: Donovan et al. (2019)/	
van der Laan et al. (2020)	35
d. Superlearner estimated optimal surrogate	35
CoP: Correlates of VE	50
CoP: Controlled VE	50
CoP: Stochastic Interventional VE	50
CoP: Mediators of VE	50

6.1 Timeline of Statistical Analysis Reports

We summarize the plans for analysis reports over the whole period of the study. When the Day 1, 57 antibody data from the immunogenicity subcohort are available, the first immunogenicity report will be produced. When Day 1, 57 antibody data on COVID cases are also available, the first correlates of risk report will be produced, focusing on Stage 1 data only. When there is enough follow-up to measure antibody markers at the later time points (i.e., Day 209, Day 394, possibly Day 759), additional immunogenicity and correlates reports will be made, including those that assess outcome-proximal correlates of risk and protection based on Stage 2 data. The initial correlates reports will likely only include the symptomatic infection/COVID study endpoint; as data sets become available for the other endpoints the reports will add correlates analyses against the secondary endpoints.

7 General Statistical Issues in Immune Correlates Assessment

Throughout this section, we define the asymptomatic infection endpoint as seroconversion without prior occurrence of the COVID endpoint.

Issue 1: Timing of endpoint definition, accounting for diagnosis at presentation (i.e., date of virological confirmation of symptomatic COVID – COVID diagnosis) or during post-COVID-19 diagnosis follow-up.

- **COV-INF:** Defined at presentation (if COVID endpoint) or at first positive serotest visit, whichever occurs first
- **COVID:** Defined at presentation/virologic confirmation
- **Asymptomatic infection:** Defined at first positive serotest (without prior COVID endpoint)
- **Non-severe COVID:** Ascertained by post-COVID diagnosis follow-up, where the failure time could be defined by the time of resolution of symptoms

- **Severe COVID:** Occurs at presentation or at any time during post-COVID diagnosis follow-up

At COVID endpoint diagnosis, participants roll over onto a post-diagnosis follow-up track (Figure 2). This is irrelevant for analysis of the first three endpoints listed above, but for the non-severe COVID endpoint and the severe COVID endpoint special considerations are needed for proper correlates analyses. Survival analysis theory typically requires *predictable processes*, such that non-severe COVID and severe COVID would have failure times defined when the classification of the endpoint is known. However, alternatively, the analysis could be simplified by defining the failure time for all three endpoints COVID, severe COVID, and non-severe COVID to be the date of presentation, even though at that time one needs to look into the future to determine whether the COVID endpoint is severe or non-severe. Such an approach could be justified by thinking of the data as a competing risks data structure, where one observes the time to COVID, and each COVID endpoint has an associated binary endpoint "type", severe or non-severe. The analyses will use this simplified approach. A justification of this simplified approach is that severe COVID is a very rare event among vaccine recipients, and it is the fact of having the event that is important, not whether it happened at or 9 days post COVID diagnosis, such that using a more refined failure time would be unlikely to carry additional meaningful information. If greater than 10% of COVID endpoint cases are missing the endpoint type, then methods accounting for missing endpoint types will be used (e.g., Heng et al., 2020).

Issue 2: Is the endpoint appropriately analyzed using ordinary survival analysis or competing risks survival analysis?

For this issue, we consider use of a time-to-event method to assess vaccine efficacy. In general, a competing risk of a given endpoint of interest is an endpoint that, once it occurs, precludes the possibility of future occurrence of the other endpoint.

1. COVID is a competing risk for asymptomatic infection
2. Severe COVID is a competing risk for non-severe COVID

Therefore, the asymptomatic infection and non-severe COVID endpoints may be best analyzed by competing risks methods. For example, instead of estimating cumulative incidence $P(T \leq t|A = a)$ for a given randomization arm $A = a$, where T is the time from enrollment until the endpoint, we analyze cumulative incidence $P(T \leq t, J = 1|A = a)$, where T is the time to the first event of $J = 1$ (event of interest) or $J = 2$ (competing event), and cumulative $\text{VE}(t)$ may be assessed using the parameter

$$\text{VE}(t) = 1 - \frac{P(T \leq t, J = 1|A = 1)}{P(T \leq t, J = 1|A = 0)}.$$

In addition, hazard-ratio-based VE may be defined as one minus the cause $(J = 1)$-specific hazard ratio (Prentice et al., 1978; Gilbert, 2000).

It is also worth noting that:

1. Asymptomatic infection is not a competing risk for COVID, because participants experiencing the asymptomatic infection endpoint continue follow-up for the COVID endpoint (such that at asymptomatic infection diagnosis it is not known whether the infection is truly asymptomatic or pre-symptomatic), and it is not certain that seroconversion prevents future COVID (if future knowledge supports this conclusion, then asymptomatic infection could be treated as a competing risk).

2. Non-severe COVID is not a competing risk for severe COVID. At presentation, if the COVID event does not qualify as severe, then post-COVID diagnosis follow-up is required to determine whether the endpoint registers as non-severe or severe. One will only know the endpoint is not severe after post-COVID diagnosis follow-up is completed (symptoms resolve), such that the failure time is not known until the end of post COVID diagnosis follow-up. Therefore, non-severe COVID is not a competing risk for severe COVID, and the severe COVID endpoint can be analyzed using ordinary survival analysis ignoring the non-severe COVID endpoint.

In sum, the COV-INF, COVID, and severe COVID endpoints will be analyzed by ordinary survival analysis methods, whereas the asymptomatic infection

and non-severe COVID endpoints will be analyzed using competing risks methods. Moreover, adding nomenclature precision, for the parent infection endpoint, the daughter endpoints COVID and asymptomatic infection are semi-competing risks data (nomenclature in the survival analysis literature), and for the COVID parent endpoint, the daughter endpoints severe COVID and non-severe COVID are semi-competing risks data.

In addition, one non-clinical endpoint may be important for correlates assessment: SARS-CoV-2 viral load at COVID diagnosis (VL Dx) (e.g., measured by nasal swab), or alternatively area under the viral load curve (AUC-VL) from the COVID diagnosis date through to undetectable viral load, or to an alternative threshold indicating low viral load. Viral load endpoints are putative surrogates of disease progression and severity for the individual, and are also putative surrogates for secondary transmission; moreover the quantitative nature of viral load endpoints may afford an opportunity to increase statistical power.

Issue 3: Coarseness level of the failure time variable

1. **COVID:** Event time defined in 'continuous time' on the day of virological confirmation.
2. **Asymptomatic infection:** Event time defined only at fixed infrequent visits (e.g., Month 6, 12, 18, 24).
3. **COV-INF:** Event time defined as 'mixed continuous and discrete', equal to the day of virological confirmation (if COVID) and by the first seropostive visit (if asymptomatic infection).
4. **Non-severe COVID:** Event time may be defined in continuous time, as the number of days from enrollment to COVID diagnosis plus the number of additional days until the COVID event is known to be non-severe. However, following the decision made for Issue 1, we simplify and define the event time at COVID diagnosis.
5. **Severe COVID:** Event time may be defined in continuous time, as the number of days from enrollment to COVID diagnosis plus the number

of additional days until the COVID event is known to be severe (which may be zero days). However, following the decision made for Issue 1, we simplify and define the event time at COVID diagnosis.

Issue 4: Binary endpoint vs. failure time endpoint

In general, in phase 3 trials with prospective follow-up for event occurrence where right-censoring occurs (either due to administrative censoring or loss to follow-up), it can be advantageous to conduct data analysis in a survival analysis paradigm. Many of the correlates analyses are specified as such. However, because the endpoints are rare, and the rate of loss to follow-up is anticipated to be very low, reliable and interpretable answers may be obtained based on simpler methods that use binary endpoints, and deal with loss to follow-up in a cruder way. If retention is very high, such that bias and precision may be minimally impacted by use of a binary endpoint, some of the correlates analyses may use a binary endpoint. In settings with competing risks, such analyses would treat the endpoint as multinomial and utilize methodology accordingly.

In sum, correlates methods are needed that consider time-to-event or binary endpoints, with or without accounting for a competing risk. In addition, the methods need to be able to handle continuous, discrete, and mixed continuous/discrete failure times.

8 Case-cohort Sampling Design for Measuring Antibody Markers

Figure 3 illustrates the case-cohort (Prentice, 1986) sampling design that is used for measuring Day 1, 57 antibody markers (and the later time points at a later point in time) in a random sample of trial participants. The random sample is stratified by the key baseline covariates: assigned randomization arm, baseline SARS-CoV-2 status (defined by serostatus and possibly also NAAT and/or RNA PCR testing), any additional important demographic factors such as the randomization strata (e.g., defined by age and/or comorbidities), and underrepresented minority status within the U.S. Because

the design uses a stratified random sample instead of the simple random sample proposed by Prentice (1986), the design may also be referred to as a "two-phase sampling design" (Breslow et al., 2009b,a), where "phase one" refers to variables measured in all participants and "phase two" refers to variables only measured in a subset (thus the "case-cohort sample" constitutes the phase-two data).

The case-cohort design enables obtaining marker data (Day 1, 57) for the immunogenicity subcohort during early trial follow-up in real-time batches, thereby accelerating the time until final data set creation and hence data analysis and results on Day 57 marker correlates. The design allows using the same immunogenicity subcohort to assess correlates for multiple endpoints, relevant for the COVID-19 VE trials with multiple endpoints (Figure 1). This makes the design operationally simpler than a case-control sampling design.

8.1 Prototype USG COVID-19 Response Team immunogenicity subcohort

Table 5 summarizes the size of the prototype USG COVID-19 Response Team immunogenicity subcohort, by baseline factors used to stratify the random sampling. In this subcohort 6 baseline demographic strata are used; if a trial has a different number of baseline demographic strata, then the table would be modified, holding the total sample size of the subcohort approximately fixed. For U.S. strata, all USG COVID-19 Response Team trials specify 50:50 balance by underrepresented minority status Yes:No. The subcohort sampling is implemented to create representative sampling across the entire period of enrollment. Non-USG COVID-19 Response Team trial sampling designs would likely be similar, with different baseline sampling strata perhaps even the simplest case with no sampling strata and use of simple random sampling. 'At-risk' refers to participants considered to be at heightened risk of severe COVID-19 illness based on a specified list of conditions.

Table 5: Planned Immunogenicity Subcohort Sample Sizes by Baseline Strata for Antibody Marker Measurement

	Baseline SARS-CoV-2 Negative[2]						Baseline SARS-CoV-2 Positive[3]					
Bas. Cov. Strata[1]	1	2	3	4	5	6	1	2	3	4	5	6
Vaccine	150	150	150	150	150	150	50	50	50	50	50	50
Placebo	20	20	20	20	20	20	50	50	50	50	50	50

[1]This schema specifies 6 baseline covariate strata for stratified sampling, for example 1 = U.S. Age 18-64 Minority; 2 = U.S. Age 18-64 non-Minority; 3 = U.S. Age $\geq$ 65 Minority; 4 = U.S Age $\geq$ 65 non-Minority; 5 = non-U.S. Age 18-64; 6 = non-U.S. Age $\geq$ 65.

[2]The vaccine group baseline negative strata are assigned large sample sizes because the correlates of risk analysis focuses on baseline negative vaccine recipients. The placebo group baseline negative strata are assigned small sample sizes given the expectation that almost all Day 57 bAb and nAb readouts will be negative/zero given the absence of prior exposure to SARS-CoV-2 antigens.

[3]Equal stratum sizes are assigned for the vaccine and placebo groups in order to compare bAb and nAb responses in previously infected persons, studying potential differences in natural+vaccine-elicited responses vs. natural-elicited responses.

If certain strata do not have enough eligible participants available for sampling, then additional sampling is done from other strata to keep the total immunogenicity subcohort sample size to close to 1620. A separate USG COVID-19 Response Team Antibody Marker sampling plan describes the algorithm, available upon request.

8.1.1 Additional sampling of participants missing the second dose

For trials with two doses of vaccine, the set of baseline negative strata for vaccine recipients is expanded to also include the subgroup that misses the second vaccination and minimally has available samples at Day 57, and similarly the corresponding subgroup to placebo recipients is added; these subgroups are defined regardless of baseline demographic factors. For the vaccine subgroup, a random sample of 150 (or the number available, whichever is smaller) participants is sampled, and for the placebo subgroup, a random sample of 20 (or the number available, whichever is smaller) participants is sampled. This additional sample may not be drawn until the last participant reaches 1 or 2 years of follow-up, to be able to ensure that sampled participants have many samples available.

The two objectives of the additional sampling are: (1) To compare long-term antibody responses one dose vs. two doses; and (2) to increase power of the Stage 2 analysis of outcome-proximal correlates.

8.2 Correlates Objectives Addressed in Two Stages

Figure 4 depicts the two stages of the immune correlates analyses. Stage 1 includes antibody marker data from all COVID and COV-INF cases diagnosed through to the last date of: (1) the time that at least 25 evaluable vaccine breakthrough COVID endpoint cases are available for analysis; and (2) the time of a data-cut at or after the primary analysis used to define the data base for the first correlates analysis. Only Day 1, 57 antibody markers, and COVID and COV-INF diagnosis time point antibody markers, are measured in Stage 1. The objectives of Stage 1 correlates analyses focus on Day 57 markers, which are the objectives listed in Table 1. Stage 1 focuses on Day 57 markers because in general validated or non-validated surrogate endpoints for approved vaccines are based on the peak antibody time point, and this approach fits the priority to develop a validated or non-validated surrogate endpoint as rapidly as possible.

Stage 2 includes antibody marker data from all COVID and COV-INF cases diagnosed after the Stage 1 cases through to the end of the trial, including all available sampling time points (6–7 time points). For immunogenicity subcohort participants, the antibody markers at all available time points other than Day 1, 57 are measured for Stage 2 correlates analyses (4–5 additional time points). The Stage 2 clinical endpoint data and antibody marker data enable assessment of longitudinal antibody markers as outcome-proximal correlates of instantaneous endpoint risk and as various types of outcome-proximal correlates of protection.

The Stage 1 immunogenicity subcohort sampling plan is finalized prior to or shortly after study start. The Stage 2 sampling plan is not made until after the results on vaccine efficacy at the primary analysis are known. The Study Oversight Group may modify the scope of the set of samples for immunoassay measurements in Stage 2 based on analysis results. The essential distinguishing mark of Stage 1 vs. Stage 2 is assessment of Day 57 marker

correlates that can be done using antibody data only from Day 1, 57 markers vs. assessment of outcome-proximal correlates that requires antibody data longitudinally including at endpoint diagnosis dates.

8.2.1 Prioritize antibody marker measurement at COVID and COV-INF diagnosis sampling time points

Conduct of the immunologic assays on diagnosis date samples for all COVID and all COV-INF endpoint samples is of the highest priority, equal to the priority of conducting the assays on the Day 1, 57 samples.

8.2.2 [With Day 29 markers]

Moreover, if Day 29 markers are included, then Stage 1 also focuses on Day 29 markers because if a correlate based on this time point is found to perform as well as a Day 57 correlate, then it may be preferred given the practical advantage to be measured earlier and to not require a Day 57 post-vaccination visit and blood draw. Another advantage of an earlier measurement is providing opportunity to include additional breakthrough COVID endpoint cases (intercurrent endpoints) in the correlates analyses.

9 Unsupervised Feature Engineering of Antibody Markers (Stage 1: Day 1, 57)

9.1 Descriptive Tables and Graphics

9.1.1 Antibody marker data

Binding antibody titers to full length SARS-CoV-2 Spike protein, to the RBD domain of the Spike protein, and to the Nucleocapsid (N) protein will be measured in all participants in the immunogenicity subcohort (augmented with infected cases). N-specific binding antibody titers are not used for correlates analyses or for graphical reporting; these data are only used for tabular reporting. Binding antibody IgG Spike, IgG RBD, IgG N, as well as fold-rise in these three markers from baseline, are measured at each pre-defined time point. Indicators of 2-fold rise and 4-fold rise in IgG concentration (fold rise [post/pre] ≥ 2 and ≥ 4, 2FR and 4FR) are measured at each pre-defined

post-vaccination timepoint. Binding antibody responders to a given antigen at each pre-defined timepoint are defined as participants with value above the antigen-specific positivity cut-off. Binding antibody IgG 2FR (4FR) at each pre-defined timepoint to a given antigen are defined as participants who had baseline values below the LLOQ with IgG concentration at least 2 times (4 times) above the assay LLOQ, or as participants with baseline values above the LLOQ with at least a 2-fold (4-fold) increase in IgG concentration.

Pseudovirus neutralizing antibody ID50 and cID80 titers, as well as fold-rise in ID50 and cID80 titers from baseline, are measured at each pre-defined time point. Indicators of 2-fold rise and 4-fold rise in ID50 titer (fold rise [post/pre] ≥ 2 and ≥ 4, 2FR and 4FR) are measured at each pre-defined post-vaccination timepoint.

Neutralization responders at each pre-defined timepoint are defined as participants who had baseline values below the LLOD with detectable ID50 neutralization titer above the assay LLOD, or as participants with baseline values above the LLOD with a 4-fold increase in neutralizing antibody titer. Neutralization 2FR (4FR) at each pre-defined timepoint are defined as participants who had baseline values below the LLOQ with ID50 at least 2 times (4 times) above the assay LLOQ, or as participants with baseline values above the LLOQ with at least a 2-fold (4-fold) increase in neutralizing antibody titer. While quantitative fold-rise is shown for both ID50 and cID80, response above LLOD, 2FR and 4FR responder status are shown only for ID50. (However, for superlearner analysis of multivariable CoRs, 2FR and 4FR responder status variables are included for each of pseudovirus-nAb ID50 and cID80, given the objectives of more comprehensive analysis in building the estimated optimal surrogate.)

For WT live virus-nAb MN50, the same types of variables are analyzed/reported as for pseudovirus-nAb ID50.

Note that for defining positive response, 2FR, and 4FR, a reason why values below the LLOD are set to half the LLOD before calculating the indicator of response, is to ensure that a vaccine recipient that has an unusually low antibody readout at baseline and a post-vaccination value below or near the

LLOD is not erroneously counted as a responder.

The following list describes the antibody variables that are measured from immunogenicity subcohort and infection case participants.

1. Individual anti-Spike antibody concentration at each pre-defined time point
2. Individual anti-Spike antibody fold-rise concentration post-vaccination relative to baseline at each pre-defined post-vaccination time point
3. Individual anti-RBD antibody concentration at each pre-defined time point
4. Individual anti-RBD antibody fold-rise post-vaccination relative to baseline at each pre-defined post-vaccination time point
5. Individual anti-N antibody concentration at each pre-defined time point
6. Individual anti-N antibody fold-rise post-vaccination relative to baseline at each pre-defined post-vaccination time point
7. 2-fold-rise and 4-fold rise (fold rise in anti-Spike antibody concentration [post/pre] ≥ 2 and ≥ 4, 2FR and 4FR) at each pre-defined post-vaccination time point
8. 2-fold-rise and 4-fold rise (fold rise in anti-RBD antibody concentration [post/pre] ≥ 2 and ≥ 4, 2FR and 4FR) at each pre-defined post-vaccination time point
9. 2-fold-rise and 4-fold rise (fold rise in anti-N antibody concentration [post/pre] ≥ 2 and ≥ 4, 2FR and 4FR) at each pre-defined post-vaccination time point
10. Pseudovirus-nAb responders, at each pre-defined timepoint defined as participants who had baseline values below the LLOQ with detectable pseudovirus-nAb ID50 titers above the assay LLOQ or as participants with baseline values above the LLOQ with a 4-fold increase in pseudovirus-nAb ID50 titers
11. Wild type live-virus-nAb responders, at each pre-defined timepoint de-

fined as participants who had baseline values below the LLOQ with detectable WT live virus-nAb MN50 titers above the assay LLOQ or as participants with baseline values above the LLOQ with a 4-fold increase in WT live virus-nAb MN50 titers

Summaries of the immunogenicity data will be reported in tables. In particular, the tables will include, for each pre-defined post-baseline time point:

1. For each binding antibody marker, the estimated percentage of participants defined as responders, and with concentrations $\geq$ 2x LLOQ or $\geq$ 4 x LLOQ, will be provided with the corresponding 95% CIs using the Clopper-Pearson method.

   In addition, the estimated percentage of participants defined as responders, participants with 2-fold rise (2FR), and participants with 4-fold rise (4FR) will be provided with the corresponding 95% CIs using the Clopper-Pearson method.

2. For the ID50 pseudo-virus neutralization antibody marker, the estimated percentage of participants defined as responders, participants with 2-fold rise (2FR), and participants with 4-fold rise (4FR) will be provided with the corresponding 95% CIs using the Clopper-Pearson method

3. For the MN50 WT live virus neutralization antibody marker, the estimated percentage of participants defined as responders, participants with 2-fold rise (2FR), and participants with 4-fold rise (4FR) will be provided with the corresponding 95% CIs using the Clopper-Pearson method

4. Geometric mean titers (GMTs) and geometric mean concentrations (GMCs) will be summarized along with their 95% CIs using the t-distribution approximation of log-transformed concentrations/titers (for each of the 5 Spike-targeted marker types including pseudovirus-nAb ID50 and cID80 and WT live virus-nAb MN50, as well as for binding Ab to N).

5. Geometric mean titer ratios (GMTRs) or geometric mean concentration ratios (GMCRs) are defined as geometric mean of individual titers/concentration ratios (post-vaccination/pre-vaccination for each injection)

6. GMTRs/GMCRs will be summarized with 95% CI (t-distribution ap-

proximation) for any post-baseline values compared to baseline, and post-Day 57 values compared to Day 57

7. The ratios of GMTs/GMCs will be estimated between groups with the two-sided 95% CIs calculated using t-distribution approximation of log-transformed titers/concentrations [the groups compared are vaccine recipient Non-Cases vs. vaccine recipient breakthrough cases used for Day 57 marker correlates analyses (Primary cases) and vaccine recipient Non-Cases vs. vaccine recipient breakthrough cases used for Day 29 marker correlates analyses (Intercurrent cases and Primary cases)].

8. The differences in the responder rates, 2FRs, 4FRs between groups will be computed along with the two-sided 95% CIs by the Wilson-Score method without continuity correction (Newcombe, 1998) (the groups for comparison are as described in the previous bullet).

All of the above point and confidence interval estimates will use inverse probability of antibody marker sampling weighting in order that estimates and inferences are for the population from which the whole study cohort was drawn. In two-phase sampling data analysis nomenclature, the "phase 1 ptids" are the per-protocol individuals excluding individuals with a COVID failure event or any other evidence of SARS-CoV-2 infection < 7 days post Day 57 visit. The "phase 2 ptids" are then the subset of these phase 1 ptids in the immunogenicity subcohort with Day 1 and 57 Ab marker data available. Thus, marker data for the COVID endpoint cases outside the subcohort will not be used in immunogenicity analyses; these cases are excluded from immunogenicity analyses. Thus again, marker data for the COVID endpoint cases outside the subcohort will not be used in immunogenicity analyses; these cases are excluded from immunogenicity analyses.

The estimated weight $\hat{w}_{subcohort.57x}$ is the inverse sampling probability weight, calculated as the empirical fraction (No. phase 1 ptids / No. phase 2 ptids) within each of the baseline strata [(vaccine, placebo) $\times$ (baseline negative, baseline positive) $\times$ (demographic strata)]. For individuals outside the phase 1 ptids, $\hat{w}_{subcohort.57x}$ is assigned the missing value code NA. All other individuals have a positive value for $\hat{w}_{subcohort.57x}$, including cases not in the subcohort.

This weight is only used for case outcome-status blinded immunogenicity inferential analyses. Note that $\hat{w}_{subcohort.57x}$ is used for all immunogenicity analyses, which are based solely on the immunogenicity subcohort, for Day 1 and Day 57 markers, and Day 29 markers if included (not used for correlates analyses).

Tables will be provided separately for (1) baseline negative individuals, (2) baseline positive individuals, (3) baseline negative individuals by subgroup defined as in Table 6, and (4) baseline positive individuals by the same subgroups as in (3). Each table will show data for all available time points and for each of the vaccine and placebo arms.

Table 6: Baseline Subgroups that are Analyzed (May Vary Slightly by Protocol)[1].

Age: < 65, ≥ 65
Heightened Risk for Severe COVID: At risk, Not at risk
Age x Risk for Severe COVID:
< 65 At risk, < 65 Not at risk, ≥ 65 At risk, ≥ 65 Not at risk
Sex Assigned at Birth: Male, Female
Age x Sex Assigned at Birth:
< 65 Male, < 65 Male, ≥ 65 Female, ≥ 65 Female
Hispanic or Latino Ethnicity: Hispanic or Latino, Not Hispanic or Latino
Race or Ethnic Group:
White Non-Hispanic[2], Black, Asian, American Indian or Alaska Native (NatAmer)
Native Hawaiian or Other Pacific Islander (PacIsl), Multiracial,
Other, Not reported, Unknown
Underrepresented Minority Status in the U.S.:
Communities of color (Comm. of color), White[2]
Age x Underrepresented Minority Status in the U.S.:
Age ≥ 65 Comm. of color, Age < 65 Comm. of color, Age ≥ 65 White, Age ≥ 65 White

[1]All analyses are done within strata defined by randomization arm and baseline positive/negative status, such that these variables are not listed here as subgroups for analysis.

[2]White Non-Hispanic is defined as Race=White and Ethnicity=Not Hispanic or Latino. All of the other Race subgroups are defined solely by the Race variable, with levels Black, Asian, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Multiracial, Other, Not reported, Unknown.

For comparing antibody levels between groups, the following groups are compared:

- Baseline negative vaccine vs. baseline negative placebo
- Baseline positive vaccine vs. baseline positive placebo
- Baseline negative vaccine vs. baseline positive vaccine
- Within baseline negative vaccine recipients, compare each of the following pairs of subgroups listed in Table 6: Age ≥ 65 vs. age < 65; risk for severe COVID: at risk vs. not at risk; age ≥ 65 at risk vs. age ≥ 65 not at risk; age < 65 at risk vs. age < 65 not at risk; male vs. female; Hispanic or Latino ethnicity: Hispanic or Latino vs. Not Hispanic or Latino; Underrepresented minority status: Communities of color vs. White non-Hispanic (within the U.S.).

The entire immunogenicity analysis is done in the per-protocol cohort with both Day 1 and Day 57 marker data available (the two-phase sample).

9.1.2 [With Day 29 markers]

If Day 29 markers are included, then participants in the immunogenicity subcohort and outcome cases will have marker data at Day 1, 29, 57. All immunogenicity analyses (graphical and tabular) of Day 57 markers are also conducted for Day 29 markers.

9.1.3 Graphical description of antibody marker data

The Day 1, 57 antibody marker data collected from the immunogenicity subcohort participants will be described graphically. These data are representative of the entire study cohort. Importantly, only antibody data from the immunogenicity subcohort are included (i.e., no data from cases outside the subcohort are included). This makes the analyses unsupervised (independent of case-control status), enabling interrogation and optimization of the antibody biomarkers prior to the inferential correlates analyses.

Plots are developed for the following purposes. All of the analyses are done separately within each of the four subgroups defined by randomization arm cross-classified with baseline negative/positive status. In addition, many of the descriptive analyses will also be done separately for each demographic

subgroup of interest listed above. For descriptive plots of individual marker data points that pool over one or more of the baseline strata subgroups, plots show all observed data points.

For each antibody marker readout, both Day 57 and baseline-subtracted Day 57 readouts are of interest. We will refer to the latter as 'delta.' All readouts, including delta, will be plotted on the $\log_{10}$ scale, with plotting labels on the natural scale. As such, delta is $\log_{10}$ fold-rise in the marker readout from baseline.

The following descriptive graphical analyses are done.

1. The distribution of each antibody marker readout at Day 1 and Day 57 will be described with plots of empirical reverse cumulative distribution functions (rcdfs) and boxplots (including individual data points) within each of the four groups defined by randomization arm (vaccine, placebo) and baseline positivity stratum (negative, positive). Inverse probability of sampling into the subcohort weights are used in the estimation of the rcdf curves; henceforth we refer to these weights ($\hat{w}_{subcohort.57x}$) are used in the estimation of the rcdf curves; henceforth we refer to these weights as "inverse probability of sampling" (IPS) weights. Analyses of Day 1 markers always pool across vaccine and placebo recipients given that the two subgroups are the same at baseline.

2. Plots are arranged to compare each Day 57 marker readout between randomization arms within each of the baseline seropositive and baseline seronegative subgroups.

3. Plots are also arranged to compare each Day 57 marker readout between baseline serostatus groups within each randomization arm.

4. The correlation of each antibody marker readout between Day 1 and Day 57, and between Day 1 and delta, is examined within each of the baseline strata subgroups, and within each randomization arm and baseline positivity stratum. Pairs plots/scatterplots will be used, annotated with baseline strata-adjusted Spearman rank correlations, implemented in the PResiduals R package available on CRAN. For calculating the correlation

within each randomization arm and baseline positivity stratum, because PResiduals does not currently handle sampling weights, the correlation estimates are computed as follows: For each re-sampled data set in the second approach to graphical plotting, the covariate-adjusted Spearman correlation is calculated. The average of the estimated correlations across re-sampled data sets is reported.

5. The correlation of each pair of Day 1 antibody marker readouts are compared within each baseline positivity stratum, pooling over the two randomization arms. Pairs plots/scatterplots and baseline-strata adjusted Spearman rank correlations are used, with covariate-adjusted Spearman rank correlations computed as described above.

6. The correlation of each pair of Day 57 and delta antibody marker readouts are compared within each randomization arm within each baseline positivity stratum. Pairs plots/scatterplots and baseline-strata adjusted Spearman rank correlations are used, with covariate-adjusted Spearman rank correlations computed as described above.

