Assessing Vaccine Durability in Randomized Trials Following Placebo Crossover

Jonathan Fintzi and Dean Follmann,
Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases
Rockville, Maryland, U.S.A.

email addresses: jon.fintzi@nih.gov and dean.follmann@nih.gov

Abstract

Randomized vaccine trials are used to assess vaccine efficacy and to characterize the durability of vaccine induced protection. There is a broad consensus that placebo volunteers in COVID-19 vaccine trials should be offered a vaccine once efficacy has been established. This will likely lead to most placebo volunteers crossing over to the vaccine arm, thus complicating the assessment of long term durability. We show how to analyze durability following placebo crossover and demonstrate that the vaccine efficacy profile that would be observed in a placebo controlled trial is recoverable in a trial with placebo crossover. This result holds no matter when the crossover occurs and with no assumptions about the form of the efficacy profile. We only require that the vaccine efficacy profile applies to the newly vaccinated irrespective of the timing of vaccination. We develop different methods to estimate efficacy within the context of a proportional hazards regression model and explore the implications of placebo crossover for estimation of vaccine efficacy under different efficacy dynamics and study designs via simulation. We apply our methods to simulated COVID-19 vaccine trials with durable and waning vaccine efficacy and a total follow-up of 2 years.

Keywords: COVID-19; Proportional hazards regression; Vaccine efficacy; Vaccine trial design.
1 Introduction

Randomized phase III clinical trials are used to definitively demonstrate the efficacy of candidate vaccines. Volunteers are randomized to receive vaccine or a placebo and followed for a period of time to assess whether the vaccine reduces the rate of disease. An important question in vaccine development is whether vaccine induced protection is durable. For COVID-19 vaccines, questions surrounding vaccine durability are particularly salient as acquired immunity against seasonal and other coronaviruses ranges from 6 months to 2 years (Poland et al., 2020, Choe et al., 2020). Clinical trials for vaccines against COVID-19 have planned to follow participants for up to two years, (Moderna, 2020).

To assess long term safety and durability, long term blinded follow-up of the original placebo and vaccine arms is ideal (World Health Organization, 2020). From an ethical perspective, placebo volunteers should be offered a vaccine once efficacy is established (Wendler et al., 2020). However, vaccination of placebo volunteers may occur before it is known whether vaccine induced protection is durable. Besides waning of efficacy, there is concern that the vaccine might eventually cause harm in subgroups. Such harm is known as vaccine associated enhanced disease (VAED) and has been observed in other contexts, such as the Dengvaxia vaccine in seronegative individuals (Sridhar et al., 2018). Finally, crossover may occur before reliable estimates of vaccine efficacy can be assessed in subgroups or for rarer events such as severe disease.

It might seem that the ability to assess vaccine durability following placebo crossover is completely lost once there is no longer an unvaccinated comparator group (World Health Organization, 2020). However, at the point of crossover the study remains a randomized trial, albeit of immediate vs deferred vaccination. In fact, the VE profile for a standard non-crossover trial can be recovered with placebo crossover under fairly mild conditions (Follmann et al., 2020). The only additional assumption we require in the placebo crossover setting, beyond the typical assumptions for a standard parallel-arm trial, is that the same VE profile applies to the newly vaccinated irrespective of the timing of vaccination.

Crossover trials for absorbing endpoints, such as infection or death, have been discussed in the literature (Nason and Follmann, 2010, Makubate and Senn, 2010). However, these methods apply to estimation of an overall intervention effect when the benefit of the intervention stops once a drug is no longer administered. Vaccination is quite different as the benefit lingers and our goal is to see how the intervention effect varies with time. Crossover has been discussed for vaccine trials, but only for the placebo arm and only to measure immune response (Follmann, 2006). Delayed vaccination has been used in an Ebola vaccine trial, but to serve as control group prior to deferred vaccination (Henao-Restrepo et al., 2017).

Statistically, VE can be estimated using the Cox proportional hazards (PH) model (Cox, 1972), where it is minus the hazard ratio for disease on vaccine compared to placebo. It is straightforward to allow VE to depend on time through use of time-varying covariates (Therneau and Grambsch, 2013). Proportional hazard models can accommodate a variety of efficacy profiles, including simple linear and piecewise–constant trends, or smooth functions that can be estimated semi–parametrically. The added flexibility of semi–parametric methods comes at the cost of greater sample size requirements, which may be compounded by the loss of the placebo arm following crossover. Hence, it is important to consider the implications of placebo crossover and other aspects of study design on the precision of VE estimates.

Our contributions in this work are threefold: first, we establish that vaccine durability can be accurately assessed following placebo crossover under fairly mild assumptions. This result has immediate implications for the conduct and analysis of ongoing COVID-19 vaccine trials.
We demonstrate how to estimate VE under placebo crossover using PH regression and develop model-based estimates for estimating the VE profile using log-linear and P-spline functions of the time since vaccination. We also provide a justification for using calendar time as the natural timescale in VE models. Second, we explore the implications of different aspects of study design and efficacy dynamics for estimation. We are particularly interested in how the pace of crossover and unspecified heterogeneity in risk affect estimation in placebo crossover and standard parallel trials. We demonstrate that placebo crossover with continued follow-up provides unambiguously better estimates than would be obtained if a trial were to be stopped early without further monitoring. This provides assurance that pre-crossover VE estimated in subgroups and for severe disease can be improved post-crossover. Finally, we apply our methods to two simulated COVID-19 vaccine trials that cross over after efficacy is established to illustrate practical issues in implementation for constant and waning VE scenarios.

2 Vaccine Efficacy Under Placebo Crossover

2.1 Conceptual Development

Consider a vaccine trial where volunteers are randomized to receive vaccine or placebo. For now, we assume that all are participants are enrolled at the same time and followed for a period of time, \( [0, 2\tau] \). Suppose that a blinded crossover occurs at time \( \tau \), at which point the original vaccinees receive placebo and original placebos receive vaccine. Following crossover, both arms are vaccinated and thus comparative efficacy might seem lost as there is no control group. However, a randomized trial remains, but now as trial of immediate vs delayed vaccination; these assignments correspond to the original vaccine and placebo arms. This ‘rebranded’ randomized trial can still provide information about safety and durability.

To illustrate, suppose we have case counts for the two randomization arms over the two periods, \( (0, \tau] \) and \( (\tau, 2\tau] \). Suppose that the vaccine–placebo case split is 20 to 100 in period one, and in period two we observe 20 cases in the Vaccine versus 12 on the placebo–turned–vaccine arm. Using a person-time analysis, and imagining the denominators are so large that they cancel out, we obtain a simple estimate of the period one VE as \( \hat{VE}_1 = 1 - 20/100 = 0.80 \). Suppose now that this VE applies to the newly vaccinated participants in the second period. The 12 placebo cases observed in period two represent 20% of the cases we would have expected had the placebo group not been vaccinated. A simple counterfactual estimate of the placebo case count, absent vaccination, divides the period two placebo case count by the case suppression rate: \( \hat{N}^{plac}_2 = 12/0.2 = 60 \). We contrast the counterfactual placebo case count of 60 to the eight observed cases in the Vaccine to obtain an estimate of VE in period two, \( \hat{VE}_2 = 1 - 20/60 \). Based on these crude estimates, we conclude that VE has waned as efficacy has dropped from 80% in period one to 66.7% in period two.

The crux of this example is that VE depends on time since vaccination. By virtue of the study randomization, the placebo can be assumed to receive the same immediate benefit from vaccination that the original vaccine group received, regardless of changes in the population attack rate. Additional considerations are discussed in Follmann et al. (2020). For our purposes, a similar analysis can be carried out in a more flexible PH framework.

In the Cox PH model, the baseline placebo hazard function for the time to disease is an arbitrary function and the hazard for participant \( i \) is proportional to the baseline hazard:

\[
h_i(t) = h_0(t) \exp(Z_i \theta),
\]
where $h_0(t)$ is the placebo hazard function, and $Z_i$ is the vaccine assignment indicator. Vaccine efficacy is defined as the change in the instantaneous risk of acquiring disease due to vaccination, and is calculated as $1 - \exp(\theta)$. Model (1) can be fit to data from all study volunteers to obtain an estimate of overall VE, and can also be applied to subgroups or to different events (e.g. infection, severe disease) to provide a more complete picture of a vaccine’s effects.

