A multiple imputation joint model to estimate COVID-19 correlates of protection, antibody decay and vaccine efficacy waning

Daniel J. Phillips

St Hugh's College

University of Oxford

Department of Statistics

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas Term, 2025

Acknowledgements

Firstly, I would like to thank my supervisors, David and Maria. I am so grateful for all your support, guidance and suggestions. Without your help, this thesis would not have been possible.

To my wonderful parents and family, thank you for your love, and for showing me the right way to live.

Thanks to the many friends who have made my time as a DPhil student a joy. Special thanks to Zhixiao and Bastian for your support through the successes and disappointments of research.

I would like to thank Josiah and Susanna for their help with proof reading.

To my colleagues at the Oxford Vaccine Group, thank you for all your hard work developing and researching the COVID-19 vaccine, and generating the data I have analysed in this thesis. To Merryn, Elaine and Andy in particular, thank you for the opportunity to be involved in the correlates of protection work, and to develop it further. And thank you for your helpful comments and input into Chapter 4.

Thanks to Geoff Nicholls, for a helpful discussion suggesting the use of Bayesian cut posteriors, which led us to developing the approximate Bayesian pooling method for multiple imputation, and overcoming the convergence issues in our Bayesian joint model.

I am so grateful to my wonderful wife Susanna, for all your love and support in these past few years. You have brought me so much joy and comfort through this time, and I love you immensely.

Above all, I thank God, who has shown such kindness to me, and who I love with all my heart.

Abstract

Estimating correlates of protection – the relationship between an immune marker and risk of infection – is an key task of vaccine research. Standard models to estimate a correlate of protection do not account for the decay of antibody levels over time. We demonstrate that when differing background infection rates cause the timing of cases to vary between trials, the estimated correlate of protection curve will not be consistent. This inconsistency is caused by antibody decay between immune marker measurement and infection. Therefore, methods are required which estimate the relationship between antibody levels at the time of exposure and subsequent protection, sometimes known as "exposure-proximal" correlates of protection. We develop a joint model of waning antibody levels and subsequent COVID-19 infection. We apply the model to data from a COVID-19 vaccine trial, to estimate correlates of protection at exposure, antibody decay, and waning vaccine efficacy. The model uses a two-stage multiple imputation approach, and pools the results via an approximate Bayesian pooling scheme. We extend our multiple imputation pooling method to other scenarios, and explore how our approach compares to Rubin's pooling rules mathematically. We compare Rubin's rules and Bayesian inference with our approach in a simulation study, and in real data examples. The proposed approximate Bayesian pooling method is able to account for skewness in the missing-data posterior, whereas Rubin's rules assumes the distribution is symmetric.

Contents

Chapter 1

Introduction

Establishing a correlate of protection, that is, an immune marker associated with protection against a disease outcome, is an important part of vaccine research. This may allow new vaccines to be licensed based on immunological data alone in scenarios where large efficacy trials are not feasible, since vaccine efficacy may be predicted from immune marker values alone. Modelling correlates of protection may further inform biological understanding of immune protection mechanisms, and allow individual dynamic predictions of risk to be made. Where correlates of protection are estimated from vaccine efficacy trials, biomarker values are commonly sampled around the time of peak response, and compared to the disease outcome in a regression model. However, this common approach may be biased, as it does not account for measurement error in the biomarker assays [1, 2]. It may also give results which do not generalise well. In a pandemic scenario, vaccine trials will be run in multiple countries, and begin at different times, meaning infection waves will hit the study populations in each trial at a different time since vaccination. Since antibody levels and vaccine efficacy decay after vaccination, this will lead to estimated protection being higher in one trial than another at the same level of antibody, even if the true protective effect of antibodies is the same. Hence models are required which relate antibody levels at the time of exposure to the risk of infection, sometimes known as "exposure-proximal" correlates of protection.

1.1 Thesis contribution

This thesis is motivated by data from an Oxford-AstraZeneca COVID-19 vaccine trial (COV002) [3, 4, 5]. A previous analysis compared observed antibody levels at the time of peak response with subsequent risk of infection [5]. We develop a joint model to estimate how antibody levels at the time of exposure relate to the risk of infection. Our joint model allows for two notions of time - the baseline hazard varies with calendar time, whereas the longitudinal antibody trajectory and subsequent effect on risk of infection vary with time-since-vaccination. We also include a link function between the longitudinal and time-to-event components, such that the exponential of the current value of the longitudinal trajectory is a covariate in the proportional hazards model.

We encountered convergence issues when trying to fit the Bayesian joint model in Stan [6]. This led us to develop a multiple imputation approach, where we impute the missing values of the true longitudinal trajectory as a covariate in the time-to-event component. The maximum likelihood estimates calculated on the different imputed datasets were not approximately Gaussian distributed. This assumption is required for the validity of Rubin's rules for pooling estimates after multiple imputation [7]. Therefore, we developed an approximate Bayesian pooling scheme, by sampling from the asymptotic distribution of the maximum likelihood estimator.

1.1.1 Thesis structure

In the rest of this chapter, we provide a literature review of statistical methods to estimate correlates of protection, and joint models of longitudinal and time-to-event data.

In Chapter 2, we develop the theoretical background for the approximate Bayesian pooling scheme for multiple imputation. We calculate the error in Rubin's pooling rules when key assumptions do not hold, and derive an approximate Bayesian pooling rule, which makes fewer assumptions. We compare our approximate Bayesian pooling rule with Rubin's rules and a Bayesian analysis, on simulated data and on three real datasets.

In Chapter 3, we review previous methods for estimating correlates of protection. We demonstrate mathematically and via simulations, that relating observed antibody levels at a fixed time after vaccination to risk of infection, may give inconsistent results between

trials where the background risk of infection only differs between the trials. We compare the results from the multiple imputation joint model developed in Chapter 4 to the extrapolation approach and a time-fixed correlates of protection method on COV002 data.
In Chapter 4, we outline a multiple imputation joint model for antibody decay and risk of infection after COVID-19 vaccination. We apply the model to COV002 data, to estimate antibody decay, correlates of protection, and vaccine efficacy waning.
Finally, we discuss the implications of this research, and possible future work in Chapter 5.

1.2 Background

1.2.1 Correlates of protection

Background and definition

The notion of a correlate of protection relates to early work on surrogate endpoints. Prentice (1989) considered the notion of a surrogate endpoint, which fully captures all the information of the true endpoint, so can be used as a substitute for it [8]. Surrogate endpoints can rarely be found in practice. A more realistic aim is an auxiliary endpoint, which strengthens the true endpoint analysis, by providing additional information [9].
Multiple different terms and definitions related to correlates of protection have been proposed in the literature [10, 11, 12]. In this thesis, we follow the definition given by Gilbert et al. (2024) [13], defining a correlate of protection to be "an immune marker measured in vaccine, and possibly also placebo, recipients that can be analysed to infer vaccine efficacy". Correlates of protection are used as a substitute for the true endpoint, for example, in decisions around licensing new vaccines. Unlike surrogate endpoints they only capture partial information about the true endpoint, since the correlation is not perfect.
Historically, a correlate of protection analysis would estimate a threshold or cut-off for immune marker levels, above which an individual would be assumed to be protected, and below which unprotected. This simple approach rarely captures the true relationship between immune marker levels and protection. In this thesis we focus on methods which estimate a continuous protection curve, where the probability of an individual being

protected depends on their immune marker levels.

Estimating correlates of protection

Various approaches have been suggested to estimate correlates of protection, using immune marker data available at a single timepoint, soon after vaccination. Dunning [14] proposed a logistic model to estimate the risk of infection, by estimating the risk of exposure and multiplying this by the probability that an individual is not protected by the vaccine, which depends on the immune marker values. This ensures vaccine efficacy is bounded between 0 and 100%. The model was extended [15] to consider other parametric forms of the correlates of protection curve, and to fit to a simple case-cohort sampling scenario. However, this extension did not include the general case-cohort scenario, where the probability of immune marker availability varies between individuals, and is unknown, but can be estimated from the data. Effects due to covariates other than the immune marker are also not included. Coudeville et al. (2010) extended the model to perform a Bayesian meta-analysis [16], pooling results from multiple studies of influenza to estimate correlates of protection. Khoury et al. (2021) applied a meta-analysis extension of the Dunning model to estimate COVID-19 correlates of protection using studies of different vaccines [17].

Causal methods for correlates of protection

Others have considered causal inference approaches to estimate correlates of protection. Gilbert et al. (2024) summarise the following four different causal methods for assessing correlates of protection [13]. The principal surrogate approach [18] measures vaccine efficacy in subgroups containing individuals whose potential immune marker would be a certain value, if they were vaccinated. Gilbert et al. (2023) [19] present a controlled effects approach. This estimates the vaccine efficacy that would be observed by comparing (i) a hypothetical intervention where all participants are vaccinated and have immune marker levels set to the same value, to (ii) all participants being assigned to placebo. The mediation approach [20] estimates the natural direct effect (or non-marker mediated vaccine efficacy) i.e. the effect due to vaccination only and not the immune marker; and the natural indirect effect (or marker-mediated vaccine efficacy) i.e. the effect of the immune

marker only and not other effects of the vaccine. This approach does not give a correlates of protection curve showing vaccine efficacy at different marker levels, but can be used to estimate the proportion of vaccine efficacy which is due to the immune marker. The stochastic interventional approach estimates the vaccine efficacy that would be observed if the immune markers were shifted by a constant value for all vaccinated participants [21]. All of the above methods assume no unmeasured confounding is present, that is, there is no unmeasured variable which is a common cause of both the immune marker levels, and also protection levels. This is a strong assumption, and is required for the estimated effect to be a true causal effect. In the controlled effects framework, Gilbert et al. (2023) [19] use E-values to estimate what level of unmeasured confounding would be required to explain away any causal effect completely, and to create conservative estimates given some plausible level of confounding.

Issues with the standard correlates of protection approach

More crucially, all of the above correlates of protection methods assume that the observed antibody levels soon after vaccination cause subsequent protection. This assumption may introduce bias for two reasons. Firstly, the observed antibody levels are measured with error, and protection depends instead on the true antibody levels. Failing to account for error in the antibody measurements may lead to bias [1, 2]. Secondly, these methods do not account for immune marker waning over time. The true causal effect is due to immune marker levels at exposure, rather than at a visit soon after vaccination. If different vaccine trials are of different lengths, immune marker levels will wane by different amounts in the two trials, giving different results. So the above methods will give different results in different trials, even when the true causal effect is the same. Hence, methods are needed which relate the antibody level at exposure to risk of infection. This is sometimes known as "exposure-proximal" correlates of protection. Accounting for measurement error will also reduce bias in the estimates.

1.2.2 Joint modelling of longitudinal and time-to-event data

Joint models

To estimate exposure-proximal correlates of protection, the immune marker levels at the time of exposure must be estimated, and compared to the risk of subsequent infection. Joint models of longitudinal data (e.g. immune markers) and time-to-event data (e.g. time-to-infection) are a commonly used method for such problems. Joint modelling has been widely applied in medical trials across a range of diseases, including HIV [22], cancer [23, 24], cystic fibrosis [25], Alzheimer's disease [26], schizophrenia [27, 28] and others. Ibrahim, Chu and Chen (2010) [29] provide an introductory review. Gould et al. (2015) [30] give a review focussing on Bayesian joint models, with more detail on implementation methods and packages. A more recent review by Papageorgiou et al. (2019) [31] explores different shapes for the longitudinal trajectory, different functional forms of the relationship between the longitudinal variable and subsequent risk, and dynamic predictions of the future trajectory and risk. Joint modelling has been shown to reduce bias in estimates of the relationship between a longitudinal biomarker and clinical event, compared to using the last observed biomarker value [29]. Further, when estimating the biomarker trajectory over time, joint models reduce bias by accounting for the informative censoring of biomarker observations at event times [32, 33].

The standard joint model uses a hierarchical linear model (or mixed effects model) for the longitudinal component, and a proportional hazards model [34] for the time-to-event component. This has been extended in various ways, including to include left-censored longitudinal observations [35], multiple longitudinal markers [24], competing risks (multiple event outcomes) [33], cure fractions [23], robust models using t-distributed measurement error [36], flexible longitudinal trajectories [37], and an accelerated failure time model for the time-to-event component [38]. Joint models can further be used to make dynamic predictions of future event probabilities [39]. Various other extensions have been developed in the literature.

Various packages have been developed to fit joint models. JMbayes2 is a flexible package using Bayesian methods to fit a wide range of joint model specifications [40]. INLAjoint uses integrated nested Laplace approximations to approximate the joint model, reducing

computation time [41, 42]. Variational inference is another approximation method which has been applied to speed up computation time in joint models [43], although no package has yet been developed.

Two-stage joint models

In full joint models, the longitudinal model and time-to-event (survival) model are fitted concurrently, and parameters in one model are allowed to affect estimation in the other. A simpler approach is a two-stage joint model, where the longitudinal model is fitted first, and the output is then included in the time-to-event model, preventing feedback from the second model to the first. In a simple (or "naïve") two-stage approach, the mean estimate of the longitudinal trajectory is used as a predictor in the time-to-event model. By fitting the models separately, computation time may be reduced [44]. However, the simple two-stage approach does not propagate the uncertainty in estimating the longitudinal trajectory into the survival model, nor accounts for the time-to-event in the longitudinal model, and so may be biased [44]. Mauff et al. (2020), and Alvares and Leiva-Yamaguchi (2023) suggested two alternate approaches which reduce the bias from fitting a two-stage model, while also benefitting from reduced computation times [45, 44]. Moreno-Bentancur et al. (2018) [46] developed a multiple imputation joint model, imputing the true value of the longitudinal trajectory at each event time in the survival model. They account for the time-to-event data in the longitudinal model, and propagate the uncertainty in the longitudinal trajectories via multiple imputation. Their approach outperformed the simple two-stage approach in scenarios with multiple longitudinal trajectories which are weakly correlated [46].

1.2.3 Exposure-proximal correlates of protection

Some methods have been proposed and applied to estimate "exposure-proximal" correlates of protection from a vaccine trial. Follmann et al. (2023) [47] use an external dataset to estimate the immune marker decay rate, applying it to a measurement soon after vaccination, in order to estimate antibody levels at exposure. They then compare estimated levels at exposure to subsequent infection in an inverse probability weighted Cox model. This approach does not account for antibody measurement error. White et al. (2014,

2015) apply a simple two-stage joint model for antibody decay and risk of infection in a malaria vaccine trial [48, 49]. Their model includes bi-phasic exponential antibody decay, and a maximum value of vaccine efficacy, estimated from the data. However the model does not account for uncertainty in the antibody predictions when estimating vaccine efficacy, so may be biased [44]. Fu and Gilbert (2017) [50] develop an inverse probability weighted joint model for longitudinal and survival data from vaccine trials. Their model estimates the immune trajectory for each individual nonparametrically, requiring the immune marker to be more densely sampled. Seekircher et al. (2024) [51] fit a joint model for COVID-19 antibody decay and subsequent infection using JMbayes2 [40]. However, their model does not account for time since last vaccination/prior infection. Papadopoulos et al. (2025) fit a joint model for antibody levels and risk of infection after a COVID-19 booster dose [52], also using JMbayes2 [40]. Their model does not account for changing rates of infection over calendar time. Huang and Follmann (2025) [53] develop a two-stage joint model to estimate exposure-proximal correlates of protection, allowing for a time-changing direct effect, and both peak and exposure-proximal immune marker effects. They compare estimated hazard between individuals in a partial likelihood method, which outperforms a simple two-stage approach. Their model shows minimal bias, although they do not consider performance in terms of coverage in their simulation study.

Chapter 2

Pooling after multiple imputation under posterior skewness: Rubin's rules and an approximate Bayesian method

Abstract

Multiple imputation (MI) is a popular method to deal with missing data. Statistical inference after MI is typically combined using a pooling method known as Rubin's rules. Statistical analysis of some datasets with missing data produces skewed posterior distributions, whereas Rubin's rules assume the observed-data posterior is symmetric, which may lead to biased inference. We consider the robustness of Rubin's pooling rules under posterior skewness.

We present an approximate Bayesian pooling scheme (ABpool) that allows for posterior skewness, while requiring substantially lower computational cost than full Bayesian inference after MI. Specifically, we sample once from a t or Gaussian approximation of the completed-data posterior distribution for each imputed dataset, and pool the samples. We mathematically derive the error in applying Rubin's rules where key assumptions do not hold.

We compare the pooling methods in a simulation study with skewed data. Results from Rubin's rules and ABpool tend to converge as the sample size increases. ABpool gives higher power for logistic regression with a skewed covariate. For linear and Cox regression, ABpool gives higher power to detect an effect in small samples when the proportion of missing data is low, but lower power when the proportion of missing data is high and sample size is moderate. We apply both pooling methods to three medical datasets, including to fit a two-stage joint model for antibody decay and COVID-19 vaccine efficacy. Rubin's rules are unable to account for skewness when applied to the medical datasets, whereas ABpool recreates the results from Bayesian inference after MI. Posterior asymmetry can be detected by examining a normal QQ-plot of the maximum likelihood estimates for the multiply imputed datasets. When applied to medical data, ABpool is able to account for posterior skewness which Rubin's rules cannot.

2.1 Introduction

Multiple imputation (MI) is a popular method to deal with the problem of missing data in statistical analyses. Missing values are filled in m times with samples from their posterior predictive distribution, to generate m completed datasets. Each imputed dataset is then analysed separately using methods for complete datasets. The results are then pooled together using a simple method known as Rubin's rules [54, 7].

Multiple imputation was originally designed as a method allowing organisations to release public datasets, which different users can subsequently analyse using complete-data methods, while still obtaining valid inferences [7]. Hence a small value of m was preferred, to reduce storage costs and computational requirements. Storage and computational capacity available to individuals and organisations have since increased considerably. The method is now widely applied outside of this setting, and is a standard approach wherever missing data occurs [55].

The combining rules developed by Rubin [7], later updated for small-samples by Barnard and Rubin [54], rely on a few key assumptions. Firstly, they assume the complete-data posterior is t- or Gaussian-distributed. Asymptotic theory suggests this should be a reasonable approximation in large enough samples. When this assumption appears inappropriate, a transformation may be applied to make the assumption more reasonable [56, 57]. In practice, researchers applying Rubin's rules rarely report whether or not they applied a transformation [56]. Zhou and Reiter [58] outline how to perform Bayesian inference after multiple imputation, which is valid even when the complete-data posterior is not well approximated by the t-distribution. This approach has not been as widely adopted in practice as Rubin's rules. This is likely due in part to the need for a method such as Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution, which may require additional computational power and training for some applied researchers.

Rubin's rules also relies on additional assumptions, which mean that the observed-data posterior is also t-distributed. One way to avoid these assumptions, is to combine multiple imputation with bootstrapping. Brand et al. [59] found that making a single imputation within each bootstrap sample outperformed Rubin's rules without bootstrapping, in simulation studies considering skewed datasets. Another way to relax these assumptions is

by approximating the marginal posterior with a mixture of estimated complete-data posterior distributions [60]. This approach was proposed by Steele, Wang and Raftery [60], who assumed the complete-data posterior to be Gaussian. This was adapted by Rashid, Mitra and Steele [61] who assumed the complete-data posterior to be t-distributed, and chose the degrees of freedom such that the posterior variance equalled the variance given by Rubin's rules [7]. Both papers consider the case where the number of imputations m is small. Rashid et al. found that both methods give under-coverage for small m, but coverage rates tend towards the nominal coverage given by Rubin's rules as m increases [61].

We propose an Approximate Bayesian pooling scheme, which approximates the method of Zhou and Reiter [58]. Our method is closely related to the mixture approach above, but differs in that we choose m to be large, and pool together samples from the mixture of complete-data posterior distributions, instead of pooling the distributions themselves. Our method relaxes the assumptions made by Rubin's rules, giving a simple pooling rule which does not suffer from the under-coverage of the previous mixture approaches due to small m.

Much research has been undertaken into different methods to generate appropriate imputations. This paper focuses on methods for pooling the results after the analysis has been performed on multiple imputed data, so we do not discuss methods for imputing the data here.

Section 2.2 introduces multiple imputation, pooling via Rubin's rules and Bayesian inference after multiple imputation. Section 2.3 details our approximate Bayesian pooling scheme, both in the general case, and for the Cox model. In Section 2.4, we approximate the error due to applying Rubin's rules inappropriately. We compare Rubin's rules with our Approximate Bayesian sampling scheme and a Bayesian approach in Sections 2.5 and 2.6, on simulated and real datasets respectively. We provide a discussion in Section 2.7.

2.2 Multiple imputation

A theoretical introduction to multiple imputation is given by Rubin (1987) [7], and a more practical introduction by Van Buuren (2012) [55]. We write the complete data as

$D = (D_{\text{obs}}, D_{\text{mis}})$, with D_{obs} the observed data, and D_{mis} the missing data. Let θ be the quantity of interest, which is usually a parameter in a statistical model for data D. Multiple imputation is motivated by the following, which is given as result 3.1 in Rubin (1987) [7],

$$P\left(\theta \mid D_{\text{obs}}\right) = \int P(\theta \mid D_{\text{obs}}, D_{\text{mis}})\, P(D_{\text{mis}} \mid D_{\text{obs}})\, dD_{\text{mis}} \tag{2.1}$$

That is, the posterior distribution for θ given the observed data D_{obs}, is the average of the complete-data posterior distributions for θ given the completed data $D = (D_{\text{obs}}, D_{\text{mis}})$, over imputations of the missing data D_{mis}. The imputations should be drawn from the posterior predictive distribution of D_{mis} given D_{obs}.

Hence to approximate the posterior distribution for θ given the observed data D_{obs}, we impute m times from the posterior for D_{mis} given D_{obs}. Let the lth imputed full dataset be $D^{(l)} = \left(D_{\text{obs}}, D_{\text{mis}}^{(l)}\right)$, for $l = 1, \ldots, m$. We then run the inference procedure for $\hat{\theta}^{*(l)} = \hat{\theta}(D^{(l)})$ on the lth imputed dataset, and pool the results from each of the m imputed datasets to approximate the posterior distribution for θ.

Various methods and software for drawing imputations have been presented in the literature; a popular method is multiple imputation using chained equations (MICE) [57, 55], which is implemented in the R package mice [62]. This paper focuses on methods for pooling the results after the imputations have been made, so we will not discuss methods for imputing the data in detail.

2.2.1 Multiple imputation and Rubin's rules

In the original formulation of Rubin's rules [7], it is assumed that with complete data, inferences for θ would be performed based on the statement

$$\theta - \hat{\theta} \sim N(0, U) \tag{2.2}$$

where $\hat{\theta}$ is an estimate for θ, with associated sampling variance estimated by U, and $N(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. For some analyses, such as logistic or Poisson regression, the sampling variance is a function of θ (given

by the Fisher information), and can be estimated by substituting in $\hat{\theta}$.

However in most scenarios the sampling variance U additionally depends on an unknown dispersion parameter, which needs to be estimated from the data. Replacing U by its estimate introduces further uncertainty to the distribution of $\theta - \hat{\theta}$. Therefore a t-distribution may be more appropriate than a Gaussian.

In this vein, Barnard and Rubin (1999) [54] assume that, with complete data, inferences for θ would be performed based on the statement

$$\theta - \hat{\theta} \sim t_{\nu_{\text{com}}}(0, U) \tag{2.3}$$

where $t_{\nu_{\text{com}}}(0, U)$ is the t-distribution with location 0, squared scale U and ν_{com} degrees of freedom. ν_{com} is often given by $\nu_{\text{com}} = n - p$, where there are n individuals present in the data, and p parameters estimated by the model. Using a t approximation gives more robust and conservative variance estimates than the normal approximation.

Both formulations of Rubin's rules further assume that the posterior distribution of θ given the observed data D_{obs} only – marginalised over the missing data – is also from a t-distribution [7, 54]. We present in this paper a method which avoids making this assumption.

Let $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(m)}$ and $U^{*(1)}, \ldots, U^{*(m)}$ be the values of $\hat{\theta}$ and U calculated for each of the m imputed complete datasets. The pooled estimate for θ and associated variance U are calculated by Rubin's rules as follows.

Rubin's rules to estimate $\hat{\theta}$ and U

The multiple imputation estimate for θ is derived in Rubin (1987) [7], and given by

$$\bar{\theta}_m = \frac{1}{m} \sum_{l=1}^{m} \hat{\theta}^{*(l)} \tag{2.4}$$

with $\hat{\theta}^{*(l)}$ the estimate calculated from the lth imputed dataset. The estimate for the associated sampling variance V_m is then given by

$$V_m = \overline{U}_m + \left(\frac{m+1}{m}\right) B_m \tag{2.5}$$

where

$$\overline{U}_m = \frac{1}{m}\sum_{l=1}^{m} U^{*(l)} \tag{2.6}$$

is the within-imputation variance – the average of the estimated variances for each imputation, and

$$B_m = \frac{1}{m-1}\sum_{l=1}^{m} \left(\hat{\theta}^{*(l)} - \bar{\theta}_m\right)\left(\hat{\theta}^{*(l)} - \bar{\theta}_m\right)^T \tag{2.7}$$

is the between-imputation variance – the variance of the estimates $\hat{\theta}^{*(l)}$ across imputations.

t approximation

Confidence intervals and hypothesis testing can then be performed based on an approximation using the t-distribution

$$\theta - \bar{\theta}_m \sim t_{\tilde{\nu}_m}\left(0, V_m\right) \tag{2.8}$$

with degrees of freedom $\tilde{\nu}_m$ derived in Barnard and Rubin [54]. An alternative degrees of freedom is given by Rubin (1987) which relies on Eq (2.2) instead of Eq (2.3), however the method given by Barnard and Rubin is generally preferred. Throughout this paper, we refer to Barnard and Rubin's pooling rule [54] as Rubin_t, and Rubin's original pooling rule [7] as Rubin_norm, since they assume a t- and Gaussian-distributed complete-data posterior respectively.

2.2.2 Bayesian inference after multiple imputation

When Bayesian inference is performed after multiple imputation, we see from Equation (2.1) that we can sample from the posterior $P\left(\theta \,|\, D_{\text{obs}}\right)$ by first imputing the missing data $D_{\text{mis}}^{(l)} \sim P(D_{\text{mis}} \,|\, D_{\text{obs}})$ then sampling from the posterior distribution for θ given the imputed complete data $\theta_j^{*(l)} \sim P(\theta \,|\, D_{\text{obs}}, D_{\text{mis}}^{(l)})$, for $j = 1, \ldots, J$. The resulting samples $\theta_j^{*(l)}$, $l = 1, \ldots, m$, $j = 1, \ldots, J$ are simply pooled across imputations to form a sample from the posterior distribution [58]. Posterior quantities such as means, medians, quantiles and credible intervals (CrIs) can be calculated from these samples.

Note when applying Rubin's rules (Section 2.2.1), as few as $m = 10$ imputations may suffice. However with Bayesian inference after multiple imputation many more imputations are needed, similar to other sampling methods such as MCMC, with m at least 100 [58]. No assumption is made on the form of the complete-data posterior when performing Bayesian inference after multiple imputation.

2.3 Approximate Bayesian pooling for multiple imputation

As well as the assumption of a t- or Gaussian-distributed complete-data posterior, Rubin's rules relies on two additional assumptions. Firstly, the distribution of the maximum likelihood estimators (MLEs) for θ across imputed datasets, $\hat{\theta}^*$, follows a Gaussian distribution (Eq. 3.3.2, Rubin (1987) [7]). Secondly, $\sqrt{\text{Var}(U^*)} \ll B = \text{Var}(\hat{\theta}^*)$, that is, the distribution of the asymptotic variance U^* has lower-order variability than the MLE $\hat{\theta}^*$ (Eq. 3.3.3, Rubin (1987) [7]). Arguments given in Rubin (1987) rely on both these assumptions to give rise to the t approximation for the marginal posterior distribution given the observed data only [7]. Both of these assumptions are testable. The first assumption can be tested by calculating $\hat{\theta}^{*(l)}$ for $l = 1, \ldots, m$, and comparing the quantiles of the samples $\hat{\theta}^{*(l)}$ with the quantiles of a standard Gaussian in a Q–Q plot. When the plot is non-linear, indicating a Gaussian approximation is not reasonable, this assumption does not hold. The second assumption can be tested by comparing the sample variance of U^* with B_m, the sample variance of $\hat{\theta}^*$. If the sample variance of U^* is non-negligible compared to B_m, the second assumption is violated. A large enough m is required for both tests to give a clear result. When either of these assumptions do not hold, Rubin's rules give a poor approximation. Bayesian inference after multiple imputation (Section 2.2.2) is a viable alternative. However, some data analysts may already be familiar with existing likelihood-based estimation procedures and software, such as linear regression or logistic regression, but not familiar with their Bayesian equivalents. In other cases a specific model may be preferred which fits more naturally with a likelihood-based estimation method. One example is the Cox model for survival analysis [34], which we expand on in Section 2.3.2. In some cases running a frequentist analysis on each multiply imputed dataset may be faster than running a Bayesian method such as Markov Chain Monte-Carlo (MCMC)

or Hamiltonian Monte Carlo (HMC) on each dataset, which may be prohibitive when a large m is required. We present an alternative Approximate Bayesian pooling method to combine maximum likelihood inference for θ on the completed datasets after multiple imputation, without requiring these two additional assumptions.

2.3.1 Pooling via approximate Bayesian sampling after multiple imputation

Similar to equations (2.2) and (2.3), we may approximate the completed-data posterior distribution with a Gaussian or t approximation. First consider a Gaussian approximation, with mean at the posterior mode, and variance given by the inverse of the observed information matrix, evaluated at the posterior mode (See Gelman et al. (1995), Chapter 4) [63]. By the Bernstein-von Mises theorem, the posterior is asymptotically Gaussian in the limit as the sample size n tends to infinity, under regularity conditions given by Gelman et al. (1995) [63]. Further, the likelihood dominates the prior distribution, and so the posterior mode converges to the mode of the likelihood function, namely, the maximum likelihood estimator (MLE). The observed information matrix will also converge to the (Fisher) expected information matrix, by the Law of Large Numbers. Hence $N(\hat{\theta}^{*(l)}, U^{*(l)})$, the asymptotic distribution of the MLE given the imputed data $D^{(l)}$, is also the limit of the posterior distribution as $n \to \infty$. So the asymptotic distribution of the MLE approximates the posterior, given the imputed data:

$$P(\theta \,|\, D_{\text{obs}}, D_{\text{mis}}^{(l)}) \approx N(\hat{\theta}^{*(l)}, U^{*(l)})\,. \tag{2.9}$$

For finite n, the approximation relies on two assumptions. Firstly, that the posterior distribution is proportional to the likelihood, and so independent of the prior distribution. Choosing an improper prior distribution $\pi(\theta) \sim 1$ will guarantee this holds, as long as the resulting posterior is proper. The assumption will also hold approximately if the prior distribution is non- or weakly-informative, or n is large enough for the likelihood to dominate the posterior. Secondly, the likelihood function is approximately Gaussian. This will generally hold for large n but, much like the asymptotic Gaussian approximation which all likelihood-based frequentist inference relies upon, may not be a good approximation

for small n.

As discussed earlier, a t approximation may be a better fit to the completed-data posterior distribution – especially in small samples, and where a dispersion parameter is present which must be marginalised over. In most scenarios, we expect the t approximation will be more appropriate, and give more robust inferences:

$$P(\theta \,|\, D_{\text{obs}}, D_{\text{mis}}^{(l)}) \approx t_{\nu_{\text{com}}}(\hat{\theta}^{*(l)}, U^{*(l)})\,. \tag{2.10}$$

where the complete-data degrees of freedom ν_{com} is given by $n-p$ in most cases, for sample size n, when estimating p parameters.

We may then closely follow the method of Zhou and Reiter [58], performing approximate Bayesian inference after multiple imputation, which allows for skewness in the observed-data posterior. We apply the above asymptotic approximations to the complete-data posterior, to reduce the computational requirement for full Bayesian inference. This suggests the following algorithm: For $l = 1, \ldots, m$, we impute the missing data from the observed data $D_{\text{mis}}^{(l)} \sim P(D_{\text{mis}} \,|\, D_{\text{obs}})$, and then estimate $\hat{\theta}^{*(l)}$ and $U^{*(l)}$ from $D^{(l)} = (D_{\text{obs}}, D_{\text{mis}}^{(l)})$ as before. We then draw J samples either from a t-distribution $\theta_j^{*(l)} \sim t_{\nu_{\text{com}}}(\hat{\theta}^{*(l)}, U^{*(l)})$, or alternatively a Gaussian $\theta_j^{*(l)} \sim N(\hat{\theta}^{*(l)}, U^{*(l)})$ for $j = 1, \ldots, J$. Finally, we pool together the samples $\theta_j^{*(l)}$, $l = 1, \ldots, m$, $j = 1, \ldots, J$, to form a sample approximately drawn from the posterior distribution of θ given the observed data $P(\theta \,|\, D_{\text{obs}})$. Posterior quantities such as means, medians, quantiles and credible intervals can be calculated from these samples. We refer to this method as Approximate Bayesian pooling (ABpool), with ABpool_t and ABpool_norm denoting the method assuming a t- and Gaussian-distributed complete-data posterior respectively. As with Bayesian inference after multiple imputation, the required number of imputations m is much larger than the required m when Rubin's rules are used. We expect the method to perform best when m is itself large, in which case $J = 1$ will suffice, so we assume this for the rest of this paper. This can be viewed as taking a finite sample m from an infinite mixture of complete-data posteriors, defined by Eq. (2.1), where the complete-data posterior is approximated by (2.9) or (2.10).

Note that by equation (2.1), if the Gaussian or t approximation (2.9, 2.10) holds exactly, and the imputation scheme is proper [7], then this sampling scheme will draw samples

from the marginal posterior distribution itself.

2.3.2 Approximate Bayesian pooling for the Cox model after multiple imputation

An example where a frequentist estimation approach may be preferred over a Bayesian method post imputation is the Cox proportional hazards model [34], which is widely used in medical research for time-to-event data. We will apply Approximate Bayesian pooling to an analysis using this model in Chapter 4, which is also given as a real data example in Section 2.6 of this chapter. Since the Cox model maximises the partial likelihood instead of the full likelihood, we must demonstrate why Approximate Bayesian pooling can be applied to this model. We first outline the model below.

The Cox model

For individual $i = 1, \ldots, n$, let T_i be the time until an event of interest or censoring. Let δ_i be the event indicator, that is δ_i is 1 if T_i represents the time of occurrence of the event of interest, and 0 if T_i is a censoring time. Let X_i be a vector of associated covariates which predict the event time, and $\lambda_0(t)$ an unknown baseline hazard function. Let $\mathcal{R}_i = \{j : T_j \geq T_i\}$ be the set of individuals at risk at time T_i. Then the hazard function at time t for individual i is given by

$$\lambda_i(t) = \lambda_0(t) \exp(X_i^T \theta) \tag{2.11}$$

and the partial likelihood is given by

$$L_P(\theta \,|\, T, \delta, X) = \prod_{i=1}^{n} \frac{\exp(X_i^T \theta)}{\sum_{j \in \mathcal{R}_i} \exp(X_j^T \theta)} \,. \tag{2.12}$$

The MLE can be shown to be equal to the value of θ which maximises the partial likelihood, $\hat{\theta} = \arg\max_\theta L_P(\theta \,|\, T, \delta, X)$.

Cox's partial likelihood method uses a non-parametric approach to marginalise over the baseline hazard, which allows the parameter estimate to be calculated independently from the baseline hazard. The method is widely used, and standard software is available. By

comparison, alternative Bayesian methods are less well-known by applied researchers, and require specification of a functional form for the baseline hazard function $\lambda_0(t)$.

Where missing values are present in a dataset to be analysed by the Cox model, multiple imputation may be applied to account for the missingness. White and Royston (2009) discuss multiple imputation of missing covariate values for the Cox model [64]. They recommend accounting for the event time in the imputation model by including the event indicator and the Nelson–Aalen (N–A) estimate of cumulative hazard as predictors of the missing covariate.

Approximate Bayesian pooling after multiple imputation for the Cox model

In order to apply approximate Bayesian pooling after multiple imputation to the Cox model, we must first justify the use of the maximum partial likelihood estimator and associated variance to approximate an appropriate Bayesian posterior. Sinha, Chen and Ibrahim (2003) [65] show that the posterior derived using the partial likelihood is an approximation of the posterior derived from the full likelihood, when using a diffuse gamma process prior on the cumulative hazard function. Then following our arguments in Section 2.3.1, approximate Bayesian pooling can be applied, to approximate Bayesian inference with a diffuse gamma process prior after multiple imputation.

2.4 Approximating the error from Rubin's rules

In this section we treat the observed data D_{obs} as fixed, and consider the distribution of the parameters (and missing data) given the observed data, from a Bayesian perspective. Hence we are conditioning on the observed data D_{obs} in all that follows, though we will not always write the conditioning explicitly, for the sake of brevity. Recall equation (2.1), namely

$$\pi\left(\theta \,|\, D_{\text{obs}}\right) = \int \pi\left(\theta \,|D\right)\, \pi\left(D_{\text{mis}} \,|\, D_{\text{obs}}\right) dD_{\text{mis}}$$

Suppose the complete-data posterior t approximation (2.10) holds exactly. Note this assumption is made both by our approximate Bayesian pooling method, and by Rubin's rules. We will then consider the error in approximating the incomplete-data posterior dis-

tribution by a t-distribution using Rubin's rules. Rubin's rules assumes the complete data MLEs $\hat{\theta}^*$ are drawn from a Gaussian distribution given the observed data, and $\sqrt{\text{Var}(U^*)} \ll B = \text{Var}(\hat{\theta}^*)$. We now calculate the error in the first four standardised moments of the posterior approximation given by Rubin's rules when these assumptions do not hold, and approximate the resulting error in an uncertainty interval.

2.4.1 t-distributed complete-data posterior

Consider the t-distributed case, where, following Barnard and Rubin (1999) [54], we assume a complete-data analysis would be based on the statement that

$$\theta - \hat{\theta}(D) \sim t_{\nu_{\text{com}}}(0, U(D)) \tag{2.13}$$

where the complete-data degrees of freedom $\nu_{\text{com}} > 4$.

Let $\hat{\theta}^*$ and U^* be random variables generated by first imputing the missing data from $\pi\left(D_{\text{mis}} \,|\, D_{\text{obs}}\right)$, then estimating the MLE $\hat{\theta}^*(D) = \hat{\theta}\left(D_{\text{mis}}, D_{\text{obs}}\right)$, and associated asymptotic variance $U^*(D)$.

We now consider the approximation

$$\pi\left(\theta \,|D\right) \approx t_{\nu_{\text{com}}}(\hat{\theta}^*(D), U^*(D)) \tag{2.14}$$

Note Rashid et al. (2015) use a t-approximation for the completed-data posteriors in the same way, except with a different degrees of freedom [61]. While this approximation is unlikely to fit well for all possible completed-datasets D, it may fit well for most imputations in analyses where equation (2.13) holds.

Assume the approximation (2.14) holds exactly. We can therefore write

$$\theta = \hat{\theta}^* + \psi\sqrt{U^*} \tag{2.15}$$

where $\psi \sim t_{\nu_{\text{com}}}(0, 1)$ independent of $\hat{\theta}^*$ and also of U^*.

Treating the observed data as fixed and taking expectations over the distribution of the missing data given the observed data, let $\bar{\theta}_\infty = E\left[\hat{\theta}^* \,|\, D_{\text{obs}}\right]$ be the expected value of the MLE $\hat{\theta}^*$ given observed data D_{obs}. Similarly let $B = \text{Var}\left[\hat{\theta}^* \,|\, D_{\text{obs}}\right]$ be the between-

imputation variance, and $\overline{U}_\infty = E\big[U^* \,|\, D_{\text{obs}}\big]$ be the expected value of the asymptotic variance U^* given observed data D_{obs}. For a random variable X with mean μ and variance σ^2, let $\gamma_1(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ be the skewness of X, and $\gamma_2(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3$ be the excess kurtosis of X. Let $\nu := \nu_{\text{com}}$, and $\lambda_\nu = \frac{B}{B+(\frac{\nu}{\nu-2})\overline{U}_\infty}$. Then, conditioning on the observed data D_{obs} on both sides of the following equations, we find that

$$\mathrm{E}[\theta \,|\, D_{\text{obs}}] = \bar{\theta}_\infty \tag{2.16}$$

$$\mathrm{Var}\big[\theta \,|\, D_{\text{obs}}\big] = B + \left(\frac{\nu}{\nu-2}\right)\overline{U}_\infty \tag{2.17}$$

$$\gamma_1(\theta \,|\, D_{\text{obs}}) = \gamma_1(\hat{\theta}^* \,|\, D_{\text{obs}})\,\lambda_\nu^{3/2} + 3\frac{\left(\frac{\nu}{\nu-2}\right)\mathrm{Cov}\left(\hat{\theta}^*, U^* \,|\, D_{\text{obs}}\right)}{\left(B + \overline{U}_\infty(\frac{\nu}{\nu-2})\right)^{3/2}} \tag{2.18}$$

$$\begin{aligned}\gamma_2\left(\theta \,|\, D_{\text{obs}}\right) = \gamma_2(\hat{\theta}^* \,|\, D_{\text{obs}})\,\lambda_\nu^2 + \frac{6}{\nu-4}(1-\lambda_\nu)^2 \\ + 3\left(\frac{\frac{\nu^2}{(\nu-2)(\nu-4)}\mathrm{Var}(U^* \,|\, D_{\text{obs}}) + 2\frac{\nu}{\nu-2}\mathrm{Cov}\left((\hat{\theta}^* - \bar{\theta}_\infty)^2, U^* \,|\, D_{\text{obs}}\right)}{\left(B + (\frac{\nu}{\nu-2})\overline{U}_\infty\right)^2}\right)\end{aligned} \tag{2.19}$$

Derivations of the skewness and kurtosis are given in Lemmas 1 and 2 in Appendix A.1, and a discussion of the implications is given later in 2.4.3.