7. Point estimates of Day 57 marker positive response rates for the vaccine arm by the baseline demographic subgroups and the baseline serostrata are provided, as well as pooled over baseline demographic strata. The point and 95% CI estimates include all of the data and use IPS weights.

9.2 Methods for Positive Response Calls for bAb and nAb Assays

As noted above, binding antibody responders at each pre-defined timepoint are defined as participants with concentration above the specified positivity cut-off, with a separate cut-off for each antigen Spike, RBD, N (10.8424, 14.0858, and 23.4711, respectively, in BAU/ml).

Pseudovirus neutralization responders at each pre-defined timepoint are defined as participants who had baseline ID50 values below the LLOD with detectable ID50 neutralization titer above the assay LLOD, or as participants with baseline values above the LLOD with a 4-fold increase in neutralizing antibody titer. Otherwise a value is negative for pseudovirus neutralization. The same approach is used based on cID80 titer. Similarly, for the WT live

virus-nAb MN50 marker, WT live virus neutralization responders at each pre-defined timepoint are defined as participants who had baseline MN50 values below the LLOQ with detectable MN50 above the assay LLOQ or as participants with baseline values above the LLOQ with a 4-fold increase in neutralizing antibody titer. Otherwise a value is negative for WT live virus neutralization.

9.3 SARS-CoV-2 Antigen Targets Used for bAb and nAb Markers

The homologous vaccine strain antigens are used for the immune correlates analyses for the bAb markers, whereas the homologous vaccine strain with D614G mutation is used for the pseudovirus nAb markers.

9.4 Score Antibody Markers Combining Information Across Individual bAb and/or nAb Readouts

Depending on the number and features of antigens that are selected for defining antibody marker variables, feature extraction/selection techniques may be employed to determine score/synthesis marker variables that are optimized according to some criterion that would reflect maximum signal-relevant diversity (e.g., He and Fong, 2019). In addition, the unsupervised dimensionality reduction techniques such as principal components analysis (PCA) and nonlinear extensions of PCA (e.g. FSDAM1 and FSDAM2; Fong and Xu, 2021) may also be used to define score variables that maximally capture the main immune response signal and to study whether there are more than one distinct signals that are associated with the outcome. If such synthesis features are defined, then they will be included as input features in the machine learning (superlearning) prediction modeling (multivariable CoR objective).

The purpose of the score markers is to seek to maximally capture the main immune response signal and to study whether score markers can provide strengthened association with outcome compared to the individual assay markers.

9.4.1 Systematic ranking of Day 57 antibody markers by signal-to-noise ratio

The signal-to-noise ratio of each Day 57 antibody marker is defined as the ratio of biological variability over technical variability. If the requisite data are available, the technical variability will be estimated as the median of the variances across two technical replicates for each test sample, and the biological variability will be estimated as the variance of the average of the two technical replicates across all test samples (without weighting for simplification) minus the technical variability (analysis done in the cohort of interest such as baseline negative vaccine recipients).

The ranking of the set of Day 57 antibody markers will be taken into account in the interpretation of results.

9.5 Decisions on Antibody Markers to Advance to Correlates or Risk and Correlates of Protection Analyses

The vaccine immunogenicity analysis characterizes SARS-CoV-2 directed antibody levels based on five antibody biomarkers measured at each blood storage time point: IgG concentration to Spike, IgG concentration to RBD, pseudovirus-nAb ID50, pseudovirus-nAb cID80, WT live virus-nAb MN50. It is likely that all five of the biomarkers at each of the Day 57 and Day 29 time points will be advanced to study as CoRs and CoPs for the initial correlates reports, given that all of the assays are validated. However, an objective of the unsupervised learning immunogenicity characterization is to determine if some markers should be prioritized (for example based on broader biologically-relevant dynamic range), such that p-values and multiplicity adjustment for tests of correlates of risk would only be done for the prioritized markers. In addition, it is possible that the unsupervised learning could lead to decisions to pare down the list of markers (e.g., eliminating markers that are very highly correlated with other markers, or eliminating markers that are revealed to have unexpected technical issues). Because the unsupervised learning is done based on immunogenicity subcohort data and is thus independent of case/non-case status, decisions made based on this learning do not compromise the validity of the CoR and CoP analyses. Decisions about the set of antibody makers to use in the CoR and CoP analyses

– and their priority level – will be made and documented in the SAP prior to implementing the CoR and CoP analyses. Unless otherwise noted, the multiplicity adjustment is applied to the full set of markers analyzed.

10 Baseline Risk Score (Proxy for SARS-CoV-2 Exposure)

A list of baseline covariates potentially relevant for SARS-CoV-2 exposure and infection risk will be specified. Based on these covariates, a baseline risk score is developed and controlled for in correlates analyses to adjust for potential confounding. The risk score is defined as the logit of the predicted outcome probability from a regression model estimated using the ensemble algorithm superlearner (i.e. stacking), where this logit predicted outcome is scaled to have empirical mean zero and empirical standard deviation one. The settings of superlearner (i.e., loss function, cross-validation technique, library of learners) that are used for implementation of superlearner for building a baseline risk score are described in Section 12.7.

The development of risk score will involve training the superlearner using placebo arm data and predictions made on vaccine arm data (CV-predictions will be made on placebo arm data). In both arms, risk score development will be restricted to baseline negative per-protocol subjects with cases as COVID endpoints starting post-enrollment. The CV-prediction performance of superlearner (CV-AUC calculation and CV-ROC curves) will be derived with cases as COVID endpoint starting post-enrollment as well. The prediction performance of superlearner (AUC calculation and ROC curve) in the vaccine arm, however, will be restricted to the same set of vaccine recipients as used in the correlates analyses with cases considered as COVID endpoints starting 7 days post second vaccination visit and non-cases as participants with follow-up beyond 7 days post second vaccination visit and never registered a COVID endpoint.

Independent of the superlearner risk score, important individual risk factors will also be specified for inclusion as adjustment factors in correlates analyses, such as age, status of having a high-risk condition, and communities of color status. For example, all or a subset of the baseline demographic strata used

in the two-phase sampling design will be adjusted for, or a coarsened categorical variable derived from the baseline strata will be adjusted for, where the amount of coarsening may depend on the number of endpoints in the vaccine arm.

Henceforth we refer to the baseline variables that are adjusted for in correlates analyses as "baseline factors" which, depending on the risk score results and performance, will consist of only the individual key risk factors, or key individual risk factors plus the baseline risk score.

If a fully automated/pre-specified approach to defining the baseline factors is required, then the following approach will be used: Advance the risk score, the at-risk indicator, and the communities of color indicator. This choice is justified by the epidemiological data showing that these two indicators are strong infection and COVID-19 risk factors, and making use of the flexibility of super learner to develop a model for how age relates to risk.

11 Correlates Analysis Descriptive Tables by Case/Non-Case Status

The key tables summarizing the distribution of each of the five antibody markers are listed below. For each table, for each time point Day 1, Day 57 separately, the positive response rate with 95% CI, and the GMT or GMC with 95% CI, is reported for each of the case and non-case groups. In addition, the point and 95% CI estimate of the difference in positive response rate (non-cases vs. cases) and the GMT or GMC ratio (non-cases/cases), is reported.

1. Antibody levels in the baseline SARS-CoV-2 negative per-protocol cohort (vaccine recipients). Cases are baseline negative per-protocol vaccine recipients with the symptomatic infection COVID-19 primary endpoint diagnosed starting 7 days after the Day 57 study visit. Non-cases/Controls are baseline negative per-protocol vaccine recipients sampled into the immunogenicity subcohort with no COVID primary endpoint up to the time of data cut and no evidence of SARS-CoV-2 infection up to six days post Day 57 visit.

2. Antibody levels in the baseline SARS-CoV-2 positive per-protocol cohort (vaccine recipients). Cases are baseline positive per-protocol vaccine recipients with the symptomatic infection COVID-19 primary endpoint diagnosed starting 7 days after the Day 57 study visit. Non-cases/Controls are baseline positive per-protocol vaccine recipients sampled into the immunogenicity subcohort with no COVID primary endpoint up to the time of data cut and no evidence of SARS-CoV-2 infection up to six days post Day 57 visit.

3. Antibody levels in baseline SARS-CoV-2 positive placebo recipients. Cases are baseline positive per-protocol placebo recipients with the symptomatic infection COVID-19 primary endpoint diagnosed any time after Day 1 or after Day 57 (by time of antibody measurement). Non-cases/Controls are baseline positive per-protocol placebo recipients sampled into the immunogenicity subcohort with no COVID primary endpoint up to the time of data cut.

4. Repeat Table 2 above for fold-rise from baseline (of interest given the analysis cohort is baseline positive).

5. Repeat Table 3 above for fold-rise from baseline (of interest given the analysis cohort is baseline positive).

The point and confidence interval estimates are computed using inverse probability sampling weights $\hat{w}_{subcohort.57x}$ defined in Section 12.3.1.

11.0.1 [With Day 29 markers]

If Day 29 markers are included, then two cases vs. non-cases comparisons are done: primary cases vs. non-cases and intercurrent+primary cases vs. non-cases. Non-cases/Controls are baseline negative per-protocol vaccine recipients sampled into the immunogenicity subcohort with no COVID primary endpoint by the time of data cut and no evidence of SARS-CoV-2 infection up to six days post Day 57 visit. Thus the same set of Non-cases are used for the two comparisons. In addition, descriptive plots show intercurrent cases and primary cases separately. For intercurrent cases the weights $\hat{w}_{intercurrent.x}$ are used, described in Section 12.3.2.

12 Correlates of Risk Analysis Plan

At first, this analysis plan for CoRs and CoPs as currently written focuses on the COVID primary endpoint, with its continuous failure times (failure time defined by the day of the event) and no competing risks. Later, it will be extended to handle the special issues with secondary endpoints.

12.1 CoR Objectives

The following CoR objectives are assessed in baseline negative per-protocol vaccine recipients:

1. **Univariable CoR** To assess each individual Day 57 antibody marker as a CoR of outcome in vaccine recipients, adjusting for baseline factors (See Section 10)
2. **Multivariable CoR** To build models predictive of outcome based on a set of Day 57 antibody marker readouts, adjusting for baseline factors (See Section 10)

12.1.1 [With Day 29 markers]

If Day 29 markers are included, then a Univariable CoR objective is added, the same as above, except using the Day 29 versions of the markers instead of the Day 57 versions.

In addition, the Multivariable CoR objective is repeated to build models predictive of outcome based on a set of Day 29 antibody marker readouts. It is also repeated to build models predictive of outcome based on a set of Day 29 and 57 antibody marker readouts used together.

12.2 Outline of the Set of CoR Analyses

The univariable CoR objective is addressed by Cox proportional hazards regression and nonparametric threshold regression. The multivariable CoR objective is addressed by superlearning. All of these analyses are implemented in automated and reproducible press-button fashion.

In addition, supportive exploratory analyses of the univariable CoR objective are conducted using flexible parametric regression modeling: hinge/threshold regression and generalized additive model regression.

12.3 Day 57 Markers Assessed as CoRs and CoPs

The following five Day 57 markers are assessed as CoRs and CoPs, usually as quantitative variables and in some analyses as ordered trinary variables or binary variables, all of which do not subtract Day 1 (baseline) values:

1. binding Ab to Spike (IgG BAU/ml)
2. binding Ab to RBD (IgG BAU/ml)
3. pseudovirus neutralization ID50 (IU)
4. pseudovirus neutralization cID80 (IU)
5. live virus neutralization MN50

For all univariable CoR analyses (first objective), the non-baseline subtracted versions of the Day 57 antibody markers are studied; the baseline-subtracted versions are not studied given that the analyses are done in the baseline negative cohort for which Day 1 readouts will generally be negative. The multivariable machine learning CoR analyses include synthesis markers that combine information across the individual markers listed above, as well as including 2FR and 4FR versions of variables.

12.3.1 Inverse probability sampling weights used in CoR analyses

In section 9.1, estimated inverse probability sampling (IPS) weights $\hat{w}_{subcohort.57x}$ were defined for per-protocol immunogenicity subcohort members, for the purpose of immunogenicity analyses. This section describes the IPS weight used for Day 57 marker correlates analyses ($\hat{w}_{57.x}$).

Consider the correlates analyses of Day 57 markers. For baseline sampling stratum x [(vaccine, placebo) $\times$ (demographic strata)], the IPS weight $w_{57.x}$ assigned to a non-case participant in stratum x is defined by $\hat{w}_{57.x} = 1/\hat{\pi}_{57}(x) = N_x/n_x$, where N_x is the number of stratum x vaccine recipient non-

cases in the Per-Protocol Baseline Negative (PPBN) cohort and n_x is the number of these participants that also have Day 1, 29, and 57 marker data available, where participants with any evidence of SARS-CoV-2 infection before 7 days post Day 57 visit are excluded from the counts N_x and n_x. For non-case participant i in the immunogenicity subcohort, $\hat{w}_{57.i} = 1/\hat{\pi}_{57}(X_i)$ denotes the weight $\hat{w}_{57.x}$ for this individual's sampling stratum. All Primary cases are assigned sampling weight N_1/n_1 where N_1 is the total number of vaccine recipient cases in the PPBN cohort restricting to cases with event time starting 7 days post Day 57, and n_1 is the number of these participants that also had the Day 1, 29, and 57 markers measured, and again participants with any evidence of SARS-CoV-2 infection $<$ 7 days post Day 57 visit are excluded from the counts N_x and n_x.

In terms of two-phase sampling data analysis nomenclature, for these Day 57 marker analyses "phase 1 ptids" are defined as the entire PPBN cohort except excluding participants with any evidence of SARS-CoV-2 infection $<$ 7 days post Day 57 visit (there are expected to be a very small number of such vaccine recipient cases, such that for vaccine recipients the phase 1 ptids is approximately representative of the target population). The "phase 2 ptids" are then the subset of these phase 1 ptids with Day 1 and Day 57 Ab marker data available. Thus the weight $\hat{w}_{57.x}$ is the inverse sampling probability weight, calculated as the empirical fraction (No. phase 1 ptids / No. phase 2 ptids) within each of the baseline negative strata (14 strata defined by PPBN vaccine group cases, PPBN placebo group cases, PPBN vaccine group non-cases divided into the 6 demographic strata, and PPBN placebo group non-cases divided into the 6 demographic strata). For baseline negative individuals outside the phase 1 ptids, $\hat{w}_{57.x}$ is assigned the missing value code NA. All other individuals have a positive value for $\hat{w}_{57.x}$.

12.3.2 [With Day 29 markers]

For baseline sampling stratum x [(vaccine, placebo) $\times$ (demographic strata)], the IPS weight $w_{29.x}$ assigned to a non-case participant in stratum x is defined by $\hat{w}_{29.x} = 1/\hat{\pi}_{29}(x) = N_x/n_x$, where N_x is the number of stratum x vaccine recipient non-cases in the PPBN cohort and n_x is the number of these

participants that also have Day 1 and Day 29 marker data available, where participants with any evidence of SARS-CoV-2 infection before 7 days post Day 29 visit are excluded from the counts N_x and n_x. For non-case participant i in the immunogenicity subcohort, $\hat{w}_{29.i} = 1/\hat{\pi}_{29}(X_i)$ denotes the weight $\hat{w}_{29.x}$ for this individual's sampling stratum. All Intercurrent and Primary cases are assigned sampling weight N_1/n_1 where N_1 is the total number of vaccine recipient cases in the PPBN cohort restricting to cases with event time starting 7 days post Day 29, and n_1 is the number of these participants that also had the Day 1 and Day 29 markers measured, and again participants with any evidence of SARS-CoV-2 infection < 7 days post Day 29 visit are excluded from the counts N_x and n_x.

In terms of two-phase sampling data analysis nomenclature, for the Day 29 marker analyses "phase 1 ptids" are defined as the entire PPBN cohort except excluding participants with any evidence of SARS-CoV-2 infection < 7 days post Day 29 visit. The "phase 2 ptids" are then the subset of these phase 1 ptids with Day 1 and Day 29 Ab marker data available. Thus the weight $\hat{w}_{29.x}$ is the inverse sampling probability weight, calculated as the empirical fraction (No. phase 1 ptids / No. phase 2 ptids) within each of the baseline negative strata (strata defined by PPBN vaccine group cases, PPBN placebo group cases, PPBN vaccine group non-cases divided into the baseline demographic covariate strata, and PPBN placebo group non-cases divided into the demographic covariate strata). For baseline negative individuals outside the phase 1 ptids, $\hat{w}_{29.x}$ is assigned the missing value code NA. All other individuals have a positive value for $\hat{w}_{29.x}$. In sum, the weights $\hat{w}_{29.x}$ are calculated in the same way as the weights $\hat{w}_{57.x}$, except the relevant time window for evidence of infection or COVID is at least 7 days post Day 29 visit instead of at least 7 days post Day 57 visit.

We refer to all COVID cases included in the Day 29 marker analyses but not in the Day 57 marker analyses as "intercurrent cases," where various graphical descriptives show data for this subgroup of cases. The variable $\hat{w}_{intercurrent.x}$ inverse sampling probability weight, calculated as the empirical fraction (No. phase 1 ptids / No. phase 2 ptids) within each of the baseline negative strata. For baseline negative participants outside the phase 1 ptids, $\hat{w}_{intercurrent.x}$ is

assigned the missing value code NA. All other individuals have a positive value for $\hat{w}_{intercurrent.x}$. The CoR inferential analyses that study both Day 29 and Day 57 Ab markers in the same analysis use the estimated weights for the Day 57 marker analyses.

12.3.3 Univariable CoR: Marginalized Cox modeling

Time-to-event methods use the Day 57 visit date as the time origin.

The IPWCC Cox regression model designed for case-cohort sampling designs will be used for estimation and inference on hazard ratios of outcomes by Day 57 marker levels, and for estimation and inference on marginalized marker-conditional cumulative incidence over time. The models will be fit using the *survey* R package available on CRAN, and will adjust for the baseline factors. We use a method from the survey package that assumes without replacement two-phase sampling and not Bernoulli sampling, which matches the sampling design and approach to weight estimation (Lumley, 2010).

For models with a single antibody marker, a two-phase Cox model with improved efficiency through calibrated weights (Breslow et al., 2009b,a) will be used if there are baseline covariates that predict the antibody marker with $R^2 > 0.4$. Based on the phase-two sample, a superlearner model will be fit (using the library specified in Table 7) to predict the antibody marker from the set of collected baseline covariates. The criterion $R^2 > 0.4$ will be checked based on the association of the fitted values from superlearner and the marker, where the association uses held-out marker values.

Based on the Cox model fit to all available data during blinded follow-up (the period during which participants are blinded to randomization arm assignment), a final time point t_F near the time of the last observed outcome will be defined. Let T be the failure time, S a Day 57 marker of interest, and X the vector of baseline factors that are adjusted for. With $S_1(t|s,x) = P(T > t|S = s, X = x, A = 1)$, the Cox model fit yields an estimate of $S_1(t|s,X_i)$ for each individual i in the phase-two sample. The marginalized conditional risk $risk_1(t|s) = E_X[P(T \leq t|s,X,A=1)]$ through time t (for all times t through t_F simultaneously) is estimated based on the

equation

$$risk_1(t|s) \;=\; \int (1 - S_1(t|s,x))dH(x) \tag{1}$$

where $H(\cdot)$ is the distribution of X in $A = 1$ individuals.

The function $risk_1(t|s)$ can be estimated by

$$\widehat{risk}_1(t|s) = \frac{1}{n}\sum_{i=1}^{n}(1 - \hat{S}_1(t|s,X_i)), \tag{2}$$

where n is the number of vaccine arm participants with phase-one data (X_i measured). The bootstrap is used to obtain 95% pointwise confidence intervals for $risk_1(t_F|s)$.

The bootstrap process will be performed by resampling with replacement the subjects within the subcohort and the subjects outside the subcohort separately within each stratum and by resampling with replacement subjects with undetermined stratification variables. Across all bootstrap samples, the number of participants in each stratum in the immunogenicity subcohort remains fixed, but the number of cases does not stay the same.

If the sampling design is case-control instead of case-cohort, then a different bootstrapping procedure is necessary. We will perform bootstrap for case control studies by resampling cases, phase 2 controls, and non-phase 2 controls separately. Across bootstrap replicates, the number of cases does not stay constant, neither do the numbers of ph2 controls by demographics strata. Specifically, the procedure will 1) sample (with replacement) the original phase 1 dataset to get dat.b. From dat.b, take only the cases, but also the counts of phase 2 and non-phase 2 controls by stratum, and 2) sample (with replacement) theses numbers of phase 2 and non-phase 2 controls by strata from the original dataset.

The results of the above Cox modeling will be output in a variety of ways:

1. Plot $\widehat{risk}_1(t_F|s)$ vs. s with 95% CIs for continuous $S = s$ varying over its whole range. Include on the plot the estimate of $\widehat{risk}_0(t_F)$ with a 95%

CI for the placebo arm (horizontal bands), computed by a Cox model marginalizing over the same baseline factors as for the analysis of the vaccine arm.

2. Based on a fit of the Cox model to a nominal categorical antibody marker defined as the tertiles of S, plot $\widehat{risk}_1(t|s)$ for each category of S values with 95% CIs, for all time points t from Day 57 through t_F. If more than 20% of vaccine recipients have S below the LLOD of the assay, then the categories instead will be (1) values $\leq$ LLOD; (2) values below the median of values $>$ LLOD; (3) values above the median of values $>$ LLOD. Include on the plot the estimated curve $\widehat{risk}_0(t)$ with 95% CIs for the placebo arm, computed by a Cox model marginalizing over the same baseline factors as for the analysis of the vaccine arm.

3. Tabular reporting of the hazard ratio per 10-fold change in the quantitative Day 57 antibody marker with 95% confidence interval and 2-sided p-value.

4. Tabular reporting of the hazard ratio for the Middle and Upper categories of the categorical Day 57 antibody marker vs. the Lower category, with 95% confidence interval and 2-sided p-value, as well as a global generalized Wald two-sided p-value for whether the hazard rate of the endpoint varies across the three categories. The table includes the attack rate (with no. of cases / no. at risk) through t_F for each of the three vaccine marker subgroups and for the placebo arm.

5. Report point and 95% CI estimates for the hazard ratio per 10-fold change in the Day 57 antibody marker, for the entire per-protocol baseline negative vaccine cohort and for each of the baseline demographic strata subgroups defined in Table 6 (reported via forest plotting).

6. Westfall-Young (1997) q-values and FWER-adjusted p-values for the generalized Wald tests are included in the table.

Grambsch and Therneau (1994) tests are applied to test the veracity of the proportional hazards assumption. This testing is done to aid interpretation of results, but not as a gateway to trigger the fitting of a more flexible version

of the Cox model, as we seek to avoid computing new confidence intervals and p-values contingent on goodness-of-fit-testing, as they would not have their correct interpretations. Other correlates of risk methods explicitly model risk flexibly as a function of the marker.

The bootstrap is used to calculate 95% pointwise CIs for $risk_1(t_F|s)$ in s. The 2-sided Wald p-value for testing the regression coefficient of the marker in the Cox model provides a valid test of the null hypothesis $H_0 : risk_1(t_F|s) = risk_1(t_F)$ for all s, and is reported.

In addition, the same Cox model analysis will be used to estimate the alternative marginalized conditional risk parameter defined by $risk_1(t|S \geq s)$ where $risk_1(t|S \geq s) = E_X[P(T \leq t|S \geq s, X, A = 1)]$, which can be estimated by

$$\widehat{risk}_1(t|S \geq s) = \frac{1}{n}\sum_{i=1}^{n}(1 - \hat{S}_1(t|S \geq s, X_i)).$$

This parameter is useful because typically subgroups of interest are defined by having marker response above a threshold. We will plot $\widehat{risk}_1(t_F|S \geq s)$ vs. s with 95% CIs for continuous S with s varying over the range of S in which the number of cases to estimate $\hat{S}_1(t|S \geq s, X_i)$ is 5 or more. This type of analysis is also included because it analyzes the same parameter as the nonparametric threshold estimation method described below, providing a way to address the threshold question both by Cox modeling and by nonparametric analysis.

12.3.4 Univariable CoR: Marginalized Cox modeling with influence-function based analytic variance estimation

The previous section described estimation and inference for $risk_1(t|s)$ via the bootstrap. This section describes how to alternatively conduct inference via the influence function.

Data structure and parameter of interest

Consider a survival distribution involving iid data units (Y, Δ, Z), in which $Y \equiv \min\{T, C\}$ is the observed minimum of failure and right-censoring time, $\Delta \equiv I\{T \leq C\}$ is the event indicator, and $Z = (W, A)$ is a vector of covariates. However, we assume a two-phase sampling structure in which we

instead observe a coarsened version of this data structure in which we first observe $(Y_1, \Delta_1, W_1), ..., (Y_n, \Delta_n, W_n) \overset{iid}{\sim} P_0$. We then construct a set of indicator variables $R_1, ..., R_n$ based on our "phase 1 sample" and observe the variable A_i for subject i if $R_i = 1$. Different methods exist to construct the indicators $R_1, ..., R_n$, but we will assume that $R_i \sim \text{Bernoulli}(\pi_0(Y_i, \Delta_i, W_i))$ for a known function π_0. Importantly, the weights partition the observations into a finite set of strata $\{1, ..., J\}$ such that i and j are in the same stratum if and only if $\pi_0(Y_i, \Delta_i, W_i) = \pi_0(Y_j, \Delta_j, W_j)$; this is one way to define two-phase sampling (Breslow et al., 2009a, 2009b). Let C_i denote the index of the stratum to which observation i belongs. This yields a coarsened observed data structure that can be summarized as follows (where P_0 is redefined to include the additional variables):

$$\mathcal{O}_1, ..., \mathcal{O}_n \equiv (Y_1, \Delta_1, W_1, R_1, R_1A_1, C_1), ..., (Y_n, \Delta_n, W_n, R_n, R_nA_n, C_n) \overset{iid}{\sim} P_0.$$

For a fixed time t, the parameter of interest is the g-computed survival curve defined as follows, where $S_0(t|w, a) = P(T > t|W = w, A = a)$ is the true conditional survival function:

$$a \mapsto \bar{S}_0(t|a) \equiv E_0[S_0(t|W, a)].$$

Our goal is to use the coarsened data structure to make pointwise inference about $\bar{S}_0(t|a)$ for a fixed value a. We proceed by assuming that the true full-data distribution follows a Cox model and fitting an inverse-probability-sampling weighted Cox model to estimate the conditional survival function S_n. We then marginalize over the observed covariates, yielding the following estimator, where β_n is the estimator of the Cox model parameter vector β_0 and Λ_n is the inverse probability sampling weighted version of the so-called Breslow estimator of the baseline cumulative hazard function Λ_0:

$$\bar{S}_n(t|a) \equiv \sum_{i=1}^{n} S_n(t|w_i, a) = \sum_{i=1}^{n} exp\left(-e^{(w_i,a)'\beta_n}\Lambda_n(t)\right).$$

Our strategy for estimating Var $\left(\bar{S}_n(t|a)\right)$ will be to derive the influence functions of β_n and $\Lambda_n(t)$, apply the delta method to find the influence function of $S_n(t|w,a)$ for fixed w, and then account for the marginalization to derive the influence function of $\bar{S}_n(t|a)$.

Inverse probability sampling weighting using estimated weights

Inverse probability sampling (IPS) weighting allows us to identify various quantities of interest in terms of the observed data structure. Although IPS weighting based on the true weights $\pi_i \equiv \pi_0(y_i, \delta_i, w_i)$ can be used, we can gain efficiency by using the following estimated weights:

$$\pi_i^* \equiv \pi_n^*(o_i) \equiv \frac{\sum_{j=1}^n I\{c_i = c_j\} r_j}{\sum_{j=1}^n I\{c_i = c_j\}}.$$

The denominator in the expression above represents the number of observations in stratum c_i (the stratum to which observation i belongs) and the denominator represents the number of observations in stratum c_i selected in the phase-two sample.