Suppose we want to assess whether VE changes at time $\tau$ post vaccination. Let $\tau^e_i$ denote the vaccination time for participant $i$. We can extend the PH model to include a time-varying vaccine parameter, $\theta(t)$, that may vary with time since vaccination:

$$h_i(t) = \lambda_0(t) \exp[Z_i(t)\theta(t)], \quad \theta(t) = \begin{cases} \theta_1, & t - \tau^e_i < \tau, \\ \theta_2, & t - \tau^e_i \geq \tau, \end{cases}$$

where $Z_i(t)$ is an indicator for whether participant $i$ is vaccinated at time $t$. We can calculate the VE in $T_1$ and $T_2$ as $1 - \exp(\theta_1)$ and $1 - \exp(\theta_2)$, respectively.

More generally, we might suppose that vaccine efficacy decays smoothly and model the decay linearly on the log-hazard scale. In this case,

$$h_i(t) = \lambda_0(t) \exp[Z_i(t)\theta(t)], \quad \theta(t) = \begin{cases} 1, & t - \tau^e_i < \tau, \\ \theta_1 + \theta_2 (t - \tau^e_i), & t - \tau^e_i \geq \tau, \end{cases}$$

and $VE(t) = 1 - \exp[\theta_1 + \theta_2 (t - \tau^e_i)]$, with $\tau^e_i = 0$ or $\tau^e_i = \tau$. The log-hazard for (3) is diagrammed in Figure for a constant baseline hazard. Again, the key insight is that the data are observations of the same decay profile.

### 2.2 General Development

We now develop this approach for the more realistic setting of a staggered entry trial and consider more general models for VE. Let $t \geq 0$ index time since study initiation, and let $\tau = \{\tau_i = (\tau^e_i, \tau^v_i); \tau^e_i \geq \tau^v_i > 0, \ i = 1, \ldots, N\}$ denote the times of study entry and vaccination for the study participants. Define $Z_i(t)$ to the vaccination indicator i.e. whether $t > \tau^v_i$. The hazard for participant $i$ is

$$h_i(t) = \begin{cases} 0, & t \leq \tau^e_i, \\ h_0(t) \exp[Z_i(t)f(t - \tau^v_i; \theta)], & t > \tau^e_i, \end{cases}$$

where $\theta$ is a vector of parameters. We calculate VE at time $t$ post-vaccination as one minus the ratio of vaccine to placebo hazards, i.e.,

$$VE(t) = 1 - \exp[f(t - \tau; \theta)].$$

The form of the general PH model, (4), allows for heterogeneity in study entry and vaccination. Furthermore, the model encompasses placebo–crossover trials, in which case $\tau^v_i > \tau^e_i$ for participants on the placebo arm, and standard trials with parallel arms, in which case $\tau^v_i = \infty$ for placebo volunteers. We can also easily simulate trial data by integrating (4) to obtain the cumulative hazard and sampling event times for each participant via the inverse survival distribution (Crowther and Lambert, 2013, Brilleman, 2019).
Figure 1: Log hazard for two study participants: \(i\), who is initially on placebo (orange line), and \(j\) who is initially given vaccine (green line). Vaccine efficacy wanes (i.e. log hazard increases) as a function of time since vaccination while the baseline attack rate is constant. At time \(\tau_i^{(v)}\), participant \(i\) is given vaccine and follows the same efficacy profile as \(j\). The hazard function for participant \(i\) is \(\lambda_0(t) = h_0(t)\) prior to crossover and \(\lambda_0(t) = h_0(t) \exp \left[\theta_1 + \theta_2 \left(t - \tau_i^{(v)}\right)\right]\) after crossover.
It is standard practice in clinical trials to index events by time on study. However, calendar time is a more natural index for the Cox model in trials with an attack rate that varies with calendar time and staggered entry, as reflected in (4). Aligning the data on study entry distorts the risk set in the Cox partial likelihood at each event time. This is diagrammed in Figure 2 where participant \( j \), who is still at risk at \( t_i \), falls out of the risk set after we align the data on study entry. By the same token, participant \( k \) is erroneously introduced into the risk set. Furthermore, indexing the model by calendar time preserves our interpretation of vaccine efficacy as the relative reduction in the hazard, *ceteris paribus*. Specifically, participants are not guaranteed to be alike in their baseline hazards at each event time, and assuming otherwise may lead to incorrect cancellation of baseline hazard terms in the partial likelihood.

Figure 2: Participant histories and baseline hazards when the data are indexed in calendar time or aligned on times of study entry. The true data generating mechanism is indexed in calendar time. (a) vs. (b): Aligning the data on study entry changes the risk set as \( k \) falls out of the risk set at \( i \)'s event time and \( j \) is incorrectly introduced into the risk set. (c) vs. (d): Baseline hazards are no longer proportional after the data are aligned on study entry if care is not taken to re-index the baseline hazards to the correct calendar time.

Suppose participant \( i \) acquires disease at calendar time \( t \) after being on study for a period
\[ s_i = t - \tau_i^{(e)}. \] Setting aside baseline covariates, the partial likelihood contribution at \( s_i \) is

\[
L_i(\theta) = \frac{h_0\left(s_i + \tau_i^{(e)}\right) \exp\left(Z_i(t)f(s_i; \theta)\right)}{\sum_{j \in R_i} h_0\left(s_j + \tau_j^{(e)}\right) \exp\left(Z_j(t)f(s_j; \theta)\right)}
\]

\[
= \frac{\exp\left(Z_i(t)f(s_i; \theta)\right)}{\sum_{j \in R_i} \exp\left(Z_j(t)f(s_j; \theta)\right)},
\]

where the sum in the denominator is over all participants still in the risk set, \( R_i \), and the cancellation in the baseline hazards no longer cancel out. However, this will not be the case if we naively fit a Cox proportional hazards model to the entry aligned data since most implementations assume proportional hazards by default.

### 2.3 Flexible Models for \( VE(t) \)

As a practical matter, we do not know the functional form of the decay profile, \( f(\cdot) \). Two simple approaches are to model \( f(\cdot) \) as piecewise constant, or to assume a flexible parametric form for the decay function. However, these approaches involve uncomfortable choices that we may want to avoid. For instance, we might not have a basis for selecting change-points in the piecewise constant approach. Or, in the parametric case, we may not want to assume that \( VE(t) \) is monotonic or to rule out the possibility of VAED, e.g., if we assume that \( f(\cdot) < 0 \).

We can model \( f(\cdot) \) semi-parametrically, and there are a variety of ways in which such a strategy could be pursued [Wood, 2017, Perperoglou et al., 2019]. An attractive options is to estimate \( f(\cdot) \) using penalised cubic P-splines [Eilers and Marx, 1996]. This is implemented in the **survival** package in \( R \), which provides users with a convenient summary method for the linear and non-linear spline effects, which is useful for testing for non-linearity in the decay profile. Let \( P_L(t; k, \delta) \) denote a P-spline basis of degree \( \delta = 3 \) with \( L \) basis terms and vector of knot locations \( k \), and let \( \gamma \) be a vector of coefficients, with \( \gamma_0 \) reserved for the log-hazard ratio immediately following vaccination. The hazard for participant \( i \) is

\[
h_i(t) = \begin{cases} 
0 & , t \leq \tau_i^{(e)} \\
h_0(t) \exp\left(Z_i(t)\{\gamma_0 + \sum_{\ell=1}^L \gamma_\ell P_\ell(t - \tau_i^{(e)}; k, \delta)\}\right) & , t > \tau_i^{(e)}.
\end{cases}
\]

In practice, we center the decay component estimated by the P-spline at zero to ensure identifiability of \( \gamma_0 \). Note that we need to evaluate the hazard for each participant at every event time,
not merely at the time when a person experiences their own event (Therneau et al., 2017). We demonstrate how to fit this model in code provided in the GitHub repository linked at the end of the manuscript.