Now in the limit as the number of imputations $m \to \infty$, Rubin's rules [54] approximates the posterior by a $t_{\nu_\infty}(\bar{\theta}_\infty, \overline{U}_\infty + B)$ distribution, where

$$\nu_\infty := \nu\Big(\frac{\nu+1}{\nu+3}\Big)\Big(\frac{\overline{U}_\infty}{\overline{U}_\infty + B}\Big) = \nu\Big(\frac{\nu+1}{\nu+3}\Big)(1-\lambda) \tag{2.20}$$

for $\lambda = B/(\overline{U}_\infty + B)$.

Hence letting θ_{Rubin} be a draw from the approximation to the posterior given by Rubin's rules,

$$\mathrm{E}[\theta_{\text{Rubin}} \,|\, D_{\text{obs}}] = \bar{\theta}_\infty \tag{2.21}$$

$$\mathrm{Var}\big[\theta_{\text{Rubin}} \,|\, D_{\text{obs}}\big] = (\overline{U}_\infty + B)\left(\frac{\nu_\infty}{\nu_\infty - 2}\right) \tag{2.22}$$

$$\gamma_1(\theta_{\text{Rubin}} \,|\, D_{\text{obs}}) = 0 \tag{2.23}$$

$$\gamma_2(\theta_{\text{Rubin}} \,|\, D_{\text{obs}}) = \left(\frac{6}{\nu_\infty - 4}\right) \tag{2.24}$$

2.4.2 Gaussian complete-data posterior

In the limit as $\nu_{\text{com}} \to \infty$, we recover the Gaussian-distributed case. In this case, following Rubin (1987) [7], we assume a complete-data analysis would be based on the statement that

$$\theta - \hat{\theta}(D) \sim N(0,\, U(D)) \tag{2.25}$$

We now consider the approximation

$$\pi\left(\theta \,|D\right) \approx N\left(\hat{\theta}^*(D),\, U^*(D)\right) \tag{2.26}$$

Steele et al. (2010) also use the same Gaussian approximation for the completed-data posteriors [60]. While this approximation is unlikely to fit well for all possible completed-datasets D, it may fit well for most imputations in analyses where equation (2.25) holds. We assume the error from the approximation (2.26) to be negligible, and drop the dependency on D in the notation for simplicity. Assuming the approximation to be exact, we can therefore write

$$\theta = \hat{\theta}^* + \psi\sqrt{U^*} \tag{2.27}$$

where $\psi \sim N(0,1)$ independent of $\hat{\theta}^*$ and also of U^*.

Recalling that $\lambda = B/(\overline{U}_\infty + B)$, we see

$$\mathrm{E}[\theta \,|\, D_{\text{obs}}] = \bar{\theta}_\infty \tag{2.28}$$

$$\mathrm{Var}\big[\theta \,|\, D_{\text{obs}}\big] = B + \overline{U}_\infty \tag{2.29}$$

$$\gamma_1\left(\theta \,|\, D_{\text{obs}}\right) = \gamma_1(\hat{\theta}^* \,|\, D_{\text{obs}})\, \lambda^{3/2} + 3\left(\frac{\mathrm{Cov}\left(\hat{\theta}^*, U^* \,|\, D_{\text{obs}}\right)}{\left(B + \overline{U}_\infty\right)^{3/2}}\right) \tag{2.30}$$

$$\gamma_2\left(\theta \,|\, D_{\text{obs}}\right) = \gamma_2(\hat{\theta}^* \,|\, D_{\text{obs}})\, \lambda^{2} + 3\left(\frac{\mathrm{Var}(U^* \,|\, D_{\text{obs}}) + 2\,\mathrm{Cov}\left((\hat{\theta}^* - \bar{\theta}_\infty)^2, U^* \,|\, D_{\text{obs}}\right)}{\left(B + \overline{U}_\infty\right)^{2}}\right) \tag{2.31}$$

Now as $\nu_{\text{com}} = \infty$, then as $m \to \infty$, Rubin's rules will estimate the distribution of θ to be $N(\bar{\theta}_\infty, \overline{U}_\infty + B)$. This is because the degrees of freedom in Rubin's t-approximation

[7, 54] tends to infinity as $m \to \infty$.

$$\mathrm{E}[\theta_{\text{Rubin}} \mid D_{\text{obs}}] = \bar{\theta}_\infty \tag{2.32}$$

$$\mathrm{Var}\left[\theta_{\text{Rubin}} \mid D_{\text{obs}}\right] = \overline{U}_\infty + B \tag{2.33}$$

$$\gamma_1(\theta_{\text{Rubin}} \mid D_{\text{obs}}) = \gamma_2(\theta_{\text{Rubin}} \mid D_{\text{obs}}) = 0 \tag{2.34}$$

2.4.3 Discussion

All lemmas mentioned in this section are stated and proved in in Appendix A.1.

We observe that Rubin's rules gives a consistent estimate of the posterior mean.

Barnard and Rubin's pooling rule [54] will always overestimate the posterior variance for finite $\nu_{\text{com}} > 3$, while the variance estimate is correct in the Gaussian case, i.e. the limit $\nu_{\text{com}} \to \infty$ (Lemmas 3 and 4). This may lead to overly conservative inference, when ν_{com} is small. The difference between the variance estimated by Rubin's rules and the true posterior variance is bounded by $(\overline{U}_\infty + B)\left(\frac{\nu_\infty}{\nu_\infty - 2} - 1\right)$ (Lemma 5). No such bound exists in terms of ν, as the variance given by Rubin's rules diverges as λ grows such that ν_∞ approaches 2. In practice such a large value of λ is unlikely to occur. See Appendix A for proofs.

We now consider how the two additional assumptions made by Rubin's rules (Eq. 3.3.2 & 3.3.3 Rubin [7]) relate to the skewness and kurtosis in the above calculations. When $\hat{\theta}^*$ is Gaussian distributed (Eq. 3.3.2, Rubin [7]), $\gamma_1(\hat{\theta}^*)$ and $\gamma_2(\hat{\theta}^*)$ are both equal to zero. Further when $\sqrt{\mathrm{Var}(U^* \mid D_{\text{obs}})}$ is negligible compared to B (Eq. 3.3.3, Rubin [7]), we show in Lemma 7 that the final terms in the skewness and excess kurtosis are also negligible. All that remains in this case, is the second term in the excess kurtosis equation (Eq. 2.19). We prove in Lemma 6 that the excess kurtosis given by Barnard and Rubin's pooling rule (Eq. 2.24) will always be larger than the second term in the posterior excess kurtosis equation (Eq. 2.19), while both terms tend to 0 in the limit as $\nu \to \infty$. That is, when both additional assumptions hold, Barnard and Rubin's pooling rule will overestimate the excess kurtosis, and correctly estimate the skewness to be 0. This means when these assumptions hold, Rubin's rules give confidence intervals (CIs) which are conservative for

finite ν, and nominal in the limit as $\nu \to \infty$.

When these assumptions do not hold, posterior skewness and excess kurtosis will be introduced, which Rubin's rules does not account for. This may firstly be due to skewness and excess kurtosis of $\hat{\theta}^*$, when λ_ν is not too small (first terms, Eq. 2.18 - 2.19).

Secondly, skewness may be introduced due to correlation between $\hat{\theta}^*$ and U^*. Two important examples of this is are Poisson and logistic regression, estimated on the log and logit scale respectively. For the simplest case of intercept-only regression, the plug-in estimator for the asymptotic variance is a deterministic function of the MLE, given by $U^* = e^{-\hat{\theta}^*}/n$ for Poisson and $U^* = \left(e^{\hat{\theta}^*} + 2 + e^{-\hat{\theta}^*}\right)/n$ for logistic regression. This tends to lead to strong negative correlation between the MLE and the asymptotic variance for intercept-only Poisson regression. For logistic regression, if the true proportion is far from 1/2 (i.e. the log proportion θ is far from 0), then the MLE and asymptotic variance will be highly correlated. Correlation may also be present in multi-parameter Poisson or logistic regression, depending on the design matrix and the true parameter values. When this correlation is non-negligible compared to $\left(B + \bar{\theta}_\infty\right)^{3/2}$, skewness will be introduced to the missing-data posterior. Thus posterior skewness may be induced by skewness in the MLEs, or by a high correlation between U^* and $\hat{\theta}^*$ given the observed data.

Additional excess kurtosis not accounted for by Rubin's rules may also be introduced, if $\mathrm{Var}(U^* \mid D_{\mathrm{obs}})$ or $\mathrm{Cov}((\hat{\theta}^* - \bar{\theta}_\infty)^2, U^* \mid D_{\mathrm{obs}})$ is large compared to $(B + \overline{U}_\infty)^2$ (last term, Eq. 2.19).

2.4.4 Cornish-Fisher approximation to estimate error in the confidence interval

We will estimate the error of the quantiles given by the approximation from Rubin's rules using a Cornish-Fisher expansion [66, 67]. Kolassa and Kuffner (2020) proved that the formal Edgeworth expansion is valid for the posterior distribution of a parameter θ, assuming certain regularity conditions hold [68]. In a missing-data scenario, we require the missingness to be ignorable. It follows that the Cornish-Fisher expansion should also hold, since this is the inverse of the Edgeworth expansion [66, 67]. Therefore the α-quantile of θ, $Q_\theta(\alpha)$, can be expressed in terms of the α-quantile of a standard normal variable z_α,

given by

$$Q_\theta(\alpha) = \mu + \sigma \left[z_\alpha + \frac{\gamma_1(\theta)}{6}(z_\alpha^2 - 1) + \frac{\gamma_2(\theta)}{24}(z_\alpha^3 - 3z_\alpha) - \frac{\gamma_1(\theta)^2}{36}(2z_\alpha^3 - 5z_\alpha) \right] + O(n^{-3/2}) \tag{2.35}$$

where μ is the mean of θ, σ the standard deviation of θ, $\gamma_1(\theta)$ the skewness, $\gamma_2(\theta)$ the excess kurtosis, and n is the sample size.

t-distributed case

Let $\sigma_\theta^2 = \mathrm{Var}[\theta \,|\, D_{\text{obs}}]$ and $\sigma_R^2 = \mathrm{Var}[\theta_{\text{Rubin}} \,|\, D_{\text{obs}}]$ and condition on the observed data D_{obs}. We may then apply the Cornish-Fisher expansion (Eq. 2.35, [66]) to the quantiles of the true posterior (Eq 2.16 - 2.19), and to the t approximation given by Rubin's rules (Eq 2.21 - 2.24), to give an approximation of the quantiles for each of these distributions in terms of z_α. Subtracting them from each other, we see

$$\begin{aligned} Q_\theta(\alpha) - Q_\theta^{Rubin}(\alpha) = {} & (\sigma_\theta - \sigma_R) z_\alpha + \frac{1}{6}\gamma_1(\theta)\,\sigma_\theta\,(z_\alpha^2 - 1) \\ & + \frac{1}{24}\left(\sigma_\theta \gamma_2(\theta) - \sigma_R \left(\frac{6}{\nu_\infty - 4}\right)\right)(z_\alpha^3 - 3z_\alpha) \\ & - \frac{1}{36}\gamma_1(\theta)^2 \sigma_\theta\,(2z_\alpha^3 - 5z_\alpha) + O(n^{-3/2}) \end{aligned} \tag{2.36}$$

where all quantities are conditioned on the observed data D_{obs}. The error in estimating quantiles will translate into an error in estimating the bounds of the confidence interval generated by Rubin's rules.

Gaussian case

This equation simplifies in the Gaussian-distributed case, which is the limit as $\nu_{\text{com}} \to \infty$. Let $V = \overline{U}_\infty + B$. Then, by the Cornish-Fisher expansion (Eq. 2.35, [66]) and Eq (2.28 - 2.31), we have that

$$Q_\theta(\alpha) - Q_\theta^{Rubin}(\alpha) = \sqrt{V}\left\{ \frac{\gamma_1(\theta)}{6}\,(z_\alpha^2 - 1) + \frac{\gamma_2(\theta)}{24}(z_\alpha^3 - 3z_\alpha) - \frac{\gamma_1(\theta)^2}{36}\,(2z_\alpha^3 - 5z_\alpha) + O(n^{-3/2}) \right\} \tag{2.37}$$

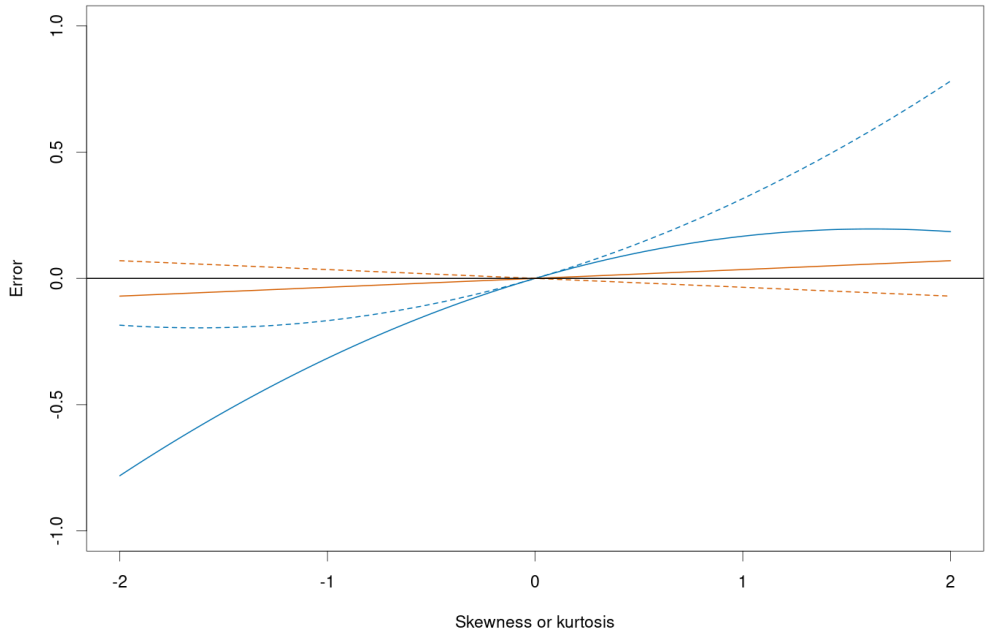


Figure 2.1: **Approximate error in confidence bounds due to posterior skewness or kurtosis**
The error in the confidence bounds estimated by Rubin's rules with infinite complete-data degrees of freedom and an infinite number of imputations. The blue lines show the error due to posterior skewness, when the excess kurtosis is zero. The orange lines show the error due to posterior excess kurtosis, when the skewness is zero. The solid lines show the error in the 2.5% quantile, and the dotted lines in the 97.5% quantile, that is, the lower and upper bound of a 95% confidence interval respectively. Error is relative to the width of the confidence interval.

where all quantities are conditioned on the observed data D_{obs}. Figure 2.1 shows the error given by equation 2.37 when V is chosen such that Rubin's Gaussian approximate confidence interval will have a width of 1. We see that posterior skewness may cause large errors in the confidence interval bounds, whereas posterior kurtosis has a much smaller effect. Hence we suggest more concern is required when the MLEs $\hat{\theta}^*$ are highly skewed and λ_ν is small, or when the MLEs $\hat{\theta}^*$ and their asymptotic variances U^* are highly correlated, than in the scenarios where posterior kurtosis may occur.

It is worth noting that throughout the above section we have assumed that the complete-data posterior distribution is correctly assumed to be t- or Gaussian-distributed by both Rubin's rules and Approximate Bayesian pooling. When this holds exactly, Approximate Bayesian pooling will perfectly approximate the posterior distribution. When this does not

hold exactly, additional error will be present for both Rubin's rules and for Approximate Bayesian pooling.

2.5 Simulations

2.5.1 Generating simulated data

We generated simulated data to compare the performance of various multiple imputation pooling methods, including Rubin's rules and approximate Bayesian pooling. We simulated data from three widely-used statistical regression models. For each model, and for a given n, we generated a binary covariate Z_i to be 0 for the first half of entries ($i = 1, \ldots, \lfloor (n+1)/2 \rfloor$) and 1 for the remaining entries. We then generated the main covariate X and the outcome Y for each setting:

(i) Linear regression: $X_i \sim N(0,1)$, $Y_i \sim N(\beta_0 + \beta_X X_i + \beta_Z Z_i, 1)$

(ii) Linear regression - lognormal covariate: $\log(X_i) \sim N(0,1)$, $Y_i \sim N(\beta_0 + \beta_X X_i + \beta_Z Z_i, 1)$

(iii) Logistic regression: $X_i \sim N(0,1)$, $\eta_i = \beta_0 + \beta_X X_i + \beta_Z Z_i$, $Y_i \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_i))$

(iv) Cox regression: $X_i \sim N(0,1)$, $Y_i \sim \text{Exponential}(h_0 \exp(\beta_X X_i + \beta_Z Z_i))$ for $h_0 = 1$.

For all settings we set $\beta_Z = 0.2$ and $\beta_0 = 0$. We set X_i to be missing for $i = 1, \ldots, n_{mis}$, so the data are missing completely at random (MCAR).

We simulated X_i from a Gaussian distribution in 3 of the 4 settings, in order to compare the pooling methods in a simple setting, where Rubin's rules seem appropriate, for a given n. We further considered a lognormal covariate X_i in linear regression (setting ii) to evaluate how the method performs when the data is skewed.

The distribution of the MLE $\hat{\beta}_X$ under the imputation scheme depends on the marginal distribution of X_{mis} via a non-linear transformation. In the simple case of linear regression with low levels of missingness in one covariate X, the transformation may be approximately linear. Hence the distribution of $\hat{\beta}_X$ may resemble the marginal posterior predictive distribution of X_{mis} given the observed X_{obs}, Y and Z. In general this cannot be expected to hold. Note also that the marginal distribution of X_{mis} given the observed data is not equal to the generative distribution of X. We might suggest that in small samples, skewness in the generative distribution of X may lead to skewness in the marginal distribution of

X_{mis}, and hence skewness in the distribution of $\hat{\beta}_X$ under the imputation scheme. In general, even if the generative distribution of X is symmetric, marginalising may introduce skewness. Similarly, even if the marginal distribution of X is symmetric, the non-linear transformation defining $\hat{\beta}_X$ given X may introduce skewness to the distribution of the MLEs under the imputation scheme. The presence of skewness in the MLEs $\hat{\beta}_X$ under repeated imputations may lead to error in the approximation performed by Rubin's rules.

Choice of β_X, n and n_{obs}

For each setting, we simulated data from up to 25 different scenarios. We varied the number of fully observed data points $n_{obs} = n - n_{mis}$, choosing $n_{obs} = 10, 20, 50, 100, 200$. In doing this we considered small samples of fully observed data, as we expect the assumptions for Rubin's rules are more likely to be violated when n is small.

For each value of n_{obs}, we chose 5 different values of the total number of observations, n. For $n_{obs} = 10$, we set $n = 11, 14, 20, 35, 100$, and for $n_{obs} = 20, 50, 100, 200$ we set n to be 2, 5, 10 and 20 times these values respectively. This corresponds to a proportion of missing X (prop_mis) of $9\%, 29\%, 50\%, 71\%$ and 90% respectively. With 5 choices of n_{obs} and 5 corresponding choices of n for each, this gives us 25 different scenarios for each setting. We chose a positive value of β_X for each value of n_{obs} such that a complete-case analysis using the corresponding model will have roughly 80% power to detect a non-zero effect (Supplementary Table A.1). Note this means for any given n_{obs}, we would expect the power to detect a non-zero effect to increase with n for the multiple imputation and Bayesian methods. This is because the number of complete observations is constant, while the number of partially missing observations is increasing, so, loosely speaking, the total information available to the model is increasing. We simulated 5000 datasets for each scenario, and compared the performance of the pooling methods in terms of power and coverage for β_X. We do not present results for Cox regression with $n_{obs} = 10$ or logistic regression with $n_{obs} = 10, 20$ due to the models not converging in a high proportion of these simulations, leaving just 20 and 15 scenarios for these settings respectively.

Imputing the missing data

We imputed the missing data from a Bayesian posterior which mirrors the data generating process. We assumed that X_i (or $\log(X_i)$) was known to be simulated from $N(0,1)$. We used this strong assumption with the aim to ensure any poor performance of the pooling rules is not due to inappropriate imputations. For the Cox scenario we also assumed that the data-generating model was exponential in the Bayesian imputation model. We set weakly-informative priors on the model parameters (including the error variance in linear regression, and log baseline hazard in the exponential model), given in Appendix A.2.1. These were intended to restrict the prior distribution to reasonable parameter values and improve stability. We then ran a model in Stan [6, 69], to sample from the observed-data posterior for each model. From this, we drew imputations for the missing X_i values from the first 10,000 iterations after warm up. We also generated 10,000 draws from the posterior for β_X in order to compare the pooling methods with a Bayesian approach (Bayes). We compared this with approximate Bayesian pooling, where the complete-data posterior is assumed to follow each of a t-distribution (ABpool_t) and a Gaussian distribution (ABpool_norm). We further compared with Rubin's rules where the complete-data posterior is assumed to follow a t-distribution (Rubin_t) [54], and a Gaussian distribution (Rubin_norm) [7]. We used 1000 imputations for Rubin's rules, thinning the Stan draws to ensure imputations are approximately independent. Due to constraints in time and computational power we did not include Bayesian inference after multiple imputation in our comparisons. Note however that for a certain choice of prior distributions, Bayesian inference after multiple imputation will generate draws from the same posterior as the above fully Bayesian approach (Bayes). We also compared with a complete-case analysis (CC), which is unbiased in these simple MCAR simulations, but in general may be biased.

2.5.2 Results

The coverage and power of the methods across the four settings are shown in Figures 2.2 and 2.3. We have displayed the results in groups of 5 scenarios with the same value of complete observations n_{obs} and β_X, separated by dotted vertical lines. Within each group, we display the results with $n = n_{obs} \times (1.1, 1.4, 2, 3.5, 10)$ from left to right. As expected,

the power often increases within each group from left to right, as the total sample size increases.

The pooling methods appear to converge towards giving the same power and coverage as n becomes larger, except when the proportion of missing X values (prop_mis) is large with a lognormal covariate. We would expect the results to converge, because for large values of n and a small proportion of missing data, the posterior distribution will be approximately Gaussian, where all methods should give the same result. A method with true coverage of 95% will have nominal coverage (which we define as observed coverage over 94.38%) for 97.5% of scenarios (83 out of 85 scenarios). Bayes, Rubin_t, ABpool_t, Rubin_norm and ABpool_norm show nominal coverage for 80, 85, 84, 75, and 77 of the 85 scenarios respectively. Hence Rubin_t and ABpool_t show appropriate coverage overall in the study, whereas the other methods give undercoverage. Some methods, Rubin_t in particular, give overcoverage in some scenarios, especially where the number of fully-observed samples is small.

Consider the results for linear regression with a Gaussian covariate (top left, Fig. 2.2 and 2.3). Rubin_norm and ABpool_norm show clear undercoverage for small $n_{obs} = 10$ and prop_mis $\leq 29\%$ (scen. 1 & 2), since these methods assume a Gaussian complete-data posterior, whereas the true complete-data posterior is t-distributed. In these scenarios ABpool_t, and more so Bayes, show slightly higher power than Rubin_t, which displays over-coverage. This difference is due to Rubin's rules over-estimating the posterior variance and kurtosis, thus giving wider confidence intervals, hence lower power and higher coverage than the other methods. We notice that the Bayesian model is biased towards smaller values of β_X, especially when n_{obs} is small. When prop_mis is very large ($\geq 71\%$), this also causes a notable bias in the multiple imputation (MI) methods. We also observe posterior skewness, which Rubin's rules does not account for, shifting the confidence intervals from Rubin's rules by roughly the same magnitude towards larger values. In these scenarios, Rubin_t gives over-coverage and higher power than the sampling methods. The higher power appears to be due to these two effects cancelling out. Rubin_norm gives the higher power of the two, and Rubin_t the higher coverage. As n_{obs} increases, the results from the different methods converge.

For linear regression with a log-normal covariate (top right, Fig. 2.2 & 2.3) we see that

the Bayesian method, followed by ABpool_t, perform best when $n_{obs} \leq 20$ is small. These methods give nominal coverage in at least 24 of 25 scenarios, and higher power than CC or Rubin_t, which again shows over-coverage. When n and prop_mis large (scenarios 15, 20, 25), both of Rubin's methods outperform ABpool and Bayes, giving nominal coverage and higher power. Much like large values of prop_mis with a Gaussian covariate above, this is caused by the shift due to not accounting for skewness negating the bias from the imputations.

Now consider logistic regression with a Gaussian covariate (bottom left, Fig. 2.2 & 2.3). The complete-case analysis shows slight under-coverage for small $n_{obs} = 50, 100$. No other methods appear to suffer from under-coverage. Bayes shows higher power for low prop_mis $\leq 29\%$ and small $n_{obs} = 50$, followed by the ABpool methods. Bayes and the ABpool methods converge as prop_mis increases. Rubin_t shows over-coverage throughout, and both Rubin methods show lower power, especially for small values of n_{obs}. The difference between ABpool and Rubin seems to be due to positive posterior skewness. For smaller values of prop_mis this is primarily caused by correlation between MLEs $\hat{\theta}^*$ and their asymptotic variances U^*, and for larger values by skewness in $\hat{\theta}^*$ (second and first term of Eq. 2.18 respectively). This biases the CIs from Rubin's rules towards smaller β_X values, decreasing power. Rubin_t rules also slightly overestimates the variance and kurtosis, which explains the higher coverage.

Finally consider Cox regression with a Gaussian covariate (bottom right, Fig. 2.2 & 2.3). All methods show nominal coverage in at least 18 of the 20 scenarios. When $n_{obs} = 20$ and prop_mis $\leq 29\%$, Bayes gives the highest power, followed by ABpool_norm and Rubin_norm giving similar, then CC, ABpool_t, and Rubin_t the lowest. Bayes likely gives the highest power because we included knowledge of a Weibull baseline hazard in the posterior, whereas the MI methods make no assumption about the baseline hazard. ABpool_norm outperforms Rubin_norm, due to a small positive posterior skewness caused by correlation between $\hat{\theta}^*$ and U^* (second term of Eq 2.18), which pushes the CI for Rubin_norm towards smaller values, decreasing power. Rubin_t is overly conservative, overestimating the variance and kurtosis, causing higher coverage and lower power, especially for smaller values of n_{obs} and prop_mis. For $n_{obs} \leq 100$ and large prop_mis $= 90\%$, Rubin_norm gives slightly higher power than Bayes and the ABpool methods. This is caused by a small neg-

ative posterior skewness due to skewness of $\hat{\theta}^*$ (first term of Eq 2.18), which pushes the CI for Rubin_norm towards larger values, and only occurs in simulations where the estimate is lower than the true value, thus increasing power. ABpool_norm and Rubin_norm give higher (or equal) power throughout than their t counterparts, while showing nominal or over-coverage.

Finally, we consider the required number of imputations m for each method. Supplementary Table A.2 shows the average power and coverage across all simulation settings and scenarios for each method, with different values of m. We see for Rubin_t that reducing m from 1000 to 100, 20 or 5 results in a loss of power of 0.3%, 2.1% or 10.1% respectively, and a slight loss of coverage, but not below nominal levels. For ABpool_t, reducing m from 10000 to 1000 or 100 results in a loss of coverage of 0.2% or 1.8% respectively, the latter giving under-coverage. A small increase in power is also observed. Similar losses in coverage and gains in power are observed for ABpool_norm and Bayes as m decreases. Hence we suggest roughly $m = 1000$ or more imputations may be required for ABpool and Bayes, and around $m = 100$ or more imputations may be required for Rubin's rules in these scenarios.

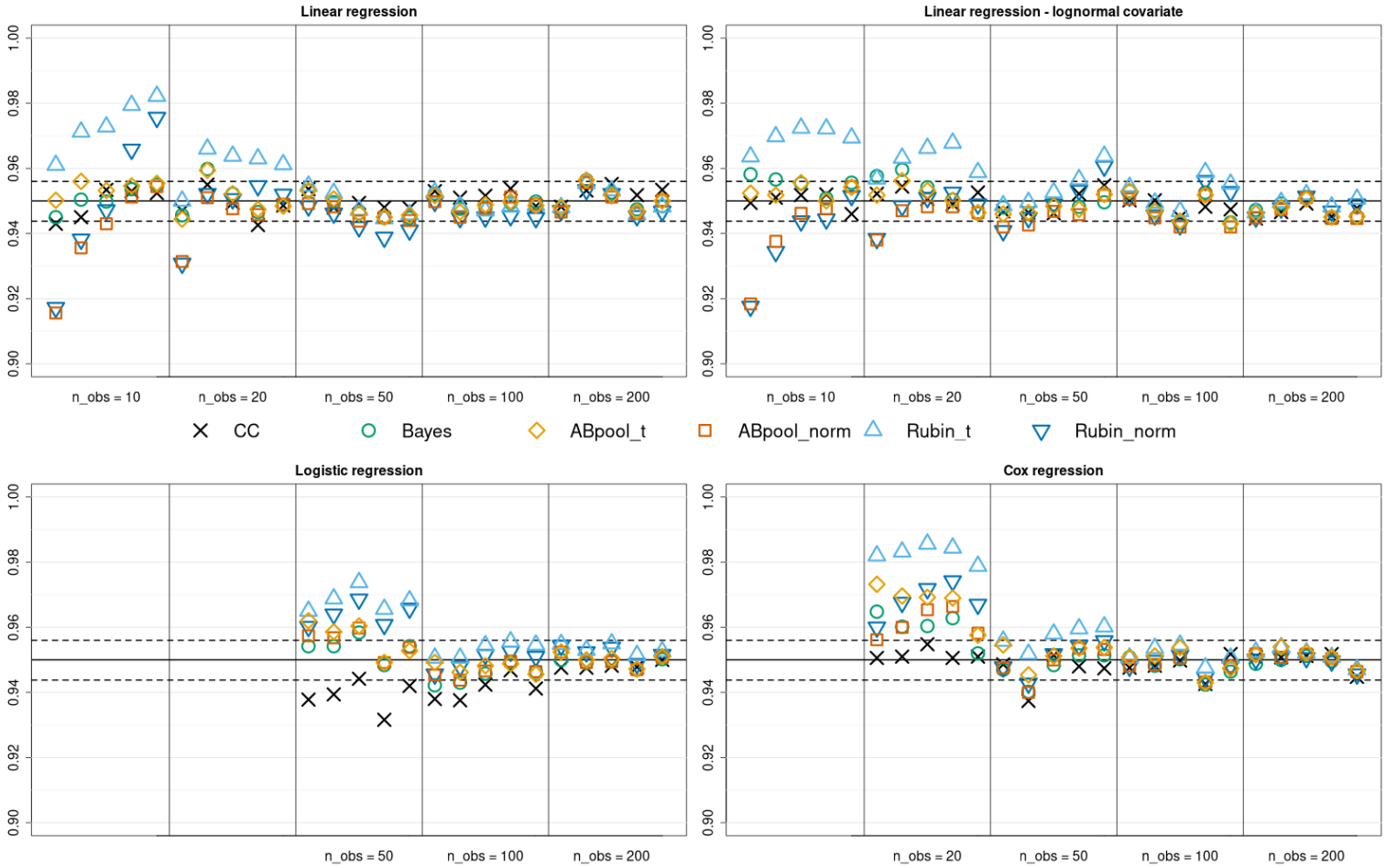


Figure 2.2: **Coverage for the pooling methods in each simulation scenario**
Coverage for a 95% uncertainty interval from the four settings: linear regression with a lognormal covariate (top right), and linear, logistic and Cox regression, each with a Gaussian covariate. For any value of n_obs, the 5 scenarios show a proportion of missing X values (prop_mis) of 9%, 29%, 50%, 71%, 90% from left to right. CC = complete-case analysis, Bayes = Bayesian analysis, ABpool = Approximate Bayesian pooling, and _t and _norm denote assuming a t- and Gaussian-distributed complete-data posterior respectively. The black horizontal line shows the nominal coverage level of 0.95. A method with nominal coverage level of 95% will lie between the dotted horizontal lines at 0.9438 and 0.956 for 95% of the scenarios on average.

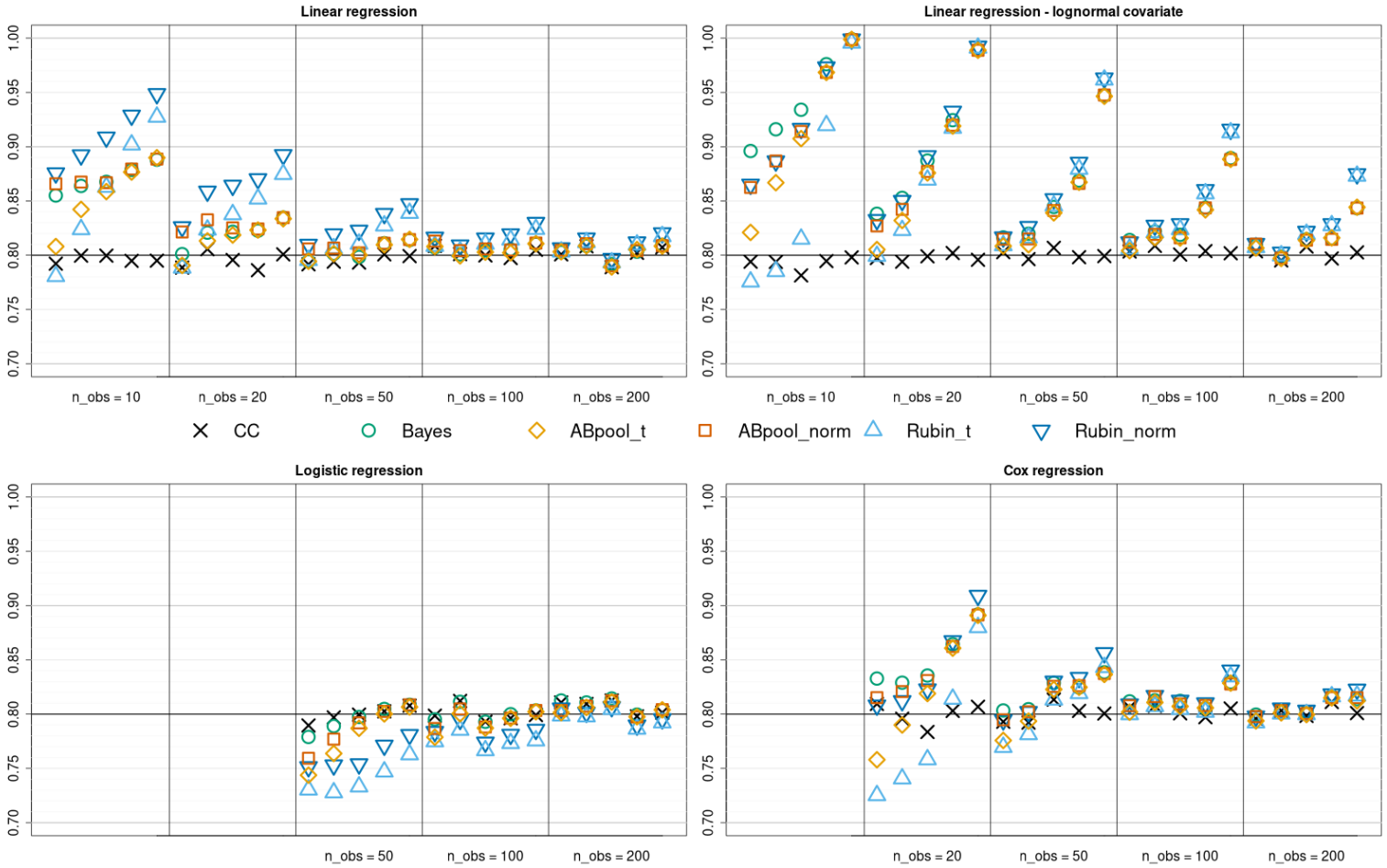


Figure 2.3: **Power for the pooling methods in each simulation scenario**
Power for a 95% uncertainty interval from the four settings: linear regression with a lognormal covariate (top right), and linear, logistic and Cox regression, each with a Gaussian covariate. For any value of n_obs, the 5 scenarios show a proportion of missing X values (prop_mis) of 9%, 29%, 50%, 71%, 90% from left to right. CC = complete-case analysis, Bayes = Bayesian analysis, ABpool = Approximate Bayesian pooling, and _t and _norm denote assuming a t- and Gaussian-distributed complete-data posterior respectively. Regression coefficients were chosen so that a complete-case analysis will have power of roughly 0.8 – denoted by the black horizontal line.

2.6 Real data

We include real data examples from three datasets, of which the first two are publicly available. The purpose of including these analyses is to give examples of real scenarios where the assumptions for Rubin's rules may be violated, and to compare the fit given by Rubin's rules, approximate Bayesian pooling, and Bayesian inference after multiple imputation in these scenarios.

2.6.1 HPV and cervical cancer correlation

Maucort-Boulch et al. (2008) investigated the correlation between human papillomavirus (HPV) prevalence and cervical cancer incidence in women across different age groups and countries [70]. The oldest age group (55-64 years) was then reanalysed by Plummer (2015) [71]. The data is available in the supplementary material of Plummer (2015) [71]. We reproduced the analysis of Plummer, considering women aged 55-64 years in 13 different countries.

We follow the notation used in Plummer [71]. In country $i = 1, \ldots, 13$, the outcome Y_i is the number of cervical cancer cases arising from T_i woman-years of follow up, and Z_i is the number of women infected with high-risk HPV in a sample size of N_i from the same country. The 13 observed values of N_i range from 37 to 696, Z_i range from 0 to 35, and Y_i from 16 to 710. Assume

$$Z_i \sim \text{Binomial}(N_i, \varphi_i) \tag{2.38}$$

$$Y_i \sim \text{Poisson}(\mu_i) \tag{2.39}$$

$$\mu_i = \lambda_i T_i \tag{2.40}$$

Further assume that the cervical cancer incidence λ_i relates to the high-risk HPV prevalence φ_i by

$$\log(\lambda_i) = \theta_1 + \theta_2 \varphi_i \tag{2.41}$$

We then focus on the slope parameter θ_2, which denotes the strength of association between HPV prevalence and cervical cancer incidence. Because the dose-response relationship is

speculative, a cut model approach (also known as modular inference) was used by Plummer [71]. The cut model prevents feedback from the cancer incidence data which may bias the estimates for the HPV prevalence parameters φ. We follow the cut method of Plummer, by first estimating the HPV prevalence φ_i using the HPV data alone, and then using the prevalence estimates in the model for cancer incidence Y_i [71]. We draw multiple imputations for the unknown HPV prevalences φ_i, which are treated as missing data [71]. We compare inference for θ_2 using Rubin's rules, approximate Bayesian pooling, and Bayesian inference after multiple imputation.

We draw imputations from a Bayesian binomial model (Eq. 2.38) in Stan [6, 69], drawing $m = 10,000$ imputations from the posterior distribution for φ. We then run a Poisson generalised linear model (GLM) regression (Eq. 2.39-2.41) given each imputation, and pool the results via Rubin's rules and approximate Bayesian pooling. Finally, we estimate via Bayesian inference after multiple imputation (MI_Bayes), sampling once from the cut posterior distribution for each of the 10,000 imputed datasets (Eq. 2.39-2.41). For each imputed dataset, we target the cut posterior in Stan for 200 warmup iterations, and take the following iteration as the sample. We pool samples across the imputed datasets, giving 10,000 samples in total.

Note though we consider both Rubin_norm and Rubin_t, Rubin_norm is more appropriate in this case, since inference on the parameters in a Poisson GLM would be based on a Gaussian approximation. However, when running the *pool* function in the R package mice [62], the default pooling method is Rubin_t with $n - p$ degrees of freedom, for sample size n when estimating p parameters ($n - p = 11$ in this case), unless the user specifies otherwise. Therefore although we would expect Rubin_norm to be preferred, corresponding to infinite degrees of freedom, we also include Rubin_t with 11 degrees of freedom in our comparisons. Q–Q plots of the MLEs for θ_1 and θ_2 after imputation indicate that the normal assumption made by Rubin's rules does not hold (Supplementary Figure A.1).

Figure 2.4 shows the results for the slope θ_2 under each pooling method. Given 10,000 imputations, we see that the approximate Bayesian pooling method (ABpool) gives results indistinguishable from Bayesian inference after multiple imputation (MI_Bayes), up to Monte-Carlo error. Generating the imputations took 11 seconds of runtime, after 1min 16sec to compile. ABpool took just 28 seconds to run, considerably faster than MI_Bayes,

Table 2.1: **Credible intervals for slope parameter θ_2**

	Posterior mean	2.5%	97.5%
MI_Bayes	13.73	9.53	19.47
ABpool_t	13.76	9.58	19.47
ABpool_norm	13.76	9.56	19.41
Rubin_t	13.76	-7892469.49	7892497.00
Rubin_norm	13.76	8.78	18.73

which took 1min 13s to compile and an additional 6min 36s to run. Rubin's rules does not match the cut posterior distribution given by MI_Bayes, as the posterior has positive skew (0.62), whereas Rubin's rules assumes a symmetric distribution. Rubin_t [54] – which is widely recommended over Rubin_norm – performs especially badly. The estimated degrees of freedom for Barnard and Rubin's t-distribution is extremely small at 0.18. This is due to very small complete-data degrees of freedom (11), and an extremely high fraction of missing information (0.99), which in combination make the degrees of freedom for Rubin_t extremely small. This gives the Rubin_t distribution extremely heavy tails, with both variance and excess kurtosis undefined.