For a fixed function h, we will need to repeatedly use IPS weighting to estimate terms of the form $E_0[h(\mathcal{O})]$ using the IPS-weighted estimator $n^{-1}\sum_{i=1}^n (r_i/\pi_i^*)h(o_i)$. However, this estimator is not asymptotically linear, since the π_i^* terms depend on the entire sample within a given stratum. Therefore, we will make use of the following equality, which can be derived using the delta method and holds under mild regularity conditions, where $p_0(c) \equiv P_0(C = c)$ and $p_0^1(c) \equiv P_0(C = c, R = 1)$:

$$\frac{1}{n}\sum_{i=1}^n \frac{r_i}{\pi_i^*} h(o_i) = \frac{1}{n}\sum_{i=1}^n \left(\frac{r_i}{\pi_i} h(o_i) + \left(1 - \frac{p_0(c_i) r_i}{p_0^1(c_i)}\right) E_0\left[h(\mathcal{O}) \,|C = c_i, R = 1\right] \right) + o_{P_0}(n^{-1/2}). \quad (3)$$

Influence function of β_n

The full-data influence function of the Cox model is given by $\tilde{\ell}_0 \equiv \mathcal{I}_0^{-1}\ell_0^*$, where $\mathcal{I}_0$ is the efficient information and ℓ_0^* is the efficient score. To pro-

vide expressions for these quantities, it is useful to first define the following expressions (where we use the notation $X^{\otimes 2} \equiv XX'$):

$$
\begin{aligned}
S_0^{(0)}(x) &\equiv E_0\left[e^{Z'\beta_0} I\{Y \geq x\}\right] \\
S_0^{(1)}(x) &\equiv E_0\left[Z e^{Z'\beta_0} I\{Y \geq x\}\right] \\
S_0^{(2)}(x) &\equiv E_0\left[Z^{\otimes 2} e^{Z'\beta_0} I\{Y \geq x\}\right] \\
m_0(x) &\equiv S_0^{(1)}(x)/S_0^{(0)}(x) \\
M_i(x) &\equiv \delta_i I\{y_i \leq x\} - e^{z_i'\beta_0} \int_0^x I\{y_i \geq u\} d\Lambda_0(u).
\end{aligned}
$$

We can then define the efficient score and information as follows (note that we assume that Y is bounded above by τ):

$$
\begin{aligned}
\ell_0^*(z_i, \delta_i, y_i) &\equiv \int_0^\tau (z_i - m_0(x))\, dM_i(x) \\
\mathcal{I} &\equiv E_0\left[e^{Z'\beta_0} \int_0^\tau (Z - m_0(x))^{\otimes 2} I\{Y \geq x\} d\Lambda_0(x).\right]
\end{aligned}
$$

According to equation (19) of Breslow and Wellner (2007) the following equality holds, given mild regularity conditions:

$$
\beta_n - \beta_0 = \frac{1}{n} \sum_{i=1}^n \frac{r_i}{\pi_i^*} \tilde{\ell}_0(o_i) + o_{P_0}(n^{-1/2}).
$$

Applying (3) to this equation, we can write the following:

$$
\beta_n - \beta_0 = \frac{1}{n} \sum_{i=1}^n \left(\frac{r_i}{\pi_i} \tilde{\ell}_0(o_i) + \left(1 - \frac{p_0(c_i) r_i}{p_0^1(c_i)}\right) E_0\left[\tilde{\ell}_0(\mathcal{O}) \,|C = c_i, R = 1\right] \right) + o_{P_0}(n^{-1/2}).
$$

Thus, the influence function of β_n under two-phase sampling with estimated weights is given by:

$$\tilde{\ell}_0^* : o_i \mapsto \frac{r_i}{\pi_i}\tilde{\ell}_0(o_i) + \left(1 - \frac{p_0(c_i)r_i}{p_0^1(c_i)}\right) E_0\left[\tilde{\ell}_0(\mathcal{O}) \,|C = c_i, R = 1\right]. \tag{4}$$

Influence function of Λ_n

The IPS-weighted Breslow estimator is given by the following:

$$\Lambda_n(t) \equiv \sum_{\{i:\delta_i=1\}} \frac{(r_i/\pi_i^*)I\{y_i \le t\}}{\sum_{j=1}^n (r_j/\pi_j^*)I\{y_j \ge y_i\}e^{z_j'\beta_n}}.$$

To study this estimator, we first define the following:

$$\begin{aligned}
S_0^{(0)}(x,\beta) &\equiv E_0[e^{Z'\beta}I\{Y \ge x\}] \\
S_n^{(0)}(x,\beta) &\equiv \sum_{i=1}^n \frac{r_i}{\pi_i^*}e^{z_i'\beta}I\{y_i \ge x\} \\
F_{1,0}(x) &\equiv P_0(Y \le x, \Delta = 1) \\
F_{1,n}(x) &\equiv \sum_{i=1}^n \frac{r_i}{\pi_i^*}\delta_i I\{y_i \le x\} \\
\Lambda_0(t,\beta) &\equiv \int_0^t \frac{dF_{1,0}(x)}{S_0^{(0)}(x,\beta)} \\
\Lambda_n(t,\beta) &\equiv \int_0^t \frac{dF_{1,n}(x)}{S_n^{(0)}(x,\beta)}.
\end{aligned}$$

It can be shown that $\Lambda_0(t,\beta_0)$ is equal to the baseline cumulative hazard function $\Lambda_0(t)$ and $\Lambda_n(t,\beta_n)$ is equal to the IPS-weighted Breslow estimator $\Lambda_n(t)$. Also note that $S_0^{(0)}(x,\beta_0) = S_0^{(0)}(x)$. We will study the asymptotic behavior of $\Lambda_n(t)$ using the following expansion, which holds given mild regularity conditions:

$$\begin{aligned}\Lambda_n(t) - \Lambda_0(t) &= \Lambda_n(t,\beta_n) - \Lambda_0(t,\beta_0) \\ &= (\Lambda_n(t,\beta_0) - \Lambda_0(t,\beta_0)) + (\Lambda_0(t,\beta_n) - \Lambda_0(t,\beta_0)) + o_{P_0}(n^{-1/2}) \\ &= (\Lambda_n(t,\beta_0) - \Lambda_0(t,\beta_0)) + \left(\frac{\partial}{\partial\beta}\Lambda_0(t,\beta)\Big|_{\beta=\beta_0}\right)(\beta_n - \beta_0) + o_{P_0}(n^{-1/2}).\end{aligned}$$

Defining $\mu_0(t) \equiv \frac{\partial}{\partial\beta}\Lambda_0(t,\beta)\Big|_{\beta=\beta_0} = \int_0^t m_0(x)d\Lambda_0(x)$, this expansion can be expressed as:

$$\Lambda_n(t) - \Lambda_0(t) = (\Lambda_n(t,\beta_0) - \Lambda_0(t,\beta_0)) + \mu_0(t)(\beta_n - \beta_0) + o_{P_0}(n^{-1/2}). \quad (5)$$

Next, we study the asymptotic behavior of $\Lambda_n(t,\beta_0)$ through the following expansion:

$$\Lambda_n(t,\beta_0) - \Lambda_0(t,\beta_0) = \left(\int_0^t \frac{dF_{1,n}(x)}{S_n^{(0)}(x,\beta)} - \int_0^t \frac{dF_{1,n}(x)}{S_0^{(0)}(x,\beta)}\right) + \left(\int_0^t \frac{dF_{1,n}(x)}{S_0^{(0)}(x,\beta)} - \int_0^t \frac{dF_{1,0}(x)}{S_0^{(0)}(x,\beta)}\right). \quad (6)$$

To study this expansion, we first apply result (3) to both $S_n^{(0)}(x,\beta)$ and $F_{1,n}(x)$ to obtain the following:

$$S_n^{(0)}(x,\beta) = \frac{1}{n}\sum_{i=1}^n \left(\frac{r_i}{\pi_i} I\{y_i \geq x\} e^{z_i'\beta} + \left(1 - \frac{p_0(c_i)r_i}{p_0^1(c_i)}\right) E_0\left[I\{Y \geq x\} e^{Z'\beta} \,|C = c_i, R = 1\right]\right) + o_{P_0}(n^{-1/2}) \quad (7)$$

$$F_{1,n}(x) = \frac{1}{n}\sum_{i=1}^n \left(\frac{r_i}{\pi_i}\delta_i I\{y_i \leq x\} + \left(1 - \frac{p_0(c_i)r_i}{p_0^1(c_i)}\right) E_0\left[\Delta I\{Y \leq x\} \,|C = c_i, R = 1\right]\right) + o_{P_0}(n^{-1/2}). \quad (8)$$

Define the following function for convenience:

$$\nu_0(o_i, x, \beta) \equiv \left(1 - \frac{p_0(c_i)r_i}{p_0^1(c_i)}\right) E_0\left[I\{Y \geq x\} e^{Z'\beta} \,|C = c_i, R = 1\right].$$

Using the delta method, we can write the following, where in the last line we have implicitly pulled out several second-order terms when substituting $dF_{1,0}$ for $dF_{1,n}$:

$$S_n^{(0)}(x,\beta) - S_0^{(0)}(x,\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i}{\pi_i}I\{y_i \geq x\}e^{z_i'\beta} + \nu_0(o_i,x,\beta) - S_0(x,\beta)\right) + o_{P_0}(n^{-1/2})$$

$$\frac{1}{S_n^{(0)}(x,\beta)} - \frac{1}{S_0^{(0)}(x,\beta)} = \frac{1}{n}\sum_{i=1}^{n}\frac{S_0(x,\beta) - (r_i/\pi_i)I\{y_i \geq x\}e^{z_i'\beta} - \nu_0(o_i,x,\beta)}{\left(S_0(x\beta)\right)^2} + o_{P_0}(n^{-1/2})$$

$$\int_0^t \frac{dF_{1,n}(x)}{S_n^{(0)}(x,\beta)} - \int_0^t \frac{dF_{1,n}(x)}{S_0^{(0)}(x,\beta)} = \int_0^t \frac{S_0(x,\beta) - \frac{1}{n}\sum_{i=1}^{n}\left((r_i/\pi_i)I\{y_i \geq x\}e^{z_i'\beta} - \nu_0(o_i,x,\beta)\right)}{\left(S_0(x\beta)\right)^2} dF_{1,0}(x) + o_{P_0}(n^{-1/2}).$$

Next, we define the following for convenience:

$$\nu_0^*(o_i,t,\beta) \equiv \left(1 - \frac{p_0(c_i)r_i}{p_0^1(c_i)}\right) E_0\left[\frac{\Delta I\{Y \leq t\}}{S_0(Y,\beta)} \,|C = c_i, R = 1\right].$$

This allows us to write the following:

$$F_{1,n}(x) - F_{1,0}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i}{\pi_i}\delta_i I\{y_i \leq x\} + \left(1 - \frac{p_0(c_i)r_i}{p_0^1(c_i)}\right) E_0\left[\Delta I\{Y \leq x\}\,|C = c_i, R = 1\right] - F_{1,0}(x)\right) + o_{P_0}(n^{-1/2})$$

$$\int_0^t \frac{d(F_{1,n} - F_{1,0})(x)}{S_0(x,\beta)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i\delta_i I\{y_i \leq t\}}{\pi_i S_0(y_i,\beta)} + \nu_0^*(o_i,t,\beta)\right) - \Lambda_0(t,\beta) + o_{P_0}(n^{-1/2}).$$

Combining this with the previous result, we have:

$$\Lambda_n(t,\beta) - \Lambda_0(t,\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i\delta_i I\{y_i \leq t\}}{\pi_i S_0(y_i,\beta)} + \nu_0^*(o_i,t,\beta) - \int_0^t \left(\frac{r_i I\{y_i \geq x\}e^{z_i'\beta} + \pi_i\nu_0(o_i,x,\beta)}{\pi_i\left(S_0(x,\beta)\right)^2}\right) dF_{1,0}(x)\right) + o_{P_0}(n^{-1/2}). \tag{9}$$

Returning to the expansion given in (6) and plugging in these equalities, we can write the following:

$$\Lambda_n(t) - \Lambda_0(t) =$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i\delta_i I\{y_i \leq t\}}{\pi_i S_0(y_i)} + \nu_0^*(o_i,t,\beta_0) - \int_0^t \left(\frac{r_i I\{y_i \geq x\}e^{z_i'\beta_0} + \pi_i\nu_0(o_i,x,\beta_0)}{\pi_i\left(S_0(x)\right)^2}\right) dF_{1,0}(x) - \mu_0(t)\tilde{\ell}_0^*(o_i)\right) + o_{P_0}(n^{-1/2}).$$

Using the above and the equality $\int_0^t h(x)d\Lambda_0(x) = \int_0^t (h(x)/S_0^{(0)}(x))dF_{1,0}(x)$, we can write the influence function of the IPS-weighted Breslow estimator as follows:

$$\varphi_0^\Lambda : o_i \mapsto \frac{r_i \delta_i I\{y_i \leq t\}}{\pi_i S_0(y_i)} + \nu_0^*(o_i, t, \beta_0) - \int_0^t \left(\frac{r_i I\{y_i \geq x\} e^{z_i' \beta_0} + \pi_i \nu_0(o_i, x, \beta_0)}{\pi_i S_0(x)} \right) d\Lambda_0(x) - \mu_0(t)\tilde{\ell}_0^*(o_i). \tag{10}$$

Influence function of $\bar{S}_n(t|a)$

Using several applications of the delta method on the influence functions given by equations (4) and (10), we can write the following, where z is a fixed covariate vector:

$$\begin{aligned} S_n(t|z) - S_0(t|z) &\equiv e^{z'\beta_n}\Lambda_n(t) - e^{z'\beta_0}\Lambda_0(t) \\ &= -\frac{1}{n}\sum_{i=1}^n S_0(t|z)\left(\Lambda_0(t)e^{z'\beta_0} z'\tilde{\ell}_0^*(o_i) + e^{z'\beta_0}\varphi_0^\Lambda(o_i)\right) + o_{P_0}(n^{-1/2}). \end{aligned}$$

Thus, the influence function of $S_n(t|z)$ is given by the following:

$$\omega_0^t : (o_i, z) \mapsto -S_0(t|z)\left(\Lambda_0(t)e^{z'\beta_0} z'\tilde{\ell}_0^*(o_i) + e^{z'\beta_0}\varphi_0^\Lambda(o_i)\right). \tag{11}$$

Next, we can study the asymptotic behavior of $\bar{S}_n(t|a)$ through the following expansion, where we use the notation $Pf \equiv E_P[f(\mathcal{O})]$ for a given probability measure P:

$$\bar{S}_n(t|a) - \bar{S}_0(t|a) = P_n S_n - P_0 S_0 = (P_n - P_0)S_0 + P_0(S_n - S_0) + (P_n - P_0)(S_n - S_0). \tag{12}$$

The first term is linear and can be written as follows:

$$(P_n - P_0)S_0 = \frac{1}{n}\sum_{i=1}^n \left(S_0(t|w_i, a) - E_0[S_0(t|W, a)]\right). \tag{13}$$

The second term can be represented as follows, which holds under mild regularity conditions on the remainder:

$$P_0(S_n - S_0) = E_0[S_n(t|W,a) - S_0(t|W,a)] = \frac{1}{n}\sum_{i=1}^{n} E_0\left[\omega_0^t(o_i,(W,a))\right] + o_{P_0}(n^{-1/2}). \tag{14}$$

Plugging these two equalities into expansion (12) and assuming regularity conditions hold such that $(P_n - P_0)(S_n - S_0) = o_{P_0}(n^{-1/2})$, we can write the following:

$$\bar{S}_n(t|a) - \bar{S}_0(t|a) = \frac{1}{n}\sum_{i=1}^{n}\left(S_0(t|w_i,a) - E_0[S_0(t|W,a)] + E_0\left[\omega_0^t(o_i,(W,a))\right]\right) + o_{P_0}(n^{-1/2}).$$

Therefore, the influence function of $\bar{S}_n(t|a)$ is given by:

$$\varphi_0^{t,a} : o_i \mapsto S_0(t|w_i,a) - E_0[S_0(t|W,a)] + E_0\left[\omega_0^t(o_i,(W,a))\right]. \tag{15}$$

The variance of $\bar{S}_n$ can be estimated by estimating the second moment of its influence function (15), as usual. However, since $\bar{S}_n \in [0,1]$, it may be useful to form confidence intervals on the $\text{logit}(\bar{S}_n)$ scale instead and then transform the confidence limits back to the $\bar{S}_n$ scale, to ensure that the entire confidence interval lies within $[0,1]$. This yields the following confidence interval estimator, where $\xi(x) \equiv \text{logit}(x)$, $\dot{\xi}(x) \equiv \frac{d}{dx}\text{logit}(x)$, $\xi^{-1}(x) \equiv \text{expit}(x)$, $Z_{\alpha/2}$ is the relevant quantile of the standard Normal distribution, and $\varphi_n^{t,a}$ is an estimator of the true influence function $\varphi_0^{t,a}$, which we will define later:

$$\xi^{-1}\left(\xi(\bar{S}_n(t|a)) \pm Z_{\alpha/2}\dot{\xi}\left(\bar{S}_n(t|a)\right)\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\varphi_0^{t,a}(o_i)\right)^2}\right). \tag{16}$$

Estimation of the influence function of β_n under iid sampling: $\tilde{\ell}_n$

First, we define the following component function estimators:

$$
\begin{aligned}
S_n^{(0)}(x) &\equiv \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi_i^*}I\{y_i \geq x\}e^{z_i'\beta_n} \\
S_n^{(1)}(x) &\equiv \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi_i^*}z_i I\{y_i \geq x\}e^{z_i'\beta_n} \\
S_n^{(2)}(x) &\equiv \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi_i^*}z_i^{\otimes 2} I\{y_i \geq x\}e^{z_i'\beta_n} \\
m_n(x) &\equiv S_n^{(1)}(x)/S_n^{(0)}(x).
\end{aligned}
$$

We can then define our estimator as $\tilde{\ell}_n \equiv \mathcal{I}_n^{-1}\tilde{\ell}_n^*$, where $\mathcal{I}_n$ and $\tilde{\ell}_n^*$ are defined as follows:

$$
\begin{aligned}
\mathcal{I}_n &\equiv \frac{1}{n}\sum_{\{i:\delta_i=1\}}\frac{r_i}{\pi_i^*}\left(\frac{S_n^{(2)}(y_i)}{S_n^{(0)}(y_i)} - (m_n(y_i))^{\otimes 2}\right) \\
\tilde{\ell}_n^* &\equiv \delta_i\left(z_i - m_n(y_i)\right) - \sum_{\{j:\delta_j=1\}}\frac{r_j e^{z_i'\beta_n} I\{y_j \leq y_i\}\left(z_i - m_n(y_j)\right)}{\pi_j^* S_n^{(0)}(y_j)}.
\end{aligned}
$$

Estimation of the influence function of inverse probability sampling weighted β_n under two-phase sampling: $\tilde{\ell}_n^*$

A consistent estimator of $\tilde{\ell}_0^*$ is given by the following, where $p_n(c) \equiv \frac{1}{n}\sum_{i=1}^{n} I\{c_i = c\}$ and $p_n^1(c) \equiv \frac{1}{n}\sum_{i=1}^{n} I\{c_i = c\}r_i$:

$$
\tilde{\ell}_n^*(o_i) \equiv \frac{r_i}{\pi_i^*}\tilde{\ell}_n(o_i) + \frac{1}{n}\sum_{\{j:c_j=c_i,r_j=1\}}\frac{\left(p_n^1(c_j) - p_n(c_j)r_i\right)\tilde{\ell}_n(o_j)}{\left(p_n^1(c_j)\right)^2}.
$$

Estimation of the influence function of the IPS-weighted Breslow estimator: φ_n^Λ

The influence function φ_0^Λ can be consistently estimated using the following:

$$\varphi_n^{\Lambda}(o_i) \equiv \frac{r_i \delta_i I\{y_i \leq t\}}{\pi_i^* S_n(y_i)} + \nu_n^*(o_i, t, \beta_n) - \frac{1}{n} \sum_{\{j:\delta_j=1\}} I\{y_j \leq t\} \left(\frac{r_i I\{y_i \geq y_j\} e^{z_i' \beta_n} + \pi_i^* \nu_n(o_i, y_j, \beta_n)}{\pi_i^* \left(S_n(y_j)\right)^2} \right) - \mu_n(t) \tilde{\ell}_n^*(o_i).$$

This involves the following nuisance estimators:

$$\nu_n^*(o_i, t, \beta_n) \equiv \left(\frac{p_n^1(c_i) - p_n(c_i) r_i}{\left(p_n^1(c_i)\right)^2} \right) \left(\frac{1}{n} \sum_{\{k:c_k=c_i, r_k=\delta_k=1\}} \frac{I\{y_k \leq t\}}{S_n(y_k)} \right)$$

$$\nu_n(o_i, y_j, \beta_n) \equiv \left(\frac{p_n^1(c_i) - p_n(c_i) r_i}{\left(p_n^1(c_i)\right)^2} \right) \left(\frac{1}{n} \sum_{\{k:c_k=c_i, r_k=1\}} I\{y_k \geq y_j\} e^{z_k' \beta_n} \right)$$

$$\mu_n(t) \equiv \frac{1}{n} \sum_{\{j:\delta_j=1\}} \frac{I\{y_j \leq t\} S_n^{(1)}(y_j)}{\left(S_n^{(0)}(y_j)\right)^2}.$$

Estimation of the influence function of conditional survival $S_n(t|z)$: ω_0^t

The influence function ω_0^t can be consistently estimated using the following:

$$\omega_n^t(o_i, z) \equiv e^{z' \beta_n} \left(\frac{\delta_i I\{y_i \leq t\}}{S_n^{(0)}(y_i)} - \frac{1}{n} \sum_{\{j:\delta_j=1\}} \frac{e^{z_i' \beta_n I\{y_j \leq t \wedge y_i\}}}{\pi_j^* \left(S_n^{(0)}(y_j)\right)^2} + \left(\frac{1}{n} \sum_{\{j:\delta_j=1\}} \frac{I\{y_j \leq t\}(z - m_n(y_j))}{\pi_j^* S_n^{(0)}(y_j)} \right)' \tilde{\ell}_n(z_i, \delta_i, y_i) \right). \tag{17}$$

Estimation of the influence function of the target parameter $\bar{S}_n(t|a)$: $\varphi_0^{t,a}$

Finally, the influence function $\varphi_0^{t,a}$ can be consistently estimated using the following:

$$\varphi_n^{t,a}(o_i) \equiv S_n(t|w_i, a) + \frac{1}{n} \sum_{j=1}^{n} \left(\omega_n^t(o_i, (w_j, a)) - S_n(t|w_j, a) \right). \tag{18}$$

12.3.5 Univariable CoR: Marginalized Cox modeling for an outcome subject to competing risks (e.g. asymptomatic infection)

If the outcome under study is subject to competing risks, then the Cox model is fit in the same way, except counting the competing risk as right-censoring. Now the parameter being estimated is the marginal conditional cumulative incidence function $risk_1(t, 1|s) = E_X[P(T \leq t, J = 1|s, X, A = 1)]$ where $J = 1$ is the outcome of interest. To estimate this parameter, we use the fact that the cause j-specific hazard $\lambda_{j1}(t|s, x)$ is linked to the conditional cumulative incidence $F_{j1}(t|s, x) = P(T \leq t, J = j|s, x, A = 1)$ via the formula

$$F_{j1}(t|s, x) = \int_0^t \lambda_{j1}(u|s, x) S_1(u|s, x)$$

where S_1 is the all-cause/overall conditional survival function for the first event of asymptomatic infection or COVID-19, whichever occurs first (i.e., the COV-INF endpoint). Therefore, after fitting the cause $J = 1$ Cox model, for any fixed t the above formula provides fitted values $\hat{F}_{j1}(t|s, X_i)$ for each participant i, and then the G-computation estimator of $risk_1(t, 1|s)$ is

$$\widehat{risk}_1(t, 1|s) = \frac{1}{n} \sum_{i=1}^{n} \hat{F}_{j1}(t|s, X_i).$$

As for the analyses without competing risks, the bootstrap is used for calculating 95% confidence intervals and for testing $H_0 : risk_1(t_F, 1|s) = risk_1(t_F, 1)$ for all s.

12.3.6 Univariate CoR: Nonparametric threshold regression modeling

The van der Laan et al. (2021) extension of the nonparametric CoR threshold estimation method of Donovan et al. (2019) is applied to each of the five non-baseline subtracted Day 57 antibody markers, using the version that defines the binary outcome Y of interest as $Y = 1$ if a COVID endpoint occurred during the blinded period of follow-up and $Y = 0$ otherwise. The analyses adjust for the same baseline factors X as used in the Cox model CoR analyses.

The extension adjusts for baseline covariates by estimating the conditional mean function $E[Y|S \geq s, X, A = 1]$ using discrete-SuperLearner and then

empirically averaging over the baseline covariates X to estimate the marginal risk $risk_1^Y(S \geq s) = E_X[P(Y = 1|S \geq s, X, A = 1)]$ for each threshold s of the the antibody marker in a specified discrete set. We do not perform pooled regression across the thresholds s, which ensures we are totally nonparametric in estimating the threshold dependence of $risk_1^Y(S \geq s)$ on s. The Super-Learner library includes a range of increasingly flexible parametric learners including logistic regression (glm), bayesian logistic regression (bayesglm), and L1-penalized logistic regression (glmnet). (Two of each learner is included in the library, one with only main-term variables and another with main-term and interaction variables.) An advantage of the nonparametric CoR threshold method compared to Cox modeling that specifies a log linear hazard ratio with the marker is that it can potentially detect a threshold of very low risk. The method is implemented with and without the monotonicity constraint that $risk_1^Y(S \geq s)$ is monotone non-increasing in s, where the results assuming monotonicity are reported unless there is evidence for violation of this assumption.

The results are reported in the same way that Donovan et al. (2019) reports results in its Figure 2, where point estimates, pointwise 95% confidence bands, and simultaneous 95% confidence bands for $risk_1^Y(S \geq s)$ are plotted for a range of threshold values. The simultaneous confidence bands cover the entire curve in s with at least 95% probability and are useful for judging whether risk varies over threshold subgroups, whereas the pointwise 95% confidence bands are useful for quantifying precision at particular threshold values. The method uses the same empirical two-phase sampling estimated weights (IPS weights) as used for the other univariable IPWCC CoR analyses. In addition, for each pre-specified risk threshold c set to take values over a grid with lowest value 0, the method is applied to estimate the inverse function $s_c = inf\{s : E_X[P(Y = 1|S \geq s, A = 1, X] \leq c\}$, where s_c is estimated by substitution of the marginal risk function estimate. Note that the substitution estimator of s_c requires that the marginal risk function is estimated for all thresholds, which is computationally infeasible. Instead, we estimate the marginal risk function on a sufficiently large discrete set and linearly interpolate to obtain marginal risk estimates for all thresholds outside

the discrete set. In order for this estimand to be well defined, we operate (for this estimand only) under the assumption that $s \mapsto risk_1^Y(S \geq s)$ is monotone. For the substitution-based estimator of the inverse function s_c to be well-defined, we require the estimate of $s \mapsto risk_1^Y(S \geq s)$ to be monotone as well. If there is evidence that the function estimate is not monotone then we replace the estimate with its monotone projection, which preserves its theoretical properties (Westling, van der Laan, Carone, 2020).

A plot of point and pointwise 95% confidence interval estimates of s_c (over the grid of c values) is provided to help indicate marker thresholds defining subgroups with very low risk of outcome. The confidence interval estimates for s_c are derived directly from the confidence interval estimates for the marginal risk function $s \mapsto risk_1^Y(S \geq s)$, and therefore its estimates are compatible with those of the marginal risk function. In addition, a plot of point and simultaneous 95% confidence interval estimates of s_c (over the grid of c values) is provided, where the simultaneous confidence interval estimates for s_c are derived directly from the simultaneous 95% confidence band estimates for the marginal risk function $s \mapsto risk_1^Y(S \geq s)$, and therefore its estimates are compatible with those of the marginal risk function. In particular, no multiple testing adjustments are needed.

The analysis is done using targeted maximum likelihood estimation (TMLE) as described in van der Laan, Zhang, and Gilbert (2020), and the pointwise and simultaneous simultaneous confidence bands are of the Wald-type, obtained from the asymptotic distribution of the TMLE.

12.4 Univariable CoR: Supportive Exploratory Flexible Parametric Risk Modeling

For each of the five non-baseline subtracted Day 57 antibody markers, flexible nonlinear modeling of outcome risk studied as a dichotomous outcome Y will be conducted, as exploratory supportive analyses. Again, the analyses adjust for the same baseline factors X as used in the Cox model CoR analyses.

The nonlinear relationship between the logit of risk and markers will be modeled using two-phase polynomial regression models (Son and Fong, 2020; Fong

et al., 2017), e.g., hinge model, or three-phase segmented models (Chen, 2020). The mean function of a q^{th} order two-phase polynomial regression model can be expressed as follows:

$$\eta(s, X) = \alpha_1 + \alpha_2^X z + \beta_{1,-}(s-e)_- + \beta_{1,+}(s-e)_+ + \beta_{2,-}(s-e)_-^2 + \beta_{2,+}(s-e)_+^2 + \cdots + \beta_{q,-}(s-e)_-^q + \beta_{q,+}(s-e)_+^q,$$

where X is the baseline covariate vector, s is a fixed value of the immunologic marker of interest, e is the threshold parameter, $(s-e)_+ = s-e$ if $s > e$ and 0 otherwise, and $(s-e)_- = 0$ if $s > e$ and $s-e$ otherwise.

In addition, a generalized additive model with degree of smoothing estimated by cross-validation is employed (Wood, 2017). Two-phase sampling designs are accounted for through inverse probability weighting and confidence intervals are obtained through the same bootstrap scheme as the Cox proportional hazard model bootstrap inference.

12.4.1 [With Day 29 markers]

For CoR analyses of Day 29 markers, the same analyses are done, except using the Day 29 visit date as the time origin and counting events starting 7 days after the Day 29 visit.

12.4.2 P-values and Multiple hypothesis testing adjustment for CoR analysis

In general, p-values are only reported from pre-specified and automated (press-button) analyses. For the CoR analyses, p-values are reported for the univariable Cox regression analyses of the five specified Day 57 antibody marker variables. Two-sided p-values for hypothesis testing of a Day 57 marker CoR are calculated both for the Cox regression of quantitative markers (two-sided Wald tests), and for the Cox regression of markers binned into tertiles (two-sided Generalized Wald tests). Therefore a total of ten 2-sided p-values for CoRs are calculated.

It is not completely clear whether to perform multiple hypothesis testing adjustment, given the expectation that the correlations among the markers

are high, and possibly very high, meaning that multiplicity correction could incur a relatively high cost on the false negative error rate.

However, given that robust evidence supporting an antibody marker as a CoR will be required for qualifying a marker, we will conduct multiplicity adjustment for CoR analysis, as the ability to make an inference that a marker passed pre-specified multiplicity adjusted criteria should aid an overall evidence package for establishing a validated or non-validated surrogate endpoint. Therefore, multiplicity adjustment is performed across the ten 2-sided p-values.

A permutation-based method (Westfall et al., 1993) will be used for both family-wise error rate (Holm-Bonferroni) and false-discovery rate (q-values; Benjamini-Hochberg) correction. 10^4 replicates of the data under the null hypotheses will be created by randomly resampling the immunologic biomarkers with replacement. For each Cox regression CoR analysis the unadjusted p-value, the FWER-adjusted p-value, and the q-value is reported for whether there is a covariate-adjusted association, where all p-values and q-values are 2-sided. The FWER-adjusted p-values and q-values are computed pooling over both the quantitative marker and tertilized marker CoR analyses. As a guideline for interpreting CoR findings, markers with FWER-adjusted p-value ≤ 0.05 are flagged as having statistical evidence for being a CoR. Additionally, markers with unadjusted p-value ≤ 0.05 and q-value ≤ 0.10 are flagged as having a hypothesis generated for being a CoR.