3 Assumptions

To recover VE under a standard trial requires that the volunteers in each arm are similar over time and that the external environment is similar over time. We elaborate below.

- **Volunteers in each arm are similar over time.** This can be violated if there is differential dropout in the two arms and dropout is related to underlying risk of disease. Relatedly, unobserved heterogeneity in risk can result in differential culling by infection of the vaccine and placebo groups. Thus, after a while, the remaining placebos tend to be a less risky group than the remaining vaccinees and the vaccine efficacy can appear to decrease, see Lipsitch (2019), Durham et al. (1998), Aalen et al. (2015). COVID-19 trials with 30,000 or more enrolled and perhaps 200-1000 cases over follow-up, any such bias should be small. In other settings, one could explicitly model the heterogeneity or integrate over the frailty distribution (Kanaan and Farrington, 2002).

- **Study environment similar.** The proportional hazards model allows for the attack rate to change with time. But if the pathogen mutates to a form that is resistant to vaccine effects, efficacy may appear to wane. Another possibility is if human behavior changes in such a way that the vaccine is less effective. For example, the viral innoculum at infection may increase over the study and overwhelm the immune response for later cases. We also want to guard against changes in assessment of endpoints that might favor one arm. Suppose symptomatic disease is the primary endpoint, the vaccine has higher efficacy against severe disease, but severe disease is assessed less vigorously later in the study. Again, vaccine efficacy could appear to wane.

The only additional assumption that we require in order to recover VE under placebo crossover is that the effect of vaccination is the same no matter when the vaccine is given. Interestingly, placebo crossover might ameliorate biases arising from violations to the assumptions listed. For example, if the circulating virus mutates following crossover and VE has different, but constant, efficacy for the two strains, we would deduce a constant VE with crossover but not with a standard trial. In the absence of crossover, we would have difficulty differentiating between a sieve effect, i.e., differential efficacy against the two strains, from equal, but waning, efficacy against both strains.

4 Simulated COVID-19 Trials

In this section, we explore how placebo crossover, the dynamics of durability, and the baseline hazard affect our estimates of vaccine efficacy and durability. Since several COVID-19 vaccine trials are powered to accrue 150 cases and follow all volunteers for 2 years, we evaluate 5 different designs: i) Quit at 150 cases, ii) Quit at 1 year, iii) crossover at 150 cases, iv) crossover at 1 year and v) a standard parallel arm trial (Moderna, 2020). We consider two settings for vaccine dynamics: constant VE of 75%, and VE waning linearly on the log-hazard scale from 85% to 35% over 1.5 years. In the crossover scenarios, placebos were crossed over during a four
week period. For each of the 10 settings we simulated 10,000 trials. Each trial enrolled 3,000 participants in a 1:1 randomization with linear accrual of participants over an initial three month period and followed participants for two years post-vaccination. While COVID-19 trials are larger, we chose 3,000 to lessen our computational burden and calibrated the event rates to match rates that would be typical of larger real-world trials. In practice, COVID-19 vaccine trials would also build interim monitoring into the study design, though we will ignore this detail. The baseline hazard was piecewise-constant and calibrated to yield an average of 50, 75, 50, and 25 cases per three month period in the placebo arm in the first year, and either the same or half the year one case rates in the second study year. The data were analyzed in each simulation using the log-linear decay model, (3), and the P-spline model, (7).

We can accurately estimate VE and the change in VE in all simulation settings using both the log-linear and P-spline model (Tables 1 and S1). Coverage probabilities of 95% confidence intervals were near their nominal levels or somewhat conservative. The P-spline model performs similarly to the log-linear model except for the estimates at year 2 where the variance becomes notably larger. Initiating crossover at one year resulted in an average accrual of 44% more cases prior to crossover compared with trials that initiated crossover at 150 cases. We found that this improved the precision of our estimates for all quantities of interest. One way to quantify the relative performance of placebo crossover and parallel arm trials is by the ratio of empirical variances. If this ratio is 2, then a crossover trial would need to be twice as large as a parallel trial to achieve the same precision. Consider the empirical variances of estimates obtained with the log-linear model in the constant VE setting where the baseline hazard in year two was half that in year one. The one year crossover versus parallel trial empirical variance ratios for estimates of VE at 0.5, 1.0, 1.5, and 2.0 years post-vaccination are 1.1, 2.5, 2.3, and 2.0. With crossover at 150 cases the analogous ratios are 2.3, 6.1, 5.7, and 5.0, respectively. This underscores the potential benefit of additional case accrual during the pre-crossover period leading into the second year when the baseline hazard was halved.

In Table 2, we provide estimates of the intercept and linear trend of the VE profile. Unsurprisingly, stopping a trial at either the 150 case mark or at one year resulted in greater uncertainty in the estimates compared with continued followup to two years. For the constant vaccine efficacy scenario, we see that continued followup after crossing over at one year reduces the empirical variance of the overall VE estimate from a constant VE model by 24%; 0.033 for stopping at one year versus 0.028 when followup is continued. Even after a crossover decision is made, we continue to accrue information about overall VE during the crossover interlude until all placebo participants have been vaccinated. The improvements in precision of the intercept and linear trend estimates under the log-linear model are even greater as the empirical variances are reduced by 67% and 88% with continued followup. Continued follow-up after placebo crossover thus offers substantial improvements in addressing questions about vaccine durability, safety, and efficacy in subgroups.

We next compare the crossover at one year design to a standard parallel trial using the log-linear model. This comparison is more of a benchmark as a standard trial is not ethically possible following vaccine approval. It is interesting that the intercept estimate is improved by crossover compared to a standard trial with a variance ratio of $0.77 = 0.040/0.052$. During the crossover interlude, the newly vaccinated placebo volunteers contribute information to the intercept. This does not happen in the standard parallel trial. The slope estimate is somewhat more variable at a crossover of 1 year compared to a standard trial. Crossing over at 150 cases results in less precise estimates of the intercept and, especially, of the linear trend compared to crossing over at 1 year. Thus delaying crossover has real statistical advantages. Results are
Table 1: Summary statistics for estimates of vaccine efficacy (VE) and change in VE for simulated trials where the baseline hazard in year two was half the baseline hazard in year one. The log-linear and P-spline models correspond to (3) and (7), respectively. We report the empirical variances and coverage of 95% confidence intervals for estimates of VE and decay in VE. We also report the average time of crossover (in years), \( \tau_x \), and the average number of events at crossover, \( N_x \), along with standard deviations beneath the crossover grouping in the design column. Time is given in years since study initiation.
similar for the P-spline model and for the waning efficacy scenario.

A design question is how estimation efficiency varies with the length of the crossover interlude. If crossover is instantaneous, the intercept term immediately factors out of the partial likelihood and additional information is not accrued. To explore this design question, we did additional simulations where we evaluated a standard parallel trial of 2 years, a trial where all placebo participants are crossed over at 1 year, and a trial where the times of vaccination for all volunteers were uniformly distributed over 2 years. The baseline hazard was constant over the 2 year period. Under the constant VE scenario, the empirical variances for the intercept term were 0.051, 0.035, 0.031, respectively while the variances for the slope were 0.039, 0.039, and 0.034 respectively (Table S3). This suggests a longer crossover interlude is somewhat better for estimation of the intercept and the slope.

Unobserved heterogeneity in the risk of disease can lead to bias in estimates of VE and complicate the task of separating time-varying efficacy from culling of the frailty distribution (Balan and Putter 2020). We simulated placebo crossover and parallel arm trials with 30,000 participants and gamma distributed frailties with mean one, and variance equal to either one or four. Crossover trials initiated vaccination of the placebo arm at one year. The baseline hazard was constant and calibrated to yield either 50 or 300 cases per six month period on the placebo arm, and VE was either constant or waned linearly on the log hazard scale, as before. The frailty distributions in the original placebo and vaccine arms at the end of followup were more similar in the placebo crossover trials than in the standard parallel trials (Table S7). In the low baseline hazard scenario, where the dominant contribution to a participant’s propensity for disease was their underlying frailty, placebo crossover trials yielded less biased estimates of VE relative to the standard parallel design (Tables S8 and S9). In the high baseline hazard scenario, the common baseline hazard dominated heterogeneity in the frailty distribution, and in this setting the bias in VE estimates under placebo crossover was comparable to the bias that was observed with parallel trials. Higher baseline hazards resulted in more differential culling of the risk set and increased bias in estimates of VE. In practice, we could mitigate biases resulting from heterogeneity in the frailty distribution by adjusting for known risk factors of disease and stratifying our analyses by site or geographic region.