The above differences can be observed in the credible intervals given by each method (Table 2.1). The Approximate Bayesian pooling methods give similar results to Bayesian inference after multiple imputation (MI_Bayes), whereas Rubin_norm produces a biased credible interval, and Rubin_t gives an extremely wide and completely uninformative credible interval.

By calculating each term from Eq. 2.30 & 2.37, we see that the posterior skewness (0.60) comes from skewness in the MLEs $\hat{\theta}^*$, and is the cause of the difference between Rubin_norm and the ABpool methods, pushing both ends of the Rubin_norm confidence interval towards higher values. Excess kurtosis and skewness due to correlation between $\hat{\theta}^*$ and U^* have a minimal effect.

Similar results, while slightly less extreme, are observed for θ_1, which is negatively skewed.

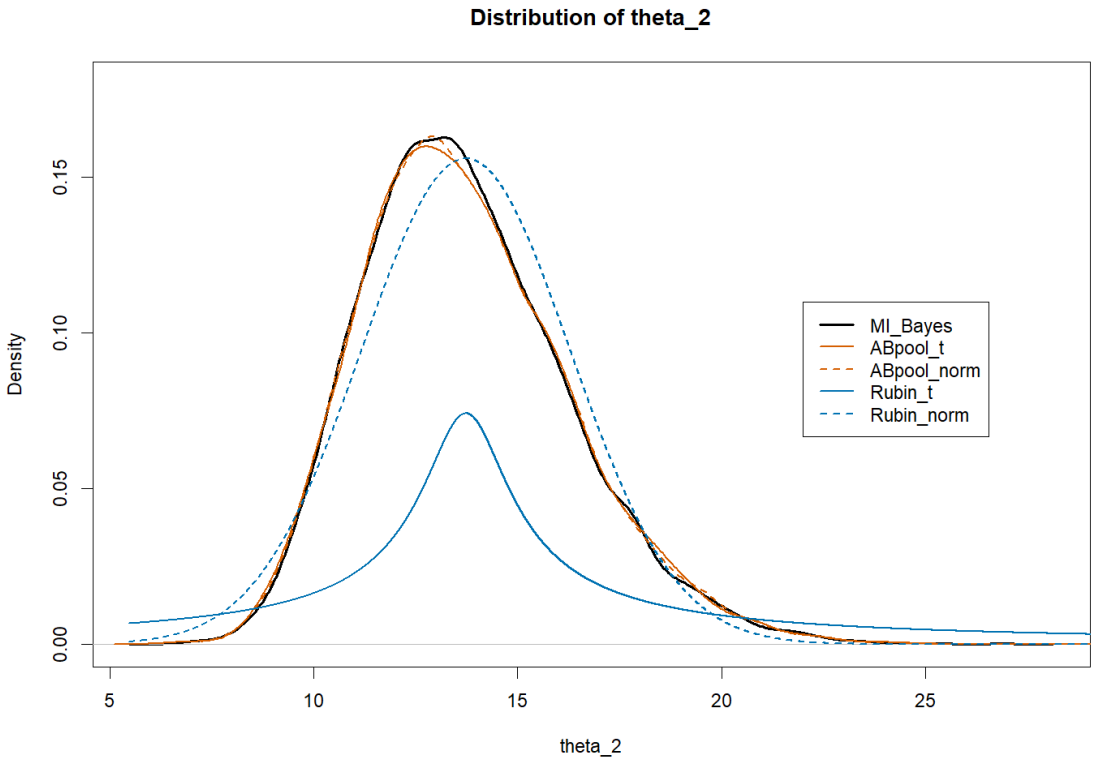


Figure 2.4: **The distribution of θ_2 estimated by Rubin's rules, approximate Bayesian pooling and Bayesian inference after multiple imputation.**
The posterior distribution for the slope θ_2. The black, orange and blue lines denote the posterior distribution estimated by Bayesian inference after multiple imputation (MI_Bayes), approximate Bayesian pooling (ABpool) and Rubin's rules (Rubin) respectively. For approximate Bayesian pooling and Rubin's rules, the solid line denotes the method assuming a t-distribution given complete data, and the dashed line denotes the method assuming a Gaussian distribution given complete data.

2.6.2 SUPPORT

We now consider data from the SUPPORT study, which aimed to improve end-of-life decision making and quality of life for dying patients [72]. A prognostic model was developed from the study to predict the risk of death in seriously ill hospitalised patients [73]. We use the support dataset which is publicly available in the Hmisc package [74], which is a random sample of 1000 patients from phases I & II of SUPPORT, of whom 668 died during follow-up. Our analysis is only intended to be used to compare multiple imputation pooling methods, not to provide clinical insight. The prognostic model used a Cox model for time-to-death including restricted cubic splines allowing non-linear effects, interaction terms, and stratification [73]. We use a simpler approach, analysing the time-to-death using the standard Cox model, including each variable untransformed as a linear predictor. The authors recommend replacing missing values with a single imputation "within the normal range" [73], imputing the same value for all individuals [75]. However, this may introduce bias, as it does not account for the added uncertainty due to the missing data [7]. We selected the variables in our model by starting with a Cox regression model containing many singly-imputed candidate variables, and performed backward deletion to include only those with a significant effect at the 5% level. Four covariates in our final model had missing values, with missingness rates of 37.8% for albumin (alb), 29.7% for bilirubin (bili), 25.3% for PaO_2/FiO_2 ratio (pafi), 0.3% for creatinine (crea). We generated multiple imputations by chained equations using the mice package in R [62, 76]. Since the covariates with missing values exhibited skewness, we generated multiple imputations of the missing values using type-1 predictive mean matching [77] with 5 candidate donors, based on an underlying linear regression model. The imputation model included all covariates present in the Cox analysis model, as well as the Nelson-Aalen estimator and event indicator, to account for the time-to-death [64]. We generated a large number of imputations ($m = 100{,}000$) to allow us to detect small differences between the approaches. We present results for the effects due to the four covariates with missing values, as they are likely to be most impacted by the imputations and choice of pooling scheme.

Q–Q plots from as few as 200 imputations indicate that the MLEs for bili and crea are not approximately Gaussian distributed (Supp. Fig. A.3 & A.4), hence there may be posterior

skewness present, which could introduce bias to Rubin's rules for these parameters. Estimating moments from the full 100,000 imputations, we see the MLEs for the effects due to bili and crea have skewness of -0.79 and -0.95, and kurtosis of 0.63 and 1.09 respectively. This is likely due to the covariates being non-Gaussian themselves, with skewness of 4.48 and 2.81, and kurtosis of 28.25 and 11.91 respectively. However the Q–Q plots for the posterior samples using Approximate Bayesian pooling with a t assumption show no clear deviation from the Gaussian distribution (Supp. Fig. A.3 & A.5). Both bili and crea effects show no noticeable kurtosis (0.03 and 0.01), nor crea a noticeable skew (-0.04), however the posterior for bili does show a slight skew (-0.15). This can be observed in the posterior density plots (Fig. 2.5). The marginal posteriors for the alb, crea and pafi effects all show a clear difference between the results given by single imputation and multiple imputation, but no noticeable difference between the ABpool methods and Rubin's rules (except that due to Monte-Carlo error). For the bili effect however, there is a slight, but noticeable negative skew in the approximate ABpool posterior, caused by the skew in the MLEs, which Rubin's rules is unable to account for (Fig. 2.5). Posterior means, medians and 95% credible intervals for these 4 variables under each method are given in Table 2.2. We compare the single imputation method recommended by the SUPPORT authors (Single_imp) with Rubin's rules using a Gaussian and t assumption (Rubin_norm [7] and Rubin_t [54]), and approximate Bayesian pooling with a Gaussian and t assumption (ABpool_norm and ABpool_t). The results show that the use of multiple imputation changes the parameter estimates slightly compared to the single imputation method suggested by the SUPPORT authors. Further, while differences are very small, the slight skewness we observe in the bili effect (Fig. 2.5) leads to the 95% credible interval using Approximate Bayesian pooling being shifted slightly towards smaller values than the credible interval given by Rubin's rules, and the posterior median being slightly larger than the posterior mean. Similarly, a negative skewness in the MLEs and causes a slight skewness in the posterior for the crea effect, which leads to the Rubin_t CI giving slightly higher values than ABpool. For the alb effect, Rubin_t rules estimates the variance and excess kurtosis to be slightly larger than ABpool_t, causing a fractionally wider confidence interval. The pafi effect shows smaller differences, which may be simply due to randomness in the Approximate Bayesian sampling.

Table 2.2: **Posterior summaries for SUPPORT**

	Method	Mean	Median	2.5%	97.5%
alb	Single_imp	-0.1778	-0.1778	-0.3077	-0.0479
	Rubin_t	-0.1945	-0.1945	-0.3446	-0.0444
	Rubin_norm	-0.1945	-0.1945	-0.3442	-0.0449
	ABpool_t	-0.1944	-0.1943	-0.3437	-0.0451
	ABpool_norm	-0.1945	-0.1944	-0.3450	-0.0446
bili	Single_imp	0.0496	0.0496	0.0325	0.0668
	Rubin_t	0.0473	0.0473	0.0275	0.0672
	Rubin_norm	0.0473	0.0473	0.0276	0.0671
	ABpool_t	0.0474	0.0476	0.0268	0.0664
	ABpool_norm	0.0473	0.0477	0.0266	0.0663
pafi	Single_imp	-0.0015	-0.0015	-0.0023	-0.0006
	Rubin_t	-0.0012	-0.0012	-0.0021	-0.0003
	Rubin_norm	-0.0012	-0.0012	-0.0021	-0.0003
	ABpool_t	-0.0012	-0.0012	-0.0021	-0.0003
	ABpool_norm	-0.0012	-0.0012	-0.0021	-0.0003
crea	Single_imp	0.1024	0.1024	0.0594	0.1453
	Rubin_t	0.1008	0.1008	0.0550	0.1466
	Rubin_norm	0.1008	0.1008	0.0551	0.1465
	ABpool_t	0.1007	0.1008	0.0547	0.1458
	ABpool_norm	0.1008	0.1009	0.0545	0.1460

Posterior mean, median and 95% credible interval for four of the parameters in the Cox model on the SUPPORT dataset. We compare the single imputation model recommended by the authors (Single_imp) with multiple imputation pooled by Rubin's rules and Approximate Bayesian pooling (ABpool), each with a Gaussian (norm) and t assumption.

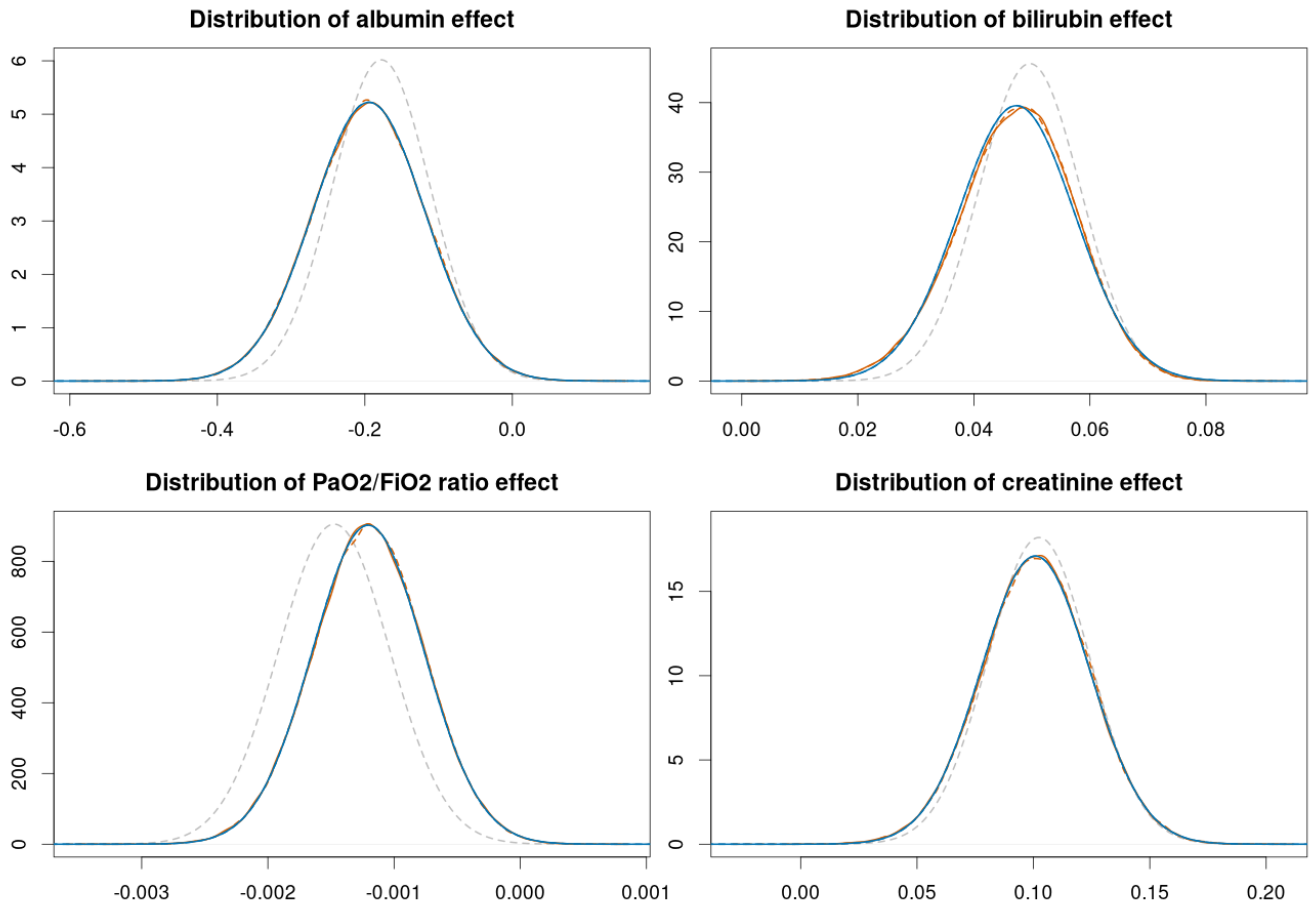


Figure 2.5: **The posterior distribution of the SUPPORT parameters.**
The posterior distribution for the alb, bili, pafi and crea parameters. The grey, orange and blue lines denote the posterior distribution estimated by single imputation, approximate Bayesian pooling and Rubin's rules respectively. Solid lines denote methods assuming a t-distribution given complete data, and dashed lines denote methods assuming a Gaussian distribution given complete data.

2.6.3 COVID-19 vaccine efficacy

Method

Our final example, which was the motivating example for our Approximate Bayesian pooling method, comes from an analysis of the COV002 trial [4]. The analysis is presented in detail in Chapter 4. We aimed to evaluate how COVID-19 vaccine efficacy relates to the vaccine-induced antibody level present at the time of exposure to the disease.

The data included 4605 individuals vaccinated with the ChAdOx1 nCoV-19 vaccine (vaccine arm) and 4423 control individuals. In the vaccine and control arms respectively, 71 (1.5%) and 217 (4.9%) individuals developed primary symptomatic COVID-19, the event of interest. We assumed antibody levels in the control arm were 0, and modelled antibody levels in the vaccine arm over time, imputing the estimated true antibody levels into a model for the risk of infection. A fully Bayesian approach would not converge, likely due to the model including many random effects, and non-linear transformations, and the sparse availability of antibody data. Therefore we employed multiple imputation to overcome this. Briefly, we assumed antibody levels decayed exponentially, and developed a Bayesian hierarchical model to predict the intercept and slope of the log-antibody levels for each individual. For each posterior draw of the log-antibody intercepts and slopes, we then derived the predicted antibody levels at each event time. Doing this across 60,000 posterior draws gave us 60,000 multiply imputed datasets.

In the post-imputation analysis, we included the predicted vaccine-induced antibody levels as a time-varying covariate in a Cox proportional hazards model (effect parameter γ), where the outcome was the time until primary symptomatic COVID-19 infection. We included COVID-19 vaccination (effect parameter ζ) and other covariates in the Cox model which may affect time-to-infection, and stratified by study site. The estimated vaccine efficacy at antibody level A was then given by

$$\text{VE}(A) = 1 - \exp(\gamma A + \zeta) \tag{2.42}$$

for slope γ and intercept ζ of the effect due to vaccine-induced antibodies. We derived estimates and uncertainty intervals for $\gamma A + \zeta$ at each value of A, which were then trans-

formed into estimates and intervals for VE(A) (Eq. 2.42), as this scale is likely to be better approximated by a Gaussian [57, 56]. We compared uncertainty intervals generated using Rubin's rules and Approximate Bayesian pooling. We did not attempt Bayesian inference after multiple imputation due to computational constraints, and the comparative simplicity of the Cox model. The imputation and analysis methods are explained in greater detail in Chapter 4. These references give results from the Approximate Bayesian pooling method assuming a Gaussian complete-data posterior (ABpool_norm) only.

Results

Supplementary Figure A.8 shows that the MLEs and approximate posterior draws for both γ and ζ are non-Gaussian, showing negative and positive skew respectively.

Results are given in Supplementary Table A.3. Figure 2.6 shows the relationship between the estimated vaccine efficacy and the antibody level at the time of exposure. At antibody levels around or above 100 BAU/mL, the pooling methods give different credible intervals. The approximate Bayesian pooling methods give a tighter credible interval, concentrating on higher estimates of vaccine efficacy. Rubin's rules gives much a wider credible interval, including the possibility of vaccine efficacy being much lower.

The posterior distribution for γ estimated by Approximate Bayesian pooling shows a strong negative skew (-0.93), which Rubin's rules is unable to account for (Supp. Fig. A.9). The posterior skewness is caused by skewness in the MLE, and covariance between the MLE and asymptotic variance, contributing -0.65 and -0.28 respectively. There is also a noticeable posterior excess kurtosis, caused by kurtosis in the MLE, and covariance between the squared deviation of the MLE from its mean, and the asymptotic variance, contributing 0.97 and 0.55 respectively to the kurtosis of 1.55. The effect of the posterior skewness on the credible interval is 4 times larger than the effect due to the posterior excess kurtosis, biasing Rubin's CI towards larger values of γ. Similar results of opposite sign (i.e. with positive skewness instead of negative), are observed for ζ, and the results for vaccine efficacy interpolate the two.

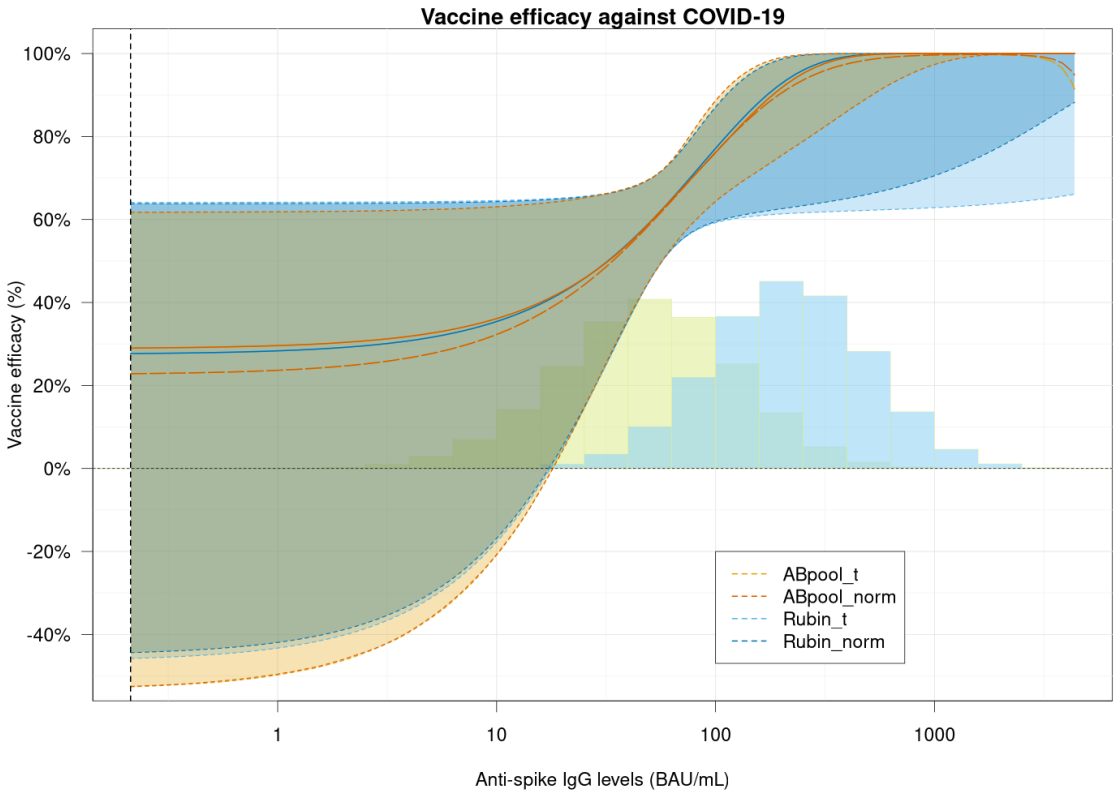


Figure 2.6: **Vaccine efficacy against primary symptomatic COVID-19 infection, plotted against antibody level at time of exposure**
Vaccine efficacy (VE) (%) plotted against anti-spike IgG antibody levels (BAU/mL) at time of exposure on a log scale. Solid curves denote posterior estimates, dotted lines denote 95% credible regions. For the ABpool approaches, the posterior median is plotted as a solid line, and the posterior mean as a long-dashed line. The orange and red curves show the results given by Approximate Bayesian pooling assuming a t and Gaussian (norm) complete-data posterior respectively. These two methods give indistinguishable results. The light and dark blue curves show results given by Rubin's rules assuming a t and Gaussian (norm) complete-data posterior respectively. For vaccine efficacy at large antibody levels, Rubin's rules shows far greater uncertainty as to the lower bound of the 95% credible interval.
The vertical dotted line shows the lower limit of quantification (LLOQ), 0.21 BAU/mL. We included histograms to show the distribution of antibody levels in the study. The blue histogram shows the distribution of the predicted anti-spike IgG antibody levels in the study at 28 days since the second dose, and the green histogram at 182 days post second dose. The two histograms overlap. The heights of the histograms are not to scale.

2.7 Discussion

In this chapter we calculated the error in confidence intervals derived by Rubin's rules, when certain key assumptions do not apply. Most problematically, the presence of posterior skewness may cause Rubin's rules to lose power and/or coverage. We presented an Approximate Bayesian pooling (ABpool) method which uses the MLEs $\hat{\theta}^*$ and their variances U^* to approximate the Bayesian posterior, while making weaker assumptions than Rubin's rules, and allowing the observed-data posterior to exhibit skewness. We compared the performance of Rubin's rules with ABpool, and with a fully Bayesian approach, on both simulated data and example datasets.

The derivation of Rubin's pooling rules relies on the assumption that the distribution of the maximum likelihood estimators (MLEs) $\hat{\theta}^*$ over imputations given the observed data are Gaussian, and that the asymptotic variance estimates U^* have negligible variability compared to B, the variance of the MLEs $\hat{\theta}^*$. These assumptions lead to a t-distributed approximation to the observed-data posterior. Our theory suggests this may be a poor approximation when the observed-data posterior is skewed. Posterior skewness may be caused by the distribution of the MLEs $\hat{\theta}^*$ exhibiting skewness, or by correlation between the MLEs $\hat{\theta}^*$ and asymptotic variances U^*. The MLEs $\hat{\theta}^*$ are more likely to be skewed when the data itself is skewed, and in small samples. In some other cases transforming the analysis to an appropriate scale may help.

For the simulation scenarios with the smallest sample sizes n, we observe that a Bayesian approach outperforms our ABpool_t, which itself outperforms Rubin_t. All three methods give valid coverage. In these scenarios Rubin_t tends to give overly conservative CIs, exhibiting over-coverage and a loss of power. We recommend a Bayesian approach in such scenarios. Where a user may choose not to apply a Bayesian method, ABpool_t performs best for linear regression, especially in the presence of skewed data. ABpool_norm and Rubin_norm perform best for logistic and Cox regression, although these pooling methods give under-coverage for linear regression.

Peculiarly, we observed that in some linear regression scenarios with a high proportion of missing data, Rubin's rules tended to give higher power than the ABpool and Bayesian approaches. Similarly in some Cox regression scenarios with large prop_mis, Rubin_norm

outperformed the other approaches. This was due to the Bayesian imputation method giving biased estimates for the parameter β_X, leading to bias in the MLEs, which was nullified by an opposite bias in the CI from Rubin's rules due to posterior skewness. This appears to be a fortunate coincidence in favour of Rubin's rules. However it is unclear whether this is due to a deeper reason, and may be expected to occur in other scenarios too.

In the small-sample HPV and cervical cancer correlation example, Rubin_t gives an extremely wide, therefore completely uninformative confidence interval, whereas the ABpool methods and Rubin_norm give results which rule out a null effect. Rubin_norm gives a biased result since it does not account for posterior skewness, so ABpool_t or Bayesian inference after multiple imputation (MI_Bayes) is preferred. ABpool_t gives a very similar result to MI_Bayes, while requiring a much shorter runtime.

Even in large-sample real data examples, we observe differences between the methods. When the data is extremely skewed, and the proportion of missing values is high, differences may still occur. In both the SUPPORT example ($n = 1000$), and the COVID-19 vaccine example ($n = 9028$), we observe a bias in the CI given by Rubin's rules. This is due to highly skewed data causing a skew in the posterior distribution, introducing bias to the Rubin_t confidence interval. The shift for the bili effect in the SUPPORT example is very small, however the shift for the antibody effect in the COVID-19 example is large – shifts of around 1% and 10% of the width of the confidence intervals in the two examples respectively. The SUPPORT example is a classic situation for multiple imputation, imputing missing values in a dataset. In both the HPV cancer correlation and COVID-19 vaccine examples, we impute an unknown latent true value, using observed levels to inform the imputations. The imputations are skewed in both cases, and the subsequent analysis has a large fraction of missing information. In the COVID-19 example, this led to a large skew in the posterior distribution for the antibody effect and vaccine efficacy. Applying Rubin's rules gives a biased CI for the antibody effect γ, which also shifted the CI for vaccine efficacy towards lower values.

An alternative method to deal with skewness in multiple imputation is combining MI with the bootstrap. Brand et al. (2018) compared bootstrap and multiple imputation methods to analyse skewed cost-effectiveness data [59]. They found making a single imputa-

tion (with appropriate uncertainty) within a bootstrap performed best in their simulation study. They suggest the standard application of Rubin's rules without a bootstrap was relatively robust and an acceptable alternative. Imputing within a bootstrap is a suitable alternative method when posterior skewness is present.

Both Rubin's rules and the Approximate Bayesian pooling method rely on the assumption that the complete-data posterior is t-distributed (or Gaussian). This is a weaker assumption than assuming the observed-data posterior is t-distributed (or Gaussian), which Rubin's rules also assumes, but Approximate Bayesian pooling does not. The observed-data posterior is generated by marginalising over the missing values in the complete-data posterior, which may introduce skewness not present in the complete-data posterior. However, there may be some small-sample cases where the complete-data posterior may also exhibit skewness. In these cases both methods may be susceptible to bias, and a Bayesian approach should be employed.

In our simulation study and real-data examples, Rubin_t was observed to be relatively robust, although not in all cases. In our simulation studies, Rubin_t still always gave appropriate coverage, even when key assumptions did not apply, for example due to skewness in the MLEs $\hat{\theta}^*$ in small-sample cases. Here Rubin_t was overly-conservative and lost power, due to small degrees of freedom. This meant despite the bias due to skewness, nominal coverage was still achieved. However, in the COVID-19 example and SUPPORT example, posterior skewness meant that the CI generated by Rubin_t did not cover the CrI generated by ABpool_t.

It is worth noting that issues with the imputation model will affect the post-imputation results. In an earlier version of the simulation study (not reported), we used predictive mean matching to impute the missing values, which is not appropriate in small samples [55]. This meant all MI methods gave dramatic undercoverage. Since Rubin_t is the most conservative, it gave the least pronounced undercoverage. Examples such as this suggest some analysts may prefer Rubin_t in general since it tends to give more conservative CIs than other methods in problematic scenarios. However, we suggest a better approach would be to fix the issue with the imputation model, and then use whichever pooling method is most appropriate. In our simulation study we similarly observed a bias in the imputation model, which led to Rubin's rules outperforming ABpool in some scenarios.

We suggest analysts consider running a small simulation study to check the performance of their imputation model and subsequent pooling when analysing data which may be problematic – in situations such as skewness, small sample size, a high proportion of missing values, imputing a latent variable, or a novel analysis method.

We have implicitly assumed throughout this chapter that the imputer and the analyst are the same person. When this is not the case, issues of uncongeniality may arise. Loosely speaking, the analysis procedure is uncongenial to the imputation model, if they differ in their assumptions about the data generating process. Meng (1994) [78] and subsequent papers have explored this issue, and noted that Rubin's rules may give invalid inference under uncongeniality. We expect approximate Bayesian pooling to have similar limitations in this case.

The main disadvantage of Approximate Bayesian pooling and Bayesian inference after multiple imputation compared to Rubin's rules is the need to generate a larger number of imputations. Previous studies have suggested Rubin's rules may require more than $m = 100$ imputations when the fraction of missing information (fmi) is very large [79, 80], although around $m = 20$ may suffice when the fmi is small. Gelman et al. [63] suggest around 100 to 2000 independent draws are required for most Bayesian posterior summaries, so a similar m may be required for ABpool and MI_Bayes. Our simulation study suggested around $m = 1000$ samples may be required for ABpool and Bayes, whereas $m = 100$ imputations may be required for Rubin's rules. Hence the number of imputations required for the sampling methods may be around 10 times that needed for Rubin's rules. Where multiply imputed datasets are shared publicly for use by many different analysts, the requirement for larger numbers of imputations when applying ABpool and MI_Bayes may be undesirable, due to the need to share many more imputed datasets. In this paper we only considered the case where $J = 1$ sample is drawn from each complete-data posterior distribution. Drawing more samples J from each posterior may allow fewer imputations m to be generated – Zhou et al. use $m = 100$, $J = 10000$ [58] – however further simulations would be required to verify this. Notably the under-coverage observed by Rashid et al. suggest that using a very small m would be unwise.

Where imputations are costly to generate or store, e.g. due to computational constraints, Rubin_t may be preferred in well-behaved scenarios – since fewer imputations are required,

and all methods will give near-identical results. In scenarios where Rubin_t may be biased, or overly conservative and lose power, a sampling approach may be used. Where a fully Bayesian approach can be applied, we would recommend this over an Approximate Bayesian pooling scheme. However Bayesian methods tend to require the analyst to code an appropriate MCMC or HMC based sampler, which may be computationally prohibitive for some applied researchers. Both Rubin's rules and Approximate Bayesian pooling allow the post-imputation analysis to be performed using maximum likelihood for completed datasets, which are familiar to many applied researchers. There are many well maintained packages for applying Rubin's rules, such as mice in R [62, 76]. For ABpool, the pooling step can be implemented by running a simple function, independent of the analysis method. In cases where Rubin's rules may give overly conservative or biased inferences, Approximate Bayesian pooling is a suitable alternative.

Chapter 3

Correlates of protection models and the need to account for antibody decay

Abstract

Establishing a correlate of protection, that is, an immune marker which is associated with the risk of a disease outcome, is crucial to vaccine development. This allows future vaccines to be licensed based on immunological data alone, avoiding the need for large, expensive, and impractical phase 3 efficacy trials. In order to be useful, such a relationship must be consistent across different vaccine trials.

Standard methods to estimate correlates of protection relate immune maker levels at a fixed date after vaccination to subsequent infection, in order to understand how the peak immune response protects against infection. We consider different trials of the same vaccine, where induced antibody levels and subsequent protection over time are the same in each trial, but the timing of cases differ, for example due to different timings of an infection wave in a pandemic. When estimating using a standard method, the results may not be consistent across vaccine trials. We mathematically derive the difference in the estimated correlate of protection curve between trials in a simplified scenario.

We further test whether the standard method gives consistent results between trials in a simulation study. We consider two scenarios: a pandemic setting where waves of infection vary in time across different trials, and an endemic setting where attack rates in the control group vary between trials, and the correlates of protection analysis is run once a certain number of cases is reached. In both settings, estimates of the correlate of protection curve for the same vaccine were inconsistent between trials. Estimated levels of protection at the same initial antibody response were lower in trials where cases occurred later on average. To overcome this problem, antibody decay should be accounted for in the model. We compare three methods to estimate correlates of protection, using data from a COVID-19 vaccine efficacy trial. The multiple imputation joint model, which accounted for both antibody decay and measurement error, estimated the largest difference between protection levels at different antibody responses. This method is later presented in Chapter 4.

3.1 Introduction

Establishing a correlate of protection, that is, an immune marker which is associated with clinical risk of a disease outcome, is a crucial part of vaccine research. In order to approve a new vaccine, large phase 3 trials are often run to measure vaccine efficacy against a clinical outcome. However, due to low prevalence of many diseases, these often require very large sample sizes and a long period of follow up, making such trials very expensive and time consuming. Further, once a vaccine has been approved for use, placebo-controlled phase 3 trials may be impractical, and considered unethical. Establishing a correlate of protection may allow new vaccines to be licensed based on immunological data alone, when large efficacy trials are not feasible. Immunological trials require a much smaller sample size and, in some cases, shorter length of follow up. The relationship between the immune marker and clinical risk can be established in one or more phase 3 trials by comparing immune marker values with occurrence of the disease outcome. Where possible, a causal relationship should be established, which is demonstrated across different trials in different populations. This can then be used to predict efficacy of a new candidate vaccine based on the induced immune marker levels.

When establishing correlates of protection from a phase 3 trial, observed immune marker values measured at a study visit around the time of peak immune response are compared with subsequent occurrence of the disease outcome. However, the results from this method may lack generalisability. In a pandemic, such as the COVID-19 pandemic, the timing of infection waves varied in different populations, causing the timings of cases to vary between trials. Similarly different trials may be of different lengths, either by design, or since different infection rates mean the threshold to run the analysis will be reached at different times. This too will cause the timings of cases to vary between trials. Since antibody levels, and thus protection, decay between the time of peak antibody response and the time of exposure, trials with later cases will observe lower vaccine efficacy at the same initial antibody level. Hence the estimated relationship may vary in different trials performed at different times or locations, due to differing background rates of infection, and the results may not be generalisable. Relating observed antibody levels to risk of infection may be subject to further bias, as it fails to account for antibody measurement

error [1, 2]. If the outcome of interest is the relationship between antibody levels at the time of exposure and protection, then failing to account for waning antibody levels may introduce further bias [29].

Some studies have related clinical risk to immune marker levels at the time of exposure. To estimate COVID-19 correlates of protection, Wei et al. (2022, 2023) related a recent antibody measurement to the risk of returning a positive COVID-19 test [81, 82]. However, they did not account for measurement error, or for decay of antibody levels between the antibody measurement and COVID-19 test. Follmann et al. (2023) applied estimates of antibody decay rates to antibody measurements at peak response, extrapolating to predict the level of antibody present at exposure [47]. However, they did not account for measurement error, or for differences in the rate of decay between individuals. Seekircher et al. (2024) use a joint model for longitudinal and time-to-event data to relate modelled COVID-19 antibody levels at the time of exposure to risk of infection [51], using the R package JMBayes2 [40]. They do not account for time-since-vaccination in the model, only calendar time. Papadopoulos et al. (2025) [52] use JMBayes2 [40] to fit a joint model to post-booster antibody levels and risk of COVID-19 infection. Their model does not account for changing background rates of risk of COVID-19 infection with calendar time, nor does the trial contain a control group, which would be needed to estimate vaccine efficacy. White et al. (2014, 2015) use a two-stage joint model to estimate correlates of protection for a malaria vaccine [48, 49]. Their model relates the mean estimate of antibody levels over time to risk of infection, and without accounting for uncertainty in the estimated antibody levels. Such simple (or "naïve") two-stage approaches have been shown to give biased estimates [44]. Fu and Gilbert (2017) develop a joint model for longitudinal and time-to-event data under case-cohort sampling using inverse probability weights [50], applied to a HIV vaccine trial. Their model requires denser samples of the immune marker, as it estimates the trajectory for each individual nonparametrically. Huang and Follmann (2025) [53] develop a two-stage joint model to estimate COVID-19 exposure-proximal correlates of protection. Their method outperforms a simple two-stage approach, and shows minimal bias, although they do not consider performance in terms of coverage in their simulation study. Their model does not account for informative censoring of immune marker data at infection when estimating the immune marker trajectories.

There are multiple challenges presented by vaccine trial data which should ideally all be accounted for in a correlates of protection model: (i) measurement error in the longitudinal variable, (ii) changes in the longitudinal variable over time, (iii) the uncertainty in estimating the true values of the longitudinal variable over time, (iv) informative censoring of the longitudinal data at infection, (v) two different time scales (baseline hazard varies with calendar times and antibody levels vary with time-since-vaccination). Failing to account for any of these challenges may lead to bias.

In this chapter we consider the impact of antibody waning and vaccine efficacy decay on the estimated correlates of protection curve. We show that where cases occur at different times since vaccination on average between different trials, the estimated correlate of protection curve will be inconsistent, if antibody waning is not accounted for in the model.

In section 3.2 we outline previous methods to estimate correlates of protection, including both methods which do not account for antibody decay and methods which do. In section 3.3 we examine the properties of different functional forms for the antibody effect. We then calculate the impact of later events in a trial on the estimated correlate of protection curve in a simplified example, when antibody decay is not accounted for. In section 3.4 we perform a simulation study to test whether standard correlates of protection methods give consistent results in trials with different baseline risk of infection. We compare the results from three correlates of protection methods on a COVID-19 vaccine trial in section 3.5, comparing a method which accounts for antibody decay and measurement error to previous approaches. We discuss our findings in section 3.6.

3.2 Previous methods

Here we consider methods which have been proposed to estimate correlates of protection from a randomised phase 3 vaccine efficacy trial, such as the recent phase 3 trials of COVID-19 vaccines. In these trials, individuals were randomised to a control arm and a vaccine arm, and followed up to observe the time until infection. Blood samples were taken from participants at various study visits, including a visit around the time of peak antibody response after vaccination (which we refer to as the peak visit), and possibly also subsequent visits. At the end of the study, blood from the peak visit (and possibly also

later visits) was analysed to measure immune marker levels, for those participants who were later infected in the study, and from a random sample of uninfected participants. In this case-cohort design, the probability of immune marker data availability depends on whether the individual was later infected. These studies aimed to relate vaccine-induced immune marker levels to the risk of subsequent infection. Since infection will likely cause an increase in immune marker levels, measurements of immune markers after the event of interest (infection) are discounted. If trajectories of immune markers are estimated, these are censored at the time of infection or dropout.

3.2.1 Time-fixed approach

One common approach is to use two-phase sampling regression to estimate covariate-adjusted vaccine efficacy, using immune marker data from the peak visit only.

First, the probability of immune marker (antibody) availability for each individual at the peak visit is estimated in a logistic regression model. This model includes the indicator of infection during the study as a binary covariate, and any other covariates which may relate to the probability that an individual's peak visit blood sample is selected for laboratory measurement of immune markers.

Second, a fixed time of interest t_0 is chosen, and a logistic or Cox regression model is used to estimate the risk of each individual experiencing the clinical outcome before t_0, where the model is weighted by the inverse probability of antibody availability, and adjusted for the observed marker value, and covariates. A separate model may be used to estimate the same risk in the control arm. Vaccine efficacy at antibody level A is then estimated as one minus the relative risk, where the relative risk is the average predicted risk in the vaccinated arm with the antibody levels all set to A, divided by the average predicted risk in the control arm. This whole process is bootstrapped, in order to account for the uncertainty in estimated inverse probability weights, as well as parameter estimates.

To formulate this mathematically, let $Z_i = 1$ if individual i is assigned to the vaccine arm, and $Z_i = 0$ if individual i is assigned to the control arm. First draw a bootstrap sample $b = 1, \ldots, B$ of the dataset. Order the bootstrap dataset such that $Z_i = 1$ for $i = 1, \ldots, n$ and $Z_j = 0$ for $j = n+1, \ldots, N$. Then run a logistic regression to estimate the probability of antibody availability for each individual in the vaccine arm. Next, run a (Cox or

logistic) regression model for the risk of infection, including observed antibody levels as a predictor, and weighting by the estimated inverse probability of antibody availability. Let $\hat{r}(X_i, z, A)$ be the predicted probability of experiencing the event before time t_0 for an individual with covariates X_i, assigned to treatment z, with antibody response A. For a vaccine trial where the control participants have no prior exposure to the pathogen of interest and hence have no antibodies present, the risk in the control arm $\hat{r}(X_i, 0, 0)$ may be estimated using a separate regression model. Alternatively, the model for the risk of infection may include both the control arm and the vaccine arm, with antibody levels set to 0 and inverse probability weights of 1 in the control arm, and a direct effect due to vaccination. Vaccine efficacy at antibody level A is then estimated by

$$\text{VE}(A) = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{r}(X_i, 1, A)}{\frac{1}{N-n}\sum_{j=n+1}^{N} \hat{r}(X_j, 0, 0)} \tag{3.1}$$

The mean estimate for vaccine efficacy can be calculated on the original (non-bootstrapped) dataset, and bootstrap confidence intervals can be generated from the bootstrap estimates. This method has been widely used to estimate correlates of protection for COVID-19 vaccine trials. Feng et al. (2021) [5] use this method with a logistic regression model on the COV002 data, and multiple COVID-19 vaccine trials [83, 84, 85, 86, 87] have used the same method with a Cox model.