As described in this SAP, the FWER adjustment is done for all advanced Day 57 markers among bAb Spike, bAb RBD, PsV nAb ID50, PsV nAb cID80, and WT LV MN50. If the antibody data set available for correlates analysis does not yet contain the WT LV MN50 data (due to a longer time horizon on performing the assay), then the multiplicity adjustment will be performed for the available 4 markers.

12.4.3 [With Day 29 markers]

If Day 29 markers are included, the same multiplicity adjustment approach is used as for Day 57 markers. The multiplicity adjustment is done separately

for Day 29 markers and for Day 57 markers, given the high degree of correlation of the analysis results (given that all endpoint cases starting 7 days post Day 57 are common among the two analyses).

12.5 Univariate CoR: evaluating markers as endpoints

12.5.1 Objective

Follmann (2018) describe a method for comparing two endpoints in terms of the sample size required to power a future study. In this approach, we imagine that our goal is to compare two (or more) markers in terms of their standardized effect size. Markers with stronger correlates signals will have higher standardized effect sizes. We then present the comparison of effect sizes in terms of a sample size ratio. For example, if the ratio of standardized effect size for Day 57 binding Ab to Spike compared to Day 57 pseudovirus neutralization ID50 is 2, then a future correlates study would need to enroll twice as many participants to achieve a similar power to reject the correlates null hypothesis for the inferior marker. The method provides, in essence, a more interpretable means of comparing the magnitude of p-values for different markers.

12.5.2 Approach

We will apply the approach to the univariate Cox models. The required output for the method of Follmann (2018) is the same as for the IPWCC Cox regression described above. For each marker, the Wald statistic for the associated log hazard ratio is divided by the square root of the harmonic mean of cases and non-cases in the phase 2 dataset to obtain the standardized association. Denote by $\hat{\Delta}_m$ the standardized association for the m-th marker. We define the comparison of the ℓ-th and m-th marker's standardized association to be $\hat{\omega}_{\ell,m} = \hat{\Delta}_\ell / \hat{\Delta}_m$. This is mapped into an estimated ratio of sample sizes

$\theta_{\ell,m}$ via the following relationship

$$\hat{\theta}_{\ell,m} = \begin{cases} \hat{\omega}^2_{\ell,m} & \text{if } \hat{\Delta}_\ell < 0 \text{ and } \hat{\Delta}_m < 0 \\ 0 & \text{if } \hat{\Delta}_\ell > 0 \text{ and } \hat{\Delta}_m < 0 \\ \text{undefined} & \text{if } \hat{\Delta}_\ell > 0 \text{ and } \hat{\Delta}_m > 0 \\ \infty & \text{if } \hat{\Delta}_\ell < 0 \text{ and } \hat{\Delta}_m > 0 \end{cases}$$

A 95% confidence interval for $\theta_{\ell,m}$ can derived based on the bootstrap, though some cared is needed to handle edge cases. Follmann (2018) describe the following approach. For the b-th bootstrap sample (generated, as described above), we compute $\hat{\Delta}^{(b)}_\ell$, $\hat{\Delta}^{(b)}_m$, and $\hat{\omega}^{(b)}_{\ell,m}$ as described above. We then define separate "versions" of $\hat{\theta}^{(b)}_{\ell,m}$,

$$\begin{aligned} \hat{\theta}^{(b)}_{\ell,m}(L) &= I(\hat{\Delta}_\ell < 0, \hat{\Delta}_m < 0)(\hat{\omega}^{(b)}_{\ell,m})^2 + I(\hat{\Delta}_\ell < 0, \hat{\Delta}_m > 0) \times \infty \\ \hat{\theta}^{(b)}_{\ell,m}(U) &= I(\hat{\Delta}_\ell < 0, \hat{\Delta}_m < 0)(\hat{\omega}^{(b)}_{\ell,m})^2 + I(\hat{\Delta}_\ell > 0, \hat{\Delta}_m > 0) \times \infty \\ &\quad + I(\hat{\Delta}_\ell < 0, \hat{\Delta}_m > 0) \times \infty \end{aligned}$$

The lower bound of the bootstrap 95% confidence interval is the 2.5-th percentile of the bootstrap distribution of $\hat{\theta}^{(b)}_{\ell,m}(L)$; the upper bound of the bootstrap 95% confidence interval is the 97.5-th percentile of the bootstrap distribution of $\hat{\theta}^{(b)}_{\ell,m}(U)$.

We will compare ratios of sample sizes of the following markers:

1. binding Ab RBD vs. pseudovirus neutralization ID50
2. binding Ab RBD vs. live virus neutralization MN50
3. pseudovirus neutralization ID50 vs. live virus neutralization MN50

12.5.3 Multi-variable extension

The above method extends naturally to multivariable models described below, where a standardized association is computed by multiplying the generalized Wald statistic by the number of "phase 2 ptids".

12.6 Multivariable CoR: Superlearning of Optimal Risk Prediction Models

12.6.1 Objectives

The multivariable CoR objective is addressed through two sub-objectives: first to build an 'estimated optimal surrogate' (Price et al., 2018), a model that best predicts the outcome from Day 57 antibody markers and baseline factors. The second sub-objective is estimation and inference on variable importance measures for each Day 57 antibody marker, for ranking of antibody markers by their importance/influence on predicting risk. The analysis plan is patterned off of the analysis of the HVTN 505 HIV-1 vaccine efficacy trial (Neidich et al., 2019). For these analyses both baseline-subtracted and non-baseline subtracted versions of the Day 57 markers are used, in a broader unbiased analysis to build models most predictive of outcome.

12.6.2 Input variable sets

Day 57 antibody markers are classified into the following four antibody marker variable sets, with individual variables listed within categories:

1. Binding antibody anti-Spike (S-bAb)
    a Day 57 anti-Spike IgG concentration
    b delta (Day 57 - Day 1) anti-Spike IgG concentration
    c indicator 2FR anti-Spike IgG concentration
    d indicator 4FR anti-Spike IgG concentration
2. Binding antibody anti-RBD (RBD-bAb)
    a Day 57 anti-RBD concentration
    b delta (Day 57 - Day 1) anti-RBD concentration
    c indicator 2FR anti-RBD concentration
    d indicator 4FR anti-RBD concentration
3. Pseudovirus neutralizing antibody anti-Spike (pseudovirus-nAb)
    a Day 57 anti-Spike ID50

b Day 57 anti-Spike cID80

c delta (Day 57 - Day 1) anti-Spike ID50

d delta (Day 57 - Day 1) anti-Spike cID80

e indicator 2FR anti-Spike ID50

f indicator 4FR anti-Spike ID50

g indicator 2FR anti-Spike cID80

h indicator 4FR anti-Spike cID80

4. Wild Type Live virus neutralizing antibody anti-Spike (WT live virus-nAb)

   a Day 57 anti-Spike MN50

   b delta (Day 57 - Day 1) anti-Spike MN50

   c indicator 2FR anti-Spike MN50

   d indicator 4FR anti-Spike MN50

The baseline factors without any marker data constitutes another set of variables to include in the superlearner modeling.

12.6.3 Missing data

We expect a very small amount of missing data from the five antibody marker types (bAb Spike, RBD; pseudovirus-nAb ID50, cID80; WT live virus-nAb MN50). However, there may be a small amount of missing data, with possibly different participants missing data for different markers. We take the following approach to handle any missing data that occurs.

First, we define the two-phase sampling indicator ϵ as taking value of one if a participant has Day 1 and Day 57 bAb data for both Spike and RBD, where here we assume that the MSD platform is highly robust such that it will have nearly 100% complete data for sampled participants. Second, for the other three marker types (pseudovirus-nAb ID50, cID80; WT live virus-nAb MN50), for participants with $\epsilon = 1$ but the Day 1 and/or Day

57 marker value is missing, we use single imputation to fill in any missing values, ignoring the uncertainty in the imputations in the analysis, because it should have negligible impact on results given the (very) small amount of missing data. Multiple linear regression will be used to impute missing values, separately for each antibody marker, based on the set of individuals with that antibody marker measured at Day 1 and 57. Accurate imputations are possible given the high correlations of the markers, especially between ID50 and cID80 within the same immunoassay. This process means that the two-phase data set has a simple 'all-or-nothing' missing data pattern where participants with $\epsilon = 1$ have all markers with Day 1 and Day 57 data, and are included in IPWCC analyses, and participants with $\epsilon = 0$ have some or all markers missing and are excluded from IPWCC analyses. This means that all IPWCC data analyses can use the same empirical frequency (IPS) sampling weights.

For analysis methods that use the whole cohort (phase-one plus phase-two data), the same phase-two data as described above are used. If some of the phase-one baseline factors that are adjusted for variables are missing with only a small amount of missing values, then single imputation will be used to fill in the values, and, as for the immunologic marker imputations, the uncertainty in the imputations will be ignored in the analyses. Simple average values will be used to fill in baseline covariate missing values of the baseline factors.

12.6.4 Implementation of superlearner

For baseline risk score development, Superlearner is applied to the placebo arm only, as mentioned in Section 10. For multivariable immune correlates of risk/estimated optimal surrogate development, Superlearner is applied to the vaccine arm only. The following details are used in the implementation of superlearner of the vaccine arm only:

- Pre-scale each quantitative and ordinal variable to have empirical mean 0 and standard deviation 1.
- For the immune correlates analysis, the final library of learners is selected

accounting for the number of phase-two endpoint cases in the vaccine arm. If the number of cases is limited, at or near 25 evaluable endpoint cases, then the modeling will only allow learning algorithms to have a maximum of 5 Day 57 antibody marker variables, and will use leave-one-out cross-validation and the negative log-likelihood loss function, a combination that tends to provide good performance in small sample size settings. On the other hand, if there are larger numbers of endpoint cases in the vaccine arm, then 5-fold cross-validation will be used, and no more than floor($n_v/6$) input variables will be used in the model where n_v is the number of evaluable vaccine endpoint cases. The choices will be finalized prior to case/control unblinded analysis.

- Include learning algorithms with and without screening of variables. Screens used will be: 1) glmnet (lasso) pre-screening (with default tuning parameter selection), 2) logistic regression univariate 2-sided p-value screening (at level $p < 0.10$), and 3) high-correlation variable screening (described below). The adaptive algorithms (SL.randomForest, SL.xgboost, SL.gam, SL.polymars) are only used with these screens, given that the limited number of endpoint cases may challenge use of these methods with no variable screening. Moreover, the adaptive algorithms are not used if there are only 25 (or close to it) endpoint cases. All of the selected learners are coded into the SuperLearner R package available on CRAN.
- Include high-correlation variable screening, not allowing any pair of input variables to have Spearman rank correlation $r > 0.9$. When a pair of variables has $r > 0.9$, the variable with the highest ranked signal-to-noise ratio (i.e., biological dynamic range) is selected; if these data are not available or there is a tie then variables are selected in the following order of priority: first WT live virus-nAb, second pseudovirus-nAb, third bAb.
- The superlearner is conducted averaging over 10 random seeds, to make results less dependent on random number generator seed.
- All of the learners are implemented with IPS weighting, using the weights $\hat{w}_{57.x}$ defined in Section 12.3.1 to account for the two-phase sampling

design.

- Two levels of cross-validation are used:
  - Outer level: CV-AUC computed over 5-fold cross-validation repeated 10 times to improve stability
  - Inner level: leave-one-out CV used to estimate ensemble weights (if n_v is near 25) and 5-fold CV if n_v is larger.
- Results for comparing classification accuracy of different models are based on point and 95% confidence interval estimates of cross-validated area under the ROC curve (CV-AUC) and difference in CV-AUC as a predictiveness metric (Hubbard et al., 2016; Williamson et al., 2020). Results are presented as forest plots of point and 95% confidence interval estimates similar to those used in Neidich et al. (2019) (Figure 3) and Magaret et al. (2019). CV-AUC is estimated using the R package *vimp* available on CRAN, including the IPS weights that are used for other data analyses.

For the baseline risk score SuperLearner analysis of the placebo arm (Section 10), the same approach is used, with the following modifications: (1) 5-fold cross-validation will be used with no more than max(20,floor(n_p/20)) input variables included in each model, where n_p is the number of evaluable placebo arm cases; (2) no IPS weighting is needed; (3) the adaptive learning algorithms are included.

Table 7 lists the learning algorithms that are applied to estimate the conditional probability of the outcome based on the input variable sets considered above. Most of the algorithms are non-data-adaptive type learning algorithms, such as parametric regression models (e.g., generalized linear models [glms]), which are simple, stable, and advantageous for an application with a limited number of endpoint events. Data-adaptive type algorithms are also included if the number of endpoint events is high enough, for increasing flexibility of modeling and reducing the risk of model misspecification: SL.ranger, SL.gam, SL.polymars, and SL.xgboost. All of the selected learners are coded into the SuperLearner R package.

Before fitting the superlearner models to the vaccine arm data, a decision will be made on how to define the "baseline risk factors" input variable set, based on prediction-accuracy results of the Superlearner analysis that built the baseline behavioral risk score based on the placebo arm, as well as on external knowledge of important individual risk factors. The set of baseline risk factors will include a subset of individual risk factors and/or the baseline risk score itself.

For the immune correlates objective the superlearner model is fit to each of the following 12 variable sets, with immunological variables listed in Section 12.6.2:

1. Baseline risk factors
2. Baseline risk factors and the Day 57 bAb anti-Spike markers
3. Baseline risk factors and the Day 57 bAb anti-RBD markers
4. Baseline risk factors and the Day 57 pseudovirus-nAb ID50 markers
5. Baseline risk factors and the Day 57 pseudovirus-nAb cID80 markers
6. Baseline risk factors and the Day 57 live virus-nAb MN50 markers
7. Baseline risk factors and the Day 57 bAb markers and the Day 57 pseudovirus-nAb ID50 markers
8. Baseline risk factors and the Day 57 bAb markers and the Day 57 pseudovirus-nAb cID80 markers
9. Baseline risk factors and the Day 57 bAb markers and the Day 57 live virus-nAb MN50 markers
10. Baseline risk factors and the Day 57 bAb markers and the combination scores across the five markers [PCA1, PCA2, FSDAM1/FSDAM2 (the first two components of nonlinear PCA), and the maximum signal diversity score He and Fong (2019)].
11. Baseline risk factors and all individual Day 57 marker variables
12. Baseline risk factors and all individual Day 57 marker variables and all combination scores (full model)

Therefore in total, 12 variable sets are studied. The reason to include the first variable set is to investigate how much incremental improvement in predicting outcome is obtained by adding antibody marker variables on top of baseline demographic/exposure factors. The other variable sets are designed to compare the four immunoassay types by their predictiveness, to compare the two pseudovirus neutralization readouts ID50 and cID80 for their predictiveness, and to investigate incremental predictive value in using multiple immunoassays. The final variable set is included as the full model that considers all variables together, which serves as another reference model.

Table 7: Learning Algorithms in the Superlearner Library of Estimators of the Conditional Probability of Outcome, for Building the Baseline Risk Score Based on the Placebo Arm[1].

Algorithms	Screens/ Tuning Parameters
SL.mean	None
SL.glm	Low-collinearity and (All, Lasso, LR)[2]
SL.bayesglm	Low-collinearity and (All, Lasso, LR)
SL.glm.interaction	Low-collinearity and (All, Lasso, LR)
SL.glmnet	(alpha=1; All)
SL.gam	Low-collinearity and (Lasso, LR)
SL.ksvm	Low-collinearity and (kernel="rbfdot","polydot") and (Lasso, LR)
SL.polymars	Low-collinearity and (Lasso, LR)
SL.xgboost[3]	All and (maxdepth,shrinkage,balance)= (4, 0.1, no)
SL.ranger[3]	All and balance = no

[1]All continuous and ordinal covariates are pre-standardized to have empirical mean 0 and standard deviation 1.

[2]**All** = include all variables; **Lasso** = include variables with non-zero coefficients in the standard implementation of SL.glmnet that optimizes the lasso tuning parameter via cross-validation; **Low-collinearity** = do not allow any pairs of quantitative variables with Spearman rank correlation > 0.90; **LR** = Univariate logistic regression Wald test 2-sided p-value < 0.10.

[3]Covariate balancing (if requested) is done using option `scale_pos_weight` in SL.xgboost and option `case.weights` in SL.ranger.

Table 8: Learning Algorithms in the Superlearner Library of Estimators of the Conditional Probability of Outcome: Simplified Library in the Event of Fewer than 50 Vaccine Breakthrough Cases for an Analysis, for Use in Multivariable CoR Analysis of Moderna COVE[1].

Algorithms	Screens/ Tuning Parameters
SL.mean	None
SL.glm	Low-collinearity and (All, Lasso, LR)[2]
SL.glmnet	(All, alpha=0) (All, alpha = 1)
SL.xgboost	(maxdepth,shrinkage,balance[3])= (2, 0.1, yes) (2, 0.1, no) (4, 0.1, yes) (4, 0.1, no)
SL.ranger	balance = (yes, no)

[2]**All** = include all variables; **Lasso** = include variables with non-zero coefficients in the standard implementation of SL.glmnet that optimizes the lasso tuning parameter via cross-validation; **Low-collinearity** = do not allow any pairs of quantitative variables with Spearman rank correlation > 0.90; **LR** = Univariate logistic regression Wald test 2-sided p-value < 0.10.
[3]Covariate balancing (if requested) is done using option `scale_pos_weight` in SL.xgboost and option `case.weights` in SL.ranger.

Given the class-imbalance issue, with many more non-case than case records, all of the cross-validation for the multivariable immune CoR objective is done stratified by case/non-case status.

In order to evaluate the relative performance of the superlearner estimated models for each of the 12 variable sets, derived using the learning algorithms specified in Table 7, the CV-AUC is estimated with a 95% confidence interval (Hubbard et al., 2016; Williamson et al., 2020). The point and 95% confidence interval estimates of CV-AUC are reported in a forest plot, which provide a way to discern which Day 57 antibody assays and readouts/markers provide the most information in predicting COVID or other outcomes. The specified library of learners may be modified prior to SAP finalization (before breaking the blind of case/non-case status).

As noted above CV-AUC is estimated using the R package *vimp* available on CRAN, which uses augmented inverse probability weighting to properly estimate CV-AUC accounting for the two-phase sampling design.

If there are fewer than 50 vaccine breakthrough cases included in a correlates analysis, then the libary of learners will be simplified to that specified in

Table 8.

In addition, for selected variable sets, similar forest plots will be made comparing performance of the various estimated models (e.g., by individual learning algorithm types such as lasso), including discrete superlearner and superlearner models. The plot will be examined to determine which individual learning algorithm types are performing the best. If there is an interpretable algorithm that has performance close to the best-performing algorithm (which is most likely to be the superlearner), then it will be fit on the entire data set of vaccine recipients and the estimated model presented in a table.

Cross-validated ROC curves are plotted for the superlearner estimated models for each of the input variable sets. In addition, boxplots of cross-validated estimated probabilities of outcome by case-control status (as estimated from the superlearner models) are plotted.

12.6.5 [With Day 29 markers]

The weights $\hat{w}_{57.x}$ are used even for models that include Day 29 markers but not Day 57 markers, because only primary cases (starting 7 days post Day 57 visit) are included in the multivariable CoR analyses. Thus the Superlearner analyses only used the weights $\hat{w}_{57.x}$.

Regarding the 12 variable sets listed above, additional variable sets will be included to represent Day 29 markers as input variables. In particular, variable sets 2–12 will only include Day 57 markers. These variable sets will be cloned into 11 new sets (13-23) including Day 29 markers in place of Day 57 markers. Lastly, another 11 variable sets (24–34) will be formed with the same structure, where each input variable set includes both Day 29 and Day 57 markers. This allows the Superlearning modeling to address whether including markers from both time points improves prediction of outcome. Therefore, if Day 29 markers are included, the number of input variable sets is 34 instead of 12.

12.6.6 Variable set and individual variable importance

The importance of variable sets (and individual variables) will be summarized by the estimated gain in population prediction potential (also referred to as the intrinsic importance) when comparing each variable set plus baseline risk factors to baseline risk factors alone. Prediction potential (predictiveness) will be measured using CV-AUC. Inference on the intrinsic importance will be based off sample splitting; thus, both the estimated variable importance and the estimated CV-AUC of each variable set when evaluated on independent data from the data used to evaluate the CV-AUC of the baseline risk factors will be reported. The class-balancing versions of SL.xgboost will be dropped from the Super Learner library in the variable importance computation as the regression carried out to account for the two-phase sampling will be based on a continuous outcome (so there won't be any imbalance).

12.7 Multivariable CoR: Multivariable Cox models

12.7.1 Objectives

A complementary analysis to the multivariable super learner analysis will use several multivariable Cox models to examine associations between sets of markers and the hazard of COVID-19. This approach has the benefit over super learner of targeting more interpretable measures of the relative importance of individual markers after adjustment for other markers. The motivation is to provide additional evidence as to whether and to what extent single vs. multiple markers exhibit stronger overall signal as a correlate.

12.7.2 Standardization of markers

In order for the magnitude of the estimated hazard ratios to be comparable across markers, we will standardize all markers prior to inclusion in the Cox model by scaling by their estimated standard deviation. Estimation of the standard deviation should take inverse probability weights into account.

12.7.3 Primary multi-variable Cox model

The primary multivariable models will be two models (one for Day 57 markers, one for Day 29 markers) fit as described in the univariable model section above, but adjusting for sets of markers. In particular, our primary model includes binding Ab to RBD, pseudovirus neutralization ID50, and live virus neutralization MN50. A Wald test using the robust variance estimator as implemented in the `survey` package will be used to evaluate the null hypothesis of no association between any marker and hazard of COVID-19. Hazard ratios will be reported per standard deviation increase alongside 95% confidence intervals.

12.7.4 Secondary multi-variable Cox models

We fill additionally fit a series of exploratory, bivariate Cox models. These models will be fit as above but will include only two markers. The sets of two markers that we consider are

1. binding Ab to RBD and pseudovirus neutralization ID50
2. binding Ab to RBD and live virus neutralization MN50

As above, Wald tests will be used to evaluate the null hypothesis of no association between any marker and hazard of COVID-19. Hazard ratios will be reported per standard deviation increase alongside 95% confidence intervals.

13 Correlates of Protection: Generalities

In general, for all of the correlate of protection analyses, the same antibody markers are assessed that were analysed as correlates of risk: the Day 57 antibody markers not subtracting for the Day 1 baseline readout are used. Each of the five Day 57 antibody biomarkers are separately studied as CoPs by the different analysis approaches summarized below.

13.0.1 [With Day 29 markers]

If Day 29 markers are included, then the same CoP analyses are done as for each of the five Day 57 markers, where, as for the CoR analyses, now the time origin is the Day 29 visit date and endpoint cases are counted starting 7 days after the Day 29 visit date.

14 Correlates of Protection: Correlates of Vaccine Efficacy Analysis Plan

For each of the five Day 57 antibody biomarkers, the method of Gilbert, Blette, Shepherd, and Hudgens (2020) will be used to estimate $VE(1)$, $VE(0)$, and $VE(1) - VE(0)$, each with a 95% confidence interval and a 95% estimated uncertainty interval (EUI), where $VE(1)$ is vaccine efficacy for the subgroup of vaccine recipients with Day 57 marker if assigned vaccine $S(1)$ above a specified cut-point value s_{cut}, and $VE(0)$ is vaccine efficacy for the subgroup of vaccine recipients with Day 57 marker if assigned vaccine $S(1)$ not greater than s_{cut}. That is,

$$VE(1) = 1 - \frac{P(Y(1) = 1 | S(1) > s_{cut})}{P(Y(0) = 1 | S(1) > s_{cut})}$$
$$VE(0) = 1 - \frac{P(Y(1) = 1 | S(1) \leq s_{cut})}{P(Y(0) = 1 | S(1) \leq s_{cut})}$$

The analysis will be done under the **NEH** assumption ("no early harm") of Gilbert et al. (2020). The cut point is defined as the percentile equal to one minus the estimated vaccine efficacy in the primary analysis, with logic that a maximally simple version of a perfect CoP would have binary marker with $S = 1$ corresponding to protection and $S = 0$ corresponding to no protection. If the estimated vaccine efficacy is high (say 90% or higher), it is possible that this cutpoint will not yield stable results, because of sparse cells; in this situation we will repeat the analysis using two additional cut-points that creates greater balance in frequencies of $S = 1$ and $S = 0$ in the vaccine group immunogenicity subcohort: 20th and 40th percentiles. If the estimated vaccine efficacy is moderate (between 50% and 80%), we will also use the two

additional cut-points the 20th and 40th percentiles. This analysis method does not require closeout placebo vaccination (CPV) (Follmann, 2006) or a good baseline immunogenicity predictor of the Day 57 antibody marker. The method is implemented using Bryan Blette's R package "psbinary" posted at his Github repository.

A limitation of the Gilbert et al. method is that it only assesses a binary biomarker. Other analyses will be considered to estimate $VE(s)$ over biomarker values s over the entire range, treating S as a quantitative or categorical variable, and gaining efficiency by incorporating CPV and/or putative baseline immunogenicity predictors (BIPs). Based on earlier simulation studies (Follmann, 2006; Huang et al., 2013, e.g.,), methods that only leverage CPV data tend to have low power relative to methods that leverage BIP data alone (BIP-only methods) or both BIP and CPV data (BIP+CPV methods). Therefore, the key for improving efficiency will be the availability of a BIP. VE curve analysis for continuous S will thus be conducted contingent on the availability of a BIP that satisfies the R^2 criterion outlined in Table 10. It is anticipated that post-crossover immune response marker data will not be available in early correlates analyses, and so BIP-only methods will be used in these initial analyses. When CPV data becomes available, new BIP+CPV analyses will be conducted that incorporate this new information. Details of the BIPs used can be found at the end of this section.

Let $Y(a)$ denote the potential binary outcome of interest if receiving intervention a, with $a = 1, 0$ standing for assignment to vaccine and placebo, respectively. Let $S(a)$ denote the potential biomarker value if receiving intervention a. The vaccine efficacy curve (Follmann, 2006; Gilbert and Hudgens, 2008) is defined as the curve of vaccine efficacy as a function of the immune response biomarker if assigned vaccination (i.e., $S(1)$): $VE(s) = 1 - P(Y(1) = 1|S(1) = s)/P(Y(0) = 1|S(1) = s)$. It characterizes the percentage reduction in clinical risk under vaccine assignment compared to under placebo assignment conditional on $S(1)$ and informs about the magnitude of potential immune response associated with certain levels of VE. Consider the existence of BIPs X correlated with $S(1)$ and/or a CPV component in the trial where a subset of placebo recipients free of the outcome are vaccinated and have

their immune response biomarkers measured as substitutes for $S(1)$. Under the NEE assumption and assuming the set of participants with $S(1)$ available is nested within the set of participants with BIP measures, the pseudo-score estimation method (Huang et al., 2013; Zhuang et al., 2019) based on discrete BIP measures allowing for adjustment of X will be adopted for estimating the risk model $P(Y(z) = 1|S(1), X)$ and subsequently $VE(s) = 1 - \int P(Y(1) = 1|S(1), x)dF_X(x|S(1))/\int P(Y(0) = 1|S(1), x)dF_X(x|S(1))$. Hypothesis testing will be conducted for testing the null hypothesis that the VE curve is constant (Zhuang et al., 2019). Estimated parametric (Gilbert and Hudgens, 2008), semiparametric (Huang and Gilbert, 2011), or nonparametric (Li and Luedtke, 2020) likelihood estimators of VE curves will be applied to continuous BIPs. In scenarios where some BIPs are not measured from all trial participants, VE curve estimators accounting for this monotone missingness in X and $S(1)$ will be adopted (Huang, 2018). If the data support positive vaccine efficacy before Day 57, sensitivity analysis approaches will be conducted for VE curve estimation under the NEH assumption. In the presence of multiple candidate biomarkers and when a CPV component is present, a multiple imputation approach as proposed in Dasgupta and Huang (2019) will be utilized to impute missing $S(1)$ data for selecting markers from multiple candidates and deriving a univariate marker score for VE curve estimation.

Finally, for scenarios with very rare events such that methods described above lack precision even with a CPV component but where the available BIP still satisfies the R^2 criterion outlined in Table 10, we will adopt sensitivity analysis methods that model the placebo risk conditional on the counterfactual $S(1)$ based on a sensitivity parameter that varies over some pre-specified range.

Among different strategies to identify BIPs, the following will be tried. First, for vector vaccines, we will study Day 1 bAb or nAb response to the vector as a BIP for the Day 57 markers of interest. Second, we will check whether Day 1 bAb or nAb to Nucleocapsid protein is a BIP for the anti-Spike/anti-RBD Day 57 markers of interest. The rationale for this latter analysis is that some studies have shown cross-reactive responses to Nucleocapsid protein and to common circulating human coronaviruses.

We will also evaluate using a multivariate BIP that corresponds to all of these aforementioned candidate univariate BIPs, which may help to achieve the target R^2 (see Table 10). When doing this, a separate BIP W will be used for each vaccine-induced immune response marker $S(1)$. Let $Y(a)$ be the counterfactual outcome of interest — e.g., a COVID disease endpoint by a prespecified time — if randomization assignment had been set to $A = a$. The analyses conducted will provide unbiased estimates of the estimands of interest when $Y(a) \perp W|S(1)$ for $a \in \{0, 1\}$. The BIP W will be a learned function of baseline covariates L — that is, $W = f(L)$ for a function f that will be learned based on the available data. All available baseline covariates will be considered for inclusion in L, including age, BMI, and Day 1 bAb or nAb to Nucleocapsid protein. If available, measurements of prior immune response to the vaccine vector (e.g., Day 1 bAb or nAb response to Ad26 for an Ad26 vector-based vaccine) will always be included in L.