5 Detailed Example Trials

In this section, we present a more detailed analysis of two simulated COVID-19 vaccine trials where the true VE was either constant at 75% or waned linearly on the log–hazard scale from 85% to 35% over 1.5 years. Each trial enrolled 30,000 participants with linear accrual over three months in a 1:1 randomization to vaccine or placebo. The baseline attack rate was piece–wise constant with change–points every three months, and was calibrated to yield 50, 75, 50, and 25 cases on the placebo arm in each period in the first year, and half the expected number of cases per period on the placebo arm in year two. In this example, interim analyses are planned at 150 cases, which ultimately result in crossover at the end of year one following evaluation and vetting of the efficacy by a regulatory agency. Placebo crossover occurs over a four week period. Each volunteer was followed for a total of two years.

The two simulated trials are summarized in Table 3. In the constant VE scenario, the trial reached 150 cases in 234 days, and recorded 208 events by the one year crossover time–point and 269 events overall. The case split across treatment arms declined from roughly 83% on the placebo arm at the 150 case interim look to 75.5% at the completion of the study in the constant VE scenario, and from 82% to 49% in the waning VE scenario. The overall VE estimate at
| Vaccine efficacy constant at 75% | Intercept | Linear trend |
|-------------------------------|-----------|--------------|
|                               | Emp. Var. | Covg.        | Emp. Var. | Covg.   |
| Stop at 150 cases             |           |              |           |         |
| Constant VE                   | 0.047     | 0.927        | —         | —       |
| log-linear                    | 0.185     | 0.942        | 2.534     | 0.905   |
| P-spline                      | 0.427     | 0.980        | 2.447     | 0.902   |
| Stop at 1 year                |           |              |           |         |
| Constant VE                   | 0.033     | 0.912        | —         | —       |
| log-linear                    | 0.119     | 0.920        | 0.737     | 0.836   |
| P-spline                      | 0.250     | 0.977        | 0.712     | 0.834   |
| Cross at 150 cases            |           |              |           |         |
| Constant VE                   | 0.039     | 0.952        | —         | —       |
| log-linear                    | 0.051     | 0.951        | 0.192     | 0.952   |
| P-spline                      | 0.086     | 0.973        | 0.188     | 0.956   |
| Cross at 1 year               |           |              |           |         |
| Constant VE                   | 0.028     | 0.950        | —         | —       |
| log-linear                    | 0.040     | 0.950        | 0.087     | 0.953   |
| P-spline                      | 0.104     | 0.978        | 0.084     | 0.959   |
| Parallel trial                |           |              |           |         |
| Constant VE                   | 0.018     | 0.949        | —         | —       |
| log-linear                    | 0.052     | 0.951        | 0.065     | 0.952   |
| P-spline                      | 0.126     | 0.974        | 0.063     | 0.955   |

| Vaccine efficacy wanes from 85% to 35% over 1.5 years | Intercept | Linear trend |
|-------------------------------------------------------|-----------|--------------|
|                                                       | Emp. Var. | Covg.        | Emp. Var. | Covg.   |
| Stop at 150 cases                                      |           |              |           |         |
| log-linear                                             | 0.253     | 0.934        | 2.877     | 0.874   |
| P-spline                                               | 0.634     | 0.981        | 2.752     | 0.876   |
| Stop at 1 year                                         |           |              |           |         |
| log-linear                                             | 0.164     | 0.907        | 0.858     | 0.815   |
| P-spline                                               | 0.351     | 0.980        | 0.821     | 0.817   |
| Cross at 150 cases                                     |           |              |           |         |
| log-linear                                             | 0.056     | 0.951        | 0.123     | 0.953   |
| P-spline                                               | 0.102     | 0.975        | 0.121     | 0.955   |
| Cross at 1 year                                       |           |              |           |         |
| log-linear                                             | 0.043     | 0.952        | 0.066     | 0.949   |
| P-spline                                               | 0.116     | 0.981        | 0.065     | 0.951   |
| Parallel trial                                         |           |              |           |         |
| log-linear                                             | 0.057     | 0.950        | 0.048     | 0.951   |
| P-spline                                               | 0.145     | 0.978        | 0.047     | 0.952   |

Table 2: Empirical variance and coverage for estimates of the intercept and linear trend in vaccine efficacy under the log–linear model, (3), and semi–parametric model, (7). Here, the time–varying baseline hazard in year two was half the baseline hazard in year one.
the one year crossover, estimated using a proportional hazards model without adjustment for time since vaccination, was 78% (95% CI: 68.7%, 84.5%) in the constant VE case and 76.3% (95% CI: 66.3%, 83.4%) in the waning VE scenario (the true geometric mean VE to one year post–vaccination is 75.6%).

Point estimates for immediate VE and the linear trend in log VE from the log–linear and P-spline models were close in both scenarios, although confidence intervals in the P-spline models were wider. The estimated efficacy profiles obtained with both methods were in agreement and recovered the true VE profile (Figure 3). The P-spline estimates had wider point-wise confidence intervals, but the inflation in the variance appears to be fairly modest for the period spanning the end of study enrollment through, roughly, year 1.5 post–vaccination. In practice, both the log–linear decay model and the P-spline model could be used to test a hypothesis of time–varying VE. This is straightforwardly carried out for the log–linear model via a likelihood ratio test (LRT) for the slope parameter in (3) where the test statistic is compared to a chi–square distribution with one degree of freedom. For the P-spline models, we perform a likelihood ratio test for whether all of the P-spline basis coefficients are jointly equal to zero, and compare the test statistic to a chi–square distribution with 3.1 degrees of freedom (the effective degrees of freedom for the P-splines in our models). In the waning VE scenario, we reject the null hypothesis of time-homogeneous VE, and fail to reject the null in the constant VE scenario (Table 3).

In the supplementary materials, we provide a table analogous to Table 4 but for a subgroup about 1/8 of the total cohort. The true VE for this subgroup was 0.75. We evaluated the estimate of VE(0) under the log-linear model. Using the pre-crossover data available at year 1 provides an estimated VE(0) of 0.84 with a 95% CI (0.16, 0.97) at 1 year. Following crossover at 1 year and using the full follow-up, the estimate improved to 0.74 with a slightly narrower confidence interval of (0.19,0.91).

These simulated examples show that for both the waning and constant VE scenarios, accurate inference about the behavior of the vaccine efficacy over time can be recovered as well as somewhat improved estimation in subgroups. Furthermore, these examples demonstrate how quitting the study at 1 year results in decidedly inferior estimates.