3.2.2 Extrapolation approach

Another approach extrapolates the observed antibody levels at the peak visit to predict their values at the time of exposure. The decay trajectory is estimated on an antibody decay dataset, which may be an external dataset, or a subset of the dataset of interest which has immune data available at multiple timepoints. We assume here that log-transformed antibodies decay linearly, although the method can also be applied to other decay trajectories.

For the antibody decay dataset, a Bayesian hierarchical linear model is first fit to the log antibody levels over time, including a random intercept for the peak antibody response, while assuming the same slope for all individuals. Let $\pi_D(d)$ be the resulting posterior distribution for the decay rate d, where the decay rate is the magnitude of the slope in the

linear model for log antibody decay. Then for individual i in the correlates of protection dataset with observed antibody level of $A_i(t_i)$ at time-since-vaccination t_i, the antibody level $A_i(t)$ at time-since-vaccination t is estimated as follows. First draw a bootstrap sample $b = 1, \ldots, B$ of the main dataset, draw $d^{(b)} \sim \pi_D(d)$ and estimate the extrapolated antibody levels over time for each individual $A_i^{(b)}(t) = A_i(t_i) \exp\left(-(t - t_i)\, d^{(b)}\right)$ in the bootstrapped dataset. The antibody trajectories can then be included in an inverse probability weighted (IPW) Cox model as a time-varying covariate. The Cox model should use calendar time as the time scale, with individuals being considered at risk at a time after vaccination which is consistent with the efficacy study (e.g. 28 days after vaccination), and predicted antibody levels included at each event time in the dataset, calculated by the earlier method. Bootstrapped mean estimates and confidence intervals can be calculated from means and percentiles of the estimates for the Cox coefficients in each bootstrap sample.

Follmann et al. (2023) [47] applied this method to estimate COVID-19 correlates of protection, estimating decay rates from a separate dataset. Zhang et al. (2024) [87] also applied the approach in their paper, using a subset of the correlates of protection dataset to estimate antibody decay rates.

3.2.3 Simple two-stage approach

In a simple two-stage (STS) joint model, a Bayesian linear hierarchical model (or mixed-effects model) is fit to the longitudinal antibody component. The mean estimated antibody trajectory for each individual is then included as a time-varying covariate in a time-to-infection model, usually a Cox model. This does not propagate the uncertainty in the estimated antibody trajectories, and has been shown to lead to bias in the time-to-event parameters [44]. The estimation of the antibody trajectories does not account for informative censoring at infection times, which may cause bias in the estimated antibody levels [32]. One benefit of a STS joint model is reduced computation time compared to the full joint model [44]. Improvements on the STS approach have been suggested which reduce the bias, while still requiring less computation time than the full joint model [44, 45].

White et al. (2014, 2015) used a STS approach to estimate correlates of protection, antibody kinetics and vaccine efficacy against malaria [48, 49]. They assumed a bi-phasic

exponential decay model for antibody levels, and vaccine efficacy which ranged between a maximum value estimated from the data at high antibody levels, and zero efficacy at antibody level 0.

3.2.4 Other two-stage approaches to fit a joint model

Huang and Follmann (2025) [53] develop a two-stage joint model to estimate COVID-19 exposure-proximal correlates of protection, allowing for a time-changing direct effect, and both peak and exposure-proximal immune marker effects. They compare estimated hazard between individuals in a partial likelihood method, which outperforms the simple two-stage (STS) approach. Their model shows minimal bias, although they do not consider performance in terms of coverage in their simulation study. We suggest an alternative two-stage method to fit a joint model using multiple imputation below, which may reduce the bias from a STS approach.

3.2.5 Multiple imputation joint model

In a multiple imputation joint model, we similarly fit a Bayesian hierarchical model to the longitudinal component. This is used to impute the values of the longitudinal trajectory at each event time in a Cox model. Multiple imputations are made from the longitudinal posterior, and the resulting Cox parameters are then pooled to estimate the time-to-event posterior.

Moreno-Bentancur et al. (2018) developed a multiple imputation joint model which reduce bias compared with simple two-stage approaches, which use a single imputation [46]. Their method focussed solely on estimating the time-to-event parameters, and pooled the results using Rubin's rules [54].

We developed an alternative multiple imputation joint modelling approach. In order to account for the informative censoring at infection, we include the Nelson-Aalen estimate of cumulative hazard and the indicators of asymptomatic and symptomatic infection as covariates in the longitudinal model. This was recommended for Cox regression by White and Royston (2009) [64]. For the second stage, we sample longitudinal parameters from the model for $m = 1, \ldots, M$. We then fit a Cox model for each sample, including the sampled antibody trajectories as a time-varying covariate. We combine the results by

taking one draw from the asymptotic distribution of the Cox parameters on each sample. We apply the multiple imputation pooling method developed in Chapter 2 to approximate the posterior for the Cox parameters, from which we derive the correlates of protection curve and associated uncertainty. Our model is described in more detail in Chapter 4, where we apply the multiple imputation joint model to estimate COVID-19 correlates of protection.

3.2.6 Joint model

Finally, a Bayesian joint model for longitudinal and time-to-event data could be used to estimate "exposure-proximal" correlates of protection. A Bayesian hierarchical model for antibody decay, and a time-to-event model for subsequent infection, are included together in a joint model. The parameters are estimated jointly via Markov Chain Monte-Carlo or a similar sampling scheme. A fully Bayesian model, which we adapted from the RStanArm package [88] and attempted to fit in Stan [6], failed to converge on the COV002 dataset. This was likely due to the model including many random effects, and non-linear transformations, and the sparse availability of antibody data.

The package JMbayes2 [40] can be used to fit Bayesian joint models. Seekircher et al. (2024) used a joint model to relate modelled COVID-19 antibody levels at the time of exposure to risk of infection [51], and Papadopoulos et al. (2025) [52] fit a joint model to post-booster antibody levels and risk of COVID-19 infection, both using JMbayes2 [40]. Seekircher et al. do not account for time-since-vaccination in the model, and Papadopoulos et al. do not account for the changing background risk of COVID-19 infection with calendar time.

3.3 Antibody effect on risk

3.3.1 Structure of antibody effect on risk

Note in previous applications of the above modelling approaches to estimate COVID-19 correlates of protection, antibody levels have generally been log-transformed prior to analysis. This includes various papers applying the inverse probability weighting method of section 3.2.1 [5, 83, 84, 85, 86, 87], the extrapolation method of section 3.2.2 [47, 87], and

the joint model of section 3.2.6 [51, 52].

Using log-transformed antibodies does have an undesirable property, however. Let the immune marker levels at time t be denoted by $A(t)$, and for the purpose of this section, assume antibody levels are observed exactly, without error. Consider the Cox proportional hazards model [34], where the hazard rate at time t, $h(t) = h_0(t)\exp\left(X(t)^\top\beta + Z(\gamma f(A(t)) + \zeta)\right)$ is given by the baseline hazard $h_0(t)$ multiplied by covariate effect due to covariates $X(t)$ and, for vaccinated individuals ($Z = 1$), the intercept and slope of the effect due to (transformed) immune marker variable $f(A(t))$. When log antibodies are used as the immune marker, $f(A(t)) = \log A(t)$, the hazard is given by

$$\begin{aligned} h(t, A) &= h_0(t)\exp\left(X^\top\beta + Z(\gamma \log A(t) + \zeta)\right) \\ &= h_0(t)\exp\left(X^\top\beta\right)\{\exp(\zeta)A(t)\}^{Z\gamma} \end{aligned} \tag{3.2}$$

Assuming antibody levels protect against the disease, then $\gamma < 0$, so for vaccinated individuals, as antibody levels tend to 0, the hazard rate tends to infinity. Hence, the predicted vaccine efficacy will tend to negative infinity as the antibody level tends to 0. This does not fit with prior biological understanding, and is an undesirable property of the model. This may lead to poor fit and inaccurate predictions at low antibody levels.

If instead untransformed antibody levels are used as the immune marker, $f(A(t)) = A(t)$, then

$$\begin{aligned} h(t, A) &= h_0(t)\exp\left(X^\top\beta + Z(\gamma A(t) + \zeta)\right) \\ &= h_0(t)\exp\left(X^\top\beta\right)\exp(\gamma A(t) + \zeta)^Z \end{aligned} \tag{3.3}$$

When antibody levels protect against the disease, $\gamma < 0$. Then for vaccinated individuals, as antibody levels tend to 0, the hazard rate tends to a constant, hence so does the predicted vaccine efficacy. If no intercept ζ for the antibody effect is included, then the efficacy for a vaccinated individual with no antibodies will be estimated as 0. Including an intercept ζ will mean a constant and finite vaccine efficacy will be predicted at 0 antibody levels, with the sign and size of the effect determined by ζ. This formulation does not have the undesirable property of the log antibody formulation.

It is, however, worth noting that in the above COVID-19 vaccine examples, low levels of antibody near 0 are not observed in the vaccine arm. Using log-transformed antibody levels in the model may fit the data better in the region where antibodies are observed. The biological mechanism for protection is complicated, and does not lead to a clear reason for one choice of model form over the other. It would be worthwhile to further explore how to best include antibody levels in the model (linearly, log-transformed and/or with spline effects).

This is an overly simplistic model, since it assumes the protection depends on modelled antibody levels only, plus an additional time-constant effect ζ. Other immune responses such as memory B cells and T cells likely also play a role in protection against infection, and we have not modelled them explicitly here, in order to simplify the modelling, and because we have no measurements of their levels over time. The additional effect due to vaccination, ζ, is intended to encompass such effects, however, their protective effect may not be constant over time. This may introduce bias to our model. Huang and Follmann included an additional time-varying effect in their model to include non-antibody immune responses, however they note that this extended model is difficult to fit due to correlation between the effects [53]. We assume that any protective effects not due to antibodies are included in the time-fixed effect ζ, although our model could be extended to include a time-varying effect.

3.3.2 The effect of different background infection rates

As explained in the introduction, shifting the peak background risk of infection later in time, will lead to lower estimates of vaccine efficacy. In this section, we characterise the difference mathematically.

For the mathematical analysis, we will assume that all individuals in the trial are exposed on one day. This simplifies the calculations, allowing us to compare the consistency of the results given by the standard method when the day of exposure varies between trials. While such an assumption is unrealistic in practice, we might expect to see similar results in a pandemic situation, where the background rate of infection peaks at different times between trial, since in that case most infections will occur close to the peak of the infection wave. Infection times will be more spread in real life, which will mean the true results will

differ from this mathematical exploration. We expect in reality results between trials will differ less than in these calculations, since the variation in infection times will dilute the effect of waning. However, where the mean infection times vary between trials, the calculations will still give some indication as to how this affects the direction and magnitude of the change in the estimated parameter values for the correlate of protection curve.

To simplify the notation, we assume all individuals reach their peak antibody response at time 0, after which antibody levels decay exponentially. We further assume that all individuals are vaccinated at time 0. The arguments can easily be extended to relax these assumptions.

Consider two vaccine trials, trial 1 and trial 2. We assume the effects due to vaccination, antibodies, and covariates are the same in both trials, however the baseline hazard rates differ.

Let the true hazard rate in trial k for individual i at time t, with covariates x_i, vaccine indicator Z_i and antibody levels $A_i(t)$ be given by

$$h_i^{(k)}(t) = h_0^{(k)}(t) \exp(x_i^T \beta_0 + Z_i(\gamma_0 f(A_i(t)) + \zeta_0)) \tag{3.4}$$

where $h_0^{(k)}(t)$ is the baseline hazard rate in trial k at time t, β_0 is the true effect of covariates X, ζ_0 and γ_0 give the intercept and slope respectively of the "exposure-proximal" effect due to vaccine-induced antibodies. Equivalently, ζ_0 and γ_0 give the intercept and slope of the antibody effect on infection immediately after peak antibody levels are reached. Assume an increase in antibody levels causes a reduction in risk, so $\gamma_0 < 0$. Further note that $\zeta_0 < 0$ if and only if vaccinated individuals with 0 antibody levels have a lower risk of infection than control individuals (that is, there is a protective direct vaccination effect). Assume antibody levels decay exponentially for each individual. We further assume for the purpose of these calculations that the decay rate is the same for all individuals, namely

$$A_i(t) = \exp(a_{0i} + a_1 t) = A_i(0) \exp(a_1 t) \tag{3.5}$$

where decay rate $a_1 < 0$.

We now focus on the baseline hazard $h_0^{(k)}(t)$. Consider the most extreme scenario of a

baseline hazard with a peak, namely, individuals in trial 1 have a hazard of 0 at all times except at time τ_1, where they have a non-zero probability of being infected. Mathematically, this can be written as $h_0^{(k)}(t) = \lambda\,\delta_{\tau_k}(t)$, for $h_0^{(k)}(t)$ the baseline hazard in trial k, where λ is a parameter and δ_{τ_k} is the Dirac measure at τ_k, i.e., the unit point mass at τ_k. Then $\int_s^t h_0^{(k)}(u)\,du = \lambda$ if $s < \tau_k \leq t$ and 0 otherwise. Assume that $\tau_2 > \tau_1$, that is, background infections peak later in trial 2 than trial 1. We now characterise the difference in predicted vaccine efficacy at a given peak antibody level $A(0)$ between the two trials, when fitting a Cox model which includes antibody levels at time 0 as a predictor, without accounting for antibody decay. We consider the two cases mentioned in Subsection 3.3.1, namely where log-transformed antibodies, and untransformed antibodies respectively, have a linear effect on the log hazard.

Log antibody effect

Let $f(A(t)) = \log A(t)$, so the true hazard in trial k for individual i at time t is given by

$$h_i^{(k)}(t) = h_0^{(k)}(t)\exp(x_i^T\beta_{D0} + Z_i(\gamma_0 \log A_i(t) + \zeta_0))$$

Then in trial k, we will have a true hazard $h_i^{(k)}(t) = 0$ if $t \neq \tau_k$ and

$$\begin{aligned} h_i^{(k)}(\tau_k) &= \lambda\delta_{\tau_k}\exp(x_i^T\beta_{D0} + Z_i(\gamma_0\,(\log A_i(0) + a_1\tau_k) + \zeta_0)) \\ &= \lambda\delta_{\tau_k}\exp(x_i^T\beta_{D0} + Z_i(\gamma_0 \log A_i(0) + \gamma_0 a_1\tau_k + \zeta_0)) \end{aligned}$$

Assume we now fit a Cox model to the data, with predictors given by the columns of X, vaccination Z and log-transformed antibody level at time 0, given by $Z \times \log A(0)$ (assumed to be 0 for unvaccinated individuals). Then in large samples, the maximum likelihood estimates (MLEs) for the parameters will converge

$$\begin{aligned} \hat{\gamma}^{(k)} &\to \gamma_0 \\ \hat{\zeta}^{(k)} &\to \gamma_0 a_1\tau_k + \zeta_0 \end{aligned}$$

Hence $\hat{\zeta}^{(2)} - \hat{\zeta}^{(1)} \to \gamma_0 a_1\,(\tau_2 - \tau_1)$. Thus the estimated log hazard for a vaccinated individual with antibody level $A(0)$ increases by $\gamma_0 a_1\,(\tau_2 - \tau_1)$. The estimated vaccine efficacy at

antibody level $A(0)$ will also be lower in trial 2, since the increased hazard in the vaccine arm will cause an increased relative risk.

Linear antibody effect

Now assume instead that $f(A(t)) = A(t)$, so untransformed antibodies have a linear effect on log hazard. Then the true hazard in trial k for individual i at time τ_1 is given by

$$h_i^{(k)}(\tau_k) = \lambda\delta_{\tau_k} \exp(x_i^T \beta_{D0} + Z_i(\gamma_0 A_i(0) \exp(a_1\tau_k) + \zeta_0))$$

Assume we now fit a Cox model to the data, with predictors given by the columns of X, vaccination Z and untransformed antibody level at time 0, by $Z \times A(0)$ (assumed to be 0 for unvaccinated individuals). Then in large samples, the MLEs for the parameters will converge

$$\hat{\gamma}^{(k)} \to \gamma_0 \exp(a_1\tau_k)$$
$$\hat{\zeta}^{(k)} \to \zeta_0$$

So we see that

$$\frac{\hat{\gamma}^{(2)}}{\hat{\gamma}^{(1)}} \to \exp\left(a_1(\tau_2 - \tau_1)\right)$$

Hence the estimated effect $\hat{\gamma}^{(2)}$ on the antibody level $A(0)$ is shrunk towards zero compared to $\hat{\gamma}^{(1)}$, decreasing estimated vaccine efficacy at antibody level $A(0)$ in trial 2 compared to trial 1. For the remainder of this thesis, we consider models which include the antibody level in the model linearly, rather than via a log transformation.

3.4 Comparisons on simulated data

We generated simulated data to test whether the time-fixed approach gives consistent estimates of the correlates of protection parameters, when the baseline hazard of infection varies between trials. We generated data assuming the biological mechanism for protection due to vaccination and antibody response is consistent between trials, but the baseline risk

of infection varies. We applied the two-phase sampling approach to estimate correlates of protection, using a Cox model with a linear effect of observed antibodies soon after vaccination. If the method performs well, estimated correlates of protection will be consistent across simulation settings. If the method is performing poorly, estimates will vary across settings, despite the assumed true biological mechanism being consistent.

3.4.1 Generating simulated data

Model

Antibody model We generate the data from a simple model for antibody decay and subsequent infection, detailed below.

First consider two time scales: calendar time and time-at-risk. Let there be a fixed calendar time after vaccination (e.g. 28 days), which we call S_i, from which individuals are considered at risk. Hence calendar time t after S_i corresponds to time-at-risk $t - S_i$ for individual i. The time S_i is generally the time of a study visit defined in the protocol of a vaccine trial, and chosen to correspond with peak levels of immunity after vaccination. We assume subsequent antibody levels are decaying for all individuals after in the study. As before, let Z_i be 1 if individual i is vaccinated and 0 otherwise, and order the dataset such that $Z_i = 1$ for $i = 1, \ldots, n$ and $Z_i = 0$ for $i = n+1, \ldots, N$. Let $A_i(t)$ be the true antibody level for vaccinated individual $i \in \{1, \ldots, n\}$ at time-at-risk t, and assume $A_j(t) = 0$ for control individuals $j = n+1, \ldots, N$, $\forall t$. Then assume $A_i(t)$ is given by

$$A_i(t) = \exp(a_{0i} + a_1 t) \tag{3.6}$$

where $a_1 < 0$ denotes the slope of log antibody decay, assumed to be the same for all individuals, and a_{0i} denotes the individual intercept. Assume the random intercepts are normally distributed

$$a_{0i} \sim N\left(\alpha_0 + x_{Li}^T \beta_{L0}, \tau_0^2\right) \tag{3.7}$$

for X_L a matrix of covariates and β_{L0} a vector of effects on peak antibody level, and intercept α_0 for a baseline individual with $x_{Li}^T = 0$.

We assume the observed log antibody level at time-at-risk 0 for individual i is given by

$$\log Y_i(0) = \log A_i(0) + \epsilon_i \tag{3.8}$$

where $\epsilon_i \sim t_4(0, \sigma^2)$ for $t_\nu(\mu, \sigma^2)$ denoting the location-scale t-distribution with ν degrees of freedom, location μ and scale σ.

Infection model Now assume in trial k for individual i in $1, \ldots, N$ that the individual is considered at-risk from calendar time S_i. Then the true hazard rate for infection at time-at-risk t is given by

$$h_i^{(k)}(t) = h_0^{(k)}(S_i + t) \exp\left(x_{Di}^T \beta_{D0} + Z_i \left(\gamma_0 A_i(t) + \zeta_0\right)\right) \tag{3.9}$$

where baseline hazard $h_0^{(k)}$ in trial k varies with calendar time, X_D is a matrix of covariates related to risk of infection and β_{D0} their true effect on hazard, and γ_0 and ζ_0 give the slope and intercept of the effect of vaccine-induced antibody levels at the time of infection on hazard. Note the true hazard rate depends on antibody levels at the time of infection, whereas the assumed hazard rate in the fitted model depends only on peak antibody levels at calendar time S_i, time-at-risk 0.

Choice of parameters

We aim to generate simulated data which is similar to the data observed in the COV002 vaccine trial (Chapter 4, [4, 5, 89]), while making simplifying assumptions where appropriate.

We assume there are 4500 individuals in each arm (vaccine/control).

We generate X_L to have a single column drawn independently from the standard Gaussian distribution. We then set $\alpha_0 = \log(217)$ which gives a median antibody level at peak of 217, and $\alpha_1 = \log(\frac{1}{2})/74$ to give a half-life of 74 days, matching the values estimated from the trial (Chapter 4, Table 4.2). We set $\tau_0 = 0.756$, $\sigma = 0.206$ (Supplementary Table C.4) to match COV002, and $\beta_{L0} = 0.2$ to show a weak effect.

For the time-to-event parameters, we set $\gamma_0 = \log(0.49)/100 = -0.007$ to match COV002 (Supplementary Table C.7), and $\zeta_0 = 0$ and $\beta_{D0} = 0.2$ to show no direct effect due to

vaccination and a weak effect due to covariates X_D. We generate $X_D \sim N(0,1)$ in the same way as X_L, independently. We assume in all trials that $h_0^{(k)}$ is constant for any given day, but may vary between days. We discuss the specific form of $h_0^{(k)}$ later, in Section 3.4.2.

Data generation

Generating true antibody trajectories We first generate the true antibody trajectories for each individual by drawing individual random intercepts from equation (3.7) and defining the antibody level at time t by equation (3.6).

The event times then depend on these true antibody trajectories. The probability of observing antibody levels at a given visit for an individual depends on whether the event later occurred (infection), since blood samples from the visit are retrospectively chosen to be sent to a laboratory for antibody measurement via a case-cohort scheme. The observed antibody levels themselves depend only on the true antibody trajectories.

Generating event times In order to generate the event times, we first generate the times S_i at which individuals enter the at-risk period, assuming recruitment is spread over 60 days, with 75 individuals in each arm entering per day. The S_i values range from 0 to 59.

We generate event times by calculating the probability an individual is infected on each day, given they are at-risk. We do this for the vaccine arm using Gauss-Kronrod quadrature [90] with 15 nodes, to approximate the integral in estimating the infection probability. Event times can be generated for each individual from these probabilities, by sequentially generating Bernoulli infection indicators for each day, and choosing the first infection. We use a single cut-off calendar time for all individuals to censor the data. This was chosen to be the shortest of: the calendar time at which 300 cases are observed in the trial, and 365 days (one year) after the beginning of the study. Individuals in the real COV002 dataset were censored earlier than this (Chapter 4, Supplementary Figure C.4), due to unblinding to receive a booster dose, although the total span of the study was similar at 347 days (Chapter 4). There were 290 primary symptomatic cases in our COV002 correlates analysis (Chapter 4, Supplementary Table C.1), so we rounded to use 300 as

the cut-off in the simulation study.

Generating observed antibody levels Finally, we generated observed antibody levels $Y_i(0)$ for individuals in the vaccine arm at time-at-risk 0 (calendar time S_i), using case-cohort sampling. We sampled from Equation (3.8) for every individual who was later infected, and from a random sample of 25% of individuals who were not infected.

3.4.2 Simulation scenarios

We consider the effect of changes to the baseline hazard $h_0^{(k)}$ in two different regimes. In the first regime we assume $h_0^{(k)}$ is constant over calendar time. We consider a low, medium and high baseline hazard, leading to an early, medium and late stopping of the trial. Namely for a baseline ($x_{Li} = 0$) participant in the control arm, we consider a hazard of 50, 100 and 200 infections per 1000 person-years, giving 3 different scenarios. The threshold of 300 cases will then be reached at different calendar times, giving trials of different lengths.

In the second regime, we consider an epidemic disease, and compare how the timing of an infection wave affects the results. We assume the baseline hazard follows the derivative of a logistic function, which loosely approximates the pattern of COVID-19 deaths per day in the UK [91] which peaked in December 2020 (though the pattern in the UK observed two peaks, we simplify this to one). We assume the baseline hazard on day t is given by

$$\frac{aK\exp(-a(t-t_0))}{(1+\exp(-a(t-t_0)))^2}$$

with parameters $a = 0.05$, $K = 0.04$ and t_0 denoting the timing of the peak infection rate. We vary t_0 to be 3, 4, 6, 8, and 9 months after the first individual reaches peak antibody levels, giving 5 different scenarios.

For each scenario within each regime, we generate 1000 simulated datasets.

3.4.3 Fitted model

We fitted a simplified version of the inverse probability weighting method described in section 3.2.1 to each dataset. The full method requires the inverse probability weights to

be estimated in a regression model, and a bootstrap to properly account for uncertainty. We instead assume the inverse probability weights to be known, which also means the boostrap is not required, reducing computational demands. We run an inverse probability weighted Cox model, with an inverse probability weight of 4 for a non-infected vaccine arm participant, and 1 for infected vaccine arm participants. This is because the probability of antibodies being sampled in these two groups was 1/4 and 1 respectively in the data generating process. We followed previous analyses of COVID-19 vaccine, by fitting a separate model to the vaccinated individuals from the control individuals [5, 83, 84, 85, 86, 87]. We differ from these references by using a linear antibody effect, instead of a log antibody effect, namely, the hazard for vaccinated individual i

$$h_i(t) = h_0(t) \exp(Y_i(0)\gamma + X_i\beta) \tag{3.10}$$

for $Y_i(0)$ the observed antibody level at time-at-risk 0, and covariates X_i.

This allows us to estimate γ, the effect of observed antibody levels at peak response on the log hazard. Since this parameter is important in determining the slope of the correlates of protection curve, and expected to vary as the timings of cases vary (Section 3.3), we compare the estimates and confidence intervals for γ across scenarios. The estimated risk in the control individuals would also be required to estimate vaccine efficacy, and affects the height of the correlates of protection curve. We ignore this for the sake of these comparisons, since we do not anticipate estimates of the risk in the control arm will change noticeably between scenarios for these particular simulations.

3.4.4 Results

Figures 3.1 and 3.2 show the coverage for the antibody effect parameter γ at different values, for the endemic and epidemic regimes respectively. For a given value of γ on the x-axis, the y-axis shows the proportion of 95% confidence intervals (CIs) which contained that value. The data generating model defines no "true" value of γ, the effect of antibody at time-at-risk 0 on later infection, but only of the effect of antibody at exposure on infection. Any "true" value of γ would change between the scenarios, due to vaccine efficacy waning. Hence in these plots we display the coverage at a range of possible values. We see for all

scenarios in both regimes, the highest coverage is observed at values between γ_0 and 0. γ_0 denotes the effect of antibody levels at exposure on risk of infection, and would be the true value of γ if no antibody waning were present, or all infections occurred immediately after peak antibody levels were reached. The true value of γ would be zero if all antibody levels had waned to 0, or in the limit as the time-at-risk of all infections tends to infinity. We should not expect the estimates of γ to be γ_0 in all scenarios, since we are fitting a different model to the model used to generate the data, by relating observed antibody levels to risk of infection. We should expect estimates for γ to be consistent across scenarios, since our model for vaccine-induced protection is the same in all scenarios.

However, the results are not consistent across scenarios. The later infections tend to occur in a scenario, the closer the observed value of $\hat{\gamma}$ is to 0. This is problematic, since the antibody effect used to generate the data is unchanged between trials; only the background infection rate changed.

In the endemic regime, a lower background risk of infection leads to 300 total cases being reached later, at which point the correlates of protection analysis is run. This means infections occur later on average (Supplementary Table B.2), meaning more time-at-risk has passed in which antibody levels can wane. So the trials with a lower baseline hazard observe values of γ closer to zero than the trials with a higher baseline hazard (Fig. 3.1, Supp. Fig. B.1 & B.3).

In the epidemic regime, a later peak for the baseline hazard curve means more antibody waning occurs before the period with the highest risk of infection. This leads to the trials with a later infection wave observing values of γ closer to zero than the trials with an earlier wave (Fig. 3.2, Supp. Fig. B.2 & B.4).

In both regimes, we observe scenarios with earlier infection give wider confidence intervals for γ (Supp. Fig. B.3 & B.4)). This is because vaccine efficacy is higher at earlier times-at-risk, so fewer infections are observed in the vaccine arm. Infections in the vaccine arm are crucial for the model to estimate γ, since they allow comparisons of the risk of infection between individuals with different antibody levels.

In both regimes, we see that the nominal level of 95% coverage is not reached for any value of γ, with coverage levels peaking around 91% to 92% (Supplementary Table B.2). There are two reasons which may contribute to the observed under-coverage. Firstly,

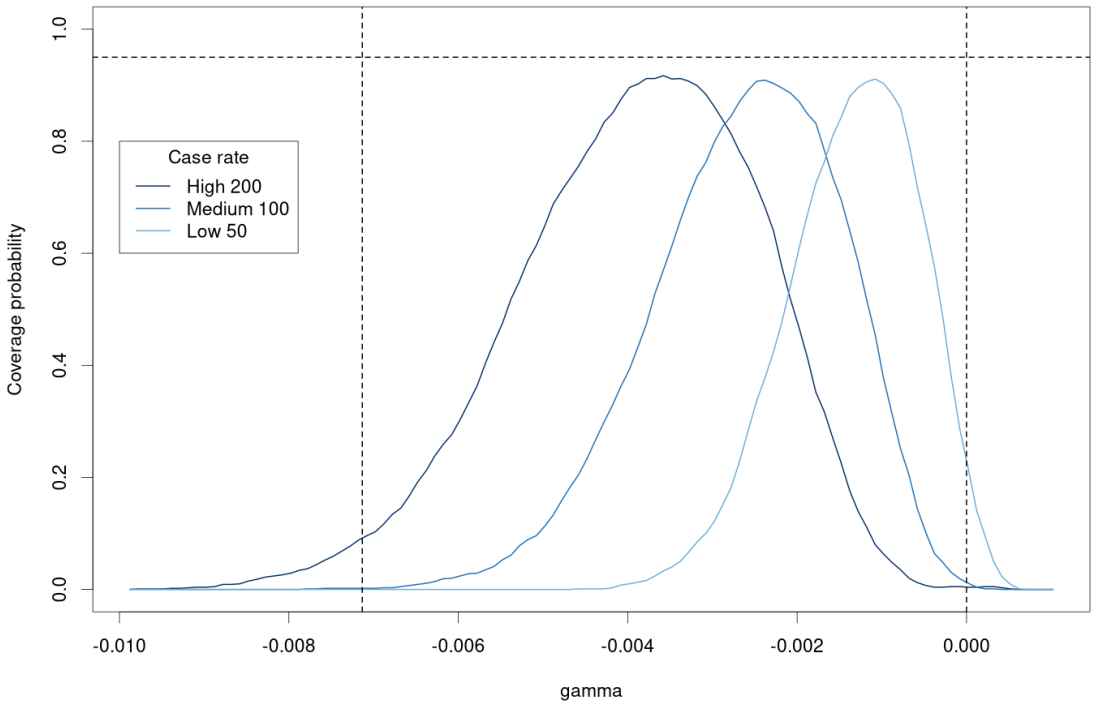


Figure 3.1: **Coverage probabilities for antibody effect parameter γ with different constant infection rates**

The coverage probability (y-axis) of a 95% confidence interval for γ associated with different values of γ (x-axis). That is, the curve shows the proportion of 95% confidence intervals which contained each value of γ. The curves show results from endemic scenarios with constant baseline infection levels of 200, 100, and 50 cases per 1000 person-years, from darkest to lightest shade of blue. The horizontal dotted line shows nominal coverage of 95%. The vertical dotted lines show the true values of γ in the extreme cases of no antibody waning ($\gamma = \gamma_0 = -0.007$) and all antibody levels waning to 0 ($\gamma = 0$).

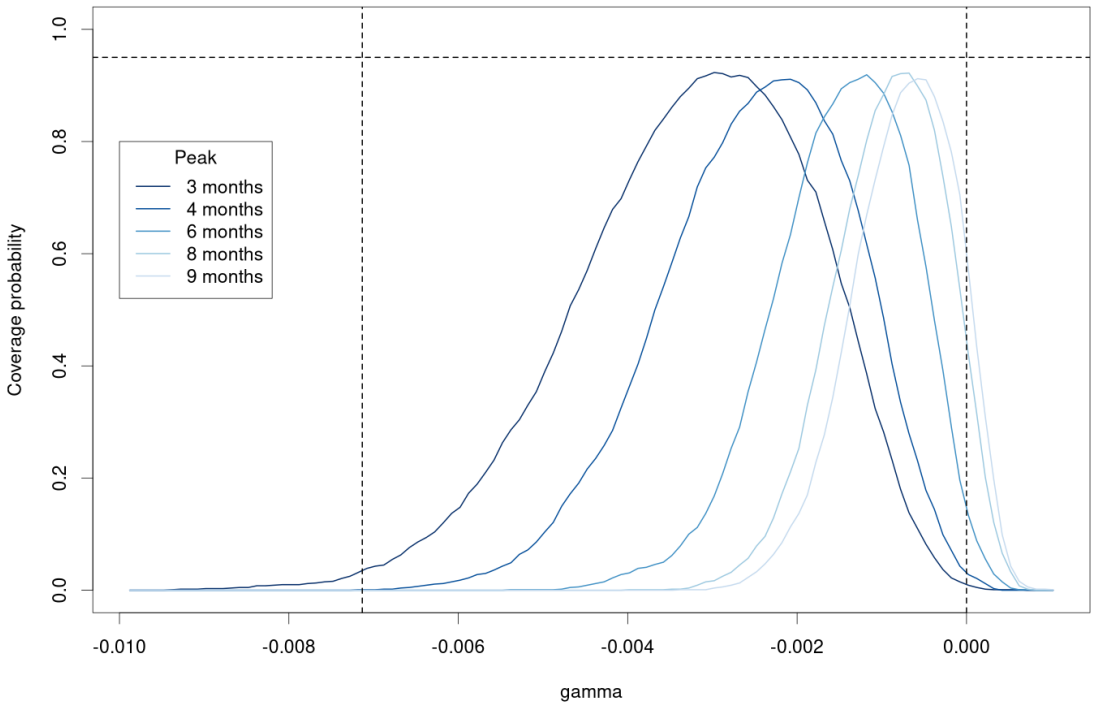


Figure 3.2: **Coverage probabilities for antibody effect parameter γ with different timings of peak infection rate**

The coverage probability (y-axis) of a 95% confidence interval for γ associated with different values of γ (x-axis). The curves show the results from epidemic scenarios where infection levels peak at 3, 4, 6, 8, and 9 months, from darkest to lightest shade of blue. The horizontal dotted line shows nominal coverage of 95%. The vertical dotted lines show the true values of γ in the extreme cases of no antibody waning ($\gamma = \gamma_0 = -0.007$) and all antibody levels waning to 0 ($\gamma = 0$).

the inverse probability weighting method does not account for the antibody measurement error. Not accounting for measurement error has been shown to lead to bias in the Cox model [2], and may also cause undercoverage, since not accounting for the uncertainty may lead to overconfidence in the results. Secondly, since a different amount of time-at-risk passes before infection for different individuals, antibody levels wane by different amounts between measurement and exposure. The data generating process assumes antibody levels at exposure have a linear effect on log hazard, however due to different levels of waning, the effect of antibody levels at time-at-risk 0 on log hazard may not be linear. This means the Cox model may be misspecified, which could also cause a decrease in coverage.
We observed in our simulation study that the time-fixed approach, which relates observed antibody levels to risk of infection, gives inconsistent results across simulation scenarios when the true mechanism of protection is unchanged.

3.5 Comparisons on real data

3.5.1 Methods

We compared the results from the Cox inverse probability weighted (IPW) model, extrapolation model, and multiple imputation joint model, fitted to data from the COV002 trial. The trial is explained in more detail in Chapter 4.2. The Cox IPW model relates observed antibody level at the time of peak immune response to the risk of subsequent infection. The extrapolation model uses observations at peak response to predict the antibody level at exposure, which is then compared to the risk of infection. Neither of these models account for the uncertainty in the antibody measurements. The model relates estimated antibody levels at the time of exposure to the risk of infection, while accounting for the uncertainty in antibody level measurements.
We used untransformed antibody levels as the immune marker in the Cox model for all three methods (Eq. (3.3)). The Cox IPW model and extrapolation model used the observed antibody level at the PB28 visit, around 28 days after receipt of the second vaccine dose. For all three methods we included both the vaccine and control arm in the Cox model, assuming antibody levels to be 0 in the control arm, and including a direct effect from vaccination.

To fit the extrapolation model, we first fit a Bayesian random intercept model to log antibody levels for the 52 vaccinated individuals who have observed antibody levels available at 3 timepoints. The estimated slope of log-antibody decay from the model was used to extrapolate the antibody levels. The estimated levels at the time of exposure were then included in a Cox IPW model. The Cox IPW model was bootstrapped, and for each bootstrap sample, a different draw from the Bayesian model for the slope of antibody decay was used.

In both the extrapolation method and the multiple imputation joint model, we relate the risk of infection to antibody levels 7 days prior, assuming a 7 day incubation period from exposure to an infection being reported in the study. Note that the Cox IPW model estimates the relationship between vaccine efficacy and antibody levels at the PB28 visit (soon after vaccination), whereas the extrapolation model and joint model relate vaccine efficacy to the antibody levels at the time of exposure.

The multiple imputation joint model is outlined in Chapter 4. Briefly, a Bayesian hierarchical random intercept and random slope model is fit to the observed antibody levels, assuming log antibody levels decay linearly. The model accounts for the time-to-infection by including both the Nelson-Aalen estimate for cumulative hazard, and indicators of any infection and symptomatic infection, as covariates predicting individual intercepts and slopes. Multiple samples of estimated intercepts and slopes for antibody decay in all individuals in the trial are drawn from the resulting posterior. For each sample, a Cox model is fit to the data, including the estimated antibody trajectory as a time-varying covariate. One draw is taken from the asymptotic distribution of the MLE for the antibody effect in the Cox model, which approximates the posterior distribution for the antibody effect, given the estimated antibody trajectories. These draws are pooled together for 60,000 iterations, to approximate a joint model for antibody decay and risk of infection. More details of the multiple imputation joint model are given in the following chapter, Section 4.3.

We estimate vaccine efficacy using the hazard ratio definition for all three models, to maximise comparability. Namely, vaccine efficacy at antibody level A, is given by 1 minus the

hazard ratio due to being vaccinated with antibody level A.

$$\mathrm{VE(A)} = 1 - \exp(\gamma \mathrm{A} + \zeta) \tag{3.11}$$

3.5.2 Results

Figure 3.3 shows the estimated correlates of protection curve. The curve is steepest for the multiple imputation joint model, followed by the extrapolation method, and then the Cox inverse probability weighted model. This is likely because the estimate for γ is diluted in the extrapolation and Cox IPW methods by not accounting for measurement error [2], and reduced further in the Cox IPW method due to antibody decay (Section 3.3.2).

3.6 Discussion

Our simulations demonstrate that standard time-fixed correlates of protection models may give inconsistent answers for the same vaccine, when run on clinical trials with different background rates of infection. This is true even when the immune response and mechanistic protection is the same. When analyses are run after cut-off at a fixed number of cases, one trial may reach this point later than another, if the background infection rate is lower. This means cases will occur later on average, so antibody levels will have waned more between measurement and exposure. Hence the predicted vaccine efficacy at a given antibody level will be lower. Similarly in a pandemic scenario such as the recent COVID-19 pandemic, the timing of the peak of an infection wave may be later in one trial than another. This could lead to lower predicted vaccine efficacy at a given level of antibody than the other trial, causing inconsistent results between trials, even if the true mechanism of protection is unchanged between populations.