If the trial of interest has more than 100 events on the vaccine arm in the subgroup of interest, then f will be chosen to be an estimate of the following population-level optimization problem:

$$\begin{aligned} &\text{minimize } E[\{S - f(L)\}^2|A = 1] \\ &\text{subject to } f(L) \perp Y|A = 1, S. \end{aligned}$$

The rationale for choosing f to (approximately) solve this optimization problem is that the BIP should be maximally predictive of S, while also satisfying the needed conditional independence assumption $Y(a) \perp W|S(1)$ when $a = 1$. Moreover, the needed conditional independence assumption $Y(a) \perp W|S(1)$ for the case that $a = 0$ is most plausible when this assumption is also satisfied for the case that $a = 1$. Also, because $W = f(L)$ for some function f, $Y(0) \perp W|S(1)$ is always more plausible than $Y(a) \perp L|S(1)$.

The solution to the above optimization problem is given by:

$$f(\ell) := \theta(\ell) - \frac{E[\theta(L)r(L)]}{E[r(L)^2]} r(\ell)$$

where $\theta(\ell) := E\{S|A = 1, L = \ell\}$, $r(\ell) := \frac{m(\ell)}{E[m(L)]} - \frac{1-m(\ell)}{1-E[m(L)]}$ and $m(\ell) := E[Y|A = 1, L = \ell]$. The following strategy is used to estimate this solution:

1. Obtain an estimate $\hat{\theta}$ of the function θ by running a Superlearner of S against L in the vaccine arm, where inverse probability of sampling weights are used to account for two-phase sampling of the marker.
2. Obtain an estimate $\hat{m}$ of m by using Superlearner to regress Y against L in the vaccine arm.
3. Obtain an estimate $\hat{r}$ via a plug-in estimator, where $E[m(L)]$ is estimated by taking the empirical mean of $\hat{m}(L)$.
4. The final estimate $\hat{f}$ of f is given by

$$\hat{f}(\ell) := \hat{\theta}(\ell) - \frac{\hat{E}[\hat{\theta}(L)\hat{r}(L)]}{\hat{E}[\hat{r}(L)^2]}\hat{r}(\ell),$$

   where $\hat{E}$ denotes an empirical expectation.

Each Superlearner will be run using the same library and settings described in Table 9. If the trial has fewer than 100 events on the vaccine arm, then the function f will be learned via Step 1 above only — that is, we will take $\hat{f} = \hat{\theta}$. All standard errors will be obtained via the bootstrap, with the above fitting of $\hat{f}$ redone within each bootstrap sample.

15 Correlates of Protection: Interventional Effects

In these analyses, we seek to understand whether, how, and to what extent Day 57 antibody markers impact vaccine efficacy in causal ways. We describe three approaches to this problem. Each involves consideration of a binary counterfactual outcome $Y(a, s)$ (e.g., indicator of the COVID disease endpoint by a pre-specified time) under a hypothetical intervention that both sets randomization assignment $A = a$ and sets the Day 57 immunologic marker S to a fixed value or based upon a random draw from a analyst-specified distribution. Below, we assume that S is scalar-valued, but some of the approaches below naturally extend to the case where a vector of immunologic markers are considered (currently such analyses are not planned). Given the central goal to develop a parsimonious surrogate endpoint based on a single immunoassay, the main analysis will use each of the methods to

assess each of the five quantitative readouts (not baseline-subtracted) separately as CoPs, adjusting for the same set of baseline covariates as used in the CoR analyses previously described in Section 12.

15.1 CoP: Controlled Vaccine Efficacy

We first describe the controlled vaccine efficacy curve defined as

$$\text{CVE}(s) = 1 - \frac{P(Y(1,s)=1)}{P(Y(0)=1)} \, .$$

The value CVE(s) takes represents the relative decrease in endpoint frequency achieved by administering vaccine and setting Day 57 immunologic marker level to s compared to the placebo control intervention. Under our approach, the value of CVE(s) is assumed to be monotone non-decreasing in s; in other words, vaccine efficacy can only potentially be improved by setting greater marker levels. The extent to which the marker plays a role in determining vaccine efficacy can be determined by the degree of flatness of the graph of CVE(s) versus s.

In addition, because the primary study cohort for correlates analysis is naive to SARS-CoV-2, each of the Day 57 markers S has no variability in the placebo arm [all values are 'negative,' below the assay lower limit of detection (LLOD)]. Therefore, advantageously in this setting $CVE(s)$ has a special connection to the mediation literature, where $CVE(s = LLOD)$ is the natural direct effect, and vaccine efficacy is 100% mediated through S if and only if $CVE(s = LLOD) = 0$. Thus inference on $CVE(s = LLOD)$ evaluates full mediation.

Since $P(Y(0)=1) = P(Y=1 \,|\, A=0)$ in view of vaccine versus placebo randomization, the controlled vaccine efficacy CVE(s) at level s can be identified using the fact that

$$P(Y(1,s)=1) = E\left[P(Y=1 \,|\, S=s, A=1, X)\right]$$

whenever $Y(1,s)$ and S are independent given $A=1$ and a vector X of covariates, and $P(S=s \,|\, A=1, X) > 0$ almost surely. In other words, identification of the controlled vaccine efficacy requires that a rich enough

set of covariates be available so that deconfounding of the relationship between endpoint Y and marker S is possible in the subpopulation of vaccine recipients, and that marker level $S = s$ may occur within each subpopulation defined by values of the covariates X (positivity).

For each s, the identified parameter corresponding to $\text{CVE}(s)$ is an irregular parameter within nonparametric models, making its estimation at root-n rate impossible; this significantly complicates estimation and inference on $\text{CVE}(s)$. Fortunately, the monotonicity of $s \mapsto \text{CVE}(s)$ provides an opportunity to circumvent these difficulties. Similarly to Westling et al. (2020a)'s approach for the causal dose-response function, we will use the general methodological template proposed in Westling and Carone (2020) to derive (i) a nonparametric Grenander-type estimator of $\text{CVE}(s)$ and (ii) a plug-in confidence interval for $\text{CVE}(s)$ based on an asymptotic Chernoff limit. This estimator will require, as an intermediate step, estimation of several nuisance functions, including the outcome regression $P(Y = 1 \,|\, S = s, A = 1, X = x)$ and the propensity score $P(S = s \,|\, X = x, A = 1)$. These nuisance functions will be estimated using the Superlearner ensembling algorithm with a rich library including both parametric regression methods as well as flexible machine learning tools.

The monotonicity-based procedure we will develop facilitates statistical inference for $\text{CVE}(s)$ for each s separately, where point estimates and 95% confidence intervals for $\text{CVE}(s)$ will be presented. However, it is also of interest to investigate whether the Day 57 marker plays a role in determining vaccine efficacy. To do so, we will formally test the null hypothesis

$$H_0 : \text{CVE}(s) \text{ is constant in } s$$

against various alternatives. We will first adapt the approach of Westling (2020) to devise a nonparametric omnibus test of this null hypothesis. We will also construct a nonparametric directional test of this hypothesis tailored to alternatives under which $\text{CVE}(s)$ is monotone in s, along the lines of Hall and Heckman (2000), for example. Leveraging the known monotonicity of the controlled vaccine efficacy will provide greater power than omnibus tests.

15.1.1 Conservative (upper bound) inference and sensitivity analysis for the Cox model correlates of risk analysis

While the above nonparametric approach is considered to be the best scientific approach because it takes the greatest care to avoid the correctness of inferences depending on parametric modeling assumptions that cannot be fully verified, we also apply the same Cox modeling approach described in Section 12.3.3, augmented with a sensitivity analysis, which harmonizes with the CoR analysis, and sensitivity analysis is generally warranted when a no unmeasured confounders assumption is made. The sensitivity analysis quantifies the rigor of evidence for a controlled VE CoP after accounting for potential bias from unmeasured confounding.

Gilbert et al. (2020a) details the sensitivity analysis approach, which was applied to the CYD14 and CYD15 dengue phase 3 data sets (Moodie et al., 2018); we plan to apply it in the same way to the COVID-19 data sets (as the structure of the problem is the same). We summarize here the essential details needed for application to the COVID-19 data sets.

We define S to be a controlled risk CoP if $P(Y(1,s)=1)$ is monotone non-increasing in s with $P(Y(1,s)=1) > P(Y(1,s')=1)$ for at least some $s < s'$, where point and 95% confidence interval estimates of $P(Y(1,s)=1)$ versus s, with built in robustness to unmeasured confounding, describe the strength of the CoP in terms of the amount and nature of decrease. Suppose the CoR analysis based on the Cox model is conducted as described in Section 12.3.3.

Let marginalized conditional risk

$$r_M(s) = risk_1(t_F|s)$$

and controlled risk

$$r_C(s) = P(Y(1,s)=1).$$

Given that CoR analysis is based on observational data — the biomarker value is not randomly assigned — a central concern is that unmeasured or uncontrolled confounding of the association between S and Y could render $r_M(s) \neq r_C(s)$, biasing estimates of the controlled risk curve $r_C(s)$ and of

controlled risk ratios of interest

$$RR_C(s_1, s_2) = r_C(s_2)/r_C(s_1) \ .$$

Because we can never be certain that confounding is adequately adjusted for, sensitivity analysis is warranted, as considered in extensive literature — see, e.g., VanderWeele and Ding (2017) and references therein. Sensitivity analysis is useful to evaluate how strong unmeasured confounding would have to be to explain away an observed causal association, that is, to determine the strength of association of an unmeasured confounder between S and Y needed for the observed exposure-outcome association to not be causal, $r_M(s) \neq r_C(s)$. We follow the recommendation of VanderWeele and Ding (2017) to report the E-value as a summary measure of the evidence of causality, or, in our application, evidence of whether S is a controlled risk CoP based on variation in the controlled risk curve. We also include other closely related measures of sensitivity.

The E-value is the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the exposure (S) and the outcome (Y) in order to fully explain away a specific observed exposure–outcome association, conditional on the measured covariates [VanderWeele and Ding (2017); VanderWeele and Mathur (2020)]. If, as in CoP analyses, the estimated marginalized risk ratio $\widehat{RR}_M(s_1, s_2) = \widehat{r}_M(s_2)/\widehat{r}_M(s_1)$ for $s_1 < s_2$ is less than one, then the E-value for $\widehat{RR}_M(s_1, s_2)$ is calculated as

$$e_{RR}(s_1, s_2) = \frac{1 + \sqrt{1 - \widehat{RR}_M(s_1, s_2)}}{\widehat{RR}_M(s_1, s_2)} \ . \tag{19}$$

We include the argument (s_1, s_2) in the notation, with $s_1 < s_2$ by convention, to be clear that the E-value depends on specification of two specific marker-level subgroups.

To illustrate the interpretation of an E-value, suppose S is binary and regression analysis yields an estimate $\widehat{RR}_M(0, 1) = \widehat{r}_M(1)/\widehat{r}_M(0) = 0.40$ with 95% confidence interval (CI) $(0.14, 0.78)$. An E-value $e(0, 1)$ of 4.4 means

that a marginalized risk ratio $RR_M(0,1)$ at the observed value 0.40 could be explained away (i.e., $RR_C(0,1) = 1.0$) by an unmeasured confounder associated with both the exposure and the outcome by a marginalized risk ratio of 4.4-fold each, after accounting for the vector X of measured confounders, but that weaker confounding could not do so.

In addition, we follow the recommendation of VanderWeele and Ding (2017) to also report the E-value $e_{UL}(s_1, s_2)$ for the upper limit $\widehat{UL}(s_1, s_2)$ of the 95% CI for the observed marginalized risk ratio $\widehat{RR}_M(s_1, s_2)$, computed as 1 if $\widehat{UL}(s_1, s_2) \geq 1$ and, otherwise, as

$$\frac{1 + \sqrt{1 - \widehat{UL}(s_1, s_2)}}{\widehat{UL}(s_1, s_2)},$$

which in the example equals $e_{UL}(0,1) = 1.88$. This E-value for the upper limit indicates, for given $s_1 < s_2$, the strength of unmeasured confounding at which statistical significance of the inference that $RR_C(s_1, s_2) < 1$ would be lost. The two E-values above are useful for judging how confident we can be that an immunologic biomarker is a controlled risk CoP, with E-values near one suggesting weak support and evidence increasing with greater E-values.

$RR_C(s_1, s_2) = (1 - CVE(s_2))/(1 - CVE(s_1))$, evidence for $RR_C(s_1, s_2) < 1$ is equivalently evidence for $CVE(s_1) < CVE(s_2)$. Thus in a placebo-controlled trial $RR_C(s_1, s_2)$ can be interpreted as the multiplicative degree of superior vaccine efficacy caused by marker level s_2 vs. marker level s_1, and E-values equivalently quantify evidence for whether $CVE(s_1)$ differs from $CVE(s_2)$.

It is also useful to provide conservative estimates of controlled risk ratios and of the controlled risk curve, accounting for unmeasured confounding. We approach these tasks based on the sensitivity analysis, or bias analysis, approach of Ding and VanderWeele (2016). We give their main result and refer readers to the paper for details. We begin by defining two (possibly context-specific) fixed sensitivity parameters. First, we set $RR_{UD}(s_1, s_2)$ to be the maximum risk ratio for the outcome Y comparing any two categories of the unmeasured confounders U, within either exposure group $S = s_1$ or $S = s_2$, conditional on the vector X of observed covariates. Second, we set

$RR_{EU}(s_1, s_2)$ to be the maximum risk ratio for any specific level of the unmeasured confounder U comparing individuals with $S = s_1$ to those with $S = s_2$, with adjustment already made for the measured covariate vector X. Thus, $RR_{UD}(s_1, s_2)$ quantifies the importance of the unmeasured confounder U for the outcome, and $RR_{EU}(s_1, s_2)$ quantifies how imbalanced the exposure/marker subgroups $S = s_1$ and $S = s_2$ are in the unmeasured confounder U. The values $RR_{UD}(s_1, s_2)$ and $RR_{EU}(s_1, s_2)$ are always specified as greater than or equal to one. We suppose that $RR_M(s_1, s_2) < 1$ for the fixed values $s_1 < s_2$ — this is the case of interest for immune correlates.

Define the bias factor

$$B(s_1, s_2) = \frac{RR_{UD}(s_1, s_2)RR_{EU}(s_1, s_2)}{RR_{UD}(s_1, s_2) + RR_{EU}(s_1, s_2) - 1}$$

for $s_1 \leq s_2$, and define $RR^U_M(s_1, s_2)$ the same way as $RR_M(s_1, s_2)$, except marginalizing over the joint distribution of X and U. Then, $RR^U_M(s_1, s_2) \leq RR_M(s_1, s_2) \times B(s_1, s_2)$, where $RR^U_M(s_1, s_2) = E\{r(s_2, X^*)\}/E\{r(s_1, X^*)\}$ with $X^* = (X, U)$ and r conditional risk defined near equation (**??**).Ding and VanderWeele (2016)

Translating this result to our problem context, under the positivity assymption, we have that $RR^U_M(s_1, s_2) = RR_C(s_1, s_2)$ and so, it follows that

$$RR_C(s_1, s_2) \leq RR_M(s_1, s_2) \times B(s_1, s_2) \ . \tag{20}$$

This inequality states that the causal risk ratio is bounded above by the marginalized risk ratio multiplied by the bias factor. It follows that a conservative (upper bound) estimate of $RR_C(s_1, s_2)$ is obtained as $\widehat{RR}_M(s_1, s_2) \times B(s_1, s_2)$, and a conservative 95% CI is obtained by multiplying each confidence limit for $RR_M(s_1, s_2)$ by $B(s_1, s_2)$. These estimates for $RR_C(s_1, s_2)$ account for the presumed-maximum plausible amount of deviation from the no unmeasured confounders assumption specified by $RR_{UD}(s_1, s_2)$ and $RR_{EU}(s_1, s_2)$. An appealing feature of this approach is that the bound (20) holds without making any assumption about the confounder vector X or the unmeasured confounder U.

The above approach does not directly provide a conservative estimate of the controlled risk curve $r_C(s)$, because additional information is needed for absolute versus relative risk estimation. To provide conservative inference for $r_C(s)$, we next select a central value s^{cent} of S such that $\widehat{r}_M(s^{cent})$ matches the observed overall risk, $\widehat{P}(Y = 1|A = 1)$. This value is a 'central' marker value at which the observed marginalized risk equals the observed overall risk. Next, we 'anchor' the analysis by assuming $r_C(s^{cent}) = r_M(s^{cent})$, where picking the central value s^{cent} makes this plausible to be at least approximately true. Under this assumption, the bound (20) implies the bounds

$$r_C(s) \leq r_M(s)B(s^{cent}, s) \quad \text{if } s \geq s^{cent} \tag{21}$$

$$r_C(s) \geq r_M(s)\frac{1}{B(s, s^{cent})} \quad \text{if } s < s^{cent}. \tag{22}$$

Therefore, after specifying $B(s^{cent}, s)$ and $B(s^{cent}, s)$ for all s, we conservatively estimate $r_c(s)$ by plugging $\widehat{r}_M(s)$ into the formulas (21) and (22). Because $B(s_1, s_2)$ is always greater than one for $s_1 < s_2$, formula (21) pulls the observed risk $\widehat{r}_M(s)$ upwards for subgroups with high biomarker values, and formula (22) pulls the observed risk $\widehat{r}_M(s)$ downwards for subgroups with low biomarker values. This makes the estimate of the controlled risk curve flatter, closer to the null curve, as desired for a sensitivity/robustness analysis.

To specify $B(s_1, s_2)$, we note that it should have greater magnitude for a greater distance of s_1 from s_2, as determined by specifying $RR_{UD}(s_1, s_2)$ and $RR_{EU}(s_1, s_2)$ increasing with $s_2 - s_1$ (for $s_1 \leq s_2$). We consider one specific approach, which sets $RR_{UD}(s_1, s_2) = RR_{EU}(s_1, s_2)$ to the common value $RR_U(s_1, s_2)$ that is specified log-linearly: $\log RR_U(s_1, s_2) = \gamma(s_2 - s_1)$ for $s_1 \leq s_2$. Then, for a user-selected pair of values $s_1 = s_1^{fix}$ and $s_2 = s_2^{fix}$ with $s_1^{fix} < s_2^{fix}$, we set a sensitivity parameter $RR_U(s_1^{fix}, s_2^{fix})$ to some value above one. It follows that

$$\log RR_U(s_1, s_2) = \left(\frac{s_2 - s_1}{s_2^{fix} - s_1^{fix}}\right) \log RR_U(s_1^{fix}, s_2^{fix}), \quad s_1 \leq s_2.$$

We anchor the sieve analysis by setting $s_1 = s_1^{fix}$ at the 15^{th} percentile of the

Day 57 antibody marker and $s_2 = s_2^{fix}$ at the 85^{th} percentile of the Day 57 antibody marker.

The sensitivity analysis is done for each of the two Cox model CoR analyses described in Section 12.3.3, first for tertiles of the Day 57 marker and second for the quantitative marker. For the former, E-values are reported for both the point estimate and the upper 95% confidence limit for $RR_C(0,1)$, where category 1 is the upper tertile, category 0 is the lower tertile, and the intermediate middle tertile subgroup of vaccine recipients is excluded from the analysis. In addition, setting $RR_{UD}(0,1) = RR_{EU}(0,1) = 2$, such that $B(0,1) = 4/3$, we report conservative estimation and inference on the causal risk ratio $RR_C(0,1)$ and equivalently on the ratio of controlled vaccine efficacy curves $(1 - CVE(1))/(1 - CVE(0))$.

Next we repeat the analysis treating S as a quantitative variable, where $P(T \leq t|S = s, X, A = 1)$ is again estimated by two-phase Cox partial likelihood regression and now $RR_M(s_1, s_2)$ is the marginalized risk ratio between s_1 and s_2. We will plot point and 95% confidence interval estimates of the observed marginalized risk and controlled risk curves, for the latter using the sensitivity analysis described in Section 15.1.1.

For validity the method requires the positivity assumption, and thus the method will only be applied if the data are reasonably supportive of the positivity assumption. To check positivity, we study the antibody marker distribution in vaccine recipients within each subgroup of the covariates X that are adjusted for. For the tertiles analysis we require evidence that within each subgroup some vaccine recipients have lower tertile responses and some vaccine recipients have upper tertile responses. For the quantitative S analysis, we look for evidence that S varies over its full range within each level of the potential confounders that are adjusted for.

15.2 CoP: Stochastic Interventional Effects on Risk and Vaccine Efficacy

Another approach to studying correlates of protection involves estimating the effect of shifting the immune response marker distribution in the vaccinated individuals (Hejazi et al., 2020a). Specifically, we can consider the effect on

risk of a given endpoint of a controlled intervention that shifts the distribution of an immune response by δ units, where δ is an analyst-specified real number. Considering a counterfactual scenario in which we are able to intervene so as to modify the immune response induced by the vaccine (e.g., a hypothetical change in dose or other re-formulation of the vaccine), we take this hypothetical intervention to lead to an improved (if $\delta > 0$) or lessened immune response (if $\delta < 0$) relative to the current vaccine (at $\delta = 0$). Using this framework, we can query the counterfactual risk of the endpoint under this hypothetical vaccine. Using notation established above, this quantity can be expressed as the mean of the counterfactual variable $Y(1, S(1) + \delta)$.

This approach is similar to the controlled effects approach described in Section 15.3, but with an important distinction. In the controlled effects approach, one assumes that it is possible to set $S = s$ for *all* individuals in the population. For high values of s, this assumption may be unrealistic if the vaccine fails to be strongly immunogenic for some subpopulations. On the other hand, with the interventional approach, it is only required that individuals' immune responses be shifted relative to their observed immune response, which may be more plausible for some vaccines.

Under assumptions (Hejazi et al., 2020a), the main two of which being no unmeasured confounding and positivity (forms of both are also required for the Controlled VE CoP analyses), the counterfactual risk of interest $E[Y(1, S(1) + \delta)]$ is identified by

$$E[P(Y = 1 \mid A = 1, S = S + \delta, X = x) \mid A = 1, X] \ .$$

Examining this quantity across a range of δ provides insight into the relative contribution of a given immune response marker in preventing the endpoint of interest.

Hejazi et al. (2020a) proposed nonparametric estimators that rely on estimates of the outcome regression (as described above) and the conditional density of the immune response marker in vaccinated participants. Their estimators efficiently account for two-phase sampling of immune responses and are implemented in the `txshift` package (Hejazi and Benkeser, 2020) for the `R` language and environment for statistical computing (R Core Team, 2020),

available via both GitHub at https://github.com/nhejazi/txshift and the Comprehensive R Archive Network at https://CRAN.R-project.org/package=txshift.

These estimators will be applied to each of the five Day 57 antibody markers (without baseline adjustment) controlling for the same set of baseline risk factors that are controlled for in other analyses previously discussed. As with the mediation analysis approach described in Section 15.3, the procedure will leverage low-dimensional risk factors alongside parametric regression strategies and flexible conditional density estimators for endpoints with fewer than 100 observed cases (pooling over the randomization arms); however, more flexible learning techniques will be employed for modeling the outcome process for endpoints with a greater number of observed cases.

In particular, conditional density estimates of immune response markers will be principally based on a nonparametric estimation strategy that reconstructs the conditional density through estimates of the conditional hazard of the discretized immune response marker values (Hejazi et al., 2020a,d,c); this approach is an extension of the proposal of Díaz and van der Laan (2011). A Super Learner ensemble (van der Laan et al., 2007) of variants of this nonparametric conditional density estimator and semiparametric conditional density estimators based on Gaussinization of residuals will be constructed using the `sl3` R package (Coyle et al., 2020). In settings with limited numbers of case endpoints, the outcome process will be modeled as a Super Learner ensemble of a library of parametric regression techniques (as recommend by Gruber and van der Laan, 2010), while the library will be augmented with flexible regression techniques — including, for example, lasso and ridge regression (Tibshirani, 1996; Tikhonov and Arsenin, 1977; Hoerl and Kennard, 1970), elastic net regression (Zou and Hastie, 2003; Friedman et al., 2009), random forests (Breiman, 2001; Wright et al., 2017), extreme gradient boosting machines (Chen and Guestrin, 2016), light and efficient gradient boosting machines (Ke et al., 2017), multivariate adaptive polynomial and regression splines (Friedman et al., 1991; Stone et al., 1994; Kooperberg et al., 1997), and the highly adaptive lasso (van der Laan, 2017; Benkeser and van der Laan, 2016; Hejazi et al., 2020b) — as the number of endpoint cases grows.

These algorithm libraries will be coordinated to match those used in other CoP analyses.

Additionally, we recall that $P(Y(0) = 1) = P(Y = 1 \mid A = 0)$ (in view of vaccine versus placebo randomization, as stated previously in Section 15.1) and may be estimated in the same way as for the analysis of controlled vaccine efficacy, thus yielding an estimate of stochastic interventional VE defined by

$$SVE(\delta) = 1 - \frac{E[P(Y = 1 \mid A = 1, S = S + \delta, X = x) \mid A = 1, X]}{P(Y(0) = 1)}.$$

Output of the analyses will be presented as point and 95% point-wise confidence interval estimates of $E[Y(1, S(1)+\delta)]$ and of $SVE(s)$ over the values of s for each of the Day 57 antibody markers, for each of a range of δ spanning -2 to 2 on the standard unit scale for each antibody marker.

Lastly, just as for the controlled VE CoP analyses, these analyses will only be performed if diagnostics support plausibility of the positivity assumption. Importantly, however, the positivity assumption for the stochastic interventional effects differs from that usually required. That is, where the positivity assumption for effects defined by static interventions requires a positive probability of treatment assignment across all strata defined by baseline factors (i.e., that a discretized immune response value be possible regardless of baseline factors), the positivity assumption of these effects is

$$s_i \in \mathcal{S} \implies s_i + \delta \in \mathcal{S} \mid A = 1, X = x$$

for all $x \in \mathcal{X}$ and $i = 1, \ldots n$. In particular, this positivity assumption does not require that the post-intervention exposure density, $q_{0,S}(S-\delta \mid A = 1, X)$, place mass across all strata defined by X. Instead, it requires that the post-intervention exposure mechanism be bounded, i.e.,

$$P\{q_{0,S}(S - \delta \mid A = 1, X)/q_{0,S}(S \mid A = 1, X) > 0\} = 1,$$

which may be readily satisfied by a suitable choice of δ.

More importantly, the static intervention approach may require consideration of counterfactual variables that are scientifically unrealistic. Namely, it may

be inconceivable to imagine a world where every participant exhibits high immune responses, given the phenotypic variability of participants' immune systems. This too may be resolved by considering an intervention $\delta(X)$, allowing the choice of δ to be a function of baseline covariates X (Hejazi et al., 2020a; Díaz and van der Laan, 2012; Haneuse and Rotnitzky, 2013; Díaz and van der Laan, 2018).

15.3 CoP: Mediation of Vaccine Efficacy

Using mediation methods, we can decompose the overall VE into so-called *natural* direct and indirect effects. We will estimate this decomposition for each Day 57 antibody marker individually (focusing on the non-baseline subtracted markers as for the other CoP analyses described above), as well as when considering all antibody markers together (although this SAP currently restricts to analysis of the individual markers).

For simplicity, as before, we describe this approach using a binary outcome, noting that extensions to time-to-event (with competing risks) are possible. The *total* effect of the vaccine can be represented by one minus the risk ratio

$$\text{RR} = \frac{P(Y(1, S(1)) = 1)}{P(Y(0, S(0)) = 1)} .$$

The natural direct and indirect effects are, respectively,

$$\text{RR}_{DE} = \frac{P(Y(1, S(0)) = 1)}{P(Y(0, S(0)) = 1)} \quad \text{and} \quad \text{RR}_{IDE} = \frac{P(Y(1, S(1)) = 1)}{P(Y(1, S(0)) = 1)} .$$

Note that $\text{RR} = \text{RR}_{DE}\text{RR}_{IDE}$, showing that the total effect decomposes into the direct times indirect effect. Another quantity of interest is the proportion mediated, which could be expressed as

$$\text{PM} = 1 - \frac{\log(\text{RR}_{DE})}{\log(\text{RR})} .$$

We note that PM=1 if and only if $\text{RR}_{DE} = 1$, i.e., no direct effect means that the marker fully mediates VE. We will estimate PM defined in this way.

As above, we must assume all confounders X of S and Y have been measured. We also assume there are no confounders of the mediator-outcome

relationship that are affected by treatment. Moreover, we require an overlap assumption that

$$P(S = s|A = 0, X = x) > 0 \text{ implies } P(S = s|A = 1, X = x) > 0 \qquad (23)$$

for all subgroups $X = x$ (i.e., a.e.). Under these assumptions, $P(Y(a, S(a') = 1)$ is identified by

$$E[P(Y = 1 \mid A = a, S, X)|A = a', X] \ .$$

In our immune CoP application it is expected that, for analyses restricting to baseline negative individuals, the conditional density of the immune response marker in the placebo arm will be a point mass at 0, that is with S below the LLOD. In other words, we do not expect any placebo recipients to have a positive value of the immune response marker. This implies the identification result that for $a = 0, 1$, $P(Y(a, S(0)) = 1) = E[P(Y = 1 \mid A = a, S = 0, X)]$. While $P(Y(0, S(1) = 1)$ is not identified, it is not necessary to estimate this term in order for estimation of the parameters of interest (natural direct effect, natural indirect effect, PM).

For a highly immunogenic vaccine, it may be the case that the needed overlap assumption (23) will be violated. This could happen, for example if each baseline negative placebo recipient has antibody marker value below the assay's LLOD (which is expected), and every vaccine recipient has antibody marker value above the LLOD. We will only include antibody markers for mediation analysis if at least 10% of vaccine recipients have marker value equal to the value in placebo recipients.