6 Discussion

Assessment of durability of vaccine induced protection is a key question in vaccine development, especially for COVID-19 vaccines. Placebo volunteers will be offered vaccine before long term follow-up has completed. In this paper we demonstrated that durability can be accurately assessed following crossover using a Cox proportional hazards model. To reflect seasonal variation in the attack rate, we use calendar time as the time index and specified flexible models for vaccine efficacy over time. Our results point out the importance of a longer crossover period and a longer crossover interlude to help improve the estimation. If crossover occurs quickly, the early VE will remain poorly estimated, no matter how many post-crossover cases occur. We show that continued crossover can help improve estimates of constant VE compared to terminating the study. Thus, VE in subgroups or for rare events can be somewhat improved following crossover. Compared with the standard trial, the estimate of VE post-crossover is less reliable, though the disadvantage remains roughly constant post-crossover. Given that a standard parallel trial is not ethically possible, such comparisons provide a kind of benchmark. However, the real question is whether crossover offers advantages compared to quitting. Here
| Time of 150 case interim look | True VE Constant at 75% Day 233 | True VE Wanes from 85% to 35% Day 265 |
|------------------------------|----------------------------------|--------------------------------------|
| **Case split by original arm** |                                  |                                      |
| **at interim look**          | Placebo = 124, Vaccine = 26     | Placebo = 123, Vaccine = 27          |
| **at 1 year crossover**      | Placebo = 175, Vaccine = 39     | Placebo = 163, Vaccine = 38          |
| **at 2 year follow-up**      | Placebo = 208, Vaccine = 68     | Placebo = 190, Vaccine = 105         |
| **Estimates at interim look**|                                  |                                      |
| *log-linear model*           |                                  |                                      |
| Intercept                    | -0.84 (95% CI: -1.6, -0.09)     | -2.16 (95% CI: -3.17, -1.16)         |
| Linear trend                 | -3.06 (95% CI: -6.05, -0.07)    | 0.81 (95% CI: -2.13, 3.75)           |
| LRT for time-varying VE      | 0.039                            | 0.589                                |
| *P-spline model*             |                                  |                                      |
| Intercept                    | -1.41 (95% CI: -2.77, -0.05)    | -2.43 (95% CI: -4.23, -0.62)         |
| Linear trend                 | -3.02 (95% CI: -6.36, 0.32)     | 0.8 (95% CI: -2.13, 3.73)            |
| LRT for time-varying VE      | 0.037                            | 0.605                                |
| **Estimates at 1 year crossover** |                                |                                      |
| *log-linear model*           |                                  |                                      |
| Intercept                    | -1.34 (95% CI: -1.98, -0.7)     | -2.36 (95% CI: -3.17, -1.55)         |
| Linear trend                 | -0.29 (95% CI: -1.74, 1.17)     | 1.8 (95% CI: 0.2, 3.4)               |
| LRT for time-varying VE      | 0.698                            | 0.027                                |
| *P-spline model*             |                                  |                                      |
| Intercept                    | -1.14 (95% CI: -2.17, -0.1)     | -2.26 (95% CI: -3.68, -0.83)         |
| Linear trend                 | -0.28 (95% CI: -1.66, 1.1)      | 1.8 (95% CI: 0.22, 3.37)             |
| LRT for time-varying VE      | 0.133                            | 0.054                                |
| **Estimates at 2 year follow-up** |                                |                                      |
| *log-linear model*           |                                  |                                      |
| Intercept                    | -1.37 (95% CI: -1.77, -0.97)    | -2.19 (95% CI: -2.62, -1.75)         |
| Linear trend                 | -0.13 (95% CI: -0.7, 0.43)      | 1.33 (95% CI: 0.82, 1.83)            |
| LRT for time-varying VE      | 0.641                            | 0.001                                |
| *P-spline model*             |                                  |                                      |
| Intercept                    | -1.33 (95% CI: -2.09, -0.58)    | -2.26 (95% CI: -3.15, -1.36)         |
| Linear trend                 | -0.13 (95% CI: -0.7, 0.44)      | 1.28 (95% CI: 0.77, 1.8)             |
| LRT for time-varying VE      | 0.178                            | 0.001                                |

Table 3: Summary of example trials simulated under constant and waning vaccine efficacy (VE) at times of interim analysis and placebo crossover. The intercept and linear trend correspond to the immediate effect of vaccination and the time–trend for VE(t) under model (3), and the true values were set to $\theta_1 = -1.39$ and $\theta_2 = 0$ in the constant VE scenario, and $\theta_1 = -1.9$ and $\theta_2 = 0.98$ in the waning VE setting. The likelihood ratio test (LRT) for waning VE compares models (3) and (7) to a PH model without adjustment for time since vaccination.
Figure 3: (Top) Number of events per quarter by treatment arm. The delayed vaccination arm consists of the original placebo participants after they have been crossed over. (Bottom) Vaccine efficacy (VE) as a function of time since vaccination. Dashed lines are the true VE(t), solid curves and ribbons are pointwise means and 95% confidence intervals.
the statistical advantages are enormous. Therefore, we strongly recommend that followup of study participants continue following vaccination of placebo volunteers.

In this work we implicitly assumed that there was no differential dropout or changes in behavior during the trial. Such an assumption is plausible for a trial with blinded crossover. If volunteers are unblinded and only placebo volunteers receive a crossover vaccination, the risk for bias due to differential behavior is substantial. Unblinded volunteers, who had originally received vaccine, will know they are protected and thus might engage in risky behavior or depart the study. Such differential behavior can, in principle, be addressed via regression adjustment or inverse probability weighting, but the conclusions from these analyses will ultimately depend on untestable assumptions. Another possibility is to censor all volunteers for a period of time around the crossover interlude. Such methods are beyond the scope of this paper. Our work applies to the setting where endpoints are assessed continuously. Certain endpoints, such as seroconversion, are assessed infrequently. While the arguments of this paper can, in principle, be applied to periodically assessed endpoints, the development is more difficult as it involves methods for interval censored data. Future work will investigate this setting. A related line of inquiry is how to combine data on seroconversion, and disease endpoints in order to improve estimation of VE for infection.

In this manuscript, we restricted ourselves to models and software with which it would be reasonable to assume that practicing statisticians are familiar. However, a more bespoke modeling approach might be beneficial in certain aspects, or even necessary. Certainly, a more complex modeling framework would be required for analyzing infrequently assessed or multivariate endpoints. Still, even within the context of the models in this manuscript there is room for more sophisticated modeling tools that better reflect the dynamics of the underlying biological processes. For instance, we used a P-spline to model the VE profile, but this allows for VE to be non-monotonic. We could enforce monotonicity in the VE profile via shape constrained splines (Pya and Wood, 2015). We also sidestepped the issue of covariate adjustment, although incorporating covariate data can mitigate bias arising from model misspecification relating to frailties, and can greatly improve precision and power in vaccine trials (Benkeser et al., 2020).

Acknowledgments

This work utilized the computational resources of the NIH HPC Biowulf computing cluster (http://hpc.nih.gov). The authors would like to thank Keith Lumbard for help with simulations, as well as Michael Fay and Anastasios Tsiatis for helpful discussions regarding this work.

Supplementary Materials

Code demonstrating how to reproduce the results in this manuscript is made available at the following GitHub repository: https://github.com/fintzij/ve_placebo_crossover.

References

Odd O Aalen, Richard J Cook, and Kjetil Røysland. Does Cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Analysis, 21:579–593, 2015.
Theodor A Balan and Hein Putter. A tutorial on frailty models. *Statistical Methods in Medical Research*, 29:3424–3454, 2020.

David Benkeser, Ivan Diaz, Alex Luedtke, Jodi Segal, Daniel Scharfstein, and Michael Rosenblum. Improving precision and power in randomized trials for covid-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *medRxiv*, 2020.

Sam Brilleman. *simsurv: Simulate Survival Data*, 2019. URL [https://CRAN.R-project.org/package=simsurv](https://CRAN.R-project.org/package=simsurv). R package version 0.2.3.

Pyoeng Gyun Choe, Chang Kyung Kang, Hyeon Jeong Suh, Jongtak Jung, Kyoung-Ho Song, Ji Hwan Bang, Eu Suk Kim, Hong Bin Kim, Sang Won Park, Nam Joong Kim, et al. Waning antibody responses in asymptomatic and symptomatic sars-cov-2 infection. *Emerging Infectious Diseases*, 27:327–329, 2020.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Michael J Crowther and Paul C Lambert. Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32:4118–4134, 2013.

L Kathryn Durham, Ira M Longini Jr, M Elizabeth Halloran, John D Clemens, Nizam Azhar, and Malla Rao. Estimation of vaccine efficacy in the presence of waning: application to cholera vaccines. *American Journal of Epidemiology*, 147(10):948–959, 1998.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, pages 89–102, 1996.

DA Follmann, J Fintzi, et al. Assessing durability of vaccine effect following blinded crossover in COVID-19 vaccine efficacy trials. *medRxiv*, 2020.

Dean Follmann. Augmented designs to assess immune response in vaccine trials. *Biometrics*, 62(4):1161–1169, 2006.

Ana Maria Henao-Restrepo, Anton Camacho, Ira M Longini, Conall H Watson, W John Edmunds, Matthias Egger, Miles W Carroll, Natalie E Dean, Ibrahima Diatta, Moussa Doumbia, et al. Efficacy and effectiveness of an rsv-vectored vaccine in preventing ebola virus disease: final results from the guinea ring vaccination, open-label, cluster-randomised trial (ebola ça suffit!). *The Lancet*, 389(10068):505–518, 2017.