It is important to note that the difference in background infection rates between trials assumed in our simulation study is entirely realistic (Supplementary Table B.1). In COVID-19 vaccine studies, the infection rates per 1,000 person years in the control arm ranged from 51.9 to 215.4, which is a wider range than we used in our simulation study. However, whether this would lead to an actual difference in cut-off times is less clear. Unlike vaccine efficacy analyses, strict criteria for the number of cases required before running a corre-

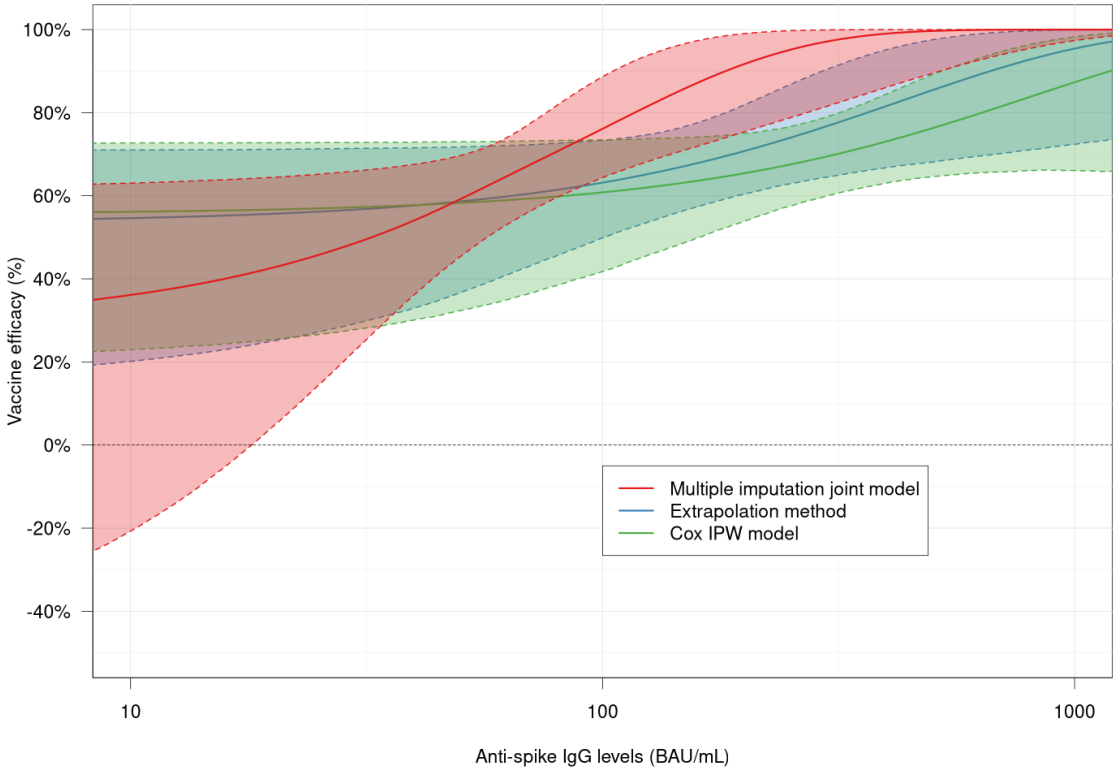


Figure 3.3: **Predicted vaccine efficacy at different antibody levels on the COV002 data, using IPW, extrapolation and a multiple imputation joint model.** VE (%) plotted against anti-spike IgG antibody levels (BAU/mL) on a log scale. The red, blue and green curves and shaded regions show the posterior median and associated 95% credible intervals for VE against primary symptomatic COVID-19 infection, using the multiple imputation joint model, extrapolation method, and Cox inverse probability weighted (IPW) model. Note the Cox IPW results show vaccine efficacy against observed antibody levels at the study visit around 28 days after vaccination, whereas the multiple imputation joint model and extrapolation method relate vaccine efficacy to predicted antibody levels at the time of exposure.

lates of protection analysis are rarely published in trial protocols. The average length of follow up in our simulated trials (Supplementary Table B.2) was longer than that observed in most COVID-19 vaccine efficacy trials. However correlates of protection analyses tend to be run at a later time after more data has accrued, and average follow-up times are rarely published. Total numbers of cases are also rarely published, although the number of vaccinated cases were lower in some studies ([85], [83]) and higher in others ([84], [89] - any infection analysis). Since the cut-off to run a correlates of protection analysis does not tend to be prespecified in the protocols of vaccine efficacy trials, nor reported after, it is difficult to judge whether a higher background rate of infection would necessarily lead to an earlier analysis.

It is clearer, however, that a difference in timings of infection waves might lead to a difference in correlates of protection results. COVID-19 vaccine trials were run in various countries (e.g. US, UK, Brazil, South Africa), where infection waves hit at different times. We have shown that a difference in the timing of peak infection rates may lead to a noticeable difference in correlates of protection parameters, even if there is no difference in the vaccine itself. This may partly explain the slight differences observed in estimated correlates of protection curves between COVID-19 vaccines [86] (although power to detect a difference was lacking).

Comparing the time-fixed Cox IPW method, extrapolation method, and multiple imputation joint model on the COV002 dataset, suggested the multiple imputation joint model may have the most power to detect a correlate of protection, followed by the extrapolation method. This is because a steeper slope is observed in the multiple imputation joint model. Not accounting for measurement error in the extrapolation and Cox IPW methods, or for antibody decay in the time-fixed Cox IPW method, will likely lead to a diluting of the estimated antibody effect ([2], Section 3.3.2). This may lead to decreased power to detect an effect, and underestimation of the strength of a correlate of protection were detected.

This work is only an initial exploration into the need to account for waning antibody levels in correlates of protection models, and could be extended in various ways. Firstly, the simulations could be extended to consider how the results might vary with different levels of initial antibody response, and different rates of antibody decay. The parameters were chosen for the simulation study to resemble those observed in a ChAdOx1 nCoV-19

vaccine trial. Using different choices of parameters would make the results more relevant to other COVID-19 vaccines, and vaccines against other diseases.

Secondly, the simulation study only considers the performance of the IPW Cox model, which does not account for antibody decay. We did not evaluate the performance of the joint modelling approach, or the extrapolation method. It would be of interest to run simulations to test whether these two methods show appropriate coverage, and whether their results depend on the timings of infection in the study. We would expect the results to be consistent across studies with different background infection rates; however, their performance in terms of coverage is as yet unknown.

It would additionally be worthwhile to further explore the most appropriate form of the correlates of protection curve. In this paper we used antibody levels in a Cox model, since this means the vaccine effect is bounded as antibody levels tend to 0. Previous papers use log antibodies in a Cox model which may fit well for large antibody levels; however, this approach means the vaccine efficacy will tend to negative infinity as antibody levels tend to 0. This is an undesirable property, and may cause poor predicted efficacy for small antibody levels. Dunning (2006) [14] and Dunning et al. (2015) [15] developed a model which does not have this undesirable property, and uses log-transformed antibody levels in a form which is likely to give a similar fit to the IPW approach for large antibody levels. The model however does not allow for covariate effects, or more general case-cohort sampling, where the sampling probability may differ between individuals. It would be interesting to extend the Dunning model further so that it can fit data from the COVID-19 vaccine studies, and compare the calibration of the model with Cox and logistic IPW models using antibody levels, log-antibody levels, and splines.

If more vaccine efficacy trials collected multiple antibody measurements at different timepoints, it would be easier to fit correlates of protection models which account for measurement error and antibody decay. Collecting samples at multiple timepoints could increase the length of time required to run a correlates of protection study, and possibly the cost, if more samples are needed. However, it is unclear whether a larger total number of antibody samples would be required — accounting for measurement error may give more precise estimates, meaning fewer measurements are required in total. The total number of samples needed, and optimal spread between visits at different times, would be interesting

to study in a future simulation study. Further, vaccine trials will often include follow-up visits where blood samples are taken from participants. With antibody samples available from just 2 or 3 different visits for some individuals in the study, the joint modelling methods discussed in this chapter can still be applied, as in Chapter 4. So these methods may still be applicable even without needing to increase cost or trial length. Some adjustments in planning may be required however, so that blood samples from multiple visits are sent for assay quantification for some individuals.

In conclusion, joint modelling is a viable approach to account for waning antibody levels measured with error. Applying joint modelling may lead to more consistent results across vaccine trials than time-fixed correlates of protection approaches, and more power to detect a correlate of protection.

Chapter 4

Joint modelling of COVID-19 correlates of protection, antibody decay and vaccine efficacy

This chapter is based on the following work [89]:

Daniel J Phillips, Maria D Christodoulou, Shuo Feng, Andrew J Pollard, Merryn Voysey, David Steinsaltz.
Improved estimates of COVID-19 correlates of protection, antibody decay and vaccine efficacy waning: a joint modelling approach.
medRxiv preprint (2024). URL https://doi.org/10.1101/2024.07.02.24309776.

Author contributions

All authors contributed to the conceptualisation of the study. SF, AJP and MV contributed to data collection. DJP, MDC and DS developed the statistical methods. DJP conducted the statistical analysis. The project was jointly supervised by MDC and DS. The first draft of the preprint was written by DJP and reviewed by MDC and DS. All authors critically reviewed and approved the final version of the original preprint. The preprint was adapted and updated by DJP for this thesis chapter.

Abstract

Reliable estimation of the relationship between COVID-19 antibody levels at the time of exposure and the risk of infection is crucial to inform policy decisions on vaccination regimes. We develop a two-stage multiple imputation joint model of anti-spike IgG antibody decay and risk of COVID-19 infection. In the first stage, we model log-transformed antibody levels over time in a Bayesian hierarchical linear model, including a random intercept and slope for log antibody decay. We take multiple draws from the antibody model, and for each draw we impute the predicted antibody levels as a time-changing covariate in a Cox model. We pool the results together using approximate Bayesian pooling, presented in Chapter 2. Our model improves upon previous analyses by accounting for measurement error, decay in antibody levels and changing rates of baseline infection in the study.

We fit the model to data from a randomized efficacy trial of the ChAdOx1 nCoV-19 vaccine to estimate correlates of protection, antibody decay, and vaccine efficacy waning. Increased anti-spike IgG antibody levels at the time of exposure correlate with increased vaccine-induced protection. We estimated vaccine efficacy against symptomatic COVID-19 infection of 88.1% (95% CrI: 77.2, 93.6) at day 35, waning to 60.4% (44.6, 71.0) at day 189 since the second dose. We report that longer intervals between the first and second vaccine dose give lasting increased protection, and observe lower efficacy in individuals aged $\geq$70 years from around 3 months after second dose. Our methods can be used in future vaccine trials to help inform the timings and priority of vaccine administration against novel diseases.

4.1 Introduction

Correlates of protection, i.e. immune markers which relate to the risk of a disease outcome, are crucial to understanding the biological mechanisms for protection against infection after a given treatment. Most studies consider how immune responses at a fixed time soon after treatment correlate with risk of a disease outcome. However, it is the immune markers present at the time of exposure which protect against infection. Modelling immune levels over time and relating these to vaccine-induced protection, allows a clearer understanding of the relationship between immune responses and protection, and avoids potential biases due to waning of immune levels between the measurement and exposure. This further provides a natural framework to investigate the effect of factors such as age, comorbidities and dose schedules on immune response and protection, and to estimate changes in levels of protection over time.

In response to the COVID-19 pandemic, the (Oxford-AstraZeneca) ChAdOx1 nCoV-19 (AZD1222) vaccine has been widely administered in the UK and worldwide, alongside other COVID-19 vaccines. Binding and neutralising antibodies have been shown to be correlates of protection for ChAdOx1 nCoV-19 [5, 86] and other COVID-19 vaccines [83, 84, 85, 92, 87]. These studies all considered how initial peak antibody responses correlate with the risk of future COVID-19 infection, as this is the timepoint most relevant for vaccine licensure. However, as shown in Chapter 3, this method may generate inconsistent estimates between studies. If used to understand how antibody levels at the time of exposure relate to risk of infection, the above approaches may be biased, as they do not account for the waning of antibody levels between the time of the antibody measurement and exposure [29]. Other studies focussed on the relationship between antibody levels at exposure and the risk of subsequent infection [81, 82, 47]. Wei et al. (2022, 2023) used observational data to estimate how risk of infection after vaccination relates to a recent antibody measurement [81, 82]. However the use of observational data may cause their results to be biased. Follmann et al. (2023) modelled antibody decay over time after vaccination, relating the risk of infection to the predicted antibody level at the time of exposure [47]. However, Follmann et al. assume the same rate of decay for all individuals, which may decrease power and lead to bias. Further, all the above studies on correlates

of protection use only one antibody measurement per individual. This means they are unable to account for measurement error, which will likely introduce bias to the results [2].

Vaccine efficacy (VE) has been shown to wane over time after two doses of the ChAdOx1 nCoV-19, BNT162b2, and mRNA-1273 vaccines in the UK and elsewhere, against the Alpha variant [93] and the Delta variant [93, 94, 95, 96, 97, 98]; as well as after a third booster dose [98, 99]. These studies all use observational data, in test-negative or cohort designs, which may introduce bias due to unmeasured confounders affecting both the probability of vaccination and of the clinical outcome (e.g. infection) [100, 101]. Randomised controlled trials are free of such bias, however sample sizes tend to be much smaller, restricting power for analysis of efficacy waning. Including antibody data in a model for vaccine efficacy waning may increase the power of the analysis [9, 29].

Joint models of longitudinal and time-to-event data are a powerful tool to investigate how longitudinal biomarker trajectories relate to the risk of health- or disease-related events [29]. The method has been applied to data on various diseases, especially disease progression in HIV/AIDS [102] and cancer trials [29]. See Ibrahim, Chu and Chen (2010) [29] for an introductory review, and Gould et al. (2015) [30] for a review with more detail on implementation methods and packages. Joint modelling has been shown to reduce bias and increase power in estimating the relationship between a longitudinal biomarker and clinical event [29, 46], when compared with methods which use observed biomarkers directly to predict risk. This is because the approach accounts for biomarker measurement error, as well as changes in the biomarker over time. Joint models have been applied to understand how viral load [103] and changes in biomarkers [104] relate to risk of mortality in COVID-19 patients, and to predict their risk of mortality from biomarker data [105]. They have also previously been used to understand how vaccine-induced immune responses relate to risk of AIDS/death in HIV patients [50], and risk of relapse/death in cancer patients [24]. White et al. (2014, 2015) applied a simple two-stage joint model to relate antibody levels to risk of infection after receipt of a malaria vaccine [48, 49]. It has however been demonstrated that such simple two-stage approaches can give biased results [45, 44]. Seekircher et al. (2024) fit a joint model to relate antibody levels to risk of COVID-19 infection [51], using the R package JMbayes2 [40]. However their approach

does not account for the time since the last infection or vaccination, assuming antibody levels instead relate to calendar time. Papadopoulos et al. (2025) fit a joint model to post-booster antibody levels and risk of COVID-19 infection [52], using JMbayes2 [40]. They use their model to make individual predictions of future antibody levels. Their model does not account for changing background rates of COVID-19 infection risk with calendar time, nor do they have a control arm available with which to estimate vaccine efficacy. Huang and Follmann (2025) [53] develop a two-stage joint model to estimate COVID-19 exposure-proximal correlates of protection, allowing for a time-changing direct effect, and both peak and exposure-proximal immune marker effects. They compare estimated hazard between individuals in a partial likelihood method, which outperforms a simple two-stage approach. Their model shows minimal bias, although they do not consider performance in terms of coverage in their simulation study.

We implement a two-stage joint model for COVID-19 antibody decay and risk of COVID-19 infection after vaccination using multiple imputation [7]. Our model accounts for both antibody decay over time and antibody measurement error. We apply our model to data from a randomised vaccine efficacy trial of two doses of the ChAdOx1 nCoV-19 vaccine [3, 4] to investigate three main outcomes (COV002). Firstly, we model the waning of anti-spike IgG antibody levels over time. Secondly, we estimate correlates of protection - the relationship between anti-spike IgG antibody levels at the time of exposure and the risk of COVID-19 infection. Thirdly, we estimate waning vaccine efficacy over time, and investigate covariate effects on antibody and vaccine efficacy waning.

4.2 COVID-19 vaccine data

4.2.1 Study description

We analysed data from COV002 (registration NCT04400838), a phase 2/3 randomized single-blind vaccine efficacy (VE) trial conducted across 19 sites in the United Kingdom. A full description of the trial including immunogenicity, efficacy, and safety data, and the protocol has been previously published [3, 4, 106, 107]. An analysis of immune correlates of protection, relating antibody measurements 28 days after the second dose with protection against infection, was previously published [5].

This study was approved in the United Kingdom by the Medicines and Healthcare products Regulatory Agency (MHRA), reference 21584/0428/001–0001, and the South-Central Berkshire Research Ethics Committee, reference 20/SC/0179. All participants provided informed consent.

Participants in the study were randomized to receive ChAdOx1 nCoV-19 (AZD1222) or a MenACWY control vaccine. We refer to the two groups as the vaccine arm and the control arm respectively. The randomization ratio (ChAdOx1 nCoV-19:MenACWY) differed by study cohort, and was either 1:1, 5:1, or 3:1. Open label groups are not included in this analysis.

4.2.2 Study endpoints and outcomes

Participants were sent weekly reminders to contact their study site if they experienced any of the primary symptoms of COVID-19 (fever $\geq$37.8 °C; cough; shortness of breath; anosmia or ageusia) and were then assessed in clinic, by taking a nose and throat swab by a nucleic acid amplification test (NAAT). In addition, participants were asked to complete a nose and throat swab at home each week.

The outcomes for this analysis were (1) primary symptomatic COVID-19 infection, that is a nucleic acid amplification test positive (NAAT+) swab with at least one qualifying symptom, (2) any COVID-19 infection, that is any NAAT+ swab.

All endpoints were evaluated by a blinded independent clinical review committee. The date of onset of infection was determined by the committee, as the earliest time at which there was evidence of infection. Evidence included NAAT+ swabs and self-reported symptoms based on telephone contact and study visits.

4.2.3 Antibody measurements

A proportion of serum samples from ChAdOx1 nCoV-19 vaccine recipients at the PB28, PB90 and PB182 study visits were tested for anti-SARS-CoV-2 Spike IgG in a single laboratory assay. For PB28 samples, all NAAT+ cases were tested if sample volume allowed, and a proportion of non-cases were tested. Samples were tested blinded to case status. The data from noncases was obtained first, and consisted mainly of the samples processed for the initial application for emergency use, which needed 15% of samples included in the

efficacy cohort to be processed on validated assays. Subsequent to this PB28 samples from NAAT+ cases were sent for testing as they occurred, if not already including the 15%. A proportion of those participants with samples tested at PB28 were additionally tested at either PB90 only, or PB90 and PB182. We assume the samples to be missing at random, i.e. the missingness depended only on factors observed in the study. To account for the missing data, factors associated with sample availability were included as covariates in the model for antibody decay (see 'Antibody decay model' below).

Anti-SARS-CoV-2 Spike IgG was measured by a multiplex immunoassay on the MSD platform at PPD Laboratories. The assay sequences were based on the ancestral sequences from Wuhan, China. Assay validation included precision and ruggedness, dilutional linearity, selectivity, and relative accuracy for each SARS-CoV-2 antigens. Post-validation studies for stability and for conversion to the WHO standard, as well as the establishment of a cut-point, were performed. The lower limit of quantification (LLOQ) for anti-spike IgG was 33 AU/mL (0.21 BAU/mL). Data points which appeared visually distant from the main cluster of antibody levels at each visit were excluded as outliers (Supplementary Fig. C.2). Outliers were excluded as they may bias the results [36]. We excluded 8 anti-spike IgG results as outliers - three results at PB28 and four at PB90 whose anti-spike IgG levels were especially low, and one results at PB90 whose levels were especially high (Supplementary Table C.3, Supplementary Fig. C.2). The individuals from whom these samples were taken, and any other antibody measurements are included in the analysis - only the outlying antibody measurements are excluded.

The assay was analysed in its original scale. Results were then converted to the WHO international standard units, binding antibody units per mL (BAU/mL), by multiplying by a conversion factor supplied by the laboratory. The anti-spike IgG conversion factor from arbitrary units per mL (AU/mL) to WHO standard binding antibody units per mL (BAU/mL) was 0.00645 (95% confidence interval (CI): 0.00594, 0.00701). We did not apply the CI, as it represents the uncertainty due to measurement error in the assay, and our antibody model already accounts for this.

4.2.4 Study design

We included a subset of participants in the COV002 trial in our study, who met the following eligibility criteria: received two doses of ChAdOx1 nCoV-19 or MenACWY control vaccine, baseline seronegative to the SARS-CoV-2 N protein at first vaccination, and followed up to at least day 21 post second vaccination with no evidence of prior infection. Vaccinated participants received a first and second dose: either two standard doses, low dose followed by standard dose, or two low doses. Nine participants who received mixed schedules in error (one dose of ChAdOx1 nCoV-19 and one dose of MenACWY control) were excluded from the analysis (Supplementary Fig. C.1). One control individual had missing BMI, we imputed this with the mean value in the study.

Participants were considered at-risk of the study event from day 21 post second dose. The at-risk period ended at the first of: the onset of COVID-19 infection, withdrawal, unblinding, or the analysis cut-off date 30 June 2021. The risk of infection on a given day was assumed to relate to antibody levels 7 days earlier, to account for the gap between first being exposed to COVID-19 until a case is reported in the trial. Thus antibody measurements are excluded if taken prior to 14 days after the second dose, after the end of the at-risk period, or less than 7 days before the end of the at-risk period. Antibody levels are assumed to decay exponentially from 14 days after the second dose.

Participants who tested NAAT+ during the at-risk period were defined as cases, those who also had at least one qualifying symptom were defined as symptomatic cases, while those who did not return a positive test during the at-risk period were defined as non-cases.

4.3 Methods

4.3.1 Statistical Analysis

We modelled antibody decay and the risk of COVID-19 infection in a two-stage joint model for longitudinal and survival data. Joint models are a popular method to estimate the relationship between a longitudinal biomarker and the risk of a health outcome, especially in studies of disease progression such as HIV and cancer trials [29]. We employed a two-stage multiple imputation approach (Chapter 2) to overcome convergence issues that

arose when attempting to fit a full joint model. The reasons for the full joint model not converging likely include the model containing many random effects and non-linear transformations, and the sparse availability of antibody data.

Antibody decay model In the first stage, we fit a log-linear Bayesian mixed effects model to post-vaccination antibody decay for ChAdOx1 nCoV-19 recipients (the vaccine arm). The Bayesian approach easily accounts for the complex structure of the data and the model, including missing antibody observations for some individuals, and unknown true antibody levels which change over time. We use the model to predict the true antibody levels, incorporating uncertainty in the predictions.

Consider a reordered dataset such that individuals $i = 1, \ldots, n$ are in the vaccine arm, and $i = n+1, \ldots, N$ are in the control arm. For individual $i = 1, \ldots, n$ in the vaccine arm, let $A_i(t)$ be the true antibody level at time t, and $Y_i(t_{ij})$ be the observed antibody level at time t_{ij}, $j = 1, \ldots, m_i$. Time refers to the time since the start of the at-risk period (21 days post second dose) unless stated otherwise. Note for a given individual i, the number of antibody observations m_i may be 0. We assume

$$\log Y_i(t_{ij}) = \log A_i(t_{ij}) + \epsilon_{ij} \tag{4.1}$$

where random error $\epsilon_{ij} \sim t_\nu(0, \sigma^2)$ follows the Student-t distribution with ν degrees of freedom, location 0 and scale σ^2. We used the Student-t distribution with $\nu = 4$ degrees of freedom as QQ-plots showed better fit than using Gaussian error (Supplementary Figure C.20), and it is more robust to outliers than a Gaussian error distribution [36].

We assume the true log antibody values $A_i(t)$ decay linearly from 14 days post second dose.

$$\log A_i(t) = b_{0i} - \exp(b_{1i})\, t \tag{4.2}$$

where $a_{0i} = b_{0i}$ and $a_{1i} = -\exp(b_{1i})$ are the random intercept and slope respectively. The negative exponential function requires the random slopes to be negative. This was included because post-vaccination antibody levels decrease after reaching their peak level (rather than increase), and to avoid convergence issues in the second part of the model.

Random effects b_{0i}, b_{1i} follow a Gaussian distribution given by

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha_0 + x_{Li}^\top \beta_{L0} \\ \alpha_1 + x_{Li}^\top \beta_{L1} \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \rho\tau_0\tau_1 \\ \rho\tau_0\tau_1 & \tau_1^2 \end{pmatrix} \right) \tag{4.3}$$

independently for each i. Here X_L is a design matrix of p_L covariates which affect the antibody trajectories with ith row $x_{Li}^\top$, and corresponding effects on the intercept and slope given by β_{L0}, β_{L1} respectively. Parameters α_0 and α_1 give the expected intercept and slope for a reference individual with $x_{Li} = 0$, τ_0^2 and τ_1^2 give the variances of the random effects, and ρ the correlation between the random effects.

The model included an effect on both the peak antibody response (intercept) and rate of decay (slope) due to the following covariates (columns of X_L): age group (18-55 years, 56-69 years, 70+ years), sex (female, male), ethnicity (white, non-white), comorbidity (none, comorbidity), BMI (<30 kg/m^2, ≥30), time interval between first and second dose (≥12 weeks, 9-11 weeks, 6-8 weeks, <6 weeks), healthcare worker (HCW) status (not a HCW, HCW facing <1 COVID-19 patient per day, HCW facing ≥1 COVID-19 patient per day), initial dose (standard dose, low dose). The probability of missing antibody data was observed to be related to whether an individual tested positive and their follow-up time. To account for this, we also included covariate effects on the intercept and slope (columns of X_L) due to: (i) returning any positive test during the study, (ii) primary symptomatic infection, and (iii) the Nelson–Aalen estimate of cumulative hazard for any COVID-19 infection in the vaccine arm. This can be seen as adapting guidance on accounting for a time-to-event outcome in multiple imputation [64] to the case of two competing risks of non-symptomatic COVID-19 infection and symptomatic COVID-19 infection.

For the individuals $i = n+1, \ldots, N$ in the control arm, we assume their antibody levels are zero, $A_i(t) \equiv 0$.

We first standardised the data; after running the model we transformed the parameters back to the scale of the original dataset. The model was fit using Hamiltonian Monte Carlo in Stan [6]. Four chains were run, each of 20,000 iterations. The first 5,000 iterations were discarded as burn-in. Convergence was checked by ensuring the Rhat convergence diagnostic (potential scale reduction factor) was less than 1.01 [108]. See the Supplementary

Information (Section C.2.3) for the reported computation time, and other additional details.

Risk of infection model In the second stage, we fit a separate Cox model [34] to the time until COVID-19 onset for each outcome (primary symptomatic COVID-19 infection or any COVID-19 infection), using the output from the antibody model to predict the risk of infection. We employed multiple imputation (Chapter 2, [7]), which accounts for the uncertainty in the predicted antibody levels in our estimation of the hazard (instantaneous risk) for infection. A previously published multiple imputation joint model was shown to reduce bias compared with single imputation simple (or "naïve") two-stage approaches [46]. We first sampled 60,000 antibody trajectories from the antibody model for each vaccinated individual, from which we imputed the predicted antibody levels at each relevant time. For each sample, we fit a Cox model for the time to infection, which depends on the time-changing individual antibody levels.

Let S_i be the calendar time when individual i enters the at-risk period and T_i the total time at risk. Let $Z_i = 1$ if individual i is in the vaccine arm, and $Z_i = 0$ if they are in the control arm. Let X_D be a design matrix of covariates which affect the risk of infection, with row $x_{Di}^\top$ for the ith individual. We stratify the baseline hazard function by study site G_i, as the rate of COVID-19 in the population varied by geographical study site during the trial. We assume a 7 day incubation period from exposure to a COVID-19 infection being reported in the study. Hence the risk of infection at a given time depends on the true antibody levels 7 days prior. Then the hazard $h_i(t)$ for individual i at time t depends on the vaccine/control arm Z_i, the true antibody level $A_i(t-7)$, and covariates x_{Di}

$$h_i(t) = \lambda_{G_i}(S_i + t) \exp\left(Z_i\left(\gamma A_i(t-7) + \zeta\right) + x_{Di}^\top \beta_D\right) \tag{4.4}$$

$\lambda_{G_i}(s)$ is the unknown baseline hazard function for study site G_i which varies with calendar time, β_D is the effect of covariates x_{Di} on the log hazard, γ is the slope of the antibody level effect of $A_i(t-7)$, and ζ is the intercept of the antibody level effect.

We included covariate effects in X_D which may affect the risk of infection: age group (18-55 years, 56-69 years, 70+ years), sex (female, male), ethnicity (white, non-white),

comorbidity (none, comorbidity), BMI (<30 kg/m^2, ≥30), healthcare worker (HCW) status (not a HCW, HCW facing <1 COVID-19 patient per day, HCW facing ≥1 COVID-19 patient per day).

For each Cox model, we then drew from the asymptotic distribution of the maximum likelihood estimators, to approximate a Bayesian posterior with diffuse improper priors (Chapter 2, [65]). We combined this draw with the corresponding antibody model parameters to form a draw from the approximate joint distribution. Quantities of interest, such as vaccine efficacy, can then be calculated from each draw. Posterior medians and 95% CrIs are calculated based on quantiles of the posterior samples.

Calculating vaccine efficacy Vaccine efficacy (VE) at a given antibody level A is then calculated as 1 minus the hazard ratio between a vaccinated individual with that level of antibody, and a control individual with matching covariates.

$$\mathrm{VE}(\mathrm{A}) = 1 - \exp(\gamma \mathrm{A} + \zeta) \tag{4.5}$$

with γ the slope of the vaccine-induced antibody effect and ζ the intercept. Antibody level at a given VE is defined by the inverse of Eq (4.5). For individuals $i = 1, \ldots, n$ in the vaccine arm, let $\mathrm{A}_i(t-7)$ be the latent antibody level at time $t-7$. Then the mean VE at time t is the mean VE among all the latent antibody levels of vaccinated individuals at time $t-7$.

$$\mathrm{VE}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{VE}(\mathrm{A}_i(t-7)) = 1 - \frac{1}{n}\sum_{i=1}^{n} \exp(\gamma \mathrm{A}_i(t-7) + \zeta) \tag{4.6}$$

The qth quantile of VE at time t, $\mathrm{VE}_q(t)$, is equal to the VE at the qth antibody quantile $A_q(t-7)$ 7 days prior, $\mathrm{VE}_q(t) = 1 - \exp(\gamma A_q(t-7) + \zeta)$.

Calculating covariate effects on vaccine efficacy To calculate covariate effects on vaccine efficacy, we predicted efficacy in new individuals with given covariate values. Note the event indicators and Nelson–Aalen (N–A) estimate of cumulative hazard for a new individual are unknown, as they depend on the infection outcome. We first built an imputation model to sequentially impute the missing event indicators (infection, symptomatic)

and N–A estimate, which appear in the longitudinal antibody model, given the known covariates (Supplementary Methods C.1.4). We imputed these missing covariates n times. For each set of imputed covariates, we drew from the predictive distribution of antibody trajectories for a new individual with the given covariate values, $\mathrm{A}_j^{\mathrm{pred}}(t), j = 1, \ldots, n$. We then averaged these to calculate the mean risk of infection given the known covariate values

$$\mathrm{VE}^{\mathrm{pred}}(t) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{VE}(\mathrm{A}_j^{\mathrm{pred}}(t)) = 1 - \frac{1}{n}\sum_{j=1}^{n} \exp(\gamma \mathrm{A}_j^{\mathrm{pred}}(t) + \zeta) \tag{4.7}$$

Software Data analysis was performed in R version 4.3.1 [76], using RStudio [109], and relying extensively on the following packages: survival [110], RStan [69], snowfall [111], ggplot2 [112]. RColorBrewer [113] and tidyr [114] were also used. Hamiltonian Monte Carlo algorithms were run in Stan [6] via the R package RStan [69].

4.3.2 Assessing model fit

We generated posterior predictive checks in order to explore how well our model fits the data [115]. Such checks compare the observed data with many replicated datasets drawn from the posterior predictive distribution. If the model fits the data well, the observed data should look like a typical draw from the replicated data. We also considered subject-specific residual plots, comparing the observed antibody levels to the posterior mean (MAP estimate) for each individual.

Generating replicated data In order to generate replicated data, we assumed the timing and availability of antibody measurements was fixed to be the same as the observed data, and drew observed antibody values at these times from the posterior predictive distribution. That is, for individual i with an antibody observation at time t_{ij}, we generated replicated data

$$\log Y_i(t_{ij})^{\mathrm{rep},k} = b_{0i}^{(k)} - \exp(b_{1i}^{(k)})t_{ij} + \sigma^{(k)}\epsilon_{ij}^{(k)} \tag{4.8}$$

using the kth draw from the posterior distribution for the intercept $b_{0i}^{(k)}$, log negative slope $b_{1i}^{(k)}$, and error standard deviation $\sigma^{(k)}$, and drawing error $\epsilon_{ij}^{(k)} \sim t_4(0, 1)$. We did this for $k = 1, \ldots, K$, where $K = 60,000$, i.e. for all posterior draws. We then compared various functions of the replicated longitudinal data $\log Y_i(t_{ij})^{\mathrm{rep},k}$ and the observed longitudinal

data $\log Y_i(t_{ij})$.

Subject-specific residuals To understand how well the model fits the data for all longitudinal datapoints, we calculated the posterior mean standardised subject-specific residual, defined by $r_{ij} = \frac{1}{K}\sum_{k=1}^{K}\left(\log Y_i(t_{ij})^{\text{rep},k} - \log Y_i(t_{ij})\right)/\sigma^{(k)}$. We plotted these against time to test how the model fits the longitudinal trajectories. We also plotted the residuals against the posterior means at each study visit, giving a plot analogous to residuals against fitted values for a frequentist model, to demonstrate how the observed values relate to the predicted underlying antibody levels.

Error distribution To test the appropriate choice of distribution for the measurement error ϵ_{ij}, we examined a QQ-plot comparing the distribution of the residuals r_{ij} to a Student t distribution with 4 degrees of freedom. Since our model assumed the error to be generated from a t_4 distribution, the QQ-plot should show roughly a straight line in the bulk of the distribution.

Median antibody level at each visit In order to test how well the model matches the observed longitudinal trajectory over time, we plotted the posterior predictive distribution of the median log antibody level among observations at each study visit. We compared this with the observed median log antibody level at each study visit.

Variance of antibody levels at each visit We further aimed to test how well the model replicates the variance in the observed antibody values at each visit. To do this, we plotted the posterior predictive distribution of the variance of the log antibody levels among observations at each study visit, and compared this with the value given by the observed data.

Curvature in log antibody decay We tested whether the model replicates any possible curvature in the log antibody decay trajectories. To do this we define the curvature estimated on 3 data points to be

$$\text{Curvature}(\log Y(t_1), \log Y(t_2), \log Y(t_3)) = \frac{\log Y(t_3) - \log Y(t_2)}{t_3 - t_2} - \frac{\log Y(t_2) - \log Y(t_1)}{t_2 - t_1} \tag{4.9}$$

which is exactly the difference in empirical slopes between the pairs $\log Y(t_3), \log Y(t_2)$ and $\log Y(t_2), \log Y(t_1)$. Since the model assumes linear decay in log antibodies, the curvature for any 3 observations of the same individual will have expectation 0 under the posterior, and be independent of the curvature for any other individual, given the error variance σ. Hence we plot the the posterior predictive distribution of the mean curvature among all individuals who have 3 antibody measurements available, one at each visit (PB28, PB90 and PB182), against the observed sample mean curvature.

4.4 Results

We used data from the COV002 trial [3, 4] to assess antibody waning, vaccine efficacy (VE) waning, and correlates of protection against any COVID-19 infection and symptomatic COVID-19 infection at the time of exposure. COV002 is a randomized single-blind vaccine efficacy trial of the ChAdOx1 nCoV-19 vaccine, conducted in the United Kingdom. Participants were considered at risk of infection from 21 days after their second dose of vaccine, over a period from 18 July 2020 to 30 June 2021. This analysis includes 4605 ChAdOx1 nCoV-19 vaccinated individuals (vaccine arm) and 4423 individuals who received a MenACWY control vaccine (control arm). Supplementary Table C.1 summarises baseline characteristics for the vaccine and control arm. Supplementary Fig. C.1 summarises exclusions for the two groups. In the vaccine and control arm, 207 (4.5%) and 377 (8.5%) tested positive for COVID-19 respectively, and 71 (1.5%) and 217 (4.9%) were primary symptomatic COVID-19 cases respectively (Supplementary Table C.1).

4.4.1 Waning anti-spike IgG antibody levels

Blood samples were taken at study visits in windows centred at 28 days, 90 days and 182 days after the second dose. We refer to these as the PB28 (post-boost + 28 days), PB90 and PB182 study visits respectively. We first consider observed antibody responses for vaccine arm participants. Samples were sent for anti-spike IgG antibody testing by case-cohort sampling. Samples taken after a reported COVID-19 infection were excluded from analysis. Observed antibody responses differed between those who did not return a positive COVID-19 test during the at risk period (non-cases), all who later tested positive

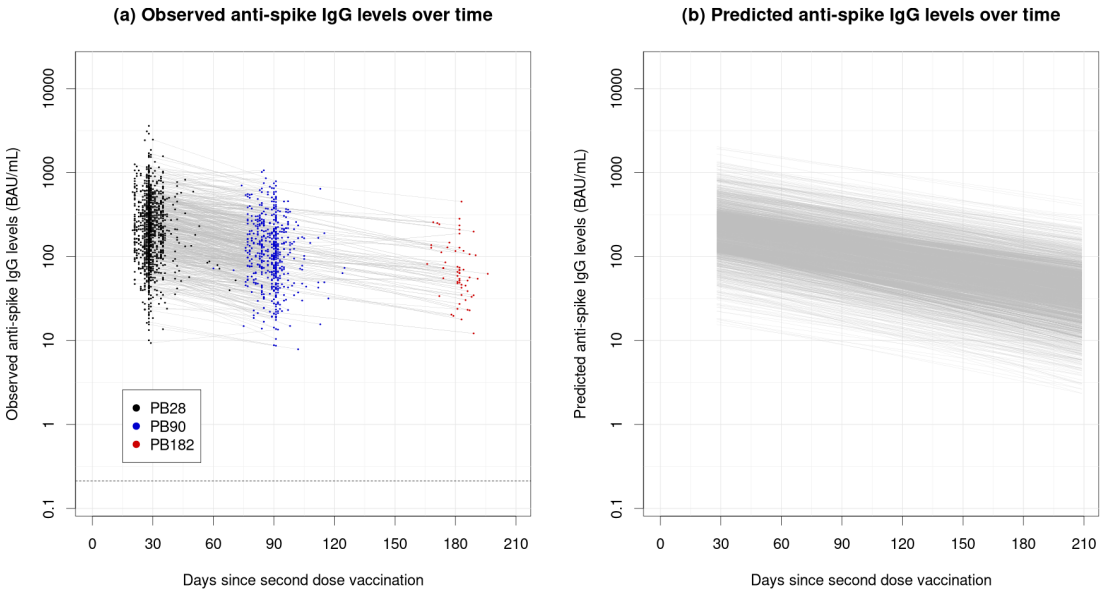


Figure 4.1: **Observed and estimated anti-spike IgG antibody levels over time since vaccination. (a)** Observed anti-spike IgG antibody levels over time since vaccination. Each point represents an observation, and consecutive observations from the same individual are connected by a line. Black, blue and red points represent observations taken at the PB28, PB90 and PB182 study visits respectively. The dashed horizontal line shows the lower limit of quantification (LLOQ), 0.21 BAU/mL. **(b)** The posterior mean estimate of the latent true antibody levels over time for each individual. Note 1b represent estimates of the latent true antibody levels which are free from measurement error, whereas the observed antibody levels in 1a are measured with error, so show higher variability.

(cases) and those who tested positive with primary symptoms (symptomatic cases). At the PB28, PB90 and PB182 study visits, antibody observations were available from 1155 (26.3%), 519 (11.8%), 58 (1.3%) non-cases, 173 (83.6%), 64 (30.9%), 1 (0.5%) cases, and 59 (83.1%), 23 (32.4%), 1 (1.4%) symptomatic cases respectively (Supplementary Table C.2). At the PB28, PB90 and PB182 study visits, observed antibody levels had a median (interquartile range (IQR)) of 219 (119, 383), 115 (66, 206), 65 (45, 117) BAU/mL among non-cases, 197 (103, 322), 93 (47, 168), 30 (NA – only one measurement) BAU/mL among cases and 165 (100, 276), 78 (62, 178), 30 (NA – only one measurement) BAU/mL among symptomatic cases respectively. Fig. 4.1a shows the observed antibody levels against time, plotted on a log scale. The figure indicates (i) log antibody levels decay linearly, (ii) different individuals have different levels of initial log antibody response and (iii) may have different rates of decay.

In total, 923 antibody observations were available from 853 (18.5%) control participants.

Among control participants, antibody levels had a median (IQR) of 0.3 (< 0.21 (lower limit of quantification (LLOQ)), 0.5) BAU/mL (Supplementary Fig. C.2, Supplementary Table C.2). In our subsequent modelling we did not model antibody levels in the control arm, nor include an antibody effect on risk for control participants.

We modelled the decay in antibody levels over time among the vaccinated participants, allowing different initial antibody responses and rates of decay for different individuals (4.3). Predicted median anti-spike IgG levels at day 28, 90 and 182 were 217 (95% credible interval (CrI): 208, 227), 119 (113, 126), 49 (45, 54) BAU/mL respectively. The median half-life of the anti-spike IgG antibody levels was 74 days (95% CrI: 69, 79). Fig. 4.1b shows the predicted antibody levels over time for each individual in the study on a log-scale.

Predicted antibody levels and rates of decay varied among individuals. The estimated 25% and 75% quartiles for the antibody response at day 28 were given by 124 (95% CrI: 117, 130) and 378 (359, 398) respectively (Supplementary Fig. C.9). The half-life of the antibody levels among individuals in the study had an estimated 25% and 75% quartile of 62 days (95% CrI: 58, 67) and 87 days (79, 96). Among non-cases, the median (IQR) predicted anti-spike IgG level at day 28 was 219 (95% CrI: 209, 229), among cases 192 (175, 209) and among symptomatic cases 169 (146, 196). The median half-life was 74 days (95% CrI: 70, 79) among non-cases, 60 (52, 69) among cases and 54 (45, 69) among symptomatic cases.

4.4.2 Correlates of protection

We modelled how the protective effect of the vaccine relates to the modelled antibody levels over time. The instantaneous risk of COVID-19 infection for a given individual in our model depends on: (i) the current background levels of COVID-19, (ii) covariates which may affect risk of COVID-19 infection (iii) the modelled level of antibody 7 days prior, and (iv) a non-antibody effect due to vaccination (4.3). We assume a 7 day gap between initial exposure to COVID-19 and a positive test being reported.

Supplementary Fig. C.3-C.6 show an estimate of the baseline hazard of infection during the trial period, the number of participants at risk over time since vaccination, the timings of antibody samples, and the timings of COVID-19 infections in the study respectively.

Differences in baseline risk of infection due to different covariates are reported in Supplementary Fig. C.7 and Supplementary Table C.7.