Benkeser et al. (2021) provide a multiply robust targeted minimum loss-based plug-in estimator of natural direct and indirect effects that is appropriate for case-cohort sampling. The estimator requires estimation of several regressions, which are used in an augmented inverse probability of treatment weighted estimator. The propensity score will be estimated by a main terms logistic regression model to account for chance imbalances across randomization arms. The sequential outcome regressions used by the approach will be based on a super learner with the 14 algorithms listed in Table 9.

Table 9: Learning Algorithms in the super learner Library for mediation methods[1].

Algorithms	Screens[2]/ Tuning Parameters
SL.mean	All
SL.glm	Low-collinearity and (All, Lasso, LR)
SL.glm.interaction	(All, Lasso, LR)
SL.gam	Low-collinearity and (Lasso, LR)
SL.glmnet	All
SL.xgboost	All
SL.ranger	All

[1] some nuisance parameters have binary outcomes, others quantitative. For the former, we used `family = binomial()` input to the `SuperLearner` function; for the latter, we used `family = gaussian()`.

[2] **All** = include all variables; **Lasso** = include variables with non-zero coefficients in the standard implementation of SL.glmnet that optimizes the lasso tuning parameter via 10-fold cross-validation; **Low-collinearity** = do not allow any pairs of quantitative variables with Spearman rank correlation > 0.90; **LR** = Univariate logistic regression Wald test 2-sided p-value < 0.10.

The estimator is implemented in the `natmed2` package available on GitHub (https://github.com/benkeser/natmed2). The baseline covariates X adjusted for are the same as for the other analyses (e.g. of CoR and of Controlled vaccine efficacy).

16 Summary of the Set of CoR and CoP Analyses and Their Requirements and Contingencies, and Synthesis of the Results, Including Reconciling Any Possible Contradictions in Results

Table 10 summarizes all of the Stage 1 / Day 57 marker correlates analyses that are done, including contingencies for whether and when each analysis is done. The quantitative version of each marker S, and the tertiles version of each marker S, is common across all of the analyses. All of the Day 57 markers are the versions that are not baseline subtracted, given that the cohort for analysis is baseline negative. Most of the analyses focus on univariate Day 57 markers. The primary reason to do this is the goal to identify a parsimonious correlate based on a single marker without needing to run the

set of assays, and secondary reasons are: (1) the assay readouts are expected to be highly correlated, especially for the ID50 and cID80 readouts from the same pseudovirus neutralization assay, and (2) there is ample precedent for univariate markers being accepted as immunological surrogate endpoints for approved vaccines (Plotkin, 2010).

Table 10: Summary of Stage 1 Day 57 Marker CoR and CoP Analyses with Requirements/Contingencies for Conduct of the Analysis (Same Considerations Apply for Day 29 Markers)

	Structure of	**Requirements/Contingencies**	
Analysis	**Day 57 Marker(s)**	**Min No. Vaccine Endpoints**	**Other**
CoR Cox Model	Tertiles of S[1]	25	None
	Quant. $S = s$[2]	25	None
	Quant. $S \geq s$[1]	25	None
CoR Nonpar. threshold	Quant. $S \geq s$[1]	35	None
CoR GAM	Quant. $S = s$[2]	35	None
CoR threshold log. regr.	Quant. $S = s$[2]	25	None
CoR Superlearner[3]	Quant. $S = s$, 2FR, 4FR	35	None
CoP: Correlates of VE	Binary S	50	None
	Quant. $S = s$	50	BIP with $R^2 \geq 0.25$
CoP: Controlled VE	Quant. $S = s$	50	Feasibility of positivity[4]
	Tertiles of $S = s$	50	Feasibility of positivity[4]
CoP: Stoch. Interv. VE	Quant. $S = s$	50	Feasibility of positivity[4]
CoP: Mediators of VE	Quant. $S = s$	50	Feasibility of positivity[4]
	Tertiles of S	50	Feasibility of positivity[4]

[1]These analyses are harmonized in addressing the same scientific question of how does endpoint risk vary over vaccinated subgroups defined by S above a threshold.

[2]These exploratory supportive analyses are harmonized in addressing the same scientific question of how does endpoint risk vary over vaccinated subgroups defined by S equal to a given marker value.

[3]Only this Superlearner analysis uses data from multiple assays and multiple readouts as input features; the other analyses consider one Day 57 biomarker at a time. [4]The positivity assumptions are as follows. Controlled VE: $P(S = s \,|\, A = 1, X) > 0$ almost surely. Stochastic Interventional VE: $s_i \in \mathcal{S} \implies s_i + \delta \in \mathcal{S} \mid A = 1, X = x$ for all $x \in \mathcal{X}$ and $i = 1, \ldots n$. Mediators of VE: $P(S = s \,|\, A = 1, X) > 0$ almost surely and
$P(S = s|A = 0, X = x) >$ implies $P(S = s|A = 1, X = x) > 0$. The quantitative analysis will require that the largest value S observed in the placebo is larger than the smallest value of S observed in the vaccine recipients. This assumption would naturally be satisified for the tertiles analysis. For quantitative S, the assumption is weaker for the Stochastic Interventional VE analysis, such that it is possible that only this analysis of the three will be done.

Some of the analyses include parametric assumptions for characterizing associations (Cox model and threshold analyses, Cox model versions of Controlled VE analyses) and others are nonparametric or approximately so (all other analyses). If parametric and nonparametric analyses of the same type (e.g., Cox model vs. nonparametric CoR analysis of the same association parameter; Controlled VE Cox model vs. nonparametric monotone dose-response) suggest contradictory results, then the interpretation from the nonparametric analysis will be prioritized, given it is more robust and less likely to be an incorrect result. The diagnostic testing of the parametric assumptions will aid this interpretation. As noted above, if the nonparametric analysis suggesting a contradictory result requires a positivity assumption, then its results will only be prioritized if diagnostics support feasibility of the positivity assumption.

16.1 Synthesis Interpretation of Results

To structure the interpretation of the whole set of CoR and CoP results, we consider the Bradford-Hill criteria for supporting causality assessments:

1. Temporal sequence of association (vaccination causes generation of antibodies, which precede occurrence of the clinical disease outcome)
2. Strength of association (CoR magnitude)
3. Consistency of association (across studies and methods)
4. Biological gradient (may be interpreted as dose-response with greater Day 57 antibody corresponding to lower risk and greater VE)
5. Specificity (that the antibody marker is induced by vaccination not natural infection, and the antibody impacts the particular clinical endpoint being analyzed)
6. Plausibility [(supported by other COVID vaccines through study in efficacy trials and challenge (animal or human) trials, and by other potential studies such as natural history re-infection studies and monoclonal antibody prevention efficacy studies that could be challenge (animal or

human) or field trials])

7. Coherence (the causality assumption does not appear to conflict with current knowledge)
8. Experimental reversibility (if VE wanes to a low level then the antibody marker also wanes coincidently; if the Day 57 marker is a strong correlate for outcome during the period of high VE, then it becomes a weaker correlate against endpoints occurring during the later period of low VE; also could be supported if vaccine breakthrough cases tend to occur very early or late in follow-up when antibody levels are known to be relatively low)
9. Analogy (supported by other respiratory virus vaccines, and natural history studies or challenge studies of other respiratory virus vaccines)

We discuss evaluation of these criteria for Day 57 markers, where the same evaluations accounting for Day 29 markers are similarly relevant.

On temporal sequence, because the analyses are done in baseline negative individuals, generally the Day 57 antibody responses must be generated by the vaccine, and if the outcome occurs well after Day 57, then there is clear temporal ordering of vaccination causing antibodies followed by outcome. The nuance is outcome cases with event times near 7 days post Day 57, some of which could have been infected with SARS-CoV-2 prior to Day 57 and have relatively long incubation periods, possibly perturbing temporal ordering by creating naturally-induced rather than vaccine-induced antibody. However, the knowledge about the distribution of the time period between SARS-CoV-2 acquisition and symptomatic COVID, and the time needed for an infection to create an adaptive immune response, suggests that this issue could only haves a minor impact, and overall the temporal sequence criterion readily holds. Yet, the correlates analysis that stringently only includes cases with documented antigen negativity at both Day 29 and Day 57 may be helpful for evaluating the temporal sequence criterion.

On strength of association, this is directly quantified in all of the analyses as a core output of each method, quantified by point estimates and confidence in-

terval estimates of covariate-adjusted association parameters or causal effect parameters.

On consistency of association, checking for similar estimates and inferences across the multiple vaccine efficacy trials will be relevant. The fact that all of the tested vaccines are designed to protect through induction of antibody to Spike protein suggest that consistency is plausible. The vaccine platform needs to be accounted for in this evaluation, where consistency may be expected for vaccines of a given type (e.g., mRNA vaccines, Spike protein vaccines, viral vector vaccines with a similar vector), whereas across types a consistent body of evidence would be very helpful, but not a requirement. FDA guidance has stipulated that a surrogate endpoint for one vaccine platform is not necessarily expected to hold for another, and that evidence for one platform would not be seen on its own as support for a surrogate endpoint for another.

In addition, we will plan to study predictiveness of the estimated optimal surrogate built on each single trial data set applied to the other trial data sets, quantified by AUC on new data sets. Moreover, consistency of association may be assessed in another sense - by studying whether the different CoR methods tend to reveal a consistent directionality and pattern of an antibody marker correlated with risk, and whether the different CoP methods tend to reveal a consistent directionality and pattern of an antibody marker connected to vaccine efficacy (as measured by the various causal effect parameters) and with different versions of vaccine efficacy. A common core element of all of the CoR and CoP methods is covariate-adjusted estimation of marker-conditional risk in vaccine recipients, e.g. of marginal conditional risk $E_X[P(T \leq t_F | S = s, A = 1, X)]$ or $E_X[P(T \leq t_F | S \geq s, A = 1, X)]$. Generally, if an estimate of this function shows strongly decreasing risk with s, then likely all of the CoR analyses will detect such a decrease, and the CoP analyses will detect a version of vaccine efficacy increasing in s. A nuance in looking for consistency of results across methods stems from the fact that different methods have different power to detect the same effect; because of this fact, consistency in magnitude (point estimate) and directionality are more important than consistency in inference/statistical significance.

The fact that all of the methods adjust for the same set of baseline covariates X will aid the ability to compare the results across methods in an interpretable manner. This discussion highlights the relevance of adjusting for the same set of baseline covariates across the different efficacy trials, although our choice to do covariate-adjustment through marginalization (rather than through conditional association parameters) lends some resilience to this issue.

Our comments on consistency of association have supposed a given study endpoint, such as COVID. Another dimension of consistency evaluation could include comparing results across endpoints. On the one hand, consistency in evidence across endpoints could strengthen the case for a CoP, especially for endpoints in the same 'class' such as moderate disease and severe disease. On the other hand, the greater the difference between endpoints, the less relevant consistency may be, because the vaccine may protect through different mechanisms against each endpoint (one potential example is prevention of asymptomatic infection vs. prevention of severe disease). Thus evidence for a CoP for a given endpoint should not necessarily be down-graded based on evidence that the same marker does not appear to be a CoP for another endpoint.

On biological gradient, many of the methods are flexible and designed to detect a dose-response pattern of antibody with risk or antibody with vaccine efficacy, with tabular and graphical output of point and confidence interval estimates designed to reveal dose-response.

On specificity, as noted above antibodies generally are almost surely vaccine-induced given the analysis is done in baseline negative individuals, although with nuance that care is needed to evaluate whether some vaccine breakthrough cases may have had SARS-CoV-2 acquisition unusually early in follow-up (e.g., prior to second vaccination). In addition, the assays are validated for measuring specific anti-SARS-CoV-2 antigen response. Moreover, the Day 57 antibody markers can be verified to be negative in all or almost all baseline negative placebo recipients. Therefore, the specificity criterion should readily hold, with the proviso of the complication of the possible in-

clusion of unusually early infections as vaccine breakthrough cases in some analyses.

On coherence, the results will be interpreted in the light of knowledge of immune correlates of protection for the same vaccine in animal challenge studies (and human challenge studies as available), where multiple studies have demonstrated that both binding and neutralizing antibodies are a correlate of protection.

The results will also be interpreted in light of any knowledge available on passively administered SARS-CoV-2 monoclonal antibodies for prevention of SARS-CoV-2 infection or COVID disease, either in challenge studies (animals or humans) or efficacy trials. In addition, the results will be interpreted in light of results on the antibody markers as correlates of re-infection in natural history studies. Note we are cautious to not use correlates studies in already-infected individuals, because the fact of infection may readily change the nature of a correlate of protection.

On experimental reversibility, we will evaluate whether the strength of association of the Day 57 CoRs and CoPs weakens when restricting to outcomes occurring more distal to vaccination. If the vaccine efficacy is found to wane over time, and the antibody marker wanes over time, then this decrease in the strength of association would be consistent with antibody as a correlate of protection. In contrast, if vaccine efficacy and antibody waned over time, but the strength of a Day 57 CoR and CoP was the same regardless of the timing of outcomes, it might call into question the role of the antibody marker as a CoP. The Stage 2 correlates analyses will also be helpful, where experimental reversibility could be supported simply by coincident waning of VE and waning antibody.

Experimental reversibility may also be supported by "population-level" correlates analyses, a term sometimes used in reference to meta-analysis that associates the level of VE with the population-level of a Day 57 marker across subgroups or trials; e.g. the population-level Day 57 marker response may be summarized by the geometric mean titer or geometric mean concentration.

On analogy, perhaps the most relevant vaccines to consider are vaccines

against other respiratory viruses, including influenza vaccine and RSV vaccines. The fact that neutralizing antibodies are a CoR and CoP for both inactivated and live virus vaccines supports that neutralizing antibodies can be a CoP for SARS-CoV-2. In addition, there is ongoing correlates of protection analysis of Novavax's Phase 3 RSV vaccine efficacy trial, that is evaluating binding antibody and neutralizing antibody CoRs and CoP correlates for severe respiratory disease in infants of vaccinated pregnant mothers (submitted). Once those results are available, they will aid in checking the analogy (and coherence) criterion.

The univariate CoR analyses essentially assess five Day 57 antibody biomarkers. The questions arise as to how do we select which biomarker seems to be the best-supported CoP, and do we need to be concerned about multiplicity adjustment issues? Given the multifactorial nature of the assessment involving biology and statistics, we for the most part avoid an approach that tries to pre-specify a quantitative ranking system; rather our approach presents the results of each marker side by side and allows human synthesis and interpretation. To guard against errors in this subjective process, we suggest that consistent results across analyses of a given trial, and consistent results (and predictive validation) across multiple trials, will provide particularly strong guidance for interpreting results. For example, if a particular Day 57 antibody marker shows remarkably consistent results in being a strong CoR and supported CoP but the other readouts do not, it may emerge as the best-supported CoP. In addition, the superlearning CoR estimated optimal surrogate objective has a special place of importance, because it includes variable importance quantification, providing some quantitative guidance on ranking the predictivneness of markers. This variable importance will be defined both internal to a given trial and based on external validation on the other efficacy trials. The metrics of CV-AUC and AUC on new trials quantifies evidence for signal in the data in a way that is protected from risk of false positive results, by virtue of having two layers of cross-validation used to estimate CV-AUC and hence avoid over-fitting. In addition, the CoR analyses use multiple hypothesis testing adjustment to help ensure clear signals and not false positive results (see Section 12.4.2). We also need a plan for

minimizing the risk of false positive results for CoP analyses, which we now address.

16.2 Multiple Hypothesis Testing Adjustment for CoP Analysis

For the univariable CoP analyses of the prioritized set of Day 57 antibody markers among the five specified marker variables, the analysis plan seeks evidence of a CoP through four different causal effect approaches. Because of this looking for evidence through different lenses, for CoP analysis we do not focus on family-wise error rate adjustment, because FWER-adjustment aims to control the risk of making even a single false rejection. Rather, in an effort to build a body of consistent evidence and to ensure that a large fraction of that evidence is reliable, for CoP analysis we focus on false discovery rate correction. To do this, we use the same permutation-based method (Westfall et al., 1993) that is used for CoR analysis. The multiplicity adjustment is performed across the Day 57 markers and across the set of CoP methods that are applied, in a single suite of hypothesis tests with calculation of q-values. As a guideline for interpreting CoP findings (but not meant to be a rigid gateway), markers with unadjusted p-value ≤ 0.05 and q-value ≤ 0.10 are flagged as having statistical evidence for being a CoP.

17 CoP: Meta-Analysis Analysis Plan

We provide a brief summary of the overall plan, where the details will be developed closer to the time that data are available for meta-analysis of multiple phase 3 vaccine efficacy trials.

Once data sets are available from the set of USG COVID-19 Response Team phase 3 trials, the data sets will be combined for additional analyses to support development of immune CoPs. Data analysis of the combined data sets provides interpretable results based on the standardization of the USG COVID-19 Response Team phase 3 trial protocols – including harmonized study endpoints, follow-up, and blood storage time points – and on the common statistical analysis plan and laboratories/immunoassays (where the use of the Duke pseudovirus assay in some USG COVID-19 Response Team tri-

als and the Monogram pseudovirus assay in other USG COVID-19 Response Team trials implies that the statistical analysis will make use of concordance testing data for making results interpretabile referenced to one of the assays.) Meta-analysis surrogate endpoint evaluation methods will be applied to the combined data sets, both for assessing Day 29 and Day 57 antibody markers (Stage 1) as surrogate endpoints for COVID and for secondary outcomes, and for assessing the antibody markers over time (Stage 2) as surrogate endpoints for COVID and for secondary outcomes.

Both individual-level and trial-level meta-analysis will be applied, where the latter studies the association of vaccine effects on an antibody marker with vaccine effects on a study outcome, for example assessing how GMT nAb cID80 titer associates with the level of vaccine efficacy against COVID. Meta-analysis has a special role in being the only correlates approach that can potentially assess immunologic markers as CoPs that are measured using sampling types that were not stored from most trial participants (e.g., PBMC for measuring T cell responses). While the current statistical analysis plan focuses on assessing antibody markers as correlates, in the future plans may be devised to incorporate T cell response data (and potentially other data types) from phase 1-2 studies into meta-analysis evaluation.

In addition to applying formal meta-analysis surrogate endpoint evaluation methods, some of the CoR and CoP statistical methods applied to the individual phase 3 trial data sets will be adapted for application to the combined data sets. This will allow addressing the following objectives: (1) to assess consistency of CoRs and CoPs across trials, subpopulations, and vaccine platforms; (2) to evaluate how well an antibody marker CoR for an outcome developed in one phase 3 trial predicts the same outcome in the other phase 3 trials (cross-validation prediction accuracy); and (3) to provide data for prediction modeling of what would be the efficacy of a new vaccine based on its distribution of antibody markers. Objective (2) provides some empirical data for considering appropriateness of use of a CoP across vaccine platforms.

We will consider two meta-analytic statistical frameworks for evaluating candidate surrogate endpoints: (1) Gabriel et al. (2016, 2019) "generalized

surrogate" approach and (2) Molenberghs et al. (2000, 2007) meta-analytic approach.

17.1 Method of Gabriel et al. (2016, 2019)

Gabriel et al. (2016, 2019) describe a non-parametric Bayesian hierarchical framework for the modeling of vaccine efficacy and one or more potential surrogate endpoint markers. The framework enables evaluation of a trial-level general surrogate (TLGS): the ability to predict the efficacy of a vaccine based on trial-level observations of a surrogate marker distribution. The model does not require individual-level data as input; instead, the effects of the vaccine on the true endpoint (i.e. vaccine efficacy) and on the surrogate endpoint (i.e. a vaccine-induced immune response) are modeled from each observed randomized trial using a bivariate normal distribution. From the Bayesian posterior it is then possible to predict the vaccine efficacy (with 95% credible interval) in a new setting based on the observed distribution of the surrogate marker. In concept the model is similar to that of locally weighted linear regression (LOESS). With this framework it is possible to evaluate the strength of surrogacy using absolute prediction error, compare multiple candidate surrogates based on relative prediction error and predict vaccine efficacy in a new setting.

The Gabriel et al. framework is implemented in R using code adapted from their publication (https://github.com/sachsmc/DPpackagemod; 2019). We will use the estimates of vaccine efficacy and candidate surrogate marker distributions from all available Phase 3 randomized, placebo controlled trials; inclusion will be contingent on normalization of the surrogate marker to the WHO or some other human convalescent serum (HCS) standard. Performance will be measured as the absolute difference between the predicted and observed vaccine efficacy (on the log relative-risk scale) with the mean taken across trials in a leave-one-out cross-validation framework (i.e. the efficacy for trial A is predicted based on the surrogate marker in trial A and the efficacy and marker distributions in the other trials). An example analysis is provided using the ratio of the virus neutralization titer in vaccine recipients to human convalescent sera (HCS) as the surrogate marker (Figure 5);

a similar analysis was also performed using the ratio of the binding IgG antibody (ELISA) to the same measure in HCS (Figure 6). In the example, vaccine efficacy was estimated from Phase 3 trials while the central value of the marker distribution were estimated (mean and 95% confidence interval) from Phase 1/2 trials in comparable populations. Details of the data sources are available in Earle at al. (2021).

17.2 Method of Molenberghs et al.

Buyse, Molenberghs et al. originally proposed a meta-analytic framework for evaluating a surrogate endpoint by building on Prentice's (1989) surrogacy criteria and generalizing to a multi-trial setting. Similar to the Prentice criteria, the approach is based on a system of three linear regressions of: (i) the treatment effect on the true endpoint, (ii) the treatment effect on the surrogate endpoint and (iii) the association of the surrogate endpoint with the true endpoint. This model is expanded to a multi-trial setting, each potentially with a distinct treatment (i.e. vaccine). Unlike Gabriel et al. (2016, 2019), the model is fit using individual-level data; the fitted parameters of the model enable assessment of the strength of a surrogate endpoint for each trial as well as overall. Originally they proposed a metric of surrogacy called the relative effect (RE), which was the ratio of the regression coefficient indicating the treatment effect on the true endpoint divided by that of the treatment effect on the surrogate. The relative effect indicates the extent to which the treatment effect on the true endpoint can be predicted by measuring the treatment effect on the surrogate.

In more recent work, Alonso and Molenberghs (2007) used information theory and developed a similar metric, R_h^2, which generalizes to settings with non-normal true and surrogate endpoints. The information theoretical framework is helpful because it provides an intuitive framing of the issue of surrogate marker validation: we want to gain information about the unobserved treatment effect on the true endpoint using the known treatment effect on the surrogate. The R_h^2 metric is similar to RE in that when $R_h^2 \approx 1$ the potential surrogate is promising and the interpretation is that once the surrogate is known, almost all of our uncertainty about the true endpoint is gone.

The information-theoretic Molenberghs et al. approach is implemented in their R package *surrogate*. The analysis takes as input the individual-level data for multiple trials, including the true and surrogate endpoints. Like implementation of the Gabriel at al. framework, inclusion of trials will be contingent on normalization of the surrogate markers to the WHO International Units. The output of the analysis will be estimates of the parameter R_h^2 for each trial and overall including a point-estimate and 95% confidence interval.

18 Estimating a Threshold of Protection Based on an Established or Putative CoP (Population-Based CoP)

For each antibody marker studied as a CoP, we will apply the Chang-Kohberger (2003) / Siber (2007) method to estimate a threshold of the antibody marker associated with the estimate of overall vaccine efficacy observed in the trial.

This method makes two simplifying assumptions: (1) that a high enough antibody marker value s^* implies that individuals with $S > s^*$ have essentially zero disease risk (perfect protection) regardless of whether they were vaccinated; and (2) $P(Y = 1|S \leq s^*, A = 1)/P(Y = 1|S \leq s^*, A = 0) = 1$ (zero vaccine efficacy if $S \leq s^*$). Based on these assumptions, s^* is calculated as the value equating $1 - \widehat{P}(S \leq s^*|A = 1)/\widehat{P}(S \leq s^*|A = 0)$ to the estimate of overall vaccine efficacy. This estimate is supplemented by estimating the reverse cumulative distribution function (RCDF) of S in baseline negative vaccine recipients and calculating a 95% confidence interval for the threshold value s^* as the points of intersection of the estimated RCDF curve with the 95% confidence interval for overall vaccine efficacy (as in the figure in Andrews and Goldblatt, 2014).

This method essentially assumes that S has already been established as a CoP, and under that assumption estimates a threshold that may be considered as a benchmark / study endpoint for future immunogenicity vaccine trial applications.

It is acknowledged that this approach makes highly simplified assumptions; nonetheless it may yield a useful benchmark and complementary information

on a threshold correlate of protection.

19 Considerations for Baseline SARS-CoV-2 Positive Study Participants

As stated above, if enough COVID cases in baseline positive vaccine and/or placebo recipients occur, then additional correlates analyses may be planned in baseline positive individuals. For example, the same or similar correlates of risk analysis plan that is used to analyze Day 57 marker correlates of risk in baseline negative vaccine recipients could be applied to assess Day 1 marker correlates of risk in baseline positive placebo recipients. In addition, analyses could be done to assess how vaccine efficacy in baseline positive participants varies with Day 1 markers. It is straightforward to make this analysis rigorous because Day 1 markers are a baseline covariate, such that regression analyses are valid based on the randomization.

20 Avoiding Bias with Pseudovirus Neutralization Analysis due to Use of Anti-HIV Antiretroviral Drugs

Because the lentivirus-based pseudovirus neutralization assay uses an HIV backbone, the presence of anti-retroviral drugs in serum will give a false positive neutralization signal. This can be easily screened for using an MuLV pseudotype control. Therefore, Day 1 and Day 57 samples of all study participants with data included in correlates analyses will be tested for presence of anti-retroviral drugs. Participants with any of the samples at Day 1 and Day 57 positive for antiretroviral use are excluded from analyses, for all analyses that include pseudovirus neutralization. Analyses that do not consider pseudovirus neutralization are unaffected by this issue.

If Day 29 markers are included, then the antiretroviral testing is applied to Day 29 samples as well as to Day 1 and Day 57 samples. And, participants with any of the samples at Day 1, 29, 57 positive for antiretroviral use are excluded from analyses, for all analyses that include pseudovirus neutralization.

21 Accommodating Crossover of Placebo Recipients to the Vaccine Arm

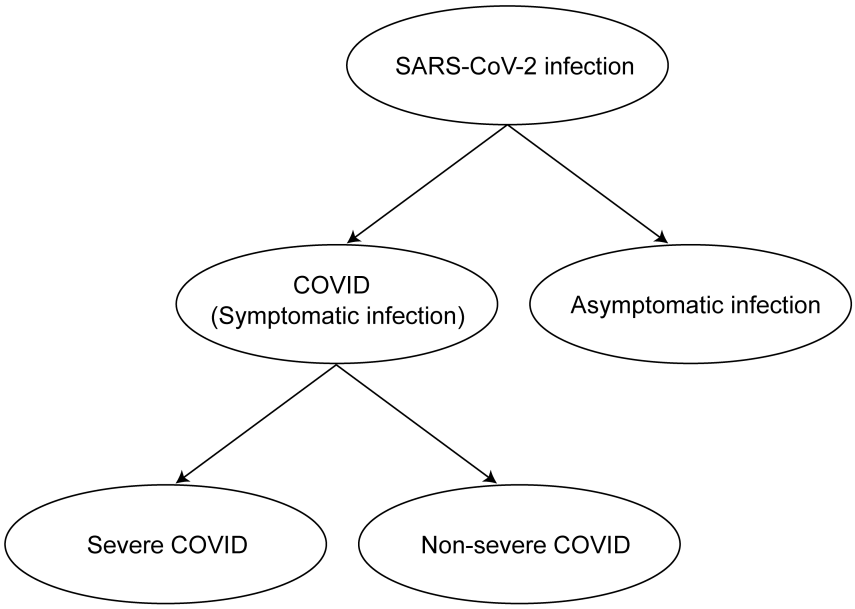
We consider how the SAP would be impacted by a scenario where at some point most placebo recipients receive the study vaccine, which has been occurring in general for the USG public-private partnership trials. The plan for assessing correlates of risk in vaccine recipients would be minimally affected, because the analysis is based on vaccine recipients alone. If crossed over placebo recipients have study visits and blood sample storage on the same schedule as if they had originally been assigned to the vaccine arm, then the new follow-up data from the crossed over placebo recipients will be included in correlates of risk analyses, which is expected to yield improved power and precision given the expanded sample size of vaccine recipients. Yet, the first CoR analyses will restrict to the blinded period of follow-up, for purity of interpretation of the results.

The plan for assessing correlates of protection, on the other hand, would be more altered based on crossover. The plan would be revised to only assess correlates of protection over follow-up through to the point that there is no longer a placebo cohort under blinded follow-up. Moreover, if immune marker data from crossed-over placebo recipients are available, then correlate of VE CoP analyses will be conducted that leverage the additional closeout placebo vaccination data.

22 COVID Correlates Analysis Report

This SAP is being implemented over time on a mock/practice COVID-19 vaccine efficacy trial data set, as discussed in the Prelude. The report is provided at the public GitHub repository CoVPN/correlates_reporting.

A



B

Clinical Endpoint	Definition
SARS-CoV-2 infection	Positive RNA PCR test or SARS-CoV-2 seroconversion*, whichever occurs first
COVID (Symptomatic infection)	Meeting a protocol-specified list of COVID-19 symptoms with virological confirmation of SARS-CoV-2 infection (symptom triggered)
Asymptomatic infection	SARS-CoV-2 seroconversion* without prior diagnosis of the COVID endpoint¶
Severe COVID	COVID endpoint with at least one protocol-specified severe disease event
Non-severe COVID	COVID endpoint with zero protocol-specified severe disease events

*Seroconversion is assessed via a validated assay that distinguishes natural vs vaccine-induced SARS-CoV-2 antibodies

¶Alternatively, the asymptomatic infection endpoint can also include an RNA PCR+ test result obtained through testing regardless of symptoms (e.g., as a requirement for travel, return to school or work, or elective medical procedures) and follow-up to confirm the participant remains asymptomatic

Figure 1: A) Structural relationships among study endpoints in a COVID-19 vaccine efficacy trial (Mehrotra et al., 2020).. B) Study endpoint definitions.