Mona N Kanaan and C Paddy Farrington. Estimation of waning vaccine efficacy. *Journal of the American Statistical Association*, 97(458):389–397, 2002.

Marc Lipsitch. Challenges of vaccine effectiveness and waning studies, 2019.

Boikanyo Makubate and Stephen Senn. Planning and analysis of cross-over trials in infertility. *Statistics in Medicine*, 29:3203–3210, 2010.

Moderna. A phase 3, randomized, stratified, observer-blind, placebo-controlled study to evaluate the efficacy, safety, and immunogenicity of mrna-1273 sars-cov-2 vaccine in adults aged 18 years and older. [https://www.modernatx.com/sites/default/files/mRNA-1273-P301-Protocol.pdf. Access date Nov 17, 2020.](https://www.modernatx.com/sites/default/files/mRNA-1273-P301-Protocol.pdf)
Martha Nason and Dean Follmann. Design and analysis of crossover trials for absorbing binary endpoints. *Biometrics*, 66(3):958–965, 2010.

Aris Perperoglou, Willi Sauerbrei, Michal Abrahamowicz, and Matthias Schmid. A review of spline function procedures in r. *BMC Medical Research Methodology*, 19, 2019.

Gregory A Poland, Inna G Ovsyannikova, and Richard B Kennedy. SARS-CoV-2 immunity: Review and applications to phase 3 vaccine candidates. *The Lancet*, 2020.

Natalya Pya and Simon N Wood. Shape constrained additive models. *Statistics and Computing*, 25(3):543–559, 2015.

Saranya Sridhar, Alexander Luedtke, Edith Langevin, Ming Zhu, Matthew Bonaparte, Tiffany Machabert, Stephen Savarino, Betzana Zambrano, Annick Moureau, Alena Khromava, et al. Effect of dengue serostatus on dengue vaccine safety and efficacy. *New England Journal of Medicine*, 379(4):327–340, 2018.

Terry Therneau, Cindy Crowson, and Elizabeth Atkinson. Using time dependent covariates and time dependent coefficients in the cox model. *Survival Vignettes*, 2017.

Terry M Therneau and Patricia M Grambsch. *Modeling survival data: Extending the Cox model*. Springer Science & Business Media, 2013.

David Wendler, Jorge Ochoa, Joseph Millum, Christine Grady, and Holly A Taylor. COVID-19 vaccine trial ethics once we have efficacious vaccines. *Science*, 2020.

Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

World Health Organization. Placebo-controlled trials of covid-19 vaccines—why we still need them, 2020.

## A Web Appendix A: Additional Simulation Results

### A.1 Trials with Year Two Baseline Hazard Equal to Year One Baseline Hazard
Table S1: Summary statistics for estimates of vaccine efficacy (VE) and change in VE for simulated trials where the baseline hazard in year two was the same as the baseline hazard in year one. The log-linear and P-spline models correspond to (3) and (7), respectively. We report the empirical variances and coverage of 95% confidence intervals for estimates of VE and decay in VE. We also report the average time of crossover (in years), $\tau_x$, and the average number of events at crossover, $N_x$, along with standard deviations beneath the crossover grouping in the design column. Time is given in years since study initiation.
| Vaccine efficacy constant at 75% | Intercept | Linear trend |
|---------------------------------|-----------|--------------|
|                                 | Emp. Var. | Covg.        | Emp. Var. | Covg.      |
| Stop at 150 cases               | Constant VE | 0.047 | 0.927 | — | — |
|                                 | log-linear | 0.186 | 0.943 | 2.542 | 0.902 |
|                                 | P-spline   | 0.426 | 0.980 | 2.455 | 0.900 |
| Stop at 1 year                  | Constant VE | 0.032 | 0.926 | — | — |
|                                 | log-linear | 0.115 | 0.937 | 0.703 | 0.881 |
|                                 | P-spline   | 0.260 | 0.977 | 0.683 | 0.880 |
| Cross at 150 cases              | Constant VE | 0.039 | 0.951 | — | — |
|                                 | log-linear | 0.046 | 0.953 | 0.111 | 0.952 |
|                                 | P-spline   | 0.076 | 0.970 | 0.109 | 0.954 |
| Cross at 1 year                 | Constant VE | 0.027 | 0.949 | — | — |
|                                 | log-linear | 0.033 | 0.952 | 0.042 | 0.953 |
|                                 | P-spline   | 0.079 | 0.977 | 0.042 | 0.956 |
| Parallel trial                  | Constant VE | 0.013 | 0.951 | — | — |
|                                 | log-linear | 0.046 | 0.950 | 0.040 | 0.950 |
|                                 | P-spline   | 0.110 | 0.977 | 0.039 | 0.952 |

| Vaccine efficacy wanes from 85% to 35% over 1.5 years |
|-------------------------------------------------------|
| Stop at 150 cases | log-linear | 0.255 | 0.932 | 2.898 | 0.874 |
|                                 | P-spline   | 0.631 | 0.980 | 2.768 | 0.875 |
| Stop at 1 year     | log-linear | 0.155 | 0.929 | 0.803 | 0.867 |
|                                 | P-spline   | 0.362 | 0.981 | 0.775 | 0.869 |
| Cross at 150 cases | log-linear | 0.050 | 0.951 | 0.064 | 0.952 |
|                                 | P-spline   | 0.086 | 0.975 | 0.063 | 0.953 |
| Cross at 1 year     | log-linear | 0.036 | 0.952 | 0.031 | 0.953 |
|                                 | P-spline   | 0.085 | 0.979 | 0.031 | 0.955 |
| Parallel trial     | log-linear | 0.049 | 0.950 | 0.031 | 0.948 |
|                                 | P-spline   | 0.124 | 0.980 | 0.031 | 0.949 |

Table S2: Empirical variance and coverage for estimates of the intercept and linear trend in vaccine efficacy under the log–linear model, (3), and semi–parametric model, (7). Here, the time–varying in year two was the same as the baseline hazard in year one.
A.2 Comparing Uniform Crossover, Crossover at One Year, and Parallel Trials

Results for a set of idealized trials with constant baseline hazard, instantaneous enrollment and crossover, and constant VE. Trials either crossed placebo participants to the vaccine arm at one year, uniformly over the two year study period, or never (standard parallel arm design).

|                                | Intercept       | Linear trend    |
|--------------------------------|-----------------|-----------------|
|                                | Emp. Var.  Covg.| Emp. Var.  Covg.|
| **Continuous crossover**       |                 |                 |
| log-linear                     | 0.031 0.950     | 0.034 0.950     |
| P-spline                       | 0.071 0.973     | 0.034 0.952     |
| **Crossover at one year**      |                 |                 |
| log-linear                     | 0.035 0.953     | 0.039 0.947     |
| P-spline                       | 0.099 0.978     | 0.038 0.952     |
| **Parallel trial**             |                 |                 |
| log-linear                     | 0.051 0.950     | 0.039 0.951     |
| P-spline                       | 0.130 0.977     | 0.039 0.952     |