We assessed how predicted antibody level at the time of exposure correlates with protection against COVID-19 infection. Predicted anti-spike IgG titer was correlated with protection. Fig. 4.2 shows the relationship between VE and anti-spike IgG antibody level predicted by our model. Anti-spike IgG values of 10, 100 and 1000 BAU/mL represent the 0.0%, 17.7% and 97.1% quantiles of predicted antibody levels at day 28 post second dose, and the 5.4%, 78.1% and 100% quantiles of predicted levels at day 182 respectively. At anti-spike IgG levels of 10, 100 and 1000 BAU/mL, the VE against symptomatic infection was 36.1% (95% CrI: −20.7, 63.0), 76.1% (64.4, 88.7), 100% (97.3, 100) respectively. We observe lower efficacy against any COVID-19 infection at the same antibody levels, 14.3% (95% CrI: −22.5, 38.1), 55.3% (44.4, 66.3), 99.9% (96.2, 100) respectively. Vaccine efficacy of 50%, 70%, and 90% against symptomatic infection was achieved at anti-spike IgG levels of 32 (95% CrI: 0, 57), 78 (51, 139), 182 (105, 507) BAU/mL respectively (Supplementary Table C.8). Against any COVID-19 infection higher antibody levels were required for the same efficacy, 84 (95% CrI: 59, 125), 156 (109, 285), 311 (194, 662) BAU/mL respectively.

4.4.3 Waning vaccine efficacy

As the antibody levels decayed over time, the vaccine efficacy waned. Fig. 4.3 shows the mean VE over time since the second dose, averaged over all individuals in the study. We report efficacy 7 days after the planned study visits. The efficacy at day 35 after the second dose was estimated to be 88.1% (95% CrI: 77.2, 93.6) against symptomatic infection and 76.0% (64.3, 83.7) against all COVID-19 infection. The efficacy waned to 77.7% (68.5, 84.1) and 60.8% (51.4, 68.4) at day 97 and 60.4% (44.6, 71.0) and 39.6% (26.1, 50.2) at day 189 against symptomatic infection and all COVID-19 infection respectively.

Levels of protection varied based on individual antibody responses. At day 28, the predicted 25%, 50% and 75% antibody quantiles were 124 (95% CrI: 117, 130), 217 (208, 227), 378 (359, 398) BAU/mL respectively (Supplementary Fig. C.9). The corresponding VE quantiles at day 35 against symptomatic infection were 81.4% (95% CrI: 68, 93.7), 93.1% (76.6, 99.4), 98.8% (85.2, 100), and against any COVID-19 infection were

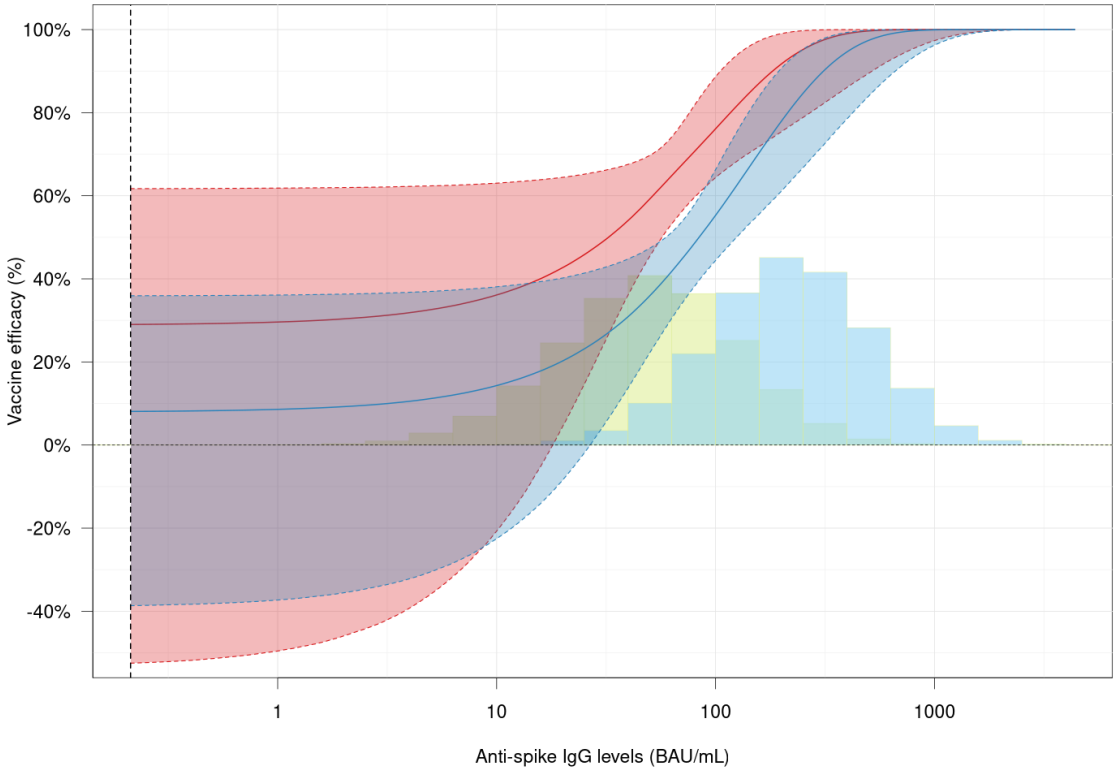


Figure 4.2: **Vaccine efficacy as a function of anti-spike IgG levels.** VE (%) plotted against anti-spike IgG antibody levels (BAU/mL) on a log scale. The red and blue curves and shaded regions show the posterior median and associated 95% credible intervals for VE against primary symptomatic COVID-19 infection, and against all NAAT+ positive tests respectively. The vertical dotted line shows the lower limit of quantification (LLOQ), 0.21 BAU/mL. The blue histogram shows the distribution of the predicted anti-spike IgG antibody levels in the study at 28 days since the second dose, and the green histogram at 182 days post second dose. The two histograms overlap. The heights of the histograms are not to scale.

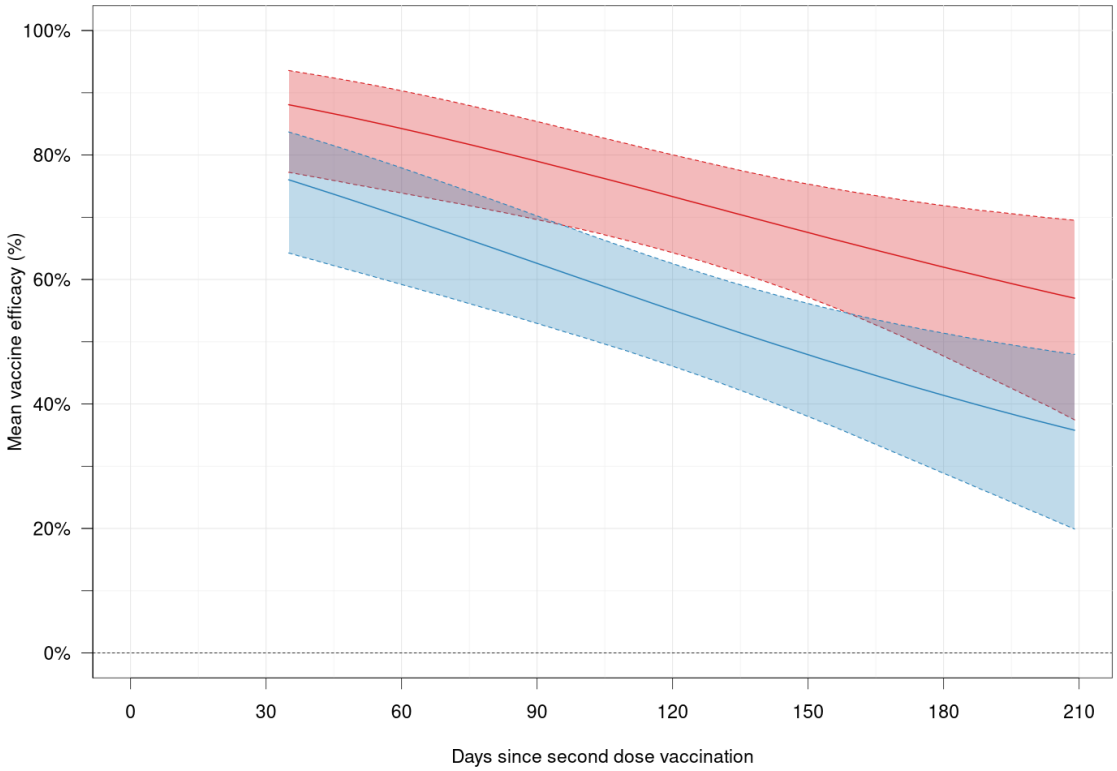


Figure 4.3: **Vaccine efficacy as a function of time since second dose vaccination.** The red curve and shaded region shows the posterior median and associated 95% credible interval for VE against symptomatic COVID-19 infection, and the blue curve and shaded region shows the posterior median and 95% credible interval for VE against all positive COVID-19 tests.

62.1% (49.7, 75.1), 80.5% (63.1, 92.7), 93.8% (77.1, 99.1) (Supplementary Fig. C.8). At day 182, the respective antibody quantiles were 26 (23, 29), 49 (45, 54), 91 (82, 102) BAU/mL, and corresponding VE quantiles were 46.3% (16.1, 65.3), 58.8% (44.9, 69.5), 73.9% (62.6, 85.6) against symptomatic infection, and 23.6% (−0.8, 41.5), 35.5% (21.6, 47.1), 52.5% (42.1, 62.3) against any COVID-19 infection.

4.4.4 Covariate effects on antibody levels and vaccine efficacy waning

We report the estimated effects of different covariates on initial antibody response, subsequent antibody decay, and vaccine efficacy over time. Our model predicts a large decrease in antibody levels for those with shorter intervals between the first and second dose (Fig. 4.4, Supplementary Table C.5). Compared with participants with dose intervals of ≥12 weeks, participants with dose intervals of 9-11 weeks, 6-8 weeks and <6 weeks have predicted anti-spike IgG antibody levels 83% (95% CrI: 74, 93), 70% (61, 81), 51% (43, 61) as high respectively, at 28 days after the second dose. This is likely because the shorter interval allows less time for the memory B-cells to mature, leading to a decreased antibody response at the second dose. We also predict a larger antibody response due to receiving a low dose as the first dose compared with a standard dose, although 95% credible intervals contain a null effect. We predict lower antibody responses at 28 days after the second dose in older age groups, and among healthcare workers. Higher antibody responses are predicted in females than males, and in non-white participants. Shorter half-lives of antibody levels are predicted for females, and individuals with a BMI ≥ 30 (Fig. 4.4, Supplementary Table C.6).

To consider covariate effects on VE, we compare with reference values of age 18-55 years, female, white ethnicity, no comorbidities, BMI <30 kg/m^2, ≥12 week interval between first and second dose, non-HCW (non-healthcare worker), and receiving a standard dose as their first dose. We then vary each of these covariates and plot VE for each covariate combination against symptomatic infection in Fig. 4.5 and against all COVID-19 infection in Supplementary Fig. C.10. We see that shorter intervals between first and second dose predict a lower VE throughout the study. VE against symptomatic infection at day 35, 97 and 189 was 93.1% (CrI: 81.6, 97.6), 85.0% (72.1, 92.1) and 68.7% (53.2, 77.4) respectively for a reference individual (with ≥12 week interval), compared with 83.2%

(71.6, 91.0), 71.7% (58.2, 80.4) and 55.0% (27.0, 69.2) respectively for an individual with <6 week interval. We also observe a lower efficacy after around 90 days in individuals aged 70 years or older compared with the 18-55 and 56-69 years age groups. VE for an individual aged 70 years or older at day 35, 97 and 189 was 90.7% (78.8, 96.4), 79.2% (67.5, 88.2) and 59.9% (35.2, 72.7) respectively, compared with a reference individual above.

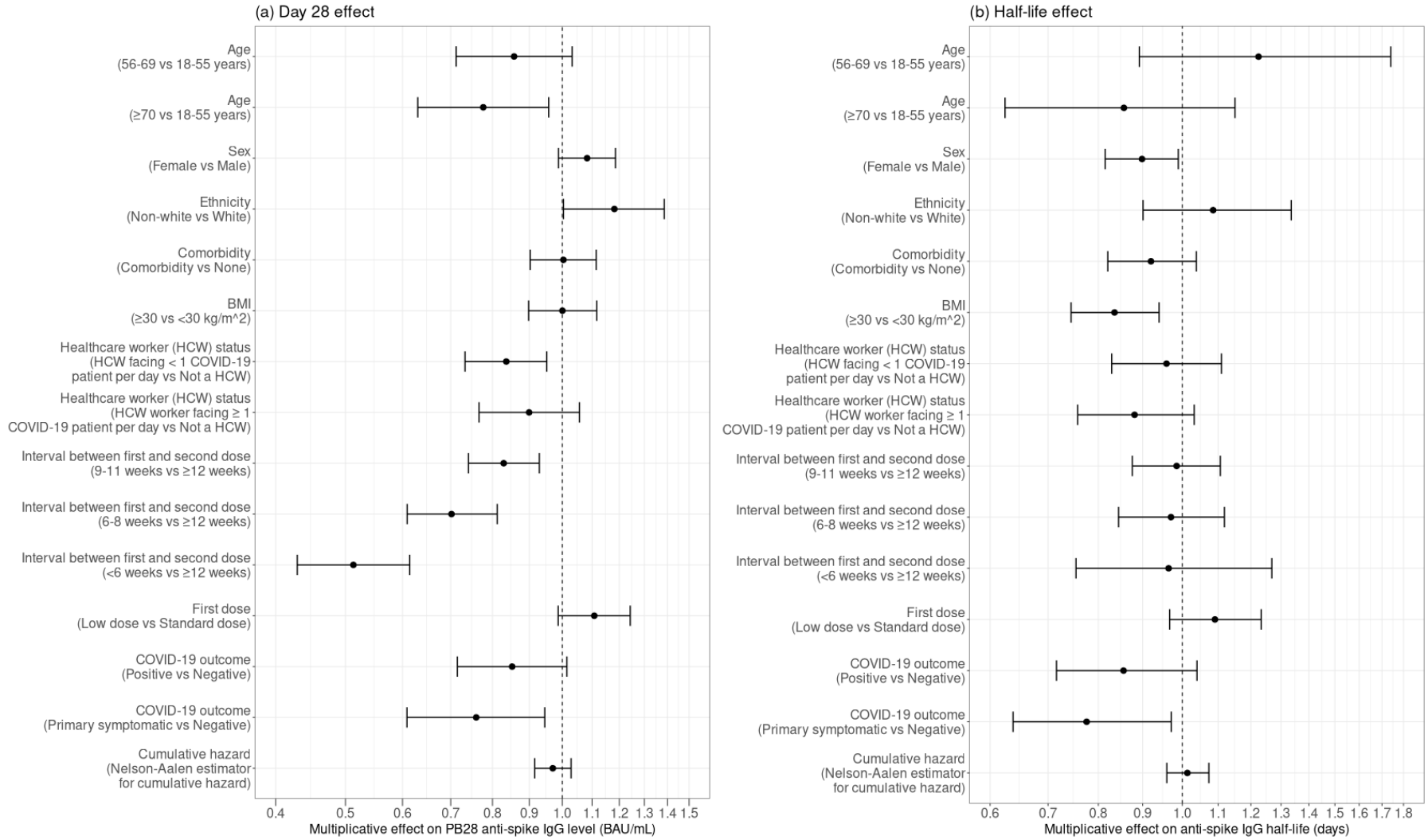


Figure 4.4: **Covariate effects on anti-spike IgG antibody levels.** The multiplicative effects on **(a)** day 28 anti-spike IgG antibody levels and **(b)** half-life of anti-spike IgG antibody decay for different covariates. The effects are plotted on a log scale. These results are also given in Supplementary Tables C.5 & C.6.

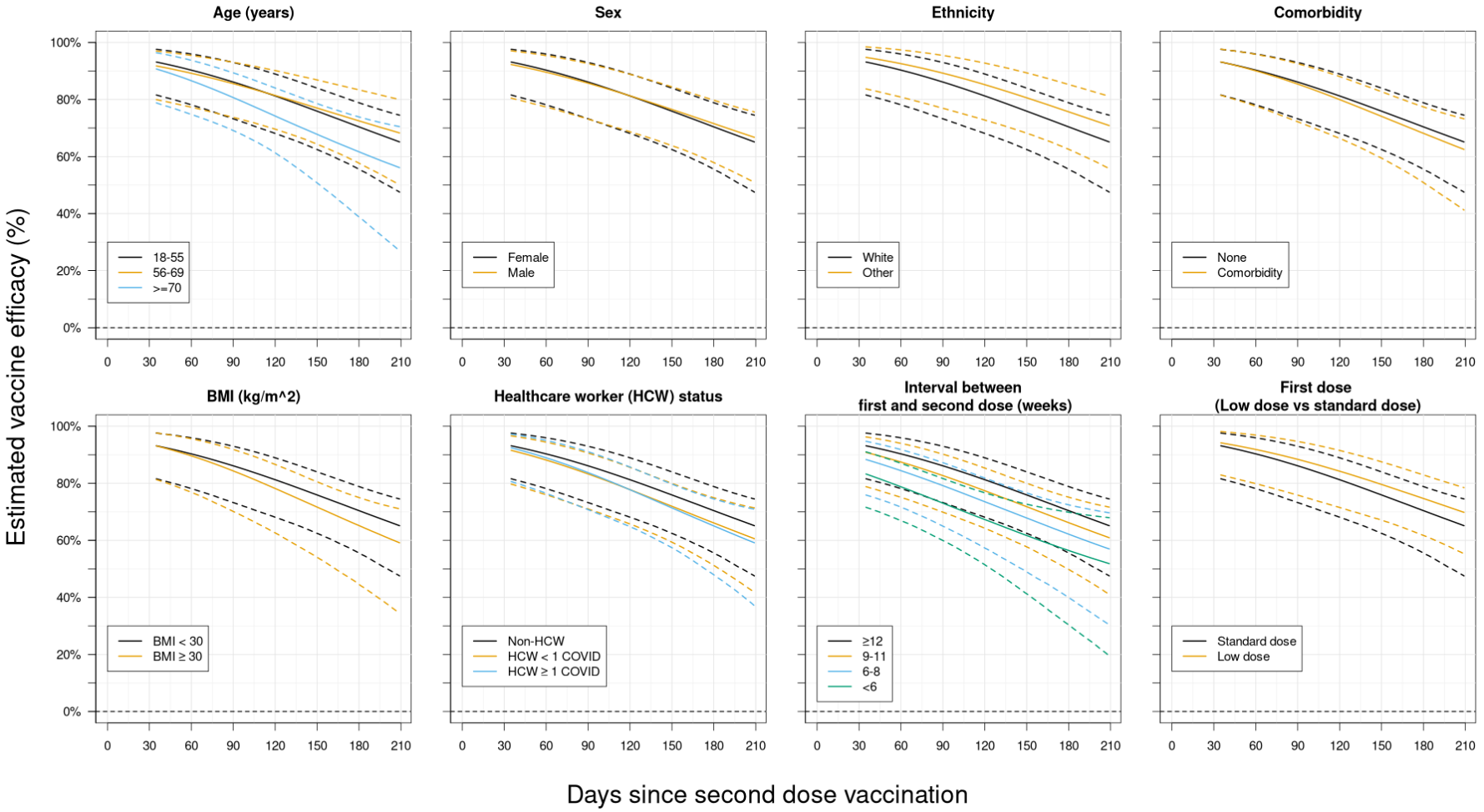


Figure 4.5: **Vaccine efficacy against symptomatic COVID-19 infection plotted against time since vaccination for different covariate combinations.** The black lines represent an individual with reference covariates: age 18-55, female sex, white ethnicity, no comorbidities, BMI<30 kg/m^2, ≥12 week interval between first and second dose, non-HCW (healthcare worker), and receiving a standard dose as their first dose. All other coloured lines represent an individual with one of the above covariates changed from these reference values. The solid lines show posterior medians and dashed lines 95% credible intervals. "HCW < 1 COVID" refers to a healthcare worker facing <1 COVID-19 patient per day, "HCW ≥ 1 COVID" refers to a healthcare worker facing ≥1 COVID-19 patient per day.

Table 4.1: **Comparison of methods and results with previous correlates of protection studies**

Paper	Antibody type	Vaccine	Doses	Dominant variant	Timing of antibody effect on risk‡	Accounts for measurement error	Accounts for antibody waning between measurement and exposure	Accounts for individual variation in slopes	Randomised vaccine /control assignment	Unvaccinated baseline comparator	Location of result in paper	Endpoint	Antibody level (BAU/mL unless otherwise stated)	Vaccine efficacy (%)	Our result
Feng et al. (2021) [5]	Spike nAb* others*	ChAdOx1†	Two doses	Pre-Delta (Alpha B.1.1.7 and B.1.177)	Peak (∼28 (14-42) days post second dose)	x	x	x	✓	✓	Table 2 (Fig. 4a)	Symptomatic Asymptomatic*	29 (NC, 83) 54 (NC, 143) 113 (NC, 245) 264 (108, 806) 899 (369, NC)§	50% 60% 70% 80% 90%	32 (0, 57) 52 (0, 82) 78 (51, 139) 117 (76, 266) 182 (105, 507) BAU/mL
Benkeser et al. (2023) [86]	Spike nAb* others*	ChAdOx1†	Two doses	Pre-Delta (B.1.2)	Peak (∼28 days post second dose)	x	x	x	✓	✓	In text and Fig. 5a	Symptomatic	21 100 1000	21.1% (−361, 79.3) 62.9% (−55.0, 85.7) 88.1% (52.0, 97.0)	43.3% (6.6, 64.6) 76.1% (64.4, 88.7) 100% (97.3, 100) VE (%)
Wei et al. (2022) [81]	Spike	ChAdOx1† BNT162b2*	Two doses	Delta (B.1.617.2)	Exposure (21-59 days prior to infection)	x	x	x	x	✓	Fig. 4c	Symptomatic Any infection* Ct<30*	25 100 400	55% (40, 65) 68% (59, 75) 82% (74, 88)	45.8% (14.5, 65.2) 76.1% (64.4, 88.7) 99.0% (86.2, 100) VE (%)
Wei et al. (2023) [82]	Spike	ChAdOx1† BNT162b2*	Two doses plus booster or infection	Omicron BA.4/5 (B.1.1.529)	Exposure (21-59 days prior to infection)	x	x	x	x	Compared with vaccine non-responders¶	Supp. Fig. 1 and Fig. 1a	Any infection Symptomatic* Ct<30*	920 (766, 1092) 1740 (1502, 2331) 3680 (2831, 4500)	50% 70% 80%	32 (0, 57) 78 (51, 139) 117 (76, 266) BAU/mL
Follmann et al. (2023) [47]	nAb	mRNA-1273	Two doses	Pre-Delta period	Exposure (7 days prior to infection)	x	✓	x	✓	✓	In text and Fig. 2	Symptomatic	100 IU50/mL 1000 IU50/mL (Neutralising antibody)	93% (91, 95) 97% (95, 98)	No comparison possible
Our analysis	Spike	ChAdOx1†	Two doses	Pre-Delta (Alpha B.1.1.7 and B.1.177)	Exposure (7 days prior to infection)	✓	✓	✓	✓	✓	In text and Fig. 4.2	Symptomatic Any infection*	10 100 1000	36% (−21, 63) 76% (64, 89) 100% (97, 100)	

We include studies of the ChAdOx1 nCoV-19 vaccine, as well as studies which relate antibody levels at the time of exposure to risk of infection, for comparison. Results are given with 95% confidence intervals or credible intervals, depending on the paper. Some results show the antibody level required for a certain vaccine efficacy, others show the vaccine efficacy associated with a certain antibody response. *We do not report these results here, see the original paper. †ChAdOx1 refers to the ChAdOx1 nCoV-19 vaccine in this table. ‡Some studies relate antibody levels at the peak response after vaccination to risk of infection, others relate antibody levels at exposure to infection. This column denotes the time at which the antibody measurement was taken which relates to risk of infection, or the time at which the modelled antibody level relates to risk of infection (for analyses which use an antibody measurement directly or model antibody levels over time respectively). §NC = not computed (beyond data range). ¶Wei et al. (2023) define vaccine non-responders to be vaccinated individuals with a recent anti-spike IgG measurement of 16 BAU/mL.

4.4.5 Assessing model fit

In order to test the robustness of our results, we performed checks to assess how well the model fits to the data. We first test the fit to the longitudinal trajectories of log antibody levels. Residual plots indicate the model shrinks the observed antibody levels towards 0, for individuals with 1 or 2 observations (Supplementary Figures C.16). This is to be expected, due to measurement error, the inclusion of random effects, and the small number of observations per individual (0, 1 or 2). Posterior predictive checks indicate the model fits the antibody trajectory over time reasonably well (Supplementary Figures C.16, C.17, C.18). There was however a noticeable curvature in the observed antibody levels over time which the model is unable to capture, indicating log antibody levels may initially decay quickly before slowing down C.19, as suggested by previous studies [116, 117]. However, this only led to a small difference between posterior median predicted antibody levels and observed median antibody levels (Supp. Fig. C.17). Fitting a model where log antibody levels decay non-linearly may require either more dense sampling of antibody levels and greater availability of samples, or strong assumptions about the rate of initial and later decay.

4.5 Discussion

We estimated correlates of protection, vaccine efficacy waning and antibody waning after two doses of the ChAdOx1 nCoV-19 (AZD1222) vaccine, against any COVID-19 infection and symptomatic COVID-19 infection. We fit a two-stage joint model using multiple imputation (Chapter 2, [7]) which accounts for antibody waning and measurement error, allowing the risk of infection to depend on the estimated anti-spike IgG level at exposure. We found antibody levels are predictive of vaccine efficacy, which wanes over time, but continues to protect against the Alpha variant 209 days after the second dose. We estimated mean VE against symptomatic infection to be 88.1% (95% CrI: 77.2, 93.6) at day 35, waning to 60.4% (44.6, 71.0) at day 189 (Results). There was noticeable variation in VE between individuals, depending on the strength of their antibody response. We reported that a longer interval between the first and second vaccine dose gives improved protection, adding to existing evidence [106, 118, 81]. VE after 3 months since second dose vaccination was also lower in individuals aged 70 years or older than in younger age groups.

We used a two-stage joint model to relate modelled antibody levels at the time of exposure to the risk of infection. Joint models have been shown to reduce bias and improve power in estimating the relationship between longitudinal outcomes (e.g. antibody levels) and the time until an event (e.g. infection) [32, 46], as well as in estimating the trajectory of the longitudinal variable (antibody levels) [32]. Moreno-Bentacur et al. (2018) previously found using a multiple imputation-based joint model reduced bias, when compared with using the longitudinal observations directly to predict risk [46]. We include multiple antibody measurements per individual where available, allowing us to account for measurement error and variation between individual rates of decay. We further account for the informative censoring of antibody data after infection, and changing background infection rates in the study.

Our approach estimates "exposure-proximal" correlates of protection by relating modelled antibody levels at the time of exposure to the risk of COVID-19 infection. This approach more closely matches the biological mechanism than relating peak antibody levels to risk of infection. Table 4.1 compares our methods and results with previous studies

of COVID-19 correlates of protection. As outlined in the introduction, previous analyses have demonstrated both binding and neutralising antibodies measured at a fixed time after vaccination are associated with VE against COVID-19, including a study analysing data from the same COV002 trial [5], an analysis of data from a different trial of the ChAdOx1 nCoV-19 vaccine [86], as well as analyses of trials of other vaccines [83, 84, 85], and booster doses [92, 87]. These studies compared peak antibody levels and risk of infection, and generally report wide confidence intervals due to a lack of power. They did not account for antibody waning between the time of measurement and exposure, which may introduce a bias [29]. Other studies have estimated correlates of protection by relating antibody levels at the time of exposure to the risk of infection, by using a recent antibody measurement [81, 82], or a model to predict antibody decay over time [47]. However the studies using a recent antibody measurement (Wei et al. 2022, 2023) used observational data from a cohort study, and did not account for antibody waning between measurement and exposure, which may introduce bias [81, 82]. The study modelling antibody levels over time (Follmann et al. 2023) assumed the rate of antibody decay to be the same for all individuals in the study, which may introduce bias if decay rates vary [47]. All the above studies only used one antibody measurement per individual, meaning they are unable to account for measurement error in the antibody levels, which may bias the analysis [2]. None of these prior studies relate protection against COVID-19 infection to antibody levels at the time of exposure while also accounting for measurement error or variation between individual rates of antibody level decay.

We observe higher vaccine efficacy with much lower uncertainty at the same level of antibody, compared with the studies relating peak antibody levels to risk of infection [5, 86] (Table 4.1). The lower efficacy these studies report is likely due to antibody waning between the measurement and exposure, meaning higher initial antibodies would be required to give the same protection at exposure. These models also do not account for measurement error, which introduces noise to the data, likely causing the observed protective effect to be diluted [2]. Our results are broadly similar to those of Wei et al. (2022) who relate antibody levels at the time of exposure to risk of infection [81] (Table 4.1). Much higher antibody levels are required for similar levels of protection against the Omicron variant [82]. We are unable to compare with the results of Follmann et al. (2023), as their paper

uses neutralising antibodies, whereas we use anti-spike IgG [47].

We estimated the peak antibody response at 28 days after ChAdOx1 nCoV-19 vaccination, the subsequent rate of decay, and associated covariate effects. Table 4.2 compares our methods and results with previous studies. Wei et al. (2022) previously used a similar approach applied to non-randomised cohort data [81]. Their model does not account for the aforementioned informative censoring of antibody measurements, which may lead to bias [32]. This is because individuals with lower antibody levels will be more likely to test positive for COVID-19, after which antibody measurements will be excluded from analysis. Thus individuals with lower antibody levels are less likely to have antibody measurements available, introducing informative missing data, which should be accounted for. However, as cases are rare in the population, this bias may be small. We estimated higher peak anti-spike IgG levels than Wei et al. study [81] (Table 4.2). This may be partly due to differences in the populations enrolled in the two studies. Our estimated half-life of antibody decay was broadly similar to previous results (Table 4.2). We observed individuals with longer dose intervals had higher peak antibody levels (Fig. 4.4, Supplementary Table C.5), as previous studies have reported [118, 106, 81]. We also observed higher peak antibody levels in younger individuals (18-55 vs $\geq$70), non-healthcare workers and non-white participants. The higher peak in non-white participants was also observed by Wei et al. (2022) [81], however there is no consensus on the effect of age and healthcare work on antibody levels (Table 4.2). We reported slower antibody decay in participants with a BMI<30kg/m^2 (Fig. 4.4, Supplementary Table C.6), and detected slightly slower antibody decay in males, however this does not translate to a noticeable difference in VE waning (Fig. 4.5, Supplementary Fig. C.10). Neither of these differences have been reported elsewhere and it is unclear if they represent true effects. We did not detect the other effects on peak antibody observed by Wei et al. (2022) [81].

We reported predicted vaccine efficacy over time since second dose from our model, calculated from randomised controlled trial data. One previous analysis considered waning after a first dose of ChAdOx1 nCoV-19 [106], however the analysis was underpowered. Apart from this, we are not aware of any studies of ChAdOx1 nCoV-19 vaccine efficacy waning evaluated on randomised data. Table 4.3 and Supplementary Table C.9 compare our methods and results to previous studies of waning vaccine efficacy, with a focus on

studies of the ChAdOx1 nCoV-19 vaccine. These studies all use non-randomised observational data, including test-negative design studies [93, 94, 95, 98], and cohort studies [119, 95, 96, 97]. The test-negative design may introduce bias due to unmeasured differences in health-seeking behaviour between the vaccinated and unvaccinated arms, likely reducing the observed VE [100, 101]. Results may also not carry over to the wider population, as those included in the study will likely be those who test more frequently. Cohort studies are similarly vulnerable to biases due to unmeasured confounding causing differences in risk of infection between the unvaccinated and vaccinated populations. Such biases may increase over time, as the unvaccinated group are increasingly unvaccinated by choice, which may correlate with fewer health-seeking behaviours. Further, accumulation of more undetected infections in the unvaccinated arm may lead to VE waning to appear greater, due to the depletion of susceptibles bias [120]. This bias may occur in randomised or observational data. Asymptomatic swabbing may negate this bias by increasing the proportion of infections which are detected. Rates of asymptomatic swabbing were high in our study [121], with around 75% of participants returning an asymptomatic test on a given week. Rates of asymptomatic swabbing were likely much lower in the other studies, where participants were not encouraged to swab weekly.

Our analysis has some advantages over these methods. The data comes from a randomised single-blinded controlled trial, avoiding the potential biases of an observational cohort study and a test-negative case-control study. We model antibody waning and vaccine efficacy waning jointly over time, which may improve the efficiency of our estimates [9]. Further, the length of follow-up in our study after second dose was relatively long at over 29 weeks. However, the sample size of our data is much smaller than most of these studies, hence we report wider CrIs.

Table 4.3 and Supplementary Table C.9 compare our methods and results to previous studies of vaccine efficacy waning against different variants. Results from previous studies varied by variant, with lower efficacy generally predicted against Delta than Alpha, and lower still against Omicron (Table 4.3, Supplementary Table C.9). Our study reports efficacy over time since second dose against the Alpha variant. Note the predicted vaccine efficacy for a reference individual reported in Table 4.3 and Figure 4.5 was slightly higher than the mean vaccine efficacy across all individuals (Fig. 4.3), primarily because

the mean vaccine efficacy averaged over all individuals including those who had a shorter dose interval than 12 weeks. We predicted a slightly higher vaccine efficacy against symptomatic infection with the Alpha variant after two doses of ChAdOx1 nCoV-19 (Table 4.3, Fig. 4.5) than Andrews et al. (2022a) in all age groups, although confidence intervals overlapped [93]. Other studies were conducted at least partly in periods where the Delta or Omicron variants were dominant, and generally reported lower vaccine efficacy. Reported efficacy varied noticeably between studies (Table 4.3, Supplementary Table C.9). Differences in results between studies may be due to different proportions of circulating variants, rates of asymptomatic swabbing, and characteristics of the study populations, as well as the depletion of susceptibles bias, and previously mentioned biases due to unmeasured confounding. Our results against the Alpha variant, calculated on data from a randomised single-blinded study, are free from bias due to unmeasured confounding.

Our data was collected in a period when the Alpha variant was dominant, and considered individuals without prior infection who received two doses of vaccine. In contrast, the Omicron variant is currently dominant, and many individuals have received three or more doses of vaccine and been previously infected. Therefore our results may not be directly applicable to current and future situation. To aid with this, we performed a supplementary analysis extrapolating our results to examine what the efficacy over time since second dose might have been, in the hypothetical scenario where the study was conducted in the Omicron BA.4/5 period. The analysis suggests a weak initial protection against Omicron BA.4/5, waning to negligible protection after 6 months. See the Supplementary Information (Sections C.1.4, C.2.1, C.3, Supplementary Fig. C.11) for further details.

We further reported covariate effects on vaccine efficacy, although this analysis lacked power. We observed lower vaccine efficacy for individuals with shorter intervals between the second dose; results from previous studies have provided varying degrees of evidence for this [122, 94, 93]. We reported lower VE in older individuals after around 3 months since the second dose (Fig. 4.5, Supplementary Fig. C.10), with overlapping credible intervals. Faster waning [93] and lower VE [119, 96] have been previously observed in older individuals. However, these effects were not observed in all groups in the Andrews et al. study, with non-significant effects in the opposite direction observed in some groups [93], and the Nordström et al. age estimates were unadjusted and likely confounded due to

different vaccines being given to different age groups [119]. Another observational study observed higher efficacy in individuals aged 80+ years than 50-79, while suggesting this could be confounded by clinical risk [122]. Hence further research on the effect of age on vaccine efficacy is necessary. We did not observe a difference in vaccine efficacy for individuals with comorbidites, although some prior studies suggest a lower VE in clinically vulnerable individuals [96, 119, 93].

Table 4.2: **Comparison of methods and results with previous studies of antibody level waning**

Paper	Vaccine	Randomised vaccine /control assignment	Account for informative censoring at infection	Random individual slopes	Antibody type	Time of reported peak antibodies	Two-dose /booster /prior infection	Peak anti-spike IgG antibody level (BAU/mL)	Antibody half-life (days)	Covariates related to increased peak antibody levels	Covariates related to slower antibody decay
Wei et al. (2022) [81]	ChAdOx1 nCoV-19 BNT162b2*	x	x	✓	Anti-spike IgG	21 days post second dose	Two-dose	184 (183, 185)	Mean 79 (78, 80)	Older age Female Non-white No LTHC Prior infection Healthcare worker Longer dose interval More deprivation	Younger age White No LTHC
Aldridge et al. (2022) [123]	ChAdOx1 nCoV-19 BNT162b2*	x	x	x	Anti-spike IgG	28 days post second dose	Two-dose	-	105.9	None	None
Wei et al. (2023) [82]	ChAdOx1 nCoV-19 BNT162b2	x	x	✓	Anti-spike IgG	21 days post second dose	Booster: BNT162b2 mRNA-1273 Infection	-	78 (72, 86) 63 (61, 66) 96 (80, 119)	Younger age Longer dose interval	Not reported
Our analysis	ChAdOx1 nCoV-19	✓	✓	✓	Anti-spike IgG	28 days post second dose	Two-dose	217 (208, 227)	Median 74 (69, 79)	Younger age (18-69) Non-white Longer dose interval Not HCW	Male BMI <30 kg/m^2

We include studies of two doses of the ChAdOx1 nCoV-19 vaccine, and studies after a subsequent booster vaccine, focussing on anti-spike IgG antibody waning. Results are given with 95% confidence intervals or 95% credible intervals, depending on the paper. We report covariate effects where the null effect is not included in the 95% interval. HCW = healthcare worker, LTHC = long-term health condition. *We do not report these results here, see the original paper.

Table 4.3: **Comparison of methods and results with previous vaccine efficacy waning studies, against the Alpha and Delta variants**

Paper	Primary course vaccine	Dominant variant	Two-dose /booster /prior infection	Study design	Randomised vaccine /control assignment	Baseline comparator	Outcome	Location of result in paper	Group	Weeks since dose	Vaccine efficacy (%)
Our analysis†	ChAdOx1 nCoV-19	Alpha	Two dose	Randomised Controlled Trial	✓	Unvaccinated	Symptomatic	Fig. 4.5	Age 18-55	5	93.1 (81.6, 97.6)
										10	89.0 (76.6, 95.1)
										20	77.7 (64.5, 85.7)
										26	70.0 (55.0, 78.5)
									Age 56-69	5	91.8 (80.0, 97.0)
										10	88.1 (76.1, 94.7)
										20	78.5 (66.3, 88.0)
										26	72.2 (57.2, 83.1)
									Age 70+	5	90.7 (78.8, 96.4)
										10	84.7 (73.0, 92.4)
										20	69.9 (54.5, 80.4)
										26	61.3 (38.1, 73.6)
							Any infection	Supp. Fig. C.10	Age 18-55	5	84.3 (70.6, 91.8)
										10	77.2 (62.7, 86.1)
										20	60.5 (46.1, 70.1)
										26	50.6 (34.8, 60.1)
									Age 56-69	5	81.8 (68.3, 90.4)
										10	75.6 (62.4, 85.3)
										20	61.6 (48.0, 73.7)
										26	53.4 (37.0, 66.5)
									Age 70+	5	79.9 (66.4, 89.0)
										10	70.4 (57.6, 80.9)
										20	50.4 (34.2, 62.8)
										26	40.1 (18.4, 53.8)
Andrews et al. (2022a) [93]	ChAdOx1 nCoV-19 BNT162b2*	Alpha Delta*	Two dose	Test-negative	x	Unvaccinated	Symptomatic	Table S11	Age 16+	2-9	82.4 (79.6, 84.7)
										10+	76.2 (49.8, 88.7)
									Age 40-64	2-9	81.7 (76.0, 86.1)
									Age 65+	2-9	88.2 (82.2, 92.1)
Nordström et al. (2022) [119]	ChAdOx1 nCoV-19 BNT162b2* mRNA-1273*	Alpha and Delta (Combined)	Two dose	Cohort	x	Unvaccinated	Any inection	Table 2	Alpha, Delta and all ages combined	2-4	68 (58, 70)
										4-9	49 (28, 64)
										9-17	41 (29, 51)
										≥ 17	−19 (−98, 28)
Skowronski et al. (2022) [94]	ChAdOx1 nCoV-19 BNT162b2* mRNA-1273* Mixed schedules*	Alpha Delta	Two dose	Test-negative	x	Unvaccinated	Any inection Hospitalisation*	Supp. Table 12	Quebec (Alpha and Delta)	4-7	90 (84, 94)
										8-11	85 (81, 88)
										20-23	70 (64, 75)
										24-27	56 (43, 65)
									British Columbia (Delta)	4-7	77 (71, 82)
										8-11	76 (73, 79)
										20-23	74 (70, 78)
										24-27	67 (48, 80)

We include studies of two doses of the ChAdOx1 nCoV-19 vaccine, prior to the Omicron variant. Results are given with 95% confidence intervals or 95% credible intervals, depending on the paper. Results are presented for different age groups, where appropriate. *We do not report these results here, see the original paper. †We report here the predicted vaccine efficacy for a reference individual in the given age group, with covariates: female, white, no comorbidities, BMI<30 kg/m^2, ≥12 week interval, not a healthcare worker, standard dose.