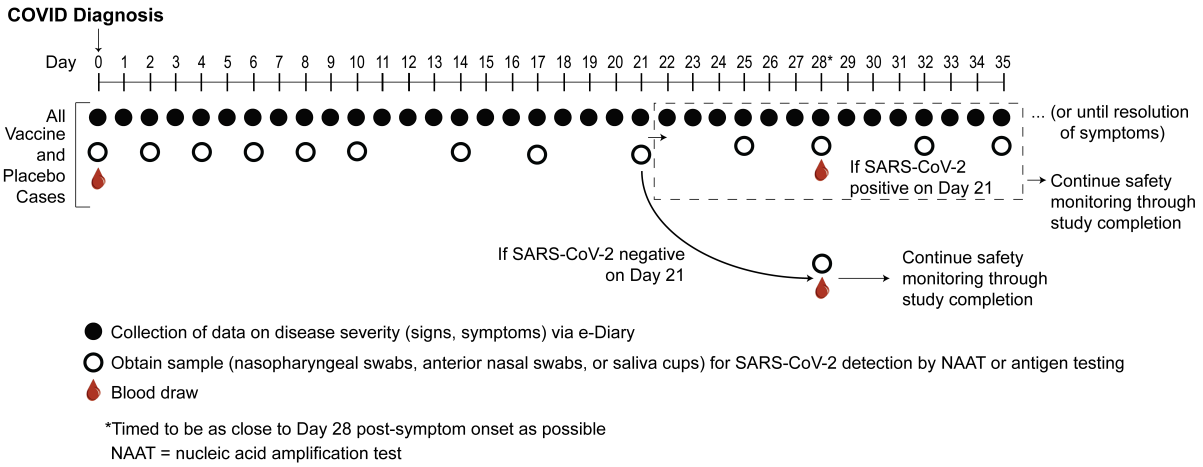


Figure 2: Example at-COVID diagnosis and post-COVID diagnosis disease severity and virologic sampling schedule, in a setting where frequent follow-up of confirmed cases can be assured. Participants diagnosed with virologically-confirmed symptomatic SARS-CoV-2 infection (COVID) enter a post-diagnosis sampling schedule to monitor viral load and COVID-related symptoms (types, severity levels, and durations).

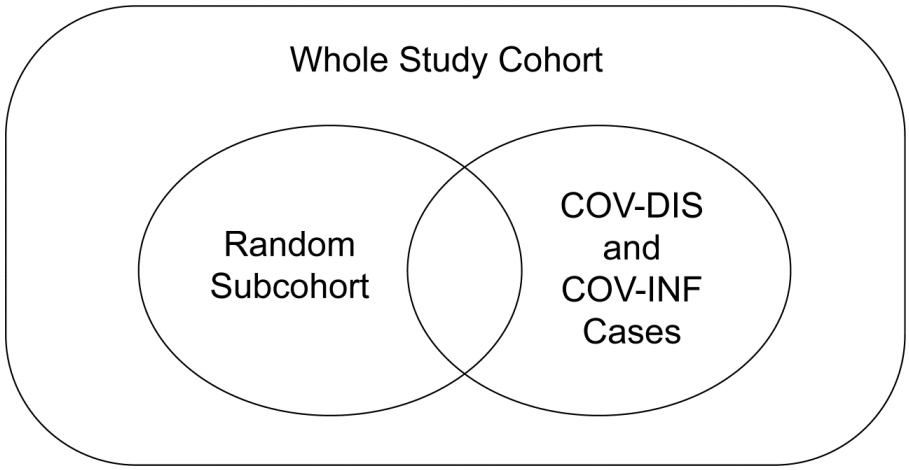


Figure 3: Case-cohort sampling design (Prentice, 1986) that measures Day 1, 57 antibody markers in all participants selected into the subcohort and in all COVID and COV-INF cases occurring outside of the subcohort.

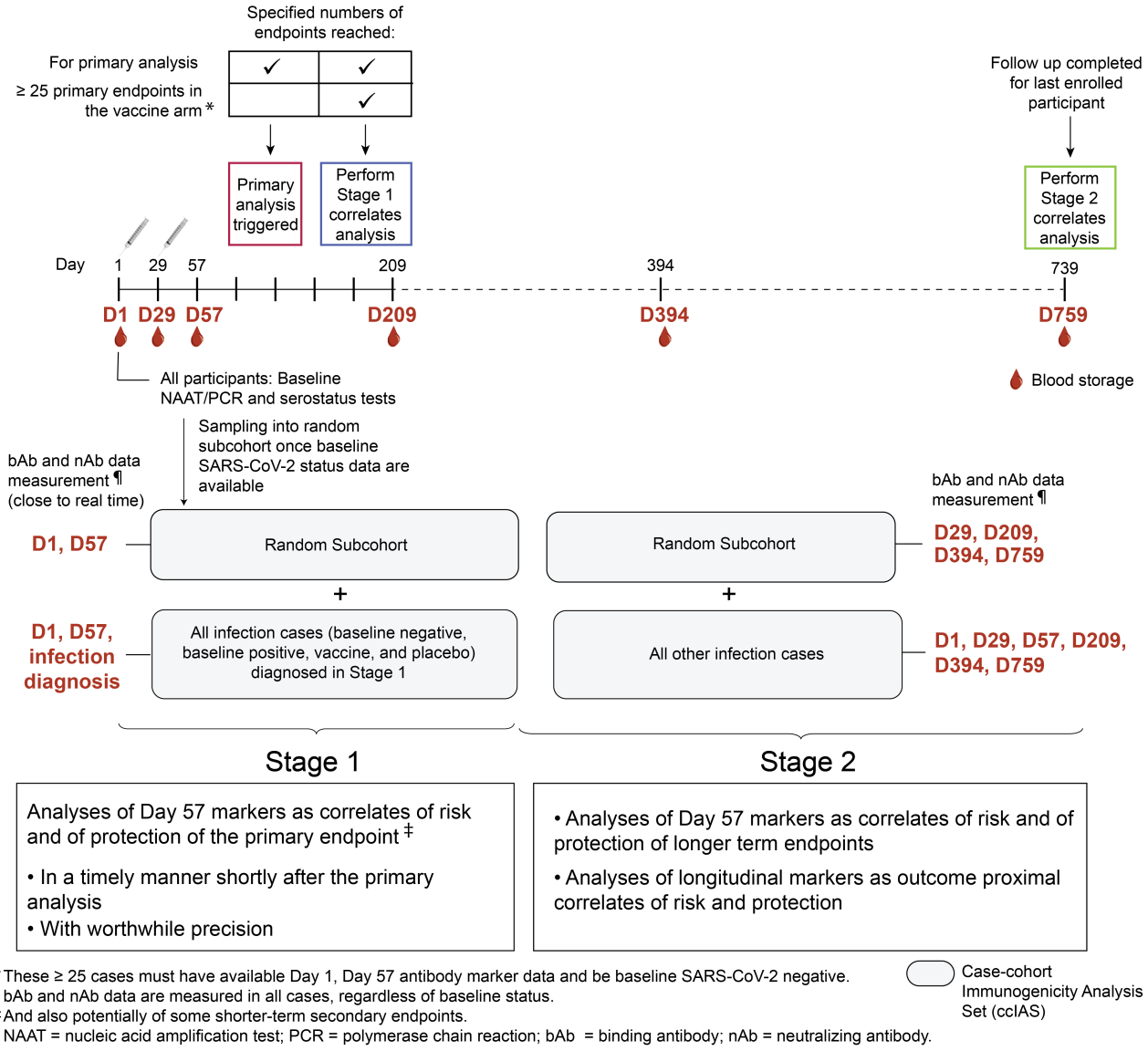


Figure 4: Two-stage correlates analysis. Stage 1 consists of analyses of Day 57 markers as correlates of risk and of protection of the primary endpoint and potentially also of some secondary endpoints, and includes antibody marker data from all COVID and SARS-CoV-2 infection cases (COV-INF) through to the time of the data lock for the first correlates analyses. Stage 2 consists of analyses of Day 57 markers as correlates of risk and of protection of longer term endpoints and analyses of longitudinal markers as outcome-proximal correlates of risk and of protection, and includes antibody marker data from all subsequent COVID and COV-INF cases. Stage 1 measures Day 1, 57 antibody markers and COV-INF and COVID diagnosis time point markers; Stage 2 measures antibody markers from all sampling time points and COV-INF plus COVID diagnosis sampling time points not yet assayed. The same immunogenicity subcohort is used for both stages. If Day 29 markers are included, then Day 29 markers are included for the same participants with Day 57 markers measured.

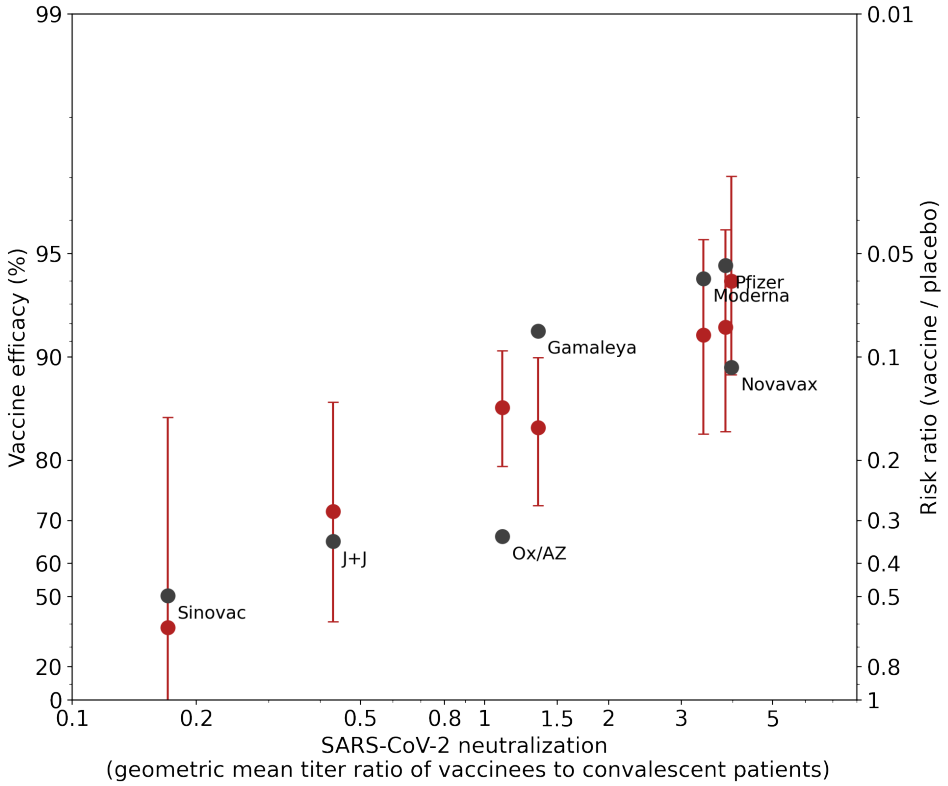


Figure 5: Randomized vaccine effect on the true endpoint (y-axis, i.e. vaccine efficacy) versus vaccine effect on a candidate surrogate endpoint (x-axis) from 7 COVID-19 vaccines (black data points). Candidate surrogate endpoint is the ratio of the geometric mean virus neutralization titer (GMT) across vaccine recipients to the GMT for human convalescent serum (HCS). Estimates of vaccine efficacy are based on Phase 3 clinical trials, while estimates of the surrogate endpoint are based on Phase 1 or 2 data in a comparable population (see Earle et al., 2021 for details). For each trial the vaccine efficacy is also predicted (red, Bayesian posterior estimate and 95% credible interval) from the observed surrogate endpoint as well as efficacy and surrogate endpoint data from each of the other six trials (a "leave-one-out" cross-validation framework)

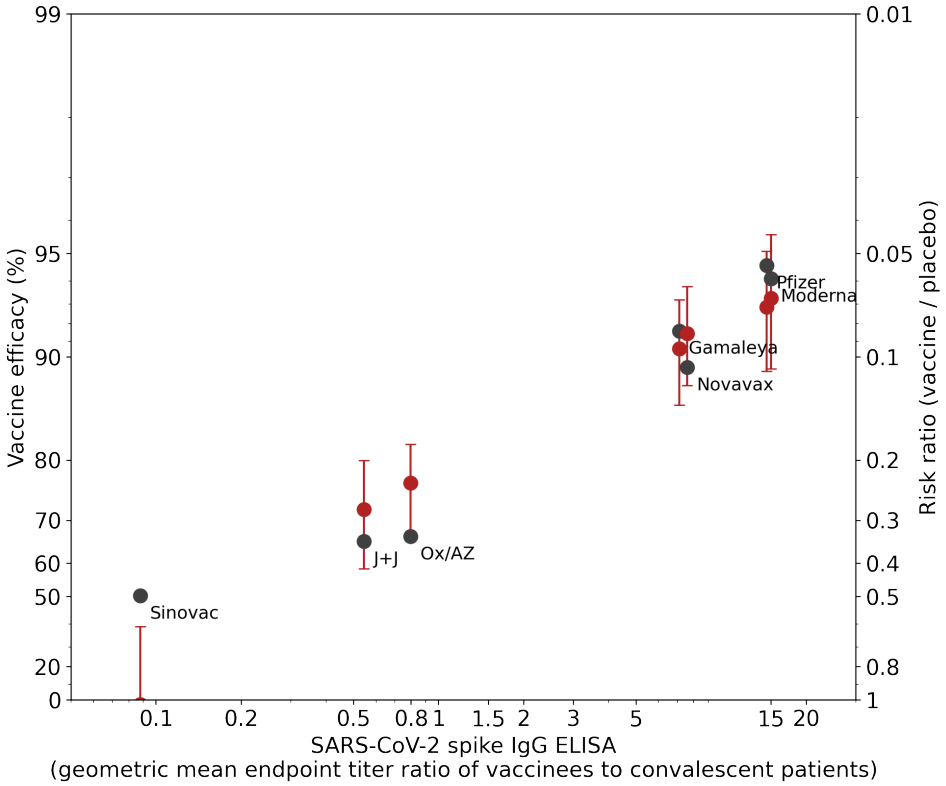


Figure 6: Randomized vaccine effect on the true endpoint (y-axis, i.e. vaccine efficacy) versus vaccine effect on a candidate surrogate endpoint (x-axis) from 7 COVID-19 vaccines (black data points). Candidate surrogate endpoint is the ratio of the binding IgG antibody (ELISA) among vaccine recipients to the same measure in human convalescent serum). Estimates of vaccine efficacy are based on Phase 3 clinical trials, while estimates of the surrogate endpoint are based on Phase 1 or 2 data in a comparable population (see Earle et al., 2021 for details). For each trial the vaccine efficacy is also predicted (red, Bayesian posterior estimate and 95% credible interval) from the observed surrogate endpoint as well as efficacy and surrogate endpoint data from each of the other six trials (a "leave-one-out" cross-validation framework)

References

Alonso, A. and Molenberghs, G. (2007), "Surrogate marker evaluation from an information theory perspective," *Biometrics*, 63, 180–186.

Andrews, N.J., Waight, P.A., Burbidge, P., Pearce, E., Roalfe, L., Zancolli, M. et al (2014), "Serotype-specific effectiveness and correlates of protection for the 13-valent pneumococcal conjugate vaccine: a postlicensure indirect cohort study," *The Lancet infectious diseases*, 14, 839–846.

Benkeser, D. and van der Laan, M.J. (2016), "The highly adaptive lasso estimator," in *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pp. 689–696, IEEE.

Benkeser, D., DĂaz, I. and Ran, J. (2021), "Inference for natural mediation effects under case-cohort sampling with applications in identifying COVID-19 vaccine correlates of protection," *arxiv*.

Breiman, L. (2001), "Random forests," *Machine learning*, 45, 5–32.

Breslow, N. and Wellner, J. (2007), "Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression." *Scandinavian Journal of Statistics*, 34, 86–102, PMCID:.

Breslow, N., Lumley, T., Ballantyne, C., Chambless, L. and Kulich, M. (2009a), "Improved Horvitz-Thompson Estimation of Model Parameters from Two-phase Stratified Samples: Applications in Epidemiology," *Statistical Biosciences*, 1, 32–49.

Breslow, N., Lumley, T., Ballantyne, C., Chambless, L. and Kulich, M. (2009b), "Using the whole cohort in the analysis of case-cohort data." *American Journal of Epidemiology*, 169, 1398–1405.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000), "The validation of surrogate endpoints in meta-analyses of randomized experiments," *Biostatistics*, 1, 49–67.

Chen, Q. (2020), "Fast Grid Search Algorithms for Multi-phase Regression Models," Ph.D. thesis, University OF Washington.

Chen, T. and Guestrin, C. (2016), "xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794, ACM.

Coyle, J.R., Hejazi, N.S., Malenica, I. and Sofrygin, O. (2020), "sl3: Modern Pipelines for Machine Learning and Super Learning," https://github.com/tlverse/sl3, R package version 1.3.7.

Dasgupta, S. and Huang, Y. (2019), "Evaluating the surrogacy of multiple vaccine-induced immune response biomarkers in HIV vaccine trials," *Biostatistics*.

Díaz, I. and van der Laan, M.J. (2011), "Super learner based conditional density estimation with application to marginal structural models," *The International Journal of Biostatistics*, 7.

Díaz, I. and van der Laan, M.J. (2012), "Population intervention causal effects based on stochastic interventions," *Biometrics*, 68, 541–549.

Díaz, I. and van der Laan, M.J. (2018), "Stochastic Treatment Regimes," in *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pp. 167–180, Springer Science & Business Media.

Ding, P. and VanderWeele, T. (2016), "Sensitivity analysis without assumptions," *Epidemiology*, 27(3), 368.

Donovan, K., Hudgens, M. and Gilbert, P.B. (2019), "Nonparametric inference for immune response thresholds of risk in vaccine studies," *Annals of Applied Statistics*, 13, 1147–1165, PMCID: PMC6613658 [Delayed release (embargo): Available on 2020-06-01].

Earle, K., Ambrosino, D., Fiore-Gartland, A., Goldblatt, D., Gilbert, P., Siber, G. et al (2021), "Evidence for antibody as a protective correlate for COVID-19 vaccines," *medRXiv*, [stat.AP] [Preprint] 27 March 2021. Cited 30 March 2021. Available from https://doi.org/10.1101/2021.03.17.20200246.

Fleming, T.R. and Powers, J.H. (2012), "Biomarkers and surrogate endpoints in clinical trials," *Statistics in medicine*, 31, 2973–2984.

Follmann, D. (2006), "Augmented designs to assess immune response in vaccine trials," *Biometrics*, 62, 1161–1169.

Follmann, D. (2018), "Reliably picking the best endpoint," *Statistics in Medicine*, 37, 4374–4385.

Fong, Y. and Xu, J. (2021), "Forward Stepwise Deep Autoencoder-based Monotone Nonlinear Dimensionality Reduction Methods," *Journal of Computational and Graphical Statistics*, pp. 1–10.

Fong, Y., Huang, Y., Gilbert, P.B. and Permar, S.R. (2017), "chngpt: threshold regression model estimation and inference," *BMC Bioinformatics*, 18, 454, PMCID: PMC5644082.

Fong, Y., Shen, X., Ashley, V., Deal, A., Seaton, K., Yu, C. et al (2018), "Modification of the Association Between T-Cell Immune Responses and Human Immunodeficiency Virus Type 1 Infection Risk by Vaccine-Induced Antibody Responses in the HVTN 505 Trial," *Journal of Infectious Diseases*, 217, 1280–1288, PMCID: PMC6018910.

Friedman, J., Hastie, T. and Tibshirani, R. (2009), "glmnet: Lasso and elastic-net regularized generalized linear models," *R package version*, 1.

Friedman, J.H. et al (1991), "Multivariate adaptive regression splines," *The annals of statistics*, 19, 1–67.

Fu, R. and Gilbert, P.B. (2017), "Joint modeling of longitudinal and survival data with the Cox model and two-phase sampling," *Lifetime Data Analysis*, 23, 136–159, PMCID: PMC5035179.

Gabriel, E.E., Sachs, M.C., Daniels, M.J. and Halloran, M.E. (2019), "Optimizing and evaluating biomarker combinations as trial-level general surrogates," *Statistics in medicine*, 38, 1135–1146.

Gabriel EE, Daniels MJ, H.M. (2016), "Comparing biomarkers as trial level general surrogates," *Biometrics*, 72, 1046–1054.

Gilbert, P., Fong, Y. and Carone, M. (2020a), "Assessment of Immune Correlates of Protection Without a Placebo Arm, with Application to COVID-19 Vaccines," *xx*, submitted.

Gilbert, P.B. (2000), "Comparison of competing risks failure time methods and time-independent methods for assessing strain variations in vaccine protection." *Statistics in Medicine*, 19, 3065–3086.

Gilbert, P.B. and Hudgens, M. (2008), "Evaluating candidate principal surrogate endpoints," *Biometrics*, 64, 1146–1154.

Gilbert, P.B., Blette, B.S., Shepherd, B.E. and Hudgens, M.G. (2020b), "Post-randomization Biomarker Effect Modification Analysis in an HIV Vaccine Clinical Trial," *Journal of Causal Inference*, 8, 54–69.

Grambsch, P. and Therneau, T. (1994), "Proportional hazards tests and diagnostics based on weighted residuals." *Biometrika*, 81, 515–526.

Gruber, S. and van der Laan, M.J. (2010), "An application of collaborative targeted maximum likelihood estimation in causal inference and genomics," *The International Journal of Biostatistics*, 6.

Hall, P. and Heckman, N.E. (2000), "Testing for monotonicity of a regression mean by calibrating for linear functions," *Annals of Statistics*, pp. 20–39.

Haneuse, S. and Rotnitzky, A. (2013), "Estimation of the effect of interventions that modify the received treatment," *Statistics in medicine*, 32, 5260–5277.

He, Z. and Fong, Y. (2019), "Maximum Diversity Weighting for Biomarkers with Application in HIV-1 Vaccine Studies," *Statistics in Medicine*, 38, 3936–3946.

Hejazi, N.S. and Benkeser, D.C. (2020), "txshift: Efficient estimation of the causal effects of stochastic interventions in R," *Journal of Open Source Software*.

Hejazi, N.S., van der Laan, M.J., Janes, H.E. and Benkeser, D.C. (2020a), "Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials," *Biometrics*, Ahead of print.

Hejazi, N.S., Coyle, J.R. and van der Laan, M.J. (2020b), "hal9001: Scalable highly adaptive lasso regression in R," *Journal of Open Source Software*.

Hejazi, N.S., Benkeser, D.C. and van der Laan, M.J. (2020c), "haldensify: Highly adaptive lasso conditional density estimation," https://github.com/nhejazi/haldensify, R package version 0.0.5.

Hejazi, N.S., Benkeser, D.C., Díaz, I. and van der Laan, M.J. (2020d), "On efficient estimation of the causal effects of stochastic interventions via the highly adaptive lasso," .

Heng, F., Sun, Y., Hyun, S. and Gilbert, P. (2020), "Analysis of the time-varying Cox model for cause-specific hazard functions with missing causes," *Lifetime Data Analysis*, 9, 1–30.

Hoerl, A.E. and Kennard, R.W. (1970), "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

Huang, Y. (2018), "Evaluating Principal Surrogate Markers in Vaccine Trials in the Presence of Multiphase Sampling," *Biometrics*, **74**, 27–39.

Huang, Y. and Gilbert, P.B. (2011), "Comparing biomarkers as principal surrogate endpoints." *Biometrics*, 67, 1442–1451, PMCID: PMC3163011.

Huang, Y., Gilbert, P.B. and Wolfson, J. (2013), "Design and estimation for evaluating principal surrogate markers in vaccine trials," *Biometrics*, 69, 301–309.

Hubbard, A.E., Khered-Pajouh, S. and van der Laan, M.J. (2016), "Statistical inference for data adaptive target parameters," *The International Journal of Biostatistics*, 12, 3–19.

Janes, H.E., Cohen, K.W., Frahm, N., De Rosa, S.C., Sanchez, B., Hural, J. et al (2017), "Higher T-cell responses induced by DNA/rAd5 HIV-1 preventive vaccine are associated with lower HIV-1 infection risk in an efficacy trial," *The Journal of Infectious Diseases*, 215, 1376–1385, PMCID: PMC5853653.

Jodar, L., Butler, J., Carlone, G., Dagan, R., Goldblatt, D., Kdž″yhty, H. et al (2003), "Serological criteria for evaluation and licensure of new pneumococcal conjugate vaccine formulations for use in infants." *Vaccine*, 21, 3265–3272.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W. et al (2017), "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146–3154.

Kooperberg, C., Bose, S. and Stone, C.J. (1997), "Polychotomous regression," *Journal of the American Statistical Association*, 92, 117–127.

Li, C. and Shepherd, B.E. (2012), "A new residual for ordinal outcomes," *Biometrika*, 99, 473–480.

Li, S. and Luedtke, A. (2020), "Nonparametric assessment of principally stratified effects in vaccine studies," *manuscript*.

Lumley, T. (2010), *Complex surveys: a guide to analysis using R*, vol. 565, John Wiley & Sons.

Magaret, C., Benkeser, D., Williamson, B., Borate, B., Carpp, L., Georgiev, I. et al (2019), "Prediction of VRC01 neutralization sensitivity by HIV-1 gp160 sequence features." *PLoS Computational Biology*, 15, e1006952, PMCID: PMC6459550.

McCallum, M., Walls, A.C., Bowen, J.E., Corti, D. and Veesler, D. (2020), "Structure-guided covalent stabilization of coronavirus spike glycoprotein trimers in the closed conformation," *Nature structural & molecular biology*, 27, 942–949.

Mehrotra, D.V., Janes, H.E., Fleming, T.R., Annunziato, P.W., Neuzil, K.M., Carpp, L.N. et al (2020), "Clinical Endpoints for Evaluating Efficacy in COVID-19 Vaccine Trials," *Annals of Internal Medicine*.

Moodie, Z., Juraska, M., Huang, Y., Zhuang, Y., Fong, Y., Carpp, L. et al (2018), "Neutralizing antibody correlates analysis of tetravalent dengue vaccine efficacy trials in Asia and Latin America." *Journal of Infectious Diseases*, 217(5), 742–753, PMCID: PMC5854020.

Neidich, S.D., Fong, Y., Li, S.S., Geraghty, D.E., Williamson, B.D., Young, W.C. et al (2019), "Antibody Fc effector functions and IgG3 associate with decreased HIV-1 risk," *Journal of Clinical Investigation*, 129, 4838–4849.

Newcombe, R. (1998), "Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods," *Statistics in Medicine*, 17, 873–90.

Plotkin, S. and Gilbert, P.B. (2018), "Correlates of protection," in *Vaccines, Seventh Edition*, eds. S. Plotkin, W. Orenstein, P. Offit, and K. Edwards, pp. 35–40, Elsevier Inc., New York.

Plotkin, S.A. (2010), "Correlates of Protection Induced by Vaccination." *Clinical Vaccine Immunology*, 17, 1055–1065, PMCID: PMC2897268.

Prentice, R. (1986), "A case-cohort design for epidemiologic cohort studies and disease prevention trials." *Biometrika*, 73, 1–11.

Prentice, R. (1989), "Surrogate endpoints in clinical trials: definition and operational criteria," *Statistics in Medicine*, 8, 431–440.

Prentice, R., Kalbfleisch, J., Peterson, A., Fluornoy, N., Farewell, V. and Breslow, N. (1978), "The analysis of failure time in the presence of competing risk." *Biometrics*, 34, 541–554.

Price, B.L., Gilbert, P.B. and van der Laan, M.J. (2018), "Estimation of the optimal surrogate based on a randomized trial," *Biometrics*, 74, 1271–1281, PMCID: PMC6393111.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Sholukh, A.M., Fiore-Gartland, A., Ford, E.S., Hou, Y., Tse, L.V., Lempp, F.A. et al (2020), "Evaluation of SARS-CoV-2 neutralization assays for antibody monitoring in natural infection and vaccine trials," *medRxiv*.

Siber, G., Chang, I., Baker, S., Fernsten, P., O'Brien, K., Santosham, M. et al (2007), "Estimating the protective concentration of anti-pneumococcal capsular polysaccharide antibodies." *Vaccine*, 25, 3816–3826.

Son, H. and Fong, Y. (2020), "Fast Grid Search and Bootstrap-based Inference for Continuous Two-phase Polynomial Regression Models," *Environmetrics*, in press.

Stone, C.J. et al (1994), "The use of polynomial splines and their tensor products in multivariate function estimation," *The Annals of Statistics*, 22, 118–171.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267–288.

Tikhonov, A.N. and Arsenin, V.I. (1977), *Solutions of ill-posed problems*, vol. 14, Winston, Washington, DC.

van der Laan, L., Zhang, W. and Gilbert, P.B. (2021), "Efficient nonparametric estimation of the covariate-adjusted threshold-response function, a support-restricted stochastic intervention." *arXiv*, arXiv:2107.11459.

van der Laan, M.J. (2017), "A Generally Efficient Targeted Minimum Loss Based Estimator based on the Highly Adaptive Lasso," *The International Journal of Biostatistics*, 13.

van der Laan, M.J., Polley, E.C. and Hubbard, A.E. (2007), "Super learner," *Statistical Applications in Genetics and Molecular Biology*, 6, number 1.

VanderWeele, T. (2013), "Surrogate measures and consistent surrogates." *Biometrics*, 69, 561–568, PMCID: PMC4221255.