Table S3: Empirical variance and coverage for estimates of the intercept and linear trend in vaccine efficacy under the log–linear model, (3), and semi–parametric model, (7), in an idealized scenario with constant baseline hazards and instantaneous crossover.
| Model                  | Time | Empir. Var. | Coverage | Empir. Var. | Coverage |
|------------------------|------|-------------|----------|-------------|----------|
| **Continuous uniform crossover** |      |             |          |             |          |
| log-linear             | 0.0  | 0.031       | 0.950    | --          | --       |
|                        | 0.5  | 0.016       | 0.951    | 0.009       | 0.950    |
|                        | 1.0  | 0.018       | 0.949    | 0.034       | 0.950    |
|                        | 1.5  | 0.037       | 0.951    | 0.077       | 0.950    |
|                        | 2.0  | 0.074       | 0.950    | 0.137       | 0.950    |
| P-spline               | 0.0  | 0.072       | 0.979    | --          | --       |
|                        | 0.5  | 0.023       | 0.981    | 0.082       | 0.983    |
|                        | 1.0  | 0.029       | 0.976    | 0.104       | 0.967    |
|                        | 1.5  | 0.044       | 0.973    | 0.103       | 0.970    |
|                        | 2.0  | 0.183       | 0.976    | 0.234       | 0.980    |
| **Crossover at one year** |      |             |          |             |          |
| log-linear             | 0.0  | 0.035       | 0.953    | --          | --       |
|                        | 0.5  | 0.026       | 0.954    | 0.010       | 0.947    |
|                        | 1.0  | 0.037       | 0.947    | 0.039       | 0.947    |
|                        | 1.5  | 0.068       | 0.947    | 0.088       | 0.947    |
|                        | 2.0  | 0.117       | 0.946    | 0.156       | 0.947    |
| P-spline               | 0.0  | 0.100       | 0.984    | --          | --       |
|                        | 0.5  | 0.032       | 0.982    | 0.112       | 0.986    |
|                        | 1.0  | 0.048       | 0.975    | 0.122       | 0.971    |
|                        | 1.5  | 0.078       | 0.970    | 0.130       | 0.981    |
|                        | 2.0  | 0.237       | 0.980    | 0.333       | 0.975    |
| **Parallel trial**     |      |             |          |             |          |
| log-linear             | 0.0  | 0.051       | 0.950    | --          | --       |
|                        | 0.5  | 0.022       | 0.950    | 0.010       | 0.951    |
|                        | 1.0  | 0.013       | 0.949    | 0.039       | 0.951    |
|                        | 1.5  | 0.024       | 0.951    | 0.088       | 0.951    |
|                        | 2.0  | 0.054       | 0.953    | 0.157       | 0.951    |
| P-spline               | 0.0  | 0.131       | 0.981    | --          | --       |
|                        | 0.5  | 0.032       | 0.982    | 0.140       | 0.986    |
|                        | 1.0  | 0.033       | 0.981    | 0.186       | 0.968    |
|                        | 1.5  | 0.034       | 0.983    | 0.165       | 0.973    |
|                        | 2.0  | 0.133       | 0.982    | 0.261       | 0.977    |

Table S4: Summary statistics for estimates of vaccine efficacy (VE) and change in VE for simulated trials in an idealized scenario with constant baseline hazards and instantaneous crossover. The log-linear and P-spline models correspond to (3) and (7), respectively.
A.3 Varying the Crossover Interlude Duration

Here, we compare placebo crossover trials with long (two months) and short (two week) crossover interlude durations and an average crossover time of one year. Simulations otherwise follow the baseline set of simulations — piecewise constant baseline hazards varying seasonally with the event rate in year two halved, staggered entry, and either waning or constant VE.

|                             | Intercept | Linear Trend |
|-----------------------------|-----------|--------------|
|                             | Empir. Var. | Covg. | Empir. Var. | Covg. |
| Vaccine efficacy constant at 75% |           |     |           |     |
| Two week crossover         |           |     |           |     |
| log-linear                  | 0.041     | 0.949 | 0.091     | 0.954 |
| P-spline                    | 0.105     | 0.975 | 0.088     | 0.959 |
| Two month crossover        |           |     |           |     |
| log-linear                  | 0.041     | 0.945 | 0.092     | 0.952 |
| P-spline                    | 0.103     | 0.975 | 0.089     | 0.957 |

Vaccine efficacy wanes from 85% to 35% over 1.5 years

|                             | Intercept | Linear Trend |
|-----------------------------|-----------|--------------|
|                             | Empir. Var. | Covg. | Empir. Var. | Covg. |
| Two week crossover         |           |     |           |     |
| log-linear                  | 0.044     | 0.950 | 0.069     | 0.949 |
| P-spline                    | 0.118     | 0.979 | 0.067     | 0.953 |
| Two month crossover        |           |     |           |     |
| log-linear                  | 0.044     | 0.949 | 0.069     | 0.951 |
| P-spline                    | 0.117     | 0.977 | 0.068     | 0.953 |

Table S5: Empirical variance and coverage for estimates of the intercept and linear trend in vaccine efficacy under the log-linear model, (3), and semi-parametric model, (7). Trials were simulated such that participants were followed for up to two years, and crossover times were uniformly distributed in either a two week or two month interval centered at one year. the baseline hazard in year 2 was half the baseline hazard in year one.
\[
\log(VE(t)) - \log(VE(0))
\]

| Design                      | Model | Time | Emp. Var. | Covg. | Emp. Var. | Covg. |
|-----------------------------|-------|------|-----------|-------|-----------|-------|
| Vaccine efficacy constant at 75% |       |      |           |       |           |       |
| Two week crossover         | log-linear | 0.0  | 0.041     | 0.949 | —         | —     |
|                             |       | 0.5  | 0.030     | 0.950 | 0.023     | 0.954 |
|                             |       | 1.0  | 0.065     | 0.950 | 0.091     | 0.954 |
|                             |       | 1.5  | 0.146     | 0.953 | 0.205     | 0.954 |
|                             |       | 2.0  | 0.272     | 0.954 | 0.364     | 0.954 |
| P-spline                    |       | 0.0  | 0.106     | 0.983 | —         | —     |
|                             |       | 0.5  | 0.042     | 0.980 | 0.147     | 0.979 |
|                             |       | 1.0  | 0.089     | 0.975 | 0.184     | 0.968 |
|                             |       | 1.5  | 0.168     | 0.968 | 0.250     | 0.974 |
|                             |       | 2.0  | 0.681     | 0.985 | 0.794     | 0.983 |
| Two month crossover         | log-linear | 0.0  | 0.041     | 0.945 | —         | —     |
|                             |       | 0.5  | 0.030     | 0.950 | 0.023     | 0.952 |
|                             |       | 1.0  | 0.065     | 0.950 | 0.092     | 0.952 |
|                             |       | 1.5  | 0.145     | 0.952 | 0.207     | 0.952 |
|                             |       | 2.0  | 0.272     | 0.955 | 0.368     | 0.952 |
| P-spline                    |       | 0.0  | 0.105     | 0.984 | —         | —     |
|                             |       | 0.5  | 0.041     | 0.980 | 0.145     | 0.977 |
|                             |       | 1.0  | 0.089     | 0.974 | 0.185     | 0.968 |
|                             |       | 1.5  | 0.166     | 0.968 | 0.251     | 0.970 |
|                             |       | 2.0  | 0.679     | 0.984 | 0.794     | 0.981 |
| Vaccine efficacy wanes from 85% to 35% over 1.5 years |       |      |           |       |           |       |
| Two week crossover         | log-linear | 0.0  | 0.044     | 0.950 | —         | —     |
|                             |       | 0.5  | 0.033     | 0.948 | 0.017     | 0.949 |
|                             |       | 1.0  | 0.057     | 0.950 | 0.069     | 0.949 |
|                             |       | 1.5  | 0.115     | 0.951 | 0.155     | 0.949 |
|                             |       | 2.0  | 0.207     | 0.949 | 0.276     | 0.949 |
| P-spline                    |       | 0.0  | 0.119     | 0.984 | —         | —     |
|                             |       | 0.5  | 0.041     | 0.975 | 0.143     | 0.983 |
|                             |       | 1.0  | 0.078     | 0.970 | 0.189     | 0.971 |
|                             |       | 1.5  | 0.126     | 0.964 | 0.217     | 0.972 |
|                             |       | 2.0  | 0.358     | 0.976 | 0.476     | 0.972 |
| Two month crossover         | log-linear | 0.0  | 0.044     | 0.949 | —         | —     |
|                             |       | 0.5  | 0.032     | 0.949 | 0.017     | 0.951 |
|                             |       | 1.0  | 0.055     | 0.951 | 0.069     | 0.951 |
|                             |       | 1.5  | 0.113     | 0.950 | 0.156     | 0.951 |
|                             |       | 2.0  | 0.205     | 0.951 | 0.277     | 0.951 |
| P-spline                    |       | 0.0  | 0.118     | 0.983 | —         | —     |
|                             |       | 0.5  | 0.040     | 0.974 | 0.141     | 0.983 |
|                             |       | 1.0  | 0.077     | 0.969 | 0.190     | 0.969 |
|                             |       | 1.5  | 0.124     | 0.963 | 0.218     | 0.972 |
|                             |       | 2.0  | 0.354     | 0.975 | 0.470     | 0.974 |