Our study has several limitations. Our results may not be directly applicable to current and future populations, for four reasons. Firstly dominant variants have changed from B.1.177 and Alpha (B.1.1.7) variants when our data was collected, to the Omicron variant today. Secondly individuals in the study were SARS-CoV-2 naïve, results will likely differ in individuals with prior infection [82]. Thirdly individuals in the study received two doses of ChAdOx1 nCoV-19, whereas three or more doses of different vaccines have been given in many populations. Fourthly the study population differs from the general population, being predominantly white, 18-55 years old and including mostly healthcare workers. For these reasons our results are unlikely to transfer directly to the general population today, although qualitative aspects of our results, and our methods, remain relevant. We further only consider the effect of anti-spike IgG binding antibodies, and do not measure neutralising antibodies or cellular immune responses. We assumed that the antibody level decays exponentially, however model fitting plots indicate a model such as biphasic exponential decay, with faster initial decay, and slower decay thereafter, may fit better. Despite this, the exponential decay model still appears to fit the antibody levels relatively well. It may be that the protection conferred by other immune responses depends on the antibody level, or changes with time, whereas we assumed a constant effect. We did not include any checks of model fit for the time-to-infection component, or the association between the antibodies and infection. We did not consider the effect on risk of hospitalisation, severe disease and death due to insufficient data. Fewer antibody observations were available at the later timepoints (PB90 and PB182 study visits), causing greater uncertainty in our estimates at later times. We assumed a 7 day gap from exposure to reported infection; deviation from this assumption may introduce bias to our analysis. We did not account for the change in dominant variant from B.1.177 to B.1.1.7 (Alpha) during the study, however we expect the impact of this on VE to be negligible [107]. The dose of virus that individuals were exposed to may have varied throughout the study, as social distancing policies changed. Further, viral load is likely to have varied between individuals depending on previous exposure to COVID-19, and the variant to which they were exposed. We have not accounted for differences in viral load or dose of virus in this analysis.

In summary, we report correlates of protection against any COVID-19 infection and

symptomatic infection from a randomised COVID-19 vaccine study, using binding antibody responses at the time of exposure. We further report vaccine efficacy waning over time, antibody waning over time, and associated covariate effects. We use a joint model which accounts for antibody measurement error, the waning of antibody levels over time, and censoring of antibody observations at COVID-19 infection. This may reduce bias and increase power in the analysis compared with previous approaches.

Data availability

Anonymized data will be made available via the Vivli platform https://vivli.org/. Source data for the figures are provided with this paper where possible.

Code availability

The code used for the analysis is available at GitHub repository https://github.com/danphillipsstats/COVID-joint-model-CoP.

Acknowledgements

We thank Geoff Nicholls for a helpful discussion suggesting the use of Bayesian cut posteriors, which led us to developing the approximate Bayesian pooling method for multiple imputation.

Author contributions

All authors contributed to the conceptualisation of the study. SF, AJP and MV contributed to data collection. DJP, MDC and DS developed the statistical methods. DJP conducted the statistical analysis. The project was jointly supervised by MDC and DS. The first draft of the report was written by DJP and reviewed by MDC and DS. All authors critically reviewed and approved the final version.

Funding

The collection of the data used in this work was supported by the UK Research and Innovation (MC_PC_19055), Engineering and Physical Sciences Research Council (EP/R013756/1), National Institute for Health Research (COV19 OxfordVacc-01), Coalition for Epidemic Preparedness Innovations (Outbreak Response To Novel Coronavirus (COVID-19)), National Institute for Health Research Oxford Biomedical Research Centre (BRC4 Vaccines Theme), Chinese Academy of Medical Sciences Innovation Fund for Medical Science, China (2018-I2M-2- 002), Thames Valley and South Midlands NIHR Clinical Research Network. DJP discloses support for the research of this work from the Engineering and Physical Sciences Research Council (EP/W523781/1). The views expressed in this publication are those of the authors and not necessarily those of the NIHR or the UK Department of Health and Social Care. AstraZeneca reviewed the final manuscript but the academic authors retained editorial control. Other funders had no role in study design, data collection and analysis, or decision to publish.

Competing interests

SF, AJP and MV are contributors to intellectual property licensed by Oxford University Innovation to AstraZeneca. AJP is Chair of the UK Department of Health and Social Care's (DHSC) Joint Committee on Vaccination & Immunisation (JCVI), but does not participate in discussions on COVID-19 vaccines. The authors report no other conflicts of interest.

Ethics approval

The COV002 study was approved in the United Kingdom by the Medicines and Healthcare products Regulatory Agency (MHRA), reference 21584/0428/001 0001, and the South-Central Berkshire Research Ethics Committee, reference 20/SC/0179. All participants provided informed consent.

Materials & Correspondence

Correspondence to Daniel J. Phillips, email: daniel.phillips@paediatrics.ox.ac.uk.

Chapter 5

Discussion

This thesis primarily focussed on three main areas: multiple imputation, joint models of longitudinal and survival data, and correlates of protection. We examine the relevant conclusions and possible future work which have arisen from this thesis in each of these three topics below.

5.1 Multiple imputation

The derivation of Rubin's rules relies on the assumption that the distribution of the post-imputation maximum likelihood estimators (MLEs) is Gaussian over the multiple imputations given the observed data, and the asymptotic variance estimates have negligible variability compared to the variability of the MLEs. We explored mathematically the error in the posterior moments given by Rubin's rules when these assumptions do not hold. We further approximated the error in the confidence intervals using a Cornish-Fisher expansion.

Most notably, when the observed-data posterior distribution is skewed, Rubin's rules lead to biased confidence intervals. In some real world and simulated examples, posterior skewness is present for analyses with small sample sizes. The confidence intervals from Rubin's t approximation for small samples [54] were also very conservative, due to a small estimate for degrees of freedom. Hence in these simulated examples (e.g. small-sample logistic regression) Rubin's rules did not show undercoverage, but did lose power. In the real-world examples this led to Rubin's t giving wider confidence intervals

on at least one side of the interval. In these scenarios, Rubin's rules appear to be giving overly conservative inference.

It is worth noting that our simulation study only considered a Gaussian missing covariate, except for a skewed lognormal covariate for linear regression. We did not test how multiple imputation pooling methods compare for logistic or Cox regression with a skewed covariate. While Rubin's rules appeared to give robust inference in our simulation study, this may not be the case in all other scenarios where the assumptions for Rubin's rules do not hold. Brand et al. [59] compared Rubin's rules with various methods combining multiple imputation and bootstrapping on skewed simulated data, to estimate the mean difference between two groups. They found Rubin's rules appeared to be robust against skewness across simulation scenarios, including for small sample sizes. Rubin's rules was outperformed by making a single imputation (with appropriate uncertainty) within each bootstrap sample. Using imputation within a bootstrap is another viable alternative under posterior skewness.

In both the COVID-19 vaccine and SUPPORT trial examples, where posterior skewness was present, Rubin's confidence interval did not fully cover the approximate Bayesian confidence interval on one side of the interval. It is entirely possible that in other situations, the skewness may be more pronounced than that observed in our simulation study, and Rubin's rules may give biased intervals that produce undercoverage.

In all simulation scenarios, we observed that the t based confidence intervals for both Rubin's rules and approximate Bayesian pooling gave valid coverage. For logistic and Cox regression, the Gaussian confidence intervals also displayed valid coverage, and tended to exhibit higher power. In linear regression a variance (or scale) parameter is estimated, which leads to a t interval being required, whereas in Cox and logistic regression this is not present, and inference is based on a Gaussian assumption. We expect this pattern to continue to other models – where a scale parameter is present, we anticipate Gaussian intervals may produce undercoverage in small samples, whereas t intervals should give valid results. Where no scale parameter is present, Gaussian intervals may give higher power.

In some scenarios, using a transformation to estimate confidence intervals on a canonical scale may produce approximately Gaussian posterior distributions, allowing Rubin's

rules to be applied. Generally, this will coincide with the scale on which the parameter can take any value on the real line. For example, predicted hazard ratios in a Cox model should be evaluated on a log scale. In some cases, there is no simple transformation which leads to approximately Gaussian posteriors, or at least, such a transformation would have to be searched for and applied after the analysis has been performed and posterior skewness observed. Where a transformation is known prior to the analysis, we join others before us in recommending it should be applied [57, 56]. Where no transformation is known prior to analysis, choosing the transformation post analysis would be reusing the same data twice, and could lead to reduced coverage. We recommend that this should be tested using simulations, prior to considering using a post-hoc transformation to perform an analysis. Where no transformation exists and the posterior distribution is skewed, Rubin's rules may give biased results. Approximate Bayesian pooling is a viable alternative.

The main benefit of Rubin's rules over approximate Bayesian pooling or Bayesian inference after multiple imputation, is the need for far fewer imputations. We suggest around 10 times as many imputations may be required for the Bayesian methods compared to Rubin's rules. Where computational constraints are relevant this may be a compelling reason to use Rubin's rules over the sampling methods in some cases. With modern computing power, the need to produce more imputations may make no difference in other cases.

Approximate Bayesian pooling relies on fewer assumptions than Rubin's rules, and provides a simple and valid alternative where Rubin's rules may give overly conservative or biased inferences.

5.2 Joint modelling of longitudinal and time-to-event data

We have developed and fit a multiple imputation joint model for longitudinal and time-to-event data. The model we apply in Chapter 4 is specifically adapted to data from a COVID-19 vaccine trial. Our multiple imputation method for fitting a joint model is more general, and may be applied in other settings. Moreno-Bentancur et al. (2018) [46] previously developed a multiple imputation joint model. Their approach was

solely focussed on estimating the time-to-event parameters, so they imputed the missing longitudinal values at each visit time, instead of modelling the values over continuous time. As both the longitudinal trajectory and time-to-event parameters were of interest in our study, we also modelled the longitudinal trajectory over time. This means that our approach accounts for the informative censoring at event times differently to theirs. Our multiple imputation joint model is an alternative two-stage approach to fitting a joint model. Like earlier approaches, we believe it is likely to improve computation time, while possibly losing accuracy, since it approximates a fully joint model. It would be interesting to compare the performance of our method with previous two-stage approaches [46, 45, 44], in terms of accuracy and computation time.

We used a simple approach to account for the time-to-event data in our longitudinal antibody model, by including the Nelson-Aalen estimate of cumulative hazard and the event indicator as covariates in both the intercept and slope model. This followed the general recommendation of White and Royston [64] for Gaussian covariates in a Cox model. However, following similar mathematical arguments to White and Royston suggests that the Nelson-Aalen estimate and event indicator should be multiplied by the event time when included as covariates affecting the individual slope. Separately, using a negative exponential function in the slope, and exponentiating the log antibody levels in the time-to-event model, are both likely to worsen the accuracy of the linear approximation that these approaches rely on. We intend to further explore the performance of alternative adjustments in simulation studies.

Researchers have tended to prefer using the longitudinal trajectory as a predictor in the time-to-event model, without transformation. This generally leads to faster and more stable computations. In some cases however, this may not be the most appropriate choice for the model (e.g. not match the biological mechanism). For example, where the longitudinal variable is a log-transformed immune marker, this modelling approach means that as the immune marker levels tend towards zero, the hazard of infection tends towards infinity. However, using log-transformed immune markers may give a better fit for higher immune marker levels, where the bulk of the observed values tend to lie. It is possible that the use of either a link function (e.g. exponentiating the log immune marker levels in the infection model) as we have used, or another method such as splines

or a changepoint may be more appropriate in some circumstances.

Fitting the longitudinal and time-to-event models concurrently as a joint model, allows the occurrence and timing of events to inform the estimation of the longitudinal trajectories. In many cases, this may be a desired property, for example as individuals who were infected early in the COV002 study may be more likely to have lower antibody levels than those who were uninfected through the whole study. In other cases, this feedback may not be desired. A researcher who is confident in their choice of longitudinal model and the reliability of their longitudinal data, may fear that feedback from the time-to-event model will contaminate their results and lead to bias in the estimated longitudinal trajectory. The cut or modular approach was developed in the Bayesian literature for such a problem, which prevents feedback from the second module to the first [71, 124, 125]. Carmona and Nicholls (2020) extended this approach to allow partial feedback from the second module to the first, where the amount of feedback is a parameter that can be tuned to the data [126]. These approaches are similar to our multiple imputation joint model. If the Nelson-Aalen estimator and event indicators are excluded as predictors in the antibody model, the resulting method is exactly the cut/modular approach. An interesting direction of further research would be to apply modular and semi-modular inference to joint models, where the time-to-event model is misspecified. These approaches, and possibly our multiple imputation approach, may outperform the standard joint model, as the misspecification in the time-to-event model will then be (partially) prevented from flowing backwards and introducing bias in the longitudinal predictions. Model misspecification has rarely been considered in the joint modelling literature, though two papers have considered misspecification in the longitudinal component [127], and in the baseline hazard, random effects distribution and longitudinal component [128].

5.3 Correlates of protection

In Chapter 3, we demonstrated via simulations that trials of the same vaccine may give inconsistent correlates of protection curves when standard approaches are used without accounting for antibody decay, and the baseline risk of infection differs between the

trials. Our simulation study generated data similar to the COV002 trial of the ChAdOx1 nCoV-19 vaccine. Other COVID-19 vaccines generated between around 8 fold higher antibody levels, and 2 fold lower than this trial [17]. It would be interesting to run similar simulations in scenarios with different initial antibody response, to explore how this affects the impact of case timings on the robustness of correlates of protection estimates. It would also be interesting to explore how the half-life of antibody decay affects the results. Vaccines against different diseases have reported vastly different rates of antibody decay [129], and even non-linear decay of log antibody levels [130, 131]. We also fitted a correlate of protection model containing a linear antibody effect; it would be worth repeating these simulations for a model containing an effect due to log-transformed antibody levels.

We developed a multiple imputation joint model for antibody decay and correlates of protection, and applied our model to a COVID-19 vaccine trial. One major limitation of our work, is the lack of a simulation study to test the performance of our model. We intend to perform simulations to explore how our multiple imputation joint model performs at estimating the correlates of protection curve in terms of coverage and power. We would also like to compare the model with other methods for estimating exposure-proximal correlates of protection, such as two-stage joint models, the model developed by Huang and Follmann (2025) [53] and a Bayesian joint model. Due to the larger computational burden of joint modelling approaches, this is unlikely to be possible in a wide range of simulation scenarios, unlike the time-fixed correlates methods. The sample size may also need to be reduced in order to test on a large number of simulated datasets; we may need to consider a smaller trial, with a high attack rate.

Many previous correlates of protection models focus on estimating a causal relationship between antibody levels at the first study visit and subsequent infection. However, our simulations in Chapter 3 demonstrate that where antibody decay is present, such estimates may differ between studies, even if the causal effect of antibody levels at exposure on protection is the same. It would be interesting to consider how joint modelling may be used to establish causal estimates for exposure-proximal correlates of protection. One paper has considered this previously, however it required strong assumptions and did not consider the presence of measurement error [132]. It is likely

that any such approach may require an assumption of no unmeasured confounding between the longitudinal and time-to-event outcomes, which may not be realistic. Hence it would be particularly interesting to develop a sensitivity analysis for estimating what the causal effect would have been, in the presence of given level of unmeasured confounding between the longitudinal and time-to-event variable. For the inverse probability weighted models comparing peak immune response with risk of infection, a sensitivity analysis using E-values was previously developed [19].

Further research into exposure-proximal correlates of protection is required, in order to balance the computational burden of these models with the desire to accurately and reliably estimate the relationship between immune marker levels and risk of infection.

Bibliography

[1] Stefanski, L. A. & Carroll, R. J. Covariate Measurement Error in Logistic Regression. *The Annals of Statistics* **13**, 1335 – 1351 (1985).

[2] Hu, P., Tsiatis, A. A. & Davidian, M. Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics* 1407–1419 (1998).

[3] Ramasamy, M. N. *et al.* Safety and immunogenicity of ChAdOx1 nCoV-19 vaccine administered in a prime-boost regimen in young and old adults (COV002): a single-blind, randomised, controlled, phase 2/3 trial. *The Lancet* **396**, 1979–1993 (2020).

[4] Voysey, M. *et al.* Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet* **397**, 99–111 (2021).

[5] Feng, S. *et al.* Correlates of protection against symptomatic and asymptomatic SARS-CoV-2 infection. *Nature Medicine* **27**, 2032–2040 (2021).

[6] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual* (2022). URL https://mc-stan.org.

[7] Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys* (Wiley, 1987).

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

[9] Fleming, T. R., Prentice, R. L., Pepe, M. S. & Glidden, D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* **13**, 955–968 (1994).

[10] Qin, L., Gilbert, P. B., Corey, L., McElrath, M. J. & Self, S. G. A framework for assessing immunological correlates of protection in vaccine trials. *The Journal of Infectious Diseases* **196**, 1304–1312 (2007).

[11] Plotkin, S. A. & Plotkin, S. A. Correlates of vaccine-induced immunity. *Clinical Infectious Diseases* **47**, 401–409 (2008).

[12] Plotkin, S. A. Correlates of protection induced by vaccination. *Clinical and Vaccine Immunology* **17**, 1055–1065 (2010).

[13] Gilbert, P. B. *et al.* Four statistical frameworks for assessing an immune correlate of protection (surrogate endpoint) from a randomized, controlled, vaccine efficacy trial. *Vaccine* **42**, 2181–2190 (2024).

[14] Dunning, A. J. A model for immunological correlates of protection. *Statistics in Medicine* **25**, 1485–1497 (2006).

[15] Dunning, A. J., Kensler, J., Coudeville, L. & Bailleux, F. Some extensions in continuous models for immunological correlates of protection. *BMC Medical Research Methodology* **15**, 107 (2015).

[16] Coudeville, L. *et al.* Relationship between haemagglutination-inhibiting antibody titres and clinical protection against influenza: development and application of a bayesian random-effects model. *BMC Medical Research Methodology* **10**, 18 (2010).

[17] Khoury, D. S. *et al.* Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection. *Nature Medicine* **27**, 1205–1211 (2021).

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

[19] Gilbert, P. B., Fong, Y., Kenny, A. & Carone, M. A controlled effects approach to assessing immune correlates of protection. *Biostatistics* **24**, 850–865 (2023).

[20] Cowling, B. J. *et al.* Influenza hemagglutination-inhibition antibody titer as a

mediator of vaccine-induced protection for influenza B. *Clinical Infectious Diseases* **68**, 1713–1717 (2019).

[21] Hejazi, N. S., van der Laan, M. J., Janes, H. E., Gilbert, P. B. & Benkeser, D. C. Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials. *Biometrics* **77**, 1241–1253 (2021).

[22] Tsiatis, A. A. & Davidian, M. Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**, 809–834 (2004).

[23] Brown, E. R. & Ibrahim, J. G. Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59**, 686–693 (2003).

[24] Ibrahim, J. G., Chen, M.-H. & Sinha, D. Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica* 863–883 (2004).

[25] Li, D., Keogh, R., Clancy, J. P. & Szczesniak, R. D. Flexible semiparametric joint modeling: an application to estimate individual lung function decline and risk of pulmonary exacerbations in cystic fibrosis. *Emerging Themes in Epidemiology* **14**, 1–13 (2017).

[26] Kang, K., Pan, D. & Song, X. A joint model for multivariate longitudinal and survival data to discover the conversion to Alzheimer's disease. *Statistics in Medicine* **41**, 356–373 (2022).

[27] Xu, J. & Zeger, S. L. Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society. Series C: Applied Statistics* **50**, 375–387 (2001).

[28] Gueorguieva, R., Rosenheck, R. & Lin, H. Joint modelling of longitudinal outcome and interval-censored competing risk dropout in a schizophrenia clinical trial. *Journal of the Royal Statistical Society Series A: Statistics in Society* **175**, 417–433 (2012).

[29] Ibrahim, J. G., Chu, H. & Chen, L. M. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology* **28**, 2796 (2010).

[30] Gould, L. A. *et al.* Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine* **34**, 2181–2195 (2015).

[31] Papageorgiou, G., Mauff, K., Tomer, A. & Rizopoulos, D. An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual Review of Statistics and Its Application* **6**, 223–240 (2019).

[32] Henderson, R., Diggle, P. & Dobson, A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480 (2000).

[33] Elashoff, R. M., Li, G. & Li, N. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* **64**, 762–771 (2008).

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

[35] Chen, Q., May, R. C., Ibrahim, J. G., Chu, H. & Cole, S. R. Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Statistics in Medicine* **33**, 4560–4576 (2014).

[36] Li, N., Elashoff, R. M. & Li, G. Robust joint modeling of longitudinal measurements and competing risks failure time data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **51**, 19–30 (2009).

[37] Andrinopoulou, E.-R., Rizopoulos, D., Takkenberg, J. J. & Lesaffre, E. Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine* **33**, 3167–3178 (2014).

[38] Tseng, Y.-K., Hsieh, F. & Wang, J.-L. Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92**, 587–603 (2005).

[39] Rizopoulos, D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67**, 819–829 (2011).

[40] Rizopoulos, D., Miranda Afonso, P. & Papageorgiou, G. *JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data* (2025). URL https://CRAN.R-project.org/package=JMbayes2. R package version 0.5-2.

[41] Rustand, D., Van Niekerk, J., Krainski, E. T., Rue, H. & Proust-Lima, C. Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested Laplace approximations. *Biostatistics* **25**, 429–448 (2024).

[42] Rustand, D., van Niekerk, J., Krainski, E. T. & Rue, H. Joint modeling of multivariate longitudinal and survival outcomes with the R package INLAjoint. *arXiv preprint arXiv:2402.08335* (2024).

[43] Tu, J. & Sun, J. Gaussian variational approximate inference for joint models of longitudinal biomarkers and a survival outcome. *Statistics in Medicine* **42**, 316–330 (2023).

[44] Alvares, D. & Leiva-Yamaguchi, V. A two-stage approach for Bayesian joint models: reducing complexity while maintaining accuracy. *Statistics and Computing* **33**, 115 (2023).

[45] Mauff, K., Steyerberg, E., Kardys, I., Boersma, E. & Rizopoulos, D. Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach. *Statistics and Computing* **30**, 999–1014 (2020).

[46] Moreno-Betancur, M. *et al.* Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). *Biostatistics* **19**, 479–496 (2018).

[47] Follmann, D. *et al.* Examining protective effects of SARS-CoV-2 neutralizing antibodies after vaccination or monoclonal antibody administration. *Nature Communications* **14**, 3605 (2023).

[48] White, M. T. *et al.* A combined analysis of immunogenicity, antibody kinetics and

vaccine efficacy from phase 2 trials of the RTS, S malaria vaccine. *BMC Medicine* **12**, 117 (2014).

[49] White, M. T. *et al.* Immunogenicity of the RTS,S/AS01 malaria vaccine and implications for duration of vaccine efficacy: secondary analysis of data from a phase 3 randomised controlled trial. *The Lancet Infectious Diseases* **15**, 1450–1458 (2015).

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

[51] Seekircher, L. *et al.* Anti-Spike IgG antibodies as correlates of protection against SARS-CoV-2 infection in the pre-Omicron and Omicron era. *Journal of Medical Virology* **96**, e29839 (2024).

[52] Papadopoulos, I. *et al.* Personalized prediction of SARS-CoV-2 vaccine-induced immunity after boost: a longitudinal analysis using joint modeling. *Frontiers in Immunology* **16**, 1619631 (2025).

[53] Huang, Y. & Follmann, D. Exposure proximal immune correlates analysis. *Biostatistics* **26**, kxae031 (2025).

[54] Barnard, J. & Rubin, D. B. Small-sample degrees of freedom with multiple imputation. *Biometrika* **86**, 948–955 (1999).

[55] Van Buuren, S. *Flexible Imputation of Missing Data*, vol. 10 (CRC press Boca Raton, FL, 2012).

[56] Marshall, A., Altman, D. G., Holder, R. L. & Royston, P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Medical Research Methodology* **9**, 57 (2009).

[57] White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* **30**, 377–399 (2011).

[58] Zhou, X. & Reiter, J. P. A note on Bayesian inference after multiple imputation. *The American Statistician* **64**, 159–163 (2010).

[59] Brand, J., van Buuren, S., le Cessie, S. & van den Hout, W. Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. *Statistics in Medicine* **38**, 210–220 (2019).

[60] Steele, R. J., Wang, N. & Raftery, A. E. Inference from multiple imputation for missing data using mixtures of normals. *Statistical Methodology* **7**, 351–365 (2010).

[61] Rashid, S., Mitra, R. & Steele, R. J. Using mixtures of t densities to make inferences in the presence of missing data with a small number of multiply imputed data sets. *Computational Statistics & Data Analysis* **92**, 84–96 (2015).

[62] van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45**, 1–67 (2011).

[63] Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* (Chapman and Hall/CRC, 1995).

[64] White, I. R. & Royston, P. Imputing missing covariate values for the Cox model. *Statistics in Medicine* **28**, 1982–1998 (2009).

[65] Sinha, D., Ibrahim, J. G. & Chen, M.-H. A Bayesian justification of Cox's partial likelihood. *Biometrika* **90**, 629–641 (2003).

[66] Cornish, E. A. & Fisher, R. A. Moments and cumulants in the specification of distributions. *Revue de l'Institut international de Statistique* 307–320 (1938).

[67] Abramowitz, M. & Stegun, I. A. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, vol. 55 (Courier Corporation, 1965).

[68] Kolassa, J. E. & Kuffner, T. A. On the validity of the formal Edgeworth expansion for posterior densities. *The Annals of Statistics* **48**, 1940 – 1958 (2020).

[69] Stan Development Team. RStan: the R interface to Stan (2024). URL https://mc-stan.org/. R package version 2.32.6.

[70] Maucort-Boulch, D., Franceschi, S., Plummer, M. & the IARC HPV Prevalence Surveys Study Group. International Correlation between Human Papillomavirus Prevalence and Cervical Cancer Incidence. *Cancer Epidemiology Biomarkers & Prevention* **17**, 717–720 (2008).

[71] Plummer, M. Cuts in Bayesian graphical models. *Statistics and Computing* **25**, 37–43 (2015).

[72] Connors, A. F. *et al.* A controlled trial to improve care for seriously ill hospitalized patients: The study to understand prognoses and preferences for outcomes and risks of treatments (SUPPORT). *JAMA* **274**, 1591–1598 (1995).

[73] Knaus, W. A. *et al.* The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine* **122**, 191–203 (1995).

[74] Harrell Jr, F. E. *Hmisc: Harrell Miscellaneous* (2025). URL https://hbiostat.org/R/Hmisc/. R package version 5.2-0.

[75] Jr, F. E. H. Support datasets description (supportdesc). https://hbiostat.org/data/repo/supportdesc (2025). Accessed: 2025-09-26.

[76] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2023). URL https://www.R-project.org/.

[77] Little, R. J. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics* **6**, 287–296 (1988).

[78] Meng, X.-L. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 538–558 (1994).

[79] Graham, J. W., Olchowski, A. E. & Gilreath, T. D. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* **8**, 206–213 (2007).

[80] Bodner, T. E. What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal* **15**, 651–675 (2008).

[81] Wei, J. *et al.* Antibody responses and correlates of protection in the general population after two doses of the ChAdOx1 or BNT162b2 vaccines. *Nature Medicine* **28**, 1072–1082 (2022).

[82] Wei, J. *et al.* Protection against SARS-CoV-2 Omicron BA.4/5 variant following booster vaccination or breakthrough infection in the UK. *Nature Communications* **14**, 2799 (2023).

[83] Gilbert, P. B. *et al.* Immune correlates analysis of the mRNA-1273 COVID-19 vaccine efficacy clinical trial. *Science* **375**, 43–50 (2022).

[84] Fong, Y. *et al.* Immune correlates analysis of the ENSEMBLE single Ad26. COV2. S dose vaccine efficacy clinical trial. *Nature Microbiology* **7**, 1996–2010 (2022).

[85] Fong, Y. *et al.* Immune correlates analysis of the PREVENT-19 COVID-19 vaccine efficacy clinical trial. *Nature Communications* **14**, 331 (2023).

[86] Benkeser, D. *et al.* Immune correlates analysis of a phase 3 trial of the AZD1222 (ChAdOx1 nCoV-19) vaccine. *npj Vaccines* **8**, 36 (2023).

[87] Zhang, B. *et al.* Omicron covid-19 immune correlates analysis of a third dose of mrna-1273 in the cove trial. *Nature Communications* **15**, 7954 (2024).

[88] Brilleman, S., Crowther, M., Moreno-Betancur, M., Buros Novik, J. & Wolfe, R. Joint longitudinal and time-to-event models via Stan. *Proceedings of StanCon 2018* 1–18 (2018).

[89] Phillips, D. J. *et al.* Improved estimates of COVID-19 correlates of protection, antibody decay and vaccine efficacy waning: a joint modelling approach. *medRxiv preprint* (2024). URL https://doi.org/10.1101/2024.07.02.24309776.

[90] Laurie, D. Calculation of Gauss-Kronrod quadrature rules. *Mathematics of Computation* **66**, 1133–1145 (1997).

[91] UK Coronavirus dashboard. https://coronavirus.data.gov.uk/. Accessed: 2022-02-16.

[92] Hertz, T. *et al.* Correlates of protection for booster doses of the SARS-CoV-2 vaccine BNT162b2. *Nature Communications* **14**, 4575 (2023).

[93] Andrews, N. *et al.* Duration of protection against mild and severe disease by Covid-19 vaccines. *New England Journal of Medicine* **386**, 340–350 (2022).

[94] Skowronski, D. M. *et al.* Two-dose severe acute respiratory syndrome coronavirus 2 vaccine effectiveness with mixed schedules and extended dosing intervals: test-negative design studies from British Columbia and Quebec, Canada. *Clinical Infectious Diseases* **75**, 1980–1992 (2022).

[95] Katikireddi, S. V. *et al.* Two-dose ChAdOx1 nCoV-19 vaccine protection against COVID-19 hospital admissions and deaths over time: a retrospective, population-based cohort study in Scotland and Brazil. *The Lancet* **399**, 25–35 (2022).

[96] Menni, C. *et al.* COVID-19 vaccine waning and effectiveness and side-effects of boosters: a prospective community study from the ZOE COVID Study. *The Lancet Infectious Diseases* **22**, 1002–1010 (2022).

[97] Horne, E. M. *et al.* Waning effectiveness of BNT162b2 and ChAdOx1 covid-19 vaccines over six months since second dose: OpenSAFELY cohort study using linked electronic health records. *BMJ* **378** (2022).

[98] Andrews, N. *et al.* Covid-19 vaccine effectiveness against the Omicron (B. 1.1. 529) variant. *New England Journal of Medicine* **386**, 1532–1546 (2022).

[99] Andrews, N. *et al.* Effectiveness of COVID-19 booster vaccines against COVID-19-related symptoms, hospitalization and death in England. *Nature Medicine* **28**, 831–837 (2022).

[100] Sullivan, S. G., Tchetgen Tchetgen, E. J. & Cowling, B. J. Theoretical basis of the test-negative study design for assessment of influenza vaccine effectiveness. *American Journal of Epidemiology* **184**, 345–353 (2016).

[101] Westreich, D. & Hudgens, M. G. Invited commentary: beware the test-negative design. *American Journal of Epidemiology* **184**, 354–356 (2016).

[102] Tsiatis, A. A. & Davidian, M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 809–834 (2004).

[103] Néant, N. *et al.* Modeling SARS-CoV-2 viral kinetics and association with mortality in hospitalized patients from the French COVID cohort. *Proceedings of the National Academy of Sciences* **118**, e2017962118 (2021).

[104] Tong-Minh, K. *et al.* Joint modeling of repeated measurements of different biomarkers predicts mortality in COVID-19 patients in the Intensive Care Unit. *Biomarker Insights* **17**, 11772719221112370 (2022).

[105] Chen, X. *et al.* A predictive paradigm for COVID-19 prognosis based on the longitudinal measure of biomarkers. *Briefings in Bioinformatics* **22**, bbab206 (2021).

[106] Voysey, M. *et al.* Single-dose administration and the influence of the timing of the booster dose on immunogenicity and efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine: a pooled analysis of four randomised trials. *The Lancet* **397**, 881–891 (2021).

[107] Emary, K. R. *et al.* Efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 variant of concern 202012/01 (B. 1.1. 7): an exploratory analysis of a randomised controlled trial. *The Lancet* **397**, 1351–1362 (2021).

[108] Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.-C. Rank-normalization, folding, and localization: an improved R for assessing convergence of MCMC (with discussion). *Bayesian Analysis* **16**, 667–718 (2021).

[109] RStudio Team. *RStudio: Integrated Development Environment for R.* RStudio, PBC., Boston, MA (2020). URL http://www.rstudio.com/.

[110] Therneau, T. M. *A Package for Survival Analysis in R* (2024). URL https://CRAN.R-project.org/package=survival. R package version 3.6-4.

[111] Knaus, J. *snowfall: Easier Cluster Computing (Based on 'snow')* (2023). URL https://CRAN.R-project.org/package=snowfall. R package version 1.84-6.3.

[112] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016). URL https://ggplot2.tidyverse.org.

[113] Neuwirth, E. *RColorBrewer: ColorBrewer Palettes* (2022). URL https://CRAN.R-project.org/package=RColorBrewer. R package version 1.1-3.

[114] Wickham, H., Vaughan, D. & Girlich, M. *tidyr: Tidy Messy Data* (2024). URL https://CRAN.R-project.org/package=tidyr. R package version 1.3.1.

[115] Rizopoulos, D., Taylor, J. M. & Kardys, I. Goodness-of-fit checks for joint models. *arXiv preprint arXiv:2601.18598* (2026).

[116] Srivastava, K. *et al.* SARS-CoV-2-infection- and vaccine-induced antibody responses are long lasting with an initial waning phase followed by a stabilization phase. *Immunity* **57**, 587–599 (2024).

[117] Deichmann, J. *et al.* Predicting antibody kinetics and duration of protection against SARS-CoV-2 following vaccination from sparse serological data. *PLoS Computational Biology* **21**, e1013192 (2025).

[118] Flaxman, A. *et al.* Reactogenicity and immunogenicity after a late second dose or a third dose of ChAdOx1 nCoV-19 in the UK: a substudy of two randomised controlled trials (COV001 and COV002). *The Lancet* **398**, 981–990 (2021).

[119] Nordström, P., Ballin, M. & Nordström, A. Risk of infection, hospitalisation, and death up to 9 months after a second dose of COVID-19 vaccine: a retrospective, total population cohort study in Sweden. *The Lancet* **399**, 814–823 (2022).

[120] Kahn, R., Schrag, S. J., Verani, J. R. & Lipsitch, M. Identifying and alleviating bias due to differential depletion of susceptible people in postmarketing evaluations of COVID-19 vaccines. *American Journal of Epidemiology* **191**, 800–811 (2022).

[121] Williams, L. R. *et al.* Implementation and adherence to regular asymptomatic testing in a covid-19 vaccine trial. *Vaccine* **42**, 126167 (2024).

[122] Amirthalingam, G. *et al.* Serological responses and vaccine effectiveness for extended COVID-19 vaccine schedules in england. *Nature Communications* **12**, 7217 (2021).

[123] Aldridge, R. W. *et al.* SARS-CoV-2 antibodies and breakthrough infections in the Virus Watch cohort. *Nature Communications* **13**, 4869 (2022).

[124] Jacob, P. E., Murray, L. M., Holmes, C. C. & Robert, C. P. Better together? Statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719* (2017).

[125] Bayarri, M. J., Berger, J. O. & Liu, F. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis* **4**, 119 – 150 (2009).

[126] Carmona, C. & Nicholls, G. Semi-modular inference: enhanced learning in multi-modular models by tempering the influence of components. In Chiappa, S. & Calandra, R. (eds.) *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, vol. 108 of *Proceedings of Machine Learning Research*, 4226–4235 (PMLR, 2020).

[127] Crowther, M. J., Andersson, T. M.-L., Lambert, P. C., Abrams, K. R. & Humphreys, K. Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification. *Statistics in Medicine* **35**, 1193–1209 (2016).

[128] Arisido, M. W., Antolini, L., Bernasconi, D. P., Valsecchi, M. G. & Rebora, P. Joint model robustness compared with the time-varying covariate Cox model to evaluate the association between a longitudinal marker and a time-to-event endpoint. *BMC Medical Research Methodology* **19**, 1–13 (2019).

[129] Young, B. *et al.* Do antibody responses to the influenza vaccine persist year-round in the elderly? A systematic review and meta-analysis. *Vaccine* **35**, 212–221 (2017).

[130] Guevara, A. *et al.* Antibody persistence and evidence of immune memory at 5

years following administration of the 9-valent HPV vaccine. *Vaccine* **35**, 5050–5057 (2017).

[131] Le, T. *et al.* Immune responses and antibody decay after immunization of adolescents and adults with an acellular pertussis vaccine: the APERT Study. *Journal of Infectious Diseases* **190**, 535–544 (2004).

[132] Zheng, C. & Liu, L. Quantifying direct and indirect effect for longitudinal mediator and survival outcome using joint modeling approach. *Biometrics* **78**, 1233–1243 (2022).

[133] stan: Prior Choice Recommendations. https://github.com/stan-dev/stan/wiki/prior-choice-recommendations. Accessed: 2026-05-25.

[134] Ghosh, J., Li, Y. & Mitra, R. On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression. *Bayesian Analysis* **13**, 359 – 383 (2018). URL https://doi.org/10.1214/17-BA1051.

[135] Falsey, A. R. *et al.* Phase 3 safety and efficacy of AZD1222 (ChAdOx1 nCoV-19) Covid-19 vaccine. *New England Journal of Medicine* **385**, 2348–2360 (2021).

[136] Dunkle, L. M. *et al.* Efficacy and safety of NVX-CoV2373 in adults in the United States and Mexico. *New England Journal of Medicine* **386**, 531–543 (2022).

[137] Polack, F. P. *et al.* Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *New England Journal of Medicine* **383**, 2603–2615 (2020).

[138] Thomas, S. J. *et al.* Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine through 6 months. *New England Journal of Medicine* **385**, 1761–1773 (2021).

[139] Sadoff, J. *et al.* Safety and efficacy of single-dose Ad26.COV2.S vaccine against Covid-19. *New England Journal of Medicine* **384**, 2187–2201 (2021).

[140] Baden, L. R. *et al.* Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine* **384**, 403–416 (2021).

[141] Hoffman, M. D., Gelman, A. *et al.* The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623 (2014).

[142] Hogan, A. B. *et al.* Estimating long-term vaccine effectiveness against SARS-CoV-2 variants: a model-based approach. *Nature Communications* **14**, 4325 (2023).

[143] Fauvernier, M., Remontet, L., Uhry, Z., Bossard, N. & Roche, L. survPen: an R package for hazard and excess hazard modelling with multidimensional penalized splines. *Journal of Open Source Software* **4**, 1434 (2019).

Appendix A

The robustness of Rubin's rules and (approximate) Bayesian pooling methods after multiple imputation

A.1 Theoretical results

Lemma 1 (Posterior skewness of θ)**.**

$$\gamma_1\left(\theta \mid D_{\text{obs}}\right) = \gamma_1(\hat{\theta}^* \mid D_{\text{obs}})\lambda_\nu^{3/2} + 3\frac{\left(\frac{\nu}{\nu-2}\right)\text{Cov}\left(\hat{\theta}^*, U^* \mid D_{\text{obs}}\right)}{\left(B + \overline{U}_\infty(\frac{\nu}{\nu-2})\right)^{3/2}} \tag{A.1}$$

and as $\nu \to \infty$ (the Gaussian case),

$$\gamma_1\left(\theta \mid D_{\text{obs}}\right) = \gamma_1(\hat{\theta}^* \mid D_{\text{obs}})\lambda^{3/2} + 3\left(\frac{\text{Cov}\left(\hat{\theta}^*, U^* \mid D_{\text{obs}}\right)}{\left(B + \overline{U}_\infty\right)^{3/2}}\right) \tag{A.2}$$

Proof. In the following let all quantities be conditioned on D_{obs}, though we drop the conditioning from the notation for the sake of space.

$\mathrm{E}\left[(\hat{\theta}^*)^2\,\psi\sqrt{U^*}\right] = 0$ since ψ is independent of U^* and $\hat{\theta}^*$, and has mean 0.

$$\begin{aligned}\mathrm{E}\left[(\hat{\theta}^* - \overline{\theta}_\infty)\psi^2 U^*\right] &= \mathrm{E}\left[\psi^2\right]\mathrm{E}\left[U^*(\hat{\theta}^* - \overline{\theta}_\infty)\right] \\ &= \left(\frac{\nu}{\nu-2}\right)\left(\mathrm{E}\left[U^*\hat{\theta}^*\right] - \mathrm{E}\left[U^*\right]\mathrm{E}\left[\hat{\theta}^*\right]\right) \\ &= \left(\frac{\nu}{\nu-2}\right)\mathrm{Cov}(\hat{\theta}^*, U^*)\end{aligned}$$

Further $\mathrm{E}\left[(\psi\sqrt{U^*})^3\right] = 0$ since ψ is independent of U^* and has third moment equal to 0. So

$$\gamma_1(\theta) = \left(\frac{1}{B + \overline{U}_\infty(\frac{\nu}{\nu-2})}\right)^{3/2}\left(B^{3/2}\,\gamma_1(\hat{\theta}^*) + 3\left(\frac{\nu}{\nu-2}\right)\mathrm{Cov}\left(\hat{\theta}^*, U^*\right)\right)$$

□

Lemma 2 (Posterior excess kurtosis of θ)**.**

$$\begin{aligned}\gamma_2\left(\theta \mid D_{\text{obs}}\right) = \gamma_2(\hat{\theta}^* \mid D_{\text{obs}})\lambda_\nu^2 \;+\; &\frac{6}{\nu-4}(1-\lambda_\nu)^2 \\ &+ 3\left(\frac{\frac{\nu^2}{(\nu-2)(\nu-4)}\mathrm{Var}(U^* \mid D_{\text{obs}}) + 2\frac{\nu}{\nu-2}\mathrm{Cov}\left((\hat{\theta}^* - \bar{\theta}_\infty)^2, U^* \mid D_{\text{obs}}\right)}{\left(B + (\frac{\nu}{\nu-2})\overline{U}_\infty\right)^2}\right)\end{aligned} \tag{A.3}$$

and in the limit as $\nu \to \infty$, the Gaussian case,

$$\gamma_2\left(\theta \mid D_{\text{obs}}\right) = \gamma_2(\hat{\theta}^* \mid D_{\text{obs}})\lambda^2 \;+\; 3\left(\frac{\mathrm{Var}(U^* \mid D_{\text{obs}}) + 2\mathrm{Cov}\left((\hat{\theta}^* - \bar{\theta}_\infty)^2, U^* \mid D_{\text{obs}}\right)}{\left(B + \overline{U}_\infty\right)^2}\right) \tag{A.4}$$

Proof. In the following let all quantities be conditioned on D_{obs}, though we drop the conditioning from the notation for the sake of space.