VanderWeele, T. and Ding, P. (2017), "Sensitivity analysis in observational research: introducing the E-value," *Annals of Internal Medicine*, 167(4), 268–74.

VanderWeele, T. and Mathur, M. (2020), "Commentary: developing best-practice guidelines for the reporting of E-values," *International Journal of Epidemiology*, Aug 2.

Westfall, P.H., Young, S.S. et al (1993), *Resampling-based multiple testing: Examples and methods for p-value adjustment*, vol. 279, John Wiley & Sons.

Westling, T. (2020), "Nonparametric tests of the causal null with non-discrete exposures," .

Westling, T. and Carone, M. (2020), "A unified study of nonparametric inference for monotone functions," *Annals of Statistics*, 48, 1001–1024.

Westling, T., Gilbert, P. and Carone, M. (2020a), "Causal isotonic regression," *Journal of the Royal Statistical Society Series B*, 82, 719–747.

Westling, T., van der Laan, M.J. and Carone, M. (2020b), "Correcting an estimator of a multivariate monotone function with isotonic regression," *Electron. J. Statist.*, 14, 3032–3069.

Williamson, B.D., Gilbert, P.B., Simon, N.R. and Carone, M. (2020), "A unified approach for inference on algorithm-agnostic variable importance," *arXiv preprint arXiv:2004.03683*.

Wood, S. (2017), *Generalized Additive Models: An Introduction with R, Second Edition*, Chapman & Hall/CRC Texts in Statistical Science, CRC Press, Boca Raton, FL.

World Health Organization (2020), "WHO Director-General's opening remarks at the media briefing on COVID-19 - 6 August 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—6-august-2020," *Aspen Security Forum/WHO Press Briefing*.

Wright, M.N., Ziegler, A. et al (2017), "`ranger`: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R," *Journal of Statistical Software*, 77.

Zheng, W. and van der Laan, M. (2017), "Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes," *Journal of Causal Inference*, 5, PMCID: PMC5459686.

Zhuang, Y., Huang, Y. and Gilbert, P.B. (2019), "Simultaneous Inference of Treatment Effect Modification by Intermediate Response Endpoint Principal Strata with Application to Vaccine Trials," *The International Journal of Biostatistics*.

Zou, H. and Hastie, T. (2003), "Regression shrinkage and selection via the elastic net, with applications to microarrays," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–20.

23 Appendix: Simulation of COVID-19 Vaccine Efficacy Trial Data Sets

24 Simulating COVID VE Trial Data Sets

24.1 Notation

1. A: randomization assignment to vaccine or placebo (1=vaccine, 0=placebo)
2. B: baseline SARS-CoV-2 status (0 if all SARS-CoV-2 diagnostic tests by Day 1 are negative and 1 if some are positive: 0=negative, 1=positive)
3. X: baseline covariate vector with components $X_1, \cdots, X_5$
   (a) X_1: Indicator At-risk for COVID
   (b) X_2: Sex assigned at birth (1=female, 0=male)
   (c) X_3: Indicator of minority
   (d) X_4: Age in years (≥ 18)
   (e) X_5: BMI
4. S_1: Vector of antibody markers measured at Day 1 (dose 1 visit), with components IgG Spike, IgG RBD, PsV ID50, PsV cID80, WT LV MN50
5. S_{29}: Vector of the same antibody markers measured at Day 29 (dose 2 visit)
6. S_{57}: Vector of the same antibody markers measured at Day 57 ($\approx$ peak immunogenicity time point)
7. R: Indicator a participant is randomly sampled into the subcohort for measurement of (S_1, S_{29}, S_{57})
8. T_{29}: Number of days from Day 29 visit until COVID endpoint starting 7 days post Day 29 visit (failure time of interest for studying Day 29

markers as correlates)

9. C_{29}: Number of days from Day 29 visit until right-censoring
10. Δ_{29}: Indicator of $T_{29} \leq C_{29}$
11. $\tilde{T}_{29} = min(T_{29}, C_{29})$
12. T_{57}: Number of days from Day 57 visit until COVID endpoint starting 7 days post Day 57 visit (failure time of interest for studying Day 57 markers as correlates)
13. C_{57}: Number of days from Day 57 visit until right-censoring
14. Δ_{57}: Indicator of $T_{57} \leq C_{57}$
15. $\tilde{T}_{57} = min(T_{57}, C_{57})$

Next, in turn we describe the three steps for simulating a data set. First, we simulate the covariates in all participants, second we simulate the Day 57 onwards failure time information in all participants, third we fill in the Day 29 to Day 57 failure time information, and fourth we define R and set (S_1, S_{29}, S_{57}) values to NAs for those with $R = 0$ and $\Delta_{29} = 0$ (non-cases).

24.2 Simulation of the covariates

First, A and B are drawn as independent Bernoulli random draws with success probabilities $P_{a=1}$ and $P_{b=1}$, respectively specified by the user. Then, for each of the four baseline strata $A = a, B = b$ with $(a, b) \in \{0, 1\} \times \{0, 1\}$, the 20-vector

$$W = (X^T, S_1^T, S_{29}^T, S_{57}^T)^T$$

is simulated. As X_1, X_2, and X_3 are the only binary variables, for simplicity we first simulate them as independent Bernoulli random variables.

1. X_1 is drawn from a Bernoulli distribution with specified success probability $P_1 = P(X_1 = 1)$
2. X_2 is drawn from a Bernoulli distribution with specified success probability $P_2 = P(X_2 = 1)$

3. X_3 is drawn from a Bernoulli distribution with specified success probability $P_3 = P(X_3 = 1)$

Next, we define a latent 17-vector variable W^L that has a multivariate normal distribution with mean vector.

$$\mu^L = (\mu_{X4}, \mu_{X5}, \mu_{S1}^T, \mu_{S29}^T, \mu_{S57}^T)^T$$

with variance elements Σ_{diag} = 17-vector of variances for the elements of W^L. The covariance elements are defined by specifying the correlation parameters $\rho[i, j]$ for all $i = 1, \cdots, 17, j = 1, \cdots, 17$ such that the $(i, j)^{th}$ element of the variance-covariance matrix Σ of W^L is

$$\rho[i, j] * \sqrt{\Sigma_{diag}[i]\Sigma_{diag}[j]}.$$

Based on specification of all of the input parameters, W^L is drawn. Then the following steps are done to attain W based on W^L:

1. X_4 is taken to be W_1^L rounded to the nearest year at enrollment
2. X_5 is taken to be W_2^L
3. S_{1j} is taken to be $W_{S1.j}^L$ for $j = 1, \cdots, 5$, and in the analysis one follows the convention that values below the LLOD are set to LLOD/2 and values above the ULOQ are set to ULOQ.
4. $S_{29.j}$ and $S_{57.j}$ are also taken to be $W_{S29.j}^L$ and $W_{S57.j}^L$, respectively, for $j = 1, \cdots, 5$, again following the convention that values below the LLOD are set to LLOD/2 and values above the ULOQ are set to ULOQ.

24.2.1 Input parameters for simulating covariates

The following lists the set of input parameters that are needed to simulate the covariate data, and indicates default values.

1. N = Total number of enrolled trial participants
2. $P_{a=1}$: $P(A = 1)$ = Probability an individual is randomized to vaccine A=1 (default 0.5)
3. $P_{b=1}$: $P(B = 1)$ = Probability an individual is baseline SARS-CoV-2 positive (default 0.10)

4. P_1: $P(X_1 = 1)$ (default 0.3)
5. P_2: $P(X_2 = 1)$ (default 0.5)
6. P_3: $P(X_3 = 1)$ (default 0.3)
7. μ^L (defaults listed below)

   Vaccine baseline negative $(a = 1, b = 0)$:

   $\mu^L = c(55, 30, 0, 0, 0, 0, 0, 4.4, 4.2, 2.2, 1.8, 2.2, 5.9, 5.7, 3.2, 2.8, 3.2)$

   Placebo baseline negative $(a = 0, b = 0)$:

   $\mu^L = c(55, 30, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$

   Vaccine baseline positive $(a = 1, b = 1)$:

   $\mu^L = c(55, 30, 3.9, 3.7, 1.8, 1.5, 1.7, 4.5, 4.3, 2.3, 1.9, 2.3, 6.0, 5.8, 3.3, 2.9, 3.3)$

   Placebo baseline positive $(a = 0, b = 1)$:

   $\mu^L = c(55, 30, 3.5, 3.3, 1.4, 1.3, 1.5, 4.1, 4.0, 2.0, 1.8, 2.0, 5.6, 5.5, 3.0, 2.7, 3.1)$
8. Σ_{diag} (defaults listed below)

   Vaccine baseline negative $(a = 1, b = 0)$:

   $\Sigma_{diag} = c(22.3^2, 7^2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.6^2, 0.7^2, 0.8^2, 0.8^2, 0.8^2,$

   $0.7^2, 0.8^2, 0.98^2, 0.98^2, 0.94^2)$

   Placebo baseline negative $(a = 0, b = 0)$:

   $\Sigma_{diag} = c(22.3^2, 7^2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2,$

   $0.2, 0.2, 0.2, 0.2, 0.2)$

   Vaccine baseline positive $(a = 1, b = 1)$:

   $\Sigma_{diag} = c(22.3^2, 7^2, 0.6^2, 0.7^2, 0.8^2, 0.8^2, 0.8^2, 0.6^2, 0.7^2, 0.8^2, 0.8^2, 0.8^2,$

   $0.7^2, 0.8^2, 0.98^2, 0.98^2, 0.94^2)$

   Placebo baseline positive $(a = 0, b = 1)$:

   $\Sigma_{diag} = c(22.3^2, 7^2, 0.6^2, 0.7^2, 0.8^2, 0.8^2, 0.8^2, 0.6^2, 0.7^2, 0.8^2, 0.8^2, 0.8^2,$

   $0.7^2, 0.8^2, 0.98^2, 0.98^2, 0.94^2)$

9. Correlation parameters ρ (defaults listed below)

   Vaccine baseline negative $(a = 1, b = 0)$:

$$\begin{aligned}
\rho[i,j] &= 0.25 + 0.1I(i = 3, j = 4) + 0.2 * I(i = 5, j = 6) : i = 1, \cdots, 6, j > i \\
\rho[i,j] &= 0.6 + 0.1I(i = 8, j = 9) + 0.2 * I(i = 10, j = 11) : i = 7, \cdots, 11, j > i \\
\rho[i,j] &= 0.7 + 0.1I(i = 13, j = 14) + 0.2 * I(i = 15, j = 16) : i = 12, \cdots, 16, j > i
\end{aligned}$$

   Placebo baseline negative $(a = 0, b = 0)$:

$$\begin{aligned}
\rho[i,j] &= 0.15 + 0.05I(i = 3, j = 4) + 0.1 * I(i = 5, j = 6) : i = 1, \cdots, 6, j > i \\
\rho[i,j] &= 0.2 + 0.05I(i = 8, j = 9) + 0.1 * I(i = 10, j = 11) : i = 7, \cdots, 11, j > i \\
\rho[i,j] &= 0.25 + 0.05I(i = 13, j = 14) + 0.1 * I(i = 15, j = 16) : i = 12, \cdots, 16, j >
\end{aligned}$$

   Vaccine baseline positive $(a = 1, b = 1)$:

$$\begin{aligned}
\rho[i,j] &= 0.55 + 0.1I(i = 3, j = 4) + 0.2 * I(i = 5, j = 6) : i = 1, \cdots, 6, j > i \\
\rho[i,j] &= 0.6 + 0.1I(i = 8, j = 9) + 0.2 * I(i = 10, j = 11) : i = 7, \cdots, 11, j > i \\
\rho[i,j] &= 0.7 + 0.1I(i = 13, j = 14) + 0.2 * I(i = 15, j = 16) : i = 12, \cdots, 16, j > i
\end{aligned}$$

   Placebo baseline positive $(a = 0, b = 1)$:

$$\begin{aligned}
\rho[i,j] &= 0.55 + 0.1I(i = 3, j = 4) + 0.2 * I(i = 5, j = 6) : i = 1, \cdots, 6, j > i \\
\rho[i,j] &= 0.6 + 0.1I(i = 8, j = 9) + 0.2 * I(i = 10, j = 11) : i = 7, \cdots, 11, j > i \\
\rho[i,j] &= 0.7 + 0.1I(i = 13, j = 14) + 0.2 * I(i = 15, j = 16) : i = 12, \cdots, 16, j > i
\end{aligned}$$

In the above correlation specification, note that extra correlation is added to IgG Spike and RBD readouts (same assay), as well as to PsV ID50 and cID80 (same assay). After $\rho[i,j]$ is define for all $i < j$, we set $\rho[j,i] = \rho[i,j]$ for all $j < i$.

24.3 Simulation of the failure time data

The failure time variables to simulate are $T_{29}, C_{29}, \tilde{T}_{29}, \Delta_{29}, T_{57}, C_{57}, \tilde{T}_{57}, \Delta_{57}$. First, in the placebo arm the event time T_{57} is simulated dependent on

A, B, X_4^*, where X_4^* is a standardized version of age X_4 that has empirical mean 0 and empirical standard deviation 1. Second, in the vaccine arm T_{57} is simulated dependent on $A, B, X_4^*, S_{57.1}^*$, where $S_{57.1}^*$ is the first Day 57 marker (IgG Spike). The positive correlations among the different marker variables implies that T_{57} depends on the other Day 57 markers as well, but for a simple simulation model we only specify dependence on $S_{57.1}^*$. Alternatively, the simulations could be designed with the failure time connected to a latent variable that is an average of the markers.

Third and fourth, we implement a parallel approach to simulate T_{29} in the placebo arm and T_{29} in the vaccine arm, now using $S_{29.1}^*$ instead of $S_{57.1}^*$. These new simulations add intercurrent failures between the Day 29 visit and the Day 57 visit. Then calculations are made to enforce structural relationships between T_{29} and T_{57}.

24.3.1 Input parameters for simulating the failure time information

Let W^* be the vector W with each of the 17 normally distributed variables (elements 4 through 20) centered and scaled to have empirical mean zero and standard deviation one. The following parameters simulate failure time information starting from the Day 57 visit.

1. τ: final time point (in days) for analysis post Day 57 visit (default 180)
2. $P_{0b}(\tau)$: $P(T_{57} \leq \tau | A = 0, B = b, W^* = 0)$ = placebo arm baseline cumulative failure probability for baseline status group b (default $P_{00}(\tau) = 0.10, P_{01}(\tau) = 0.05$)
3. $VE_b = 1 - \frac{\lambda_{1b0}}{\lambda_{0b0}}$ for $b = 0, 1$, where $\lambda_{ab0} = \lambda(t|A = a, B = b, W^* = 0)$ for $a = 0, 1$ (default $VE_0 = VE_1 = 0.90$)
4. β_{0b} = Placebo group log hazard ratio per year increase in age X_4 for baseline stratum b. Once β_{0b} and $\Sigma_{diag}[1]$ are specified, the parameter β_{0b}^* in the Cox model (24) below is calculated as $\beta_{0b}^* = \beta_{0b}\sqrt{\Sigma_{diag}[1]}$. (default: $\beta_{0b} = log10(1.1)$ for each $b = 0, 1$)
5. β_{1b} = Vaccine group log hazard ratio per year increase in age X_4 for baseline stratum b. Once β_{1b} and $\Sigma_{diag}[1]$ are specified, the parameter

β^*_{1b} in the Cox model (25) below is calculated as $\beta^*_{1b} = \beta_{1b}\sqrt{\Sigma_{diag}[1]}$. (default: $\beta_{1b} = log10(1.1)$ for each $b = 0, 1$)

6. γ_{1b} = Vaccine group log hazard ratio per unit change in the marker $S_{57.1}$ (i.e., per 10-fold change in the marker on the natural/antilog10 scale). Once γ_{1b} and $\Sigma_{diag}[13]$ are specified, the parameter γ^*_{1b} in the Cox model (25) below is calculated as $\gamma^*_{1b} = \gamma_{1b}\sqrt{\Sigma_{diag}[13]}$. (default: $\gamma_{1b} = 0$ for each $b = 0, 1$ (null case))

7. FollowupRange = range of days since enrollment until the date of data cut for the analysis, accounting for staggered enrollment (default FollowupRange = c(4*7, 6.5*7))

8. $P_{LFU}(\tau)$: Probability loss to follow-up (prematurely) by τ (default $P_{LFU}(\tau) = 0.05$)

The following parameters simulate failure time information between the Day 29 and Day 57 visits ("intercurrent" failure defined in terms of T_{29}), specified in parallel fashion to the failure time information for T_{57}. With the exception of the intercurrent vaccine efficacy parameter, the following parameters are by default defined according to the Day 57 failure time simulation parameters.

1. $P_{0b.intcur}(57)$: $P(T_{29} \leq 28|A = 0, B = b, W^* = 0)$ = placebo arm baseline cumulative intercurrent failure probability for baseline status group b (default $P_{0b.intcur}(57) = P_{0b}(\tau) * 28/\tau$, which specifies the same placebo arm incidence of failure from Day 29 to Day 57 as after Day 57)

2. $VE_{b.intcur} = 1 - \frac{\lambda_{1b0.intcur}}{\lambda_{0b0.intcur}}$ for $b = 0, 1$, where $\lambda_{ab0.intcur} = \lambda(t|A = a, B = b, W^* = 0)$ for $a = 0, 1$ (default $VE_{b.intcur} = 0.7 * VE_b$ for each $b = 0, 1$)

3. $\beta_{0b.intcur}$ = Placebo group log intercurrent hazard ratio per year increase in age X_4 for baseline stratum b. Once $\beta_{0b.intcur}$ and $\Sigma_{diag}[1]$ are specified, the parameter $\beta^*_{0b.intcur}$ in the Cox model (26) below is calculated as $\beta^*_{0b.intcur} = \beta_{0b.intcur}\sqrt{\Sigma_{diag}[1]}$. (default: $\beta_{0b.intcur} = \beta_{0b}$ for each $b = 0, 1$)

4. $\beta_{1b.intcur}$ = Vaccine group log intercurrent hazard ratio per year increase in age X_4 for baseline stratum b. Once $\beta_{1b.intcur}$ and $\Sigma_{diag}[1]$ are specified, the parameter $\beta^*_{1b.intcur}$ in the Cox model (27) below is calculated as

$\beta^*_{1b.intcur} = \beta_{1b}\sqrt{\Sigma_{diag}[1]}$. (default: $\beta_{1b.intcur} = \beta_{1b}$ for each $b = 0, 1$)

5. $\gamma_{1b.intcur}$ = Vaccine group log intercurrent hazard ratio per unit change in the marker $S_{29.1}$ (i.e., per 10-fold change in the marker on the natural/antilog10 scale). Once $\gamma_{1b.intcur}$ and $\Sigma_{diag}[8]$ are specified, the parameter $\gamma^*_{1b.intcur}$ in the Cox model (27) below is calculated as $\gamma^*_{1b.intcur} = \gamma_{1b.intcur}\sqrt{\Sigma_{diag}[8]}$. (default: $\gamma_{1b.intcur} = \gamma_{1b}$ for each $b = 0, 1$)

24.3.2 Exponential/proportional hazards models for T_{57}

For the placebo arm, we assume the following simple proportional hazards models for T_{57}, separately by baseline status:

$$\lambda_{0b}(t|W) = \lambda_{0b0}e^{\beta^*_{0b}X^*_4} \tag{24}$$

where, assuming an exponential distribution, $\lambda_{0b0} = \lambda(t|A = 0, B = 0, W^* = 0)$ is determined by the equation

$$1 - e^{-\lambda_{0b0}\tau} = P_{0b}(\tau)$$

and the parameters β^*_{00} and β^*_{01} specify how strongly X^*_4 (standardized age) associates with COVID.

For the vaccine arm, we assume the following proportional hazards models for T_{57}, again separately by baseline status:

$$\lambda_{1b}(t|W) = \lambda_{1b0}e^{\beta^*_{1b}X^*_4 + \gamma^*_{1b}S^*_{571}} \tag{25}$$

where λ_{1b0} is determined by the equation

$$VE_b = 1 - \frac{\lambda_{1b0}}{\lambda_{0b0}},$$

where VE_b (proportional hazards vaccine efficacy at central covariate level $W^* = 0$) is input by the user.

24.3.3 Simulating T_{57}

Once a participant's values $A, B, X^*_4, S^*_{S57.1}$ are generated, then we simulate the participant's T_{57} value from an exponential distribution with rate parameter defined by the input parameters and the Cox model (24) or (25).

24.3.4 Simulating C_{57} and Δ_{57}

First, a random variable C_{157} is simulated from a Uniform distribution over the range FollowupRange. Second, an exponential random variable C_{257} is simulated with rate parameter λ_{cens} determined by

$$1 - e^{-\lambda_{cens}\tau} = P_{LFU}(\tau).$$

Then, we set $C_{57} = min(C_{157}, C_{257})$, and next set $T^*_{57} = min(T_{57}, C_{57})$ and $\Delta_{57} = I(T_{57} \leq C_{57})$. Note that, because in the analysis outcomes for Day 57 correlates analyses are only counted starting 7 days post Day 57 visit, cases with $T_{57} < 7$ are excluded from the analysis (this is handled in the data analysis code, not in the data set construction code).

24.3.5 Exponential/proportional hazards models for T_{29} intercurrent failure

For the placebo arm, we assume the following simple proportional hazards models for T_{29}, separately by baseline status:

$$\lambda_{0b.intcur}(t|W) = \lambda_{0b0.intcur} e^{\beta^*_{0b.intcur} X^*_4} \tag{26}$$

where, assuming an exponential distribution, $\lambda_{0b0.intcur} = \lambda(t|A = 0, B = b, W^* = 0)$ is determined by the equation

$$1 - e^{-\lambda_{0b0.intcur} * 57} = P_{0b.intcur}(57)$$

and the parameters $\beta^*_{00.intcur}$ and $\beta^*_{01.intcur}$ specify how strongly X^*_4 (standardized age) associates with COVID. For the vaccine arm, we assume the following proportional hazards models for T_{29} intercurrently, again separately by baseline status:

$$\lambda_{1b.intcur}(t|W) = \lambda_{1b0.intcur} e^{\beta^*_{1b.intcur} X^*_4 + \gamma^*_{1b.intcur} S^*_{291}} \tag{27}$$

where $\lambda_{1b0.intcur}$ is determined by the equation

$$VE_{b.intcur} = 1 - \frac{\lambda_{1b0.intcur}}{\lambda_{0b0.intcur}},$$

where $VE_{b.intcur}$ (proportional hazards vaccine efficacy at central covariate level $W^* = 0$) is input by the user.

24.3.6 Simulating T_{29}

Once a participant's values $A, B, X_4^*, S_{S29.1}^*$ are generated, then we initially simulate the participant's T_{29} value from an exponential distribution with rate parameter defined by the input parameters and the Cox model (26) or (27). If $T_{29} < T_{57} + 28$, then the value of T_{29} is kept. If $T_{29} \geq T_{57} + 28$, then we set the final value of $T_{29} = T_{57} + 28$ plus a a draw from a random uniform distribution over -3 to 3 days rounded to the nearest day (to account for visit window variability).

24.3.7 Simulating C_{29} and Δ_{29}

For simplicity, we do not allow dropout between the Day 29 visit and the Day 57 visit. Therefore, we set $C_{29} = C_{57} + 28$. Then we set $\Delta_{29} = I(T_{29} \leq C_{29})$.

Note that, because in the analysis outcomes for Day 29 correlates analyses are only counted starting 7 days post Day 29 visit, cases with $T_{29} < 7$ are excluded from the analysis (this is handled in the data analysis code, not in the data set construction code).

24.4 Simulating the subcohort indicator R

The subcohort indicator R is one if a participant is sampled for measurement of (S_1, S_{29}, S_{57}).

24.4.1 Input parameters for simulating R

We simulate R following Table 5 in this SAP, where the six baseline demographic strata are:

1. $X_4 \geq 65$ and $X_3 = 0$ (minority)
2. $X_4 \geq 65$ and $X_3 = 1$ (non-minority)
3. $X_4 < 65$ and $X_1 = 1$ (at-risk) and $X_3 = 0$
4. $X_4 < 65$ and $X_1 = 1$ (at-risk) and $X_3 = 1$
5. $X_4 < 65$ and $X_1 = 0$ (not at-risk) and $X_3 = 0$

6. $X_4 < 65$ and $X_1 = 0$ (not at-risk) and $X_3 = 1$

For each of the 24 subgroups/cells defined by (a, b) cross-classified with the above 6 demographic subgroups (as in Table 5), define the total numbers to be sampled into the immunogenicity subcohort, $n2(a, b, c)$ for $a = 0, 1, b = 0, 1, c = 0, \cdots, 6$. Then, for each subgroup (a, b, c), R is set to 1 for a random sample of size $n2(a, b, c)$ without replacement. Lastly, all non-cases (with $\Delta = 0$) and $R = 0$ have all three values (S_1, S_{29}, S_{57}) set to NA.

The default settings for $n2(a, b, c)$ to match Table 5 in the SAP are as follows:

- $n2(1, 0, c) = 150$ for $c = 1, \cdots, 6$
- $n2(0, 0, c) = 20$ for $c = 1, \cdots, 6$
- $n2(1, 1, c) = 50$ for $c = 1, \cdots, 6$
- $n2(0, 1, c) = 50$ for $c = 1, \cdots, 6$

24.4.2 Per-protocol indicator

Lastly, we simulate the per-protocol (PP) indicator variable, which is 1 if both immunizations at Day 0, 29 were received and there were no specified protocol violations.

Input parameter:

$P_{PP=1}$ = Probability a participant is per-protocol (default = 0.99)

$P_{PP=1}$ is simulated from a Bernoulli random variable with success probability $P_{PP=1}$. The Day 57 marker correlates analyses are done in individuals with $P_{PP=1}$.

24.4.3 Variables output for the data set

The following collates all of the variables defined for the simulated data set.

- A
- B
- $X_1, \cdots, X_5$

- PP
- R
- $S_{1.1}, \cdots, S_{1.5}$
- $S_{29.1}, \cdots, S_{29.5}$
- $S_{57.1}, \cdots, S_{57.5}$
- $\tilde{T}_{29}, \Delta_{29}$
- $\tilde{T}_{57}, \Delta_{57}$

25 Appendix: Notes on Planning for Stage 2 Correlates Analyses

The paper Sun, Zhou, Gilbert (submitted) "Analysis of Cox model with Longitudinal Covariates with Measurement Errors and Partly Interval-Censored Failure Time, with Application to an AIDS Clinical Trial" may be a suitable method for assessing antibody markers over time as correlates of the SARS-CoV-2 infection endpoint, once there is follow-up data for more than a year with antibody markers measured at all time points up to at least a year. The paper extends Fu and Gilbert (2017) from right-censored failure time data to partly interval-censored failure time data, which means a composite endpoint is analyzed with one component subject to right-censoring and the other component subject to interval censoring. In our application, the COVID primary endpoint is subject to right-censoring, and seroconversion is subject to interval-censoring.

26 Appendix on Stochastic VE Analysis Project

We consider an approach that simply implements the stochastic VE analysis as specified in the SAP, using a grid of mean shifts, and then the results are interpreted by marking the mean shifts corresponding to expected shifts under each of the SARS-CoV-2 variants considered in the Montefiori panel used in the phase 1 study. With this mindset, this is a data analysis project, not a new methods project.

Let S_0 denote D57 antibody level to the D614G strain, and S_v denote D57 antibody level to variant v. Let p_{cc} be the density of baseline covariates L in the case-cohort study. We make the following assumption:

$$\textbf{A.1} \qquad E_{cc}[S_0|L] = E_{Ph1}[S_0|L].$$

Let $\hat{E}_{cc}[S_0|L]$ be an estimator of $E_{cc}[S_0|L]$ from the case-cohort data.

Suppose there are n participants in the phase 1 study with data on S_0 and L. Let $\mu_0^{std.cc}$ be the mean of S_0 in the phase 1 sample standardized/transported to the distribution of L in the case-cohort study. We estimate $\mu_0^{std.cc}$ by

$$\hat{\mu}_0^{std.cc} = \sum_{i=1}^{n} \hat{E}_{cc}[S_0|L_i]\hat{p}_{cc}(L_i),$$

where assumption **A.1** is needed for this estimate to be unbiased.

Next, we need a way to estimate $\mu_v^{std.cc}$. This is challenging given that S_v is only measured in between 10 and 28 vaccine recipients in the phase 1 trial. Because of the small sample size, we make the assumption that

$$E_{cc}[S_v|L] - E_{cc}[S_0|L] = E_{ph1}[S_v] - E_{ph1}[S_0],$$

and then we estimate $E_{cc}[S_v|L]$ by

$$\hat{E}_{cc}[S_v|L] = \hat{E}_{cc}[S_0|L] + \left(\bar{S}_v - \bar{S}_0\right),$$

where $\bar{S}_0$ is the sample average of the S_0 measurements in the phase one trial and similarly for S_v.

In conclusion, for a given variant v, the mean shift of focus for interpreting the result of the stochastic VE analysis is

$$\Delta_v = \hat{E}_{cc}[S_0|L] - \hat{E}_{cc}[S_v|L].$$

26.1 Remarks

The estimator of $\hat{\mu}_0^{std.cc}$ requires that we have data on the same baseline covariates L that are used in the case-cohort study for estimation of $E_{cc}[S_0|L]$.

If we do not, then we may need to simplify the estimator of $E_{cc}[S_0|L]$ to only include a few covariates L collected in both studies. In the most extreme case, with no L available in the phase one study, we simply take as the estimate of $\hat{\mu}_0^{std.cc}$ the sample average of S_0 values in the phase 1 study, $\bar{S}_0$. In this case Δ_v is simply taken to be $\bar{S}_0 - \bar{S}_v$.