Table S6: Empirical variance and coverage for estimates of vaccine efficacy and the decay in efficacy under the log–linear model, (3), and semi–parametric model, (7). Trials were simulated such that participants were followed for up to two years, and crossover times were uniformly distributed in either a two week or two month interval centered at one year. the baseline hazard in year 2 was half the baseline hazard in year one.
A.4 Frailty simulation results

This section presents results from simulated trials in where participants were heterogeneous in their baseline hazards. Simulation was analogous to trials simulated elsewhere in this manuscript, except that each trial consisted of 30,000 participants and the participant level hazard was \( \tilde{h}_i(t) = U_i h_i(t) \), with \( h_i(t) \) corresponding to either a constant VE or log-linear VE model. In both cases the baseline hazard was constant. We considered two settings for the baseline hazard: a low event rate scenario calibrated to yield 100 cases per year on placebo or a high event rate scenario calibrated to yield 600 cases per year on placebo. Participant frailties were drawn from a gamma distribution with mean one and a variance of either one (low frailty variance scenario) or four (high frailty variance scenario).

| Baseline Hazard | Frailty Variance Design | Original arm | Mean | SD | 25%ile | 50%ile | 75%ile |
|-----------------|-------------------------|--------------|------|----|--------|--------|--------|
| VE constant at 75% |                         |              |      |    |        |        |        |
| Low             | Low Cross at 1 year Placebo | 0.992 | 0.992 | 0.285 | 0.688 | 1.375 |
|                 |                          | Vaccine     | 0.997 | 0.996 | 0.287 | 0.691 | 1.382 |
|                 |                         Parallel trial Placebo | 0.987 | 0.987 | 0.284 | 0.684 | 1.368 |
|                 |                          Vaccine | 0.997 | 0.996 | 0.287 | 0.691 | 1.382 |
| High            | Cross at 1 year Placebo | 0.969 | 1.937 | 0.010 | 0.169 | 1.010 |
|                 |                          Vaccine | 0.987 | 1.973 | 0.010 | 0.172 | 1.028 |
|                 | Parallel trial Placebo | 0.949 | 1.897 | 0.010 | 0.166 | 0.989 |
|                 |                          Vaccine | 0.987 | 1.973 | 0.010 | 0.172 | 1.028 |
| VE wanes from 85% to 35% over 1.5 years |                         |              |      |    |        |        |        |
| Low             | Low Cross at 1 year Placebo | 0.955 | 0.955 | 0.275 | 0.662 | 1.323 |
|                 |                          Vaccine | 0.980 | 0.980 | 0.282 | 0.680 | 1.359 |
|                 |                         Parallel trial Placebo | 0.926 | 0.926 | 0.266 | 0.642 | 1.283 |
|                 |                          Vaccine | 0.980 | 0.980 | 0.282 | 0.680 | 1.359 |
| High            | Cross at 1 year Placebo | 0.841 | 1.681 | 0.009 | 0.147 | 0.876 |
|                 |                          Vaccine | 0.926 | 1.851 | 0.010 | 0.162 | 0.965 |
|                 | Parallel trial Placebo | 0.757 | 1.514 | 0.008 | 0.132 | 0.790 |
|                 |                          Vaccine | 0.926 | 1.851 | 0.010 | 0.162 | 0.965 |

Table S7: Summary statistics of the frailty distribution of participants still in the risk set at the end of two years of followup. We report geometric means of summary statistics of each frailty distribution from 10,000 simulated trials.
Table S8: Bias of estimates of VE and the decay in VE for trials simulated with constant VE at 75% and gamma distributed frailties. The low baseline hazard scenario was calibrated to yield an average of 50 cases per six month period on the placebo arm, while the high baseline hazard scenario was calibrated to yield 300 cases per six month period. The frailty distribution had mean one and a variance of either one (low variance) or four (high variance).
Table S9: Bias of estimates of VE and the decay in VE for trials simulated with VE waning from 85% to 35% linear on the log hazard scale and gamma distributed frailties. The low baseline hazard scenario was calibrated to yield an average of 50 cases per six month period on the placebo arm, while the high baseline hazard scenario was calibrated to yield 300 cases per six month period. The frailty distribution had mean one and a variance of either one (low variance) or four (high variance).
A.5  Results for Severe Disease in Example Trials

| Time of 150 case interim look | True VE Constant at 75% | True VE Wanes from 85% to 35% |
|-------------------------------|-------------------------|-------------------------------|
| Case split by original arm    | Day 233                 | Day 265                       |
| at interim look               | Placebo = 15, Vaccine = 2 | Placebo = 18, Vaccine = 2     |
| at 1 year crossover          | Placebo = 22, Vaccine = 7 | Placebo = 21, Vaccine = 5     |
| at 2 year follow-up           | Placebo = 26, Vaccine = 10 | Placebo = 23, Vaccine = 14    |
| Estimates at interim look     |                         |                               |
| log-linear model              |                         |                               |
| Intercept                     | -1.55 (95% CI: -3.48, 0.38) | -3.2 (95% CI: -5.54, -0.86)   |
| Linear trend                  | 1.57 (95% CI: -2.94, 6.08)  | 6.11 (95% CI: 1.49, 10.73)    |
| LRT for time-varying VE       | 0.487                   | 0.004                         |
| P-spline model                |                         |                               |
| Intercept                     | -0.61 (95% CI: -3.3, 2.08)   | -12.35 (95% CI: -38.95, 14.26) |
| Linear trend                  | 1.35 (95% CI: -2.69, 5.38)   | 2.64 (95% CI: -4.15, 9.44)    |
| LRT for time-varying VE       | 0.270                   |                               |
| Estimates at 1 year crossover |                         |                               |
| log-linear model              |                         |                               |
| Intercept                     | -1.85 (95% CI: -3.53, -0.17) | -3.47 (95% CI: -5.58, -1.37)  |
| Linear trend                  | 1.68 (95% CI: -0.97, 4.33)    | 5.04 (95% CI: 2.05, 8.03)     |
| LRT for time-varying VE       | 0.202                   | 0.001                         |
| P-spline model                |                         |                               |
| Intercept                     | -0.6 (95% CI: -2.98, 1.78)   | -6.14 (95% CI: -14.85, 2.58)  |
| Linear trend                  | 1.42 (95% CI: -1.16, 4)      | 4.87 (95% CI: 1.21, 8.54)     |
| LRT for time-varying VE       | 0.024                   | 0.001                         |
| Estimates at 2 year follow-up |                         |                               |
| log-linear model              |                         |                               |
| Intercept                     | -1.34 (95% CI: -2.48, -0.21) | -2.53 (95% CI: -3.93, -1.13)  |
| Linear trend                  | 0.45 (95% CI: -1.25, 2.16)    | 2.58 (95% CI: 0.56, 4.6)      |
| LRT for time-varying VE       | 0.600                   | 0.002                         |
| P-spline model                |                         |                               |
| Intercept                     | -1.55 (95% CI: -3.69, 0.59)   | -4.74 (95% CI: -9.07, -0.4)   |
| Linear trend                  | 0.45 (95% CI: -1.26, 2.16)    | 2.21 (95% CI: 0.37, 4.06)     |
| LRT for time-varying VE       | 0.167                   | 0.001                         |

Table S10: Summary of severe case subgroup analyses for example trials simulated under constant and waning vaccine efficacy (VE) at times of interim analysis and placebo crossover. Participants with disease independently developed severe disease with probability 1/8. The intercept and linear trend correspond to the immediate effect of vaccination and the time–trend for VE(t) under model (3), and the true values were set to $\theta_1 = -1.39$ and $\theta_2 = 0$ in the constant VE scenario, and $\theta_1 = -1.9$ and $\theta_2 = 0.98$ in the waning VE setting. The likelihood ratio test (LRT) for waning VE compares models (3) and (7) to a PH model without adjustment for time since vaccination.