Note ψ is independent of $\sqrt{U^*}$ and $\hat{\theta}^*$, and has odd moments equal to 0. Applying this, we see

$$\begin{aligned}\mathrm{E}\left[\left(\hat{\theta}^* - \bar{\theta}_\infty\right)^4\right] &= B^2\left(\gamma_2(\hat{\theta}^*) + 3\right) \\ \mathrm{E}\left[\left(\hat{\theta}^* - \bar{\theta}_\infty\right)^3\psi\sqrt{U^*}\right] &= 0\end{aligned}$$

as ψ is independent of $\sqrt{U^*}$ and $\hat{\theta}^*$ and has mean 0.

$$\begin{aligned}
\mathrm{E}\left[\left(\hat{\theta}^* \ - \bar{\theta}_\infty\right)^2 \left(\psi\sqrt{U^*}\right)^2\right] &= \mathrm{E}\left[(\hat{\theta}^* - \overline{\theta}_\infty)^2 \, \psi^2 \, U^*\right] \\
&= \mathrm{E}\left[\psi^2\right] \mathrm{E}\left[U^*(\hat{\theta}^* - \overline{\theta}_\infty)^2\right] \\
&= \frac{\nu}{\nu-2}\left(B\overline{U}_\infty + \mathrm{Cov}\left((\hat{\theta}^* - \bar{\theta}_\infty)^2, U^*\right)\right)
\end{aligned}$$

and

$$\mathrm{E}\left[\left(\hat{\theta}^* \ - \bar{\theta}_\infty\right)\left(\psi\sqrt{U^*}\right)^3\right] = 0$$

as ψ is independent of $\sqrt{U^*}$ and $\hat{\theta}^*$ and has third moment equal to 0.

$$\begin{aligned}
\mathrm{E}\left[\left(\psi\sqrt{U^*}\right)^4\right] &= \mathrm{E}\left[\psi^4\right] \mathrm{E}\left[U^{*2}\right] \\
&= \frac{3\nu^2}{(\nu-2)(\nu-4)}\left(\mathrm{Var}(U^*) + \overline{U}_\infty^2\right)
\end{aligned}$$

So

$$\begin{aligned}
\gamma_2(\theta) = &\left(\frac{1}{B + \frac{\nu}{\nu-2}\overline{U}_\infty}\right)^2 \left[B^2\left(\gamma_2(\hat{\theta}^*) + 3\right) + \right. \\
&\left. 6\left(\frac{\nu}{\nu-2}\right)\left(B\overline{U}_\infty + \mathrm{Cov}\left((\hat{\theta}^* - \bar{\theta}_\infty)^2, U^*\right)\right) + \frac{3\nu^2}{(\nu-2)(\nu-4)}\,\mathrm{E}[U^{*2}]\right] - 3 \\
= &\gamma_2(\hat{\theta}^*)\lambda_\nu^2 + 3\left(\frac{B^2 + 2B\overline{U}_\infty\left(\frac{\nu}{\nu-2}\right) + \frac{\nu^2}{(\nu-2)(\nu-4)}\overline{U}_\infty^2}{\left(B + \frac{\nu}{\nu-2}\overline{U}_\infty\right)^2} - 1 + \right. \\
&\left. \frac{\frac{\nu^2}{(\nu-2)(\nu-4)}\mathrm{Var}(U^*) + 2\frac{\nu}{\nu-2}\mathrm{Cov}\left((\hat{\theta}^* - \bar{\theta}_\infty)^2, U^*\right)}{\left(B + \frac{\nu}{\nu-2}\overline{U}_\infty\right)^2}\right) \\
= &\gamma_2(\hat{\theta}^*)\lambda_\nu^2 + 3\left(\frac{\frac{2\nu^2}{(\nu-2)^2(\nu-4)}\overline{U}_\infty^2 + \frac{\nu^2}{(\nu-2)(\nu-4)}\mathrm{Var}(U^*) + 2\frac{\nu}{\nu-2}\mathrm{Cov}\left((\hat{\theta}^* - \bar{\theta}_\infty)^2, U^*\right)}{\left(B + \frac{\nu}{\nu-2}\overline{U}_\infty\right)^2}\right)
\end{aligned}$$

This can be further simplified by noting that

$$1 - \lambda_\nu = \frac{\frac{\nu}{\nu-2}\overline{U}_\infty}{B + \frac{\nu}{\nu-2}\overline{U}_\infty}$$

The result then follows, and the Gaussian case follows in the limit $\nu \to \infty$. □

Lemma 3 (Rubin's rules overestimate variance)**.** *Let* $\sigma_\theta^2 = \mathrm{Var}[\theta \mid D_{\mathrm{obs}}]$ *and* $\sigma_R^2 = \mathrm{Var}[\theta_{\mathrm{Rubin}} \mid D_{\mathrm{obs}}]$.
Then for any $\nu_{com} > 3$ *and any choice of* U, B *such that* $\nu_\infty = \nu\left(\frac{\nu+1}{\nu+3}\right)(1-\lambda) > 2$, *we have that* $\sigma_R > \sigma_\theta$, *and* $\sigma_R^2 - \sigma_\theta^2$ *is strictly increasing w.r.t* λ, *for any* ν_{com}, *and any* $V = \overline{U}_\infty + B$.

Proof. Letting $\nu := \nu_{\mathrm{com}}$, $\lambda = \frac{B}{\overline{U}_\infty + B}$ and $V = \overline{U}_\infty + B$ we have

$$\sigma_R^2 - \sigma_\theta^2 = (1-\lambda)\, V \left(\frac{\nu\left(\frac{\nu+1}{\nu+3}\right)}{\nu\left(\frac{\nu+1}{\nu+3}\right)(1-\lambda) - 2} \right) - \left[(1-\lambda)\, V \left(\frac{\nu}{\nu-2} \right) + \lambda\, V \right]$$

Let

$$V(\lambda,\nu) := \frac{\sigma_R^2 - \sigma_\theta^2}{V} = (1-\lambda) \left(\frac{\nu\left(\frac{\nu+1}{\nu+3}\right)}{\nu\left(\frac{\nu+1}{\nu+3}\right)(1-\lambda) - 2} \right) - \left[(1-\lambda) \left(\frac{\nu}{\nu-2} \right) + \lambda \right]$$

$$\text{Then} \qquad \frac{\partial V(\lambda,\nu)}{\partial \lambda} = \left(\frac{2\nu\left(\frac{\nu+1}{\nu+3}\right)}{\left(\nu\left(\frac{\nu+1}{\nu+3}\right)(1-\lambda) - 2\right)^2} \right) - \left[1 - \frac{\nu}{\nu-2} \right]$$

Now for $\nu > 3$ and λ such that $\nu_\infty = \nu\left(\frac{\nu+1}{\nu+3}\right)(1-\lambda) > 2$, $\frac{\partial V(\lambda,\nu)}{\partial \lambda}$ is strictly positive. Hence $\sigma_R^2 - \sigma_\theta^2$ is strictly increasing w.r.t λ, and so maximised when $\lambda = 0$. We see that

$$V(0,\nu) = \left(\frac{\nu\left(\frac{\nu+1}{\nu+3}\right)}{\nu\left(\frac{\nu+1}{\nu+3}\right) - 2} \right) - \frac{\nu}{\nu-2} = \frac{4\nu}{(\nu-3)(\nu+2)(\nu-2)} > 0 \;\; \forall \nu > 3$$

Hence $V(\lambda,\nu) > 0 \;\; \forall \nu > 3, \lambda \in [0,1]$ such that $\nu_\infty(\lambda,\nu) > 2$ and so $\sigma_R > \sigma_\theta$.

□

Note that σ_R^2 is unbounded for λ large enough that $\nu_\infty(\lambda,\nu) \leq 2$, whereas σ_θ^2 is bounded above by $\nu/(\nu-2)$ for all λ.

Lemma 4 (Consistency of variance from Rubin's rules as $\nu_{\mathrm{com}} \to \infty$)**.** *For any choice of fixed* U, B, $\sigma_R \to \sigma_\theta$ *as* $\nu_{com} \to \infty$.

Proof. Note for fixed U and B, λ is also fixed. Then as $\nu \to \infty$ for fixed λ, $\nu_\infty \to \infty$. Hence $\lim_{\nu\to\infty} \sigma_R^2 = \lim_{\nu\to\infty} \sigma_\theta^2 = \overline{U}_\infty + B$. □

Lemma 5 (Bounding difference between posterior variance and variance from Rubin's rules)**.** $\sigma_R^2 - \sigma_\theta^2 < V\left(\frac{\nu_\infty}{\nu_\infty - 2} - 1\right)$

Proof. Fix $\nu_\infty > 2$, and write $\nu_{\text{com}} = \nu(\nu_\infty, \lambda)$ as a function of ν_∞ and λ. Also define $\tilde{V}(\lambda, \nu_\infty) = V(\lambda, \nu(\nu_\infty, \lambda))$.

$$\tilde{V}(\lambda, \nu_\infty) := \frac{\sigma_R^2 - \sigma_\theta^2}{V} = \frac{\nu_\infty}{\nu_\infty - 2} - (1-\lambda)\frac{\nu(\nu_\infty, \lambda)}{\nu(\nu_\infty, \lambda) - 2} - \lambda$$
$$\frac{\partial}{\partial\lambda}\tilde{V}(\lambda, \nu_\infty) = \frac{\nu(\nu_\infty, \lambda)}{\nu(\nu_\infty, \lambda) - 2} - 1 + 2(1-\lambda)\frac{\frac{\partial}{\partial\lambda}\nu(\nu_\infty, \lambda)}{(\nu(\nu_\infty, \lambda) - 2)^2}$$

Now note that for fixed ν_∞, as λ increases, ν must also increase, as $\nu\frac{\nu+1}{\nu+3}$ is strictly increasing for $\nu > 0$. Hence $\frac{\partial}{\partial\lambda}\nu(\nu_\infty, \lambda) > 0$. So the final term is strictly positive, and as $\nu > 2$, $\frac{\nu(\nu_\infty,\lambda)}{\nu(\nu_\infty,\lambda)-2} - 1 > 0$. Hence $\frac{\partial}{\partial\lambda}\tilde{V}(\lambda, \nu_\infty) > 0$. Now $\lim_{\lambda\to1} \nu(\nu_\infty, \lambda) = \infty$, so $\lim_{\lambda\to1} \frac{\nu(\nu_\infty,\lambda)}{\nu(\nu_\infty,\lambda)-2} = 1$. It follows that

$$\max_\lambda \tilde{V}(\lambda, \nu_\infty) = \lim_{\lambda\to1} \tilde{V}(\lambda, \nu_\infty) = \left(\frac{\nu_\infty}{\nu_\infty - 2} - 1\right)$$

Hence $\sigma_R^2 - \sigma_\theta^2 < V\left(\frac{\nu_\infty}{\nu_\infty - 2} - 1\right)$ □

Lemma 6 (Rubin's rules conservative for excess kurtosis, under certain conditions)**.** *For any $\nu_{com} > 4$ and any choice of λ such that $\nu_\infty = \nu\frac{\nu+1}{\nu+3}(1-\lambda) > 4$, we have that*

$$\frac{1}{\nu - 4}(1 - \lambda_\nu)^2 < \frac{1}{\nu_\infty - 4} \tag{A.5}$$

Proof. In all that follows, we restrict to choices of $\lambda \in (0, 1)$ and $\nu > 4$ such that $\nu_\infty > 4$. Note that

$$\frac{1}{\nu - 4}(1 - \lambda_\nu)^2 < \frac{1}{\nu_\infty - 4}$$

if and only if

$$\left(\nu\frac{\nu+1}{\nu+3}(1-\lambda) - 4\right)(1 - \lambda_\nu)^2 - (\nu - 4) < 0$$

Now note that $(1-\lambda_\nu) > (1-\lambda)$ for all λ, hence the above inequality holds if and only if

$$\nu\frac{\nu+1}{\nu+3}(1-\lambda_\nu)^3 - 4(1-\lambda_\nu)^2 - (\nu-4) < 0$$

Let

$$f_\nu(x) = \nu\frac{\nu+1}{\nu+3}(1-x)^3 - 4(1-x)^2 - (\nu-4)$$

then

$$\begin{aligned}\frac{\partial}{\partial x}f_\nu(x) =& -3\nu\frac{\nu+1}{\nu+3}(1-x)^2 + 8(1-x)\\ =& -(1-x)\left(3\nu\frac{\nu+1}{\nu+3}(1-x) - 8\right) < 0\end{aligned}$$

since $\nu\frac{\nu+1}{\nu+3}(1-x) = \nu\frac{\nu+1}{\nu+3}(1-\lambda_\nu) > \nu\frac{\nu+1}{\nu+3}(1-\lambda) > 4$, and $\lambda_\nu \in (0,1)$. Hence

$$f_\nu(x) < f_\nu(0) = \frac{-2\nu}{\nu+3} < 0$$

and the result follows. □

Lemma 7. *We condition on D_{obs} in all that follows, though we drop this from the notation to save space.*
In the limit as $\frac{\sqrt{\text{Var}(U^*)}}{B} \to 0$ with all other quantities held constant, the following three terms all tend to 0 as well:

$$3\left|\frac{\left(\frac{\nu}{\nu-2}\right)\text{Cov}\left(\hat{\theta}^*, U^*\right)}{\left(B + \overline{U}_\infty(\frac{\nu}{\nu-2})\right)^{3/2}}\right| \leq 3\frac{\nu}{\nu-2}\frac{\sqrt{\text{Var}(U^*)}}{B} \to 0 \tag{A.6}$$

$$3\left|\frac{\frac{\nu^2}{(\nu-2)(\nu-4)}\text{Var}(U^*)}{\left(B + (\frac{\nu}{\nu-2})\overline{U}_\infty\right)^2}\right| \leq 3\frac{\nu^2}{(\nu-2)(\nu-4)}\frac{\text{Var}(U^*)}{B^2} \to 0 \tag{A.7}$$

$$6\left|\frac{\frac{\nu}{\nu-2}\text{Cov}\left((\hat{\theta}^* - \bar{\theta}_\infty)^2, U^*\right)}{\left(B + (\frac{\nu}{\nu-2})\overline{U}_\infty\right)^2}\right| \leq 6\frac{\nu}{\nu-2}\sqrt{\gamma_2(\hat{\theta}^*) + 2}\frac{\sqrt{\text{Var}(U^*)}}{B} \to 0 \tag{A.8}$$

Proof. Recall the covariance inequality: for random variables A and B,

$$|\text{Cov}(A,B)| \leq \sqrt{\text{Var}(A)}\sqrt{\text{Var}(B)}$$

Noting further that B, ν and $\overline{U}_\infty$ are all positive, Eq. A.7 is trivial to show. Recalling $\text{Var}(\hat{\theta}^*) = B$, Eq. A.6 is then also simple to prove.

To show Eq. A.8, first note that

$$\begin{aligned}\text{Var}\left((\hat{\theta}^* - \bar{\theta}_\infty)^2\right) &= \text{E}\left[(\hat{\theta}^* - \bar{\theta}_\infty)^4\right] - \text{E}\left[(\hat{\theta}^* - \bar{\theta}_\infty)^2\right]^2 \\ &= B^2(\gamma_2(\hat{\theta}^*) + 3) - B^2\end{aligned}$$

and the result can be shown by applying the covariance inequality. □

A.2 Simulations

Supplementary Table A.1: **Choice of β_X in each scenario**

Setting	X distribution	$n_{\text{obs}} = 10$,	20	50	100	200
Linear regression	Gaussian	1.285	0.734	0.420	0.288	0.201
	Lognormal	1.506	0.815	0.442	0.296	0.204
Logistic regression	Gaussian			0.870	0.590	0.410
Cox regression	Gaussian		0.865	0.475	0.325	0.225

The values 10, 20, 50, 100, 200 in the top row give the value of n_{obs}, the number of completely-observed samples. The values in the table below give the value of β_X used to generate the simulated data for each setting and data-generating distribution of X, and each choice of n_{obs}.

A.2.1 Priors used in simulation study

For all scenarios, we first centred Z by subtracting 0.5, and then set priors on β_X the effect of X, β_Z the effect of $Z - 0.5$ and any other parameters. Where an intercept is present, this gives the value of the linear predictor at $Z = 0.5$, $X = 0$. We used the same priors for all choices of n_{obs} and prop_mis within each setting.

Linear regression, Gaussian X

$\beta_0 \sim N(0, 1)$

$\beta_X \sim N(0, 2)$ (note we used a wider prior because the largest true value of β_X used to

generate the data was 1.285 for $n_{\text{obs}} = 10$, to avoid a clash between the prior and the data)

$\beta_Z \sim N(0, 1)$

$\sigma \sim N(0, 1)$

Linear regression, lognormal X

$\beta_0 \sim N(0, 1)$

$\beta_X \sim N(0, 2)$ (note we used a wider prior because the largest true value of β_X used to generate the data was 1.506 for $n_{\text{obs}} = 10$, to avoid a clash between the prior and the data)

$\beta_Z \sim N(0, 1)$

$\sigma \sim N(0, 1)$

$\mu_X \sim N(0, 0.5)$ the parameter in the lognormal distribution. The true parameter used to generate the data was 0.

$\sigma_X \sim N(\log(0.5), 1)$ the parameter in the lognormal distribution. The true parameter used to generate the data was 0.551, so that the skew of the lognormal distribution is 2.

Logistic regression, Gaussian X

We used Student-t priors with 5 degrees of freedom, as recommended by the authors of the stan package [133, 134]. $\beta_0 \sim t_5(0, 1)$

$\beta_X \sim t_5(0, 1)$

$\beta_Z \sim t_5(0, 1)$

Cox/exponential regression, Gaussian X

$\log(h_0) \sim N(0, 1)$

$\beta_X \sim N(0, 1)$ (note we used a narrower prior than for linear regression, because the largest true value of β_X used to generate the data was 0.865)

$\beta_Z \sim N(0, 1)$

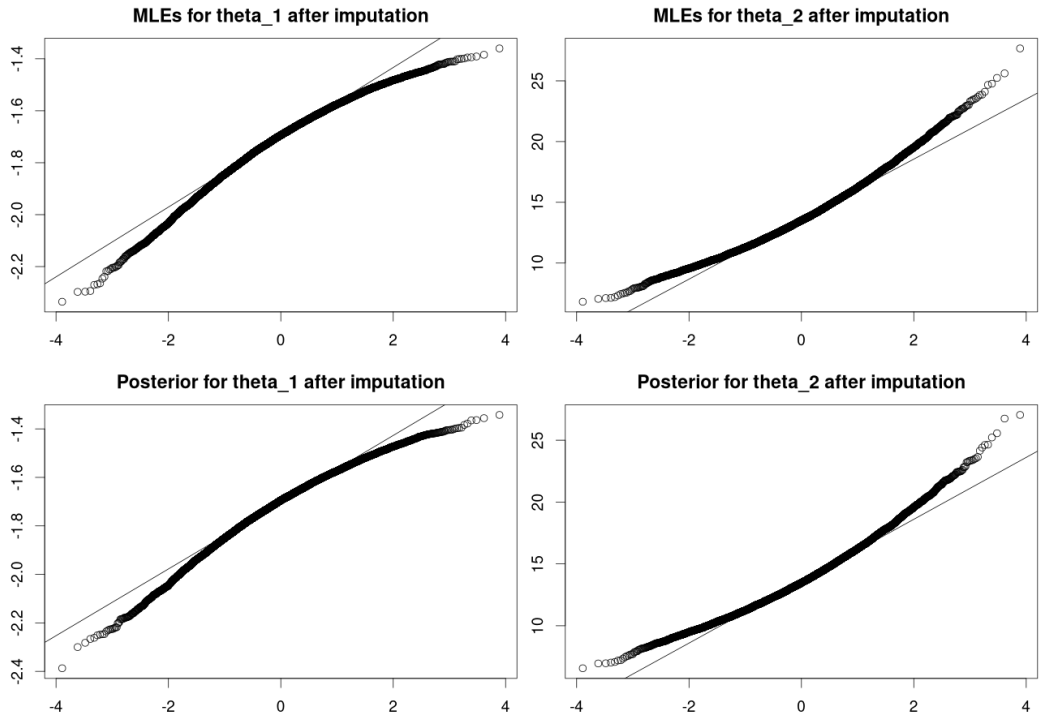
Supplementary Table A.2: **Average power and coverage across all simulation scenarios by** m

Method	m	Power	Coverage
CC		0.797	0.948
Bayes	100	0.844	0.932
Bayes	1000	0.832	0.949
Bayes	10000	0.831	0.951
Rubin_t	5	0.716	0.952
Rubin_t	20	0.796	0.957
Rubin_t	100	0.813	0.958
Rubin_t	1000	0.817	0.958
Rubin_norm	1000	0.836	0.950
AB_pool_norm	100	0.842	0.930
AB_pool_norm	1000	0.830	0.947
AB_pool_norm	10000	0.829	0.948
AB_pool_t	100	0.836	0.934
AB_pool_t	1000	0.823	0.950
AB_pool_t	10000	0.822	0.952

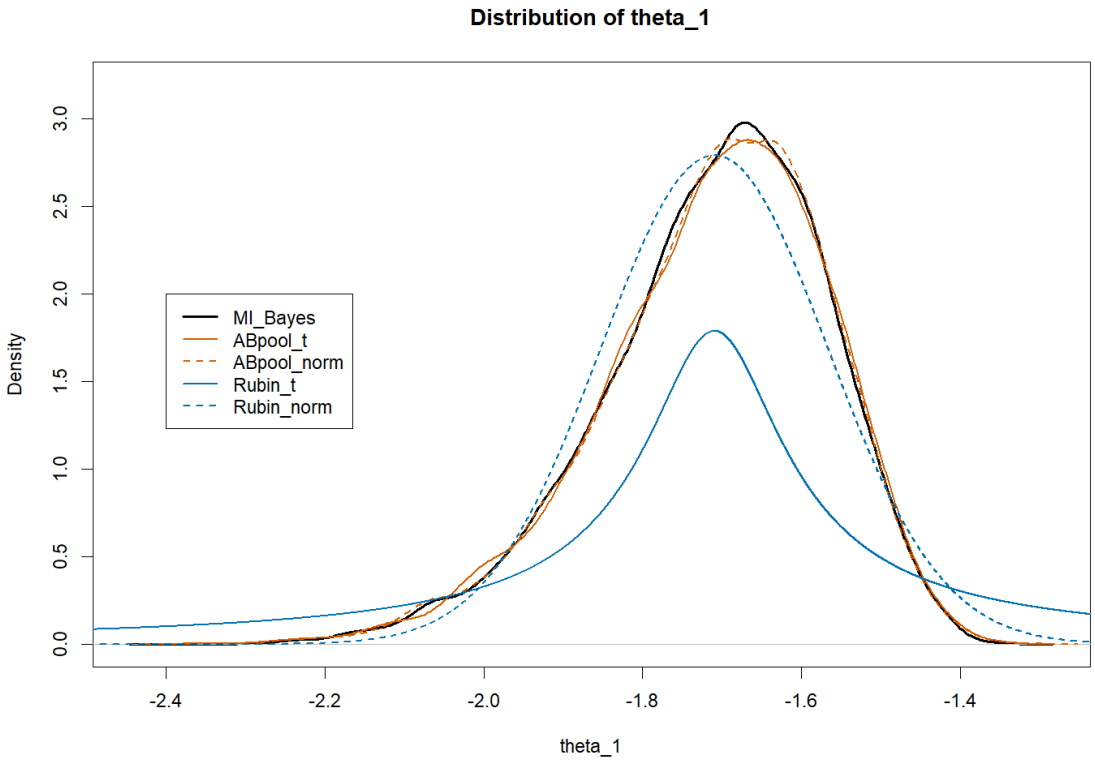
CC = complete-case analysis, Bayes = Bayesian analysis, ABpool = Approximate Bayesian pooling, and _t and _norm denote assuming a t- and Gaussian-distributed complete-data posterior respectively. Power and coverage are averaged across all scenarios and all methods – 25 linear regression scenarios with a Gaussian covariate and 25 with a lognormal covariate, 15 logistic regression scenarios, and 20 Cox regression scenarios.

A.3 Real data

A.3.1 HPV and cervical cancer correlation



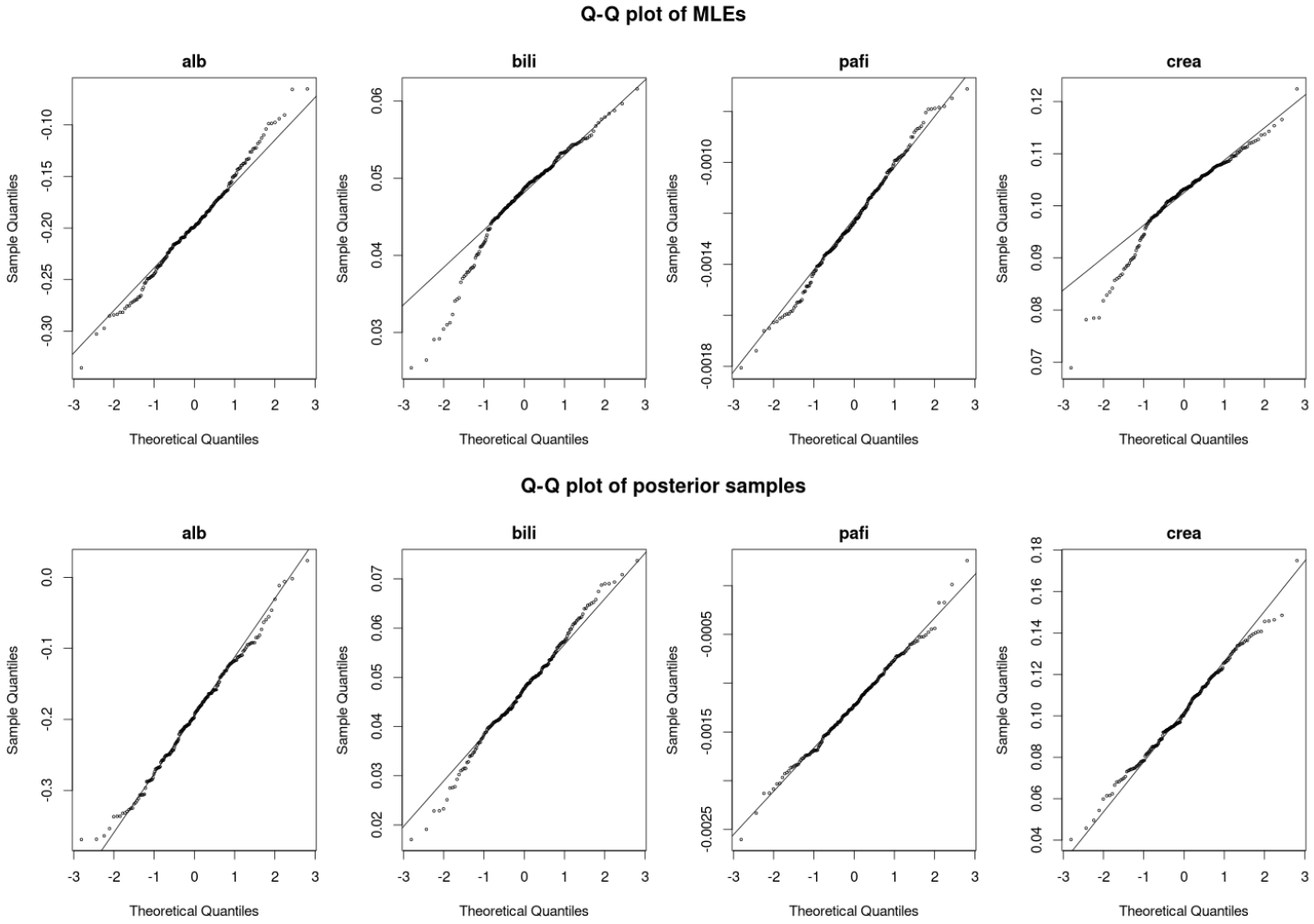
Supplementary Figure A.1: **Q–Q plots for HPV and cervical cancer correlation parameters**
Normal quantile-quantile plots for the maximum likelihood estimates and posterior draws for θ_1 and θ_2, respectively the intercept and slope in the relationship between HPV prevalence and cervical cancer incidence. Posterior samples are from the Bayesian inference after multiple imputation method to approximate the cut posterior.



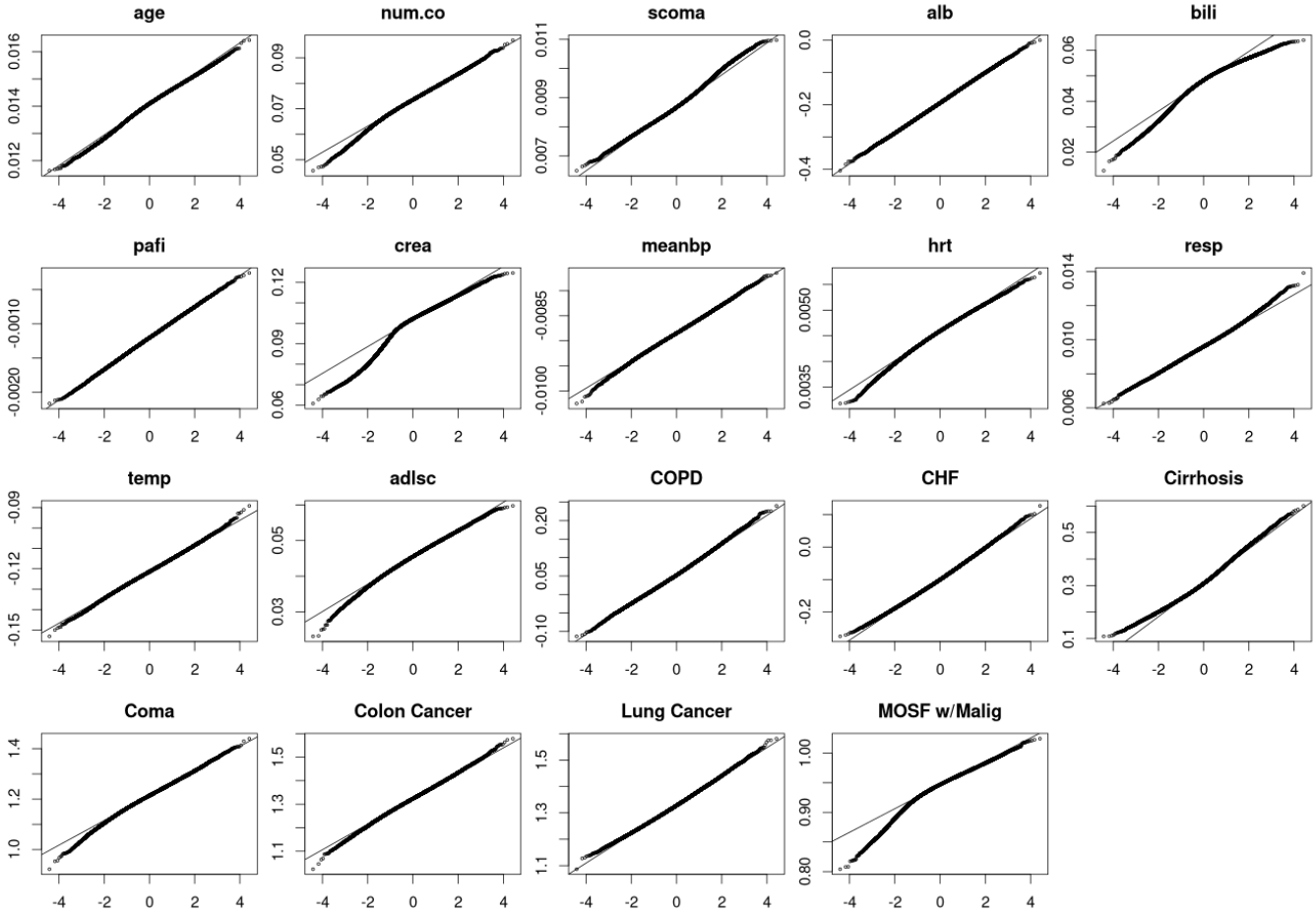
Supplementary Figure A.2: **The distribution of θ_1 estimated by Rubin's rules, approximate Bayesian pooling and Bayesian inference after multiple imputation.** The posterior distribution for the intercept θ_1. The black, orange and blue lines denote the posterior distribution estimated by Bayesian inference after multiple imputation (MI_Bayes), approximate Bayesian pooling (ABpool) and Rubin's rules (Rubin) respectively. For approximate Bayesian pooling and Rubin's rules, the solid line denotes the method assuming a t-distribution given complete data, and the dashed line denotes the method assuming a Gaussian distribution given complete data.

A.3.2 SUPPORT

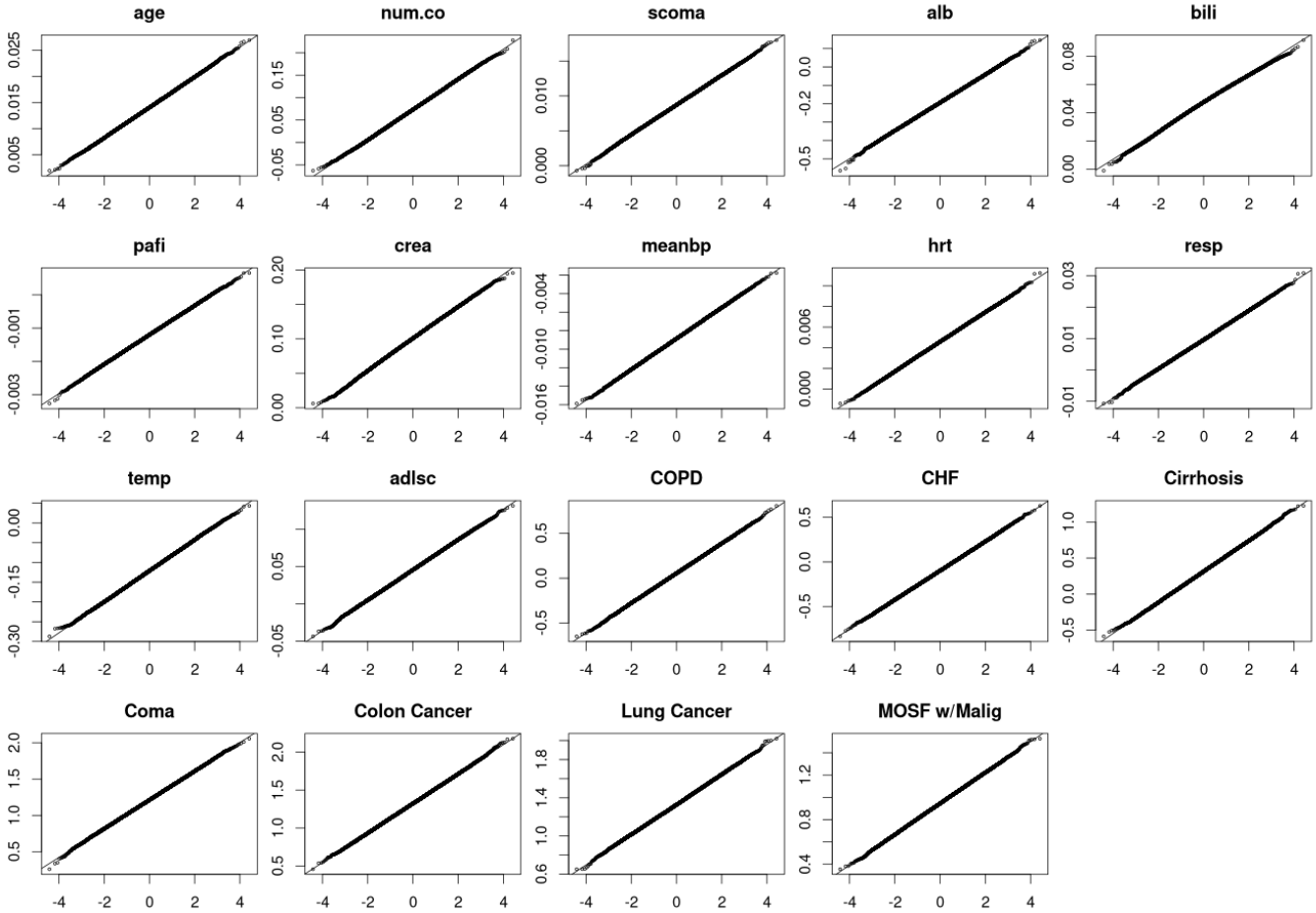
Additional figures



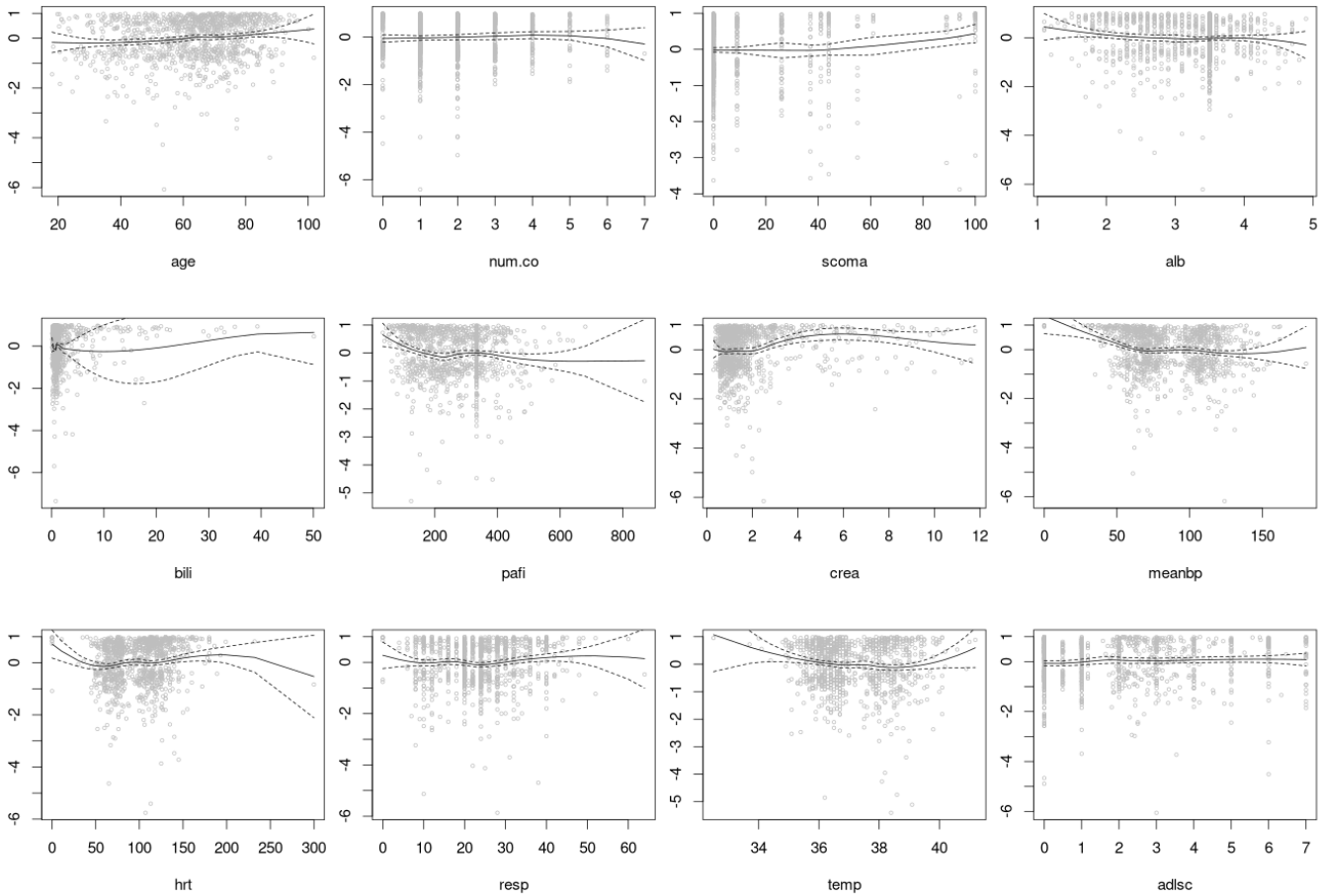
Supplementary Figure A.3: **Q–Q plots for the SUPPORT analysis**
Normal quantile-quantile plots for the maximum likelihood estimates and posterior draws for the four variables in the SUPPORT analysis which contain missing values. Only the first 200 imputations are used, to demonstrate the usefulness of testing normality with a Q–Q plot without needing to make thousands of imputations. Posterior samples are from the Approximate Bayesian pooling approach, assuming the posterior distribution under complete data is t-distributed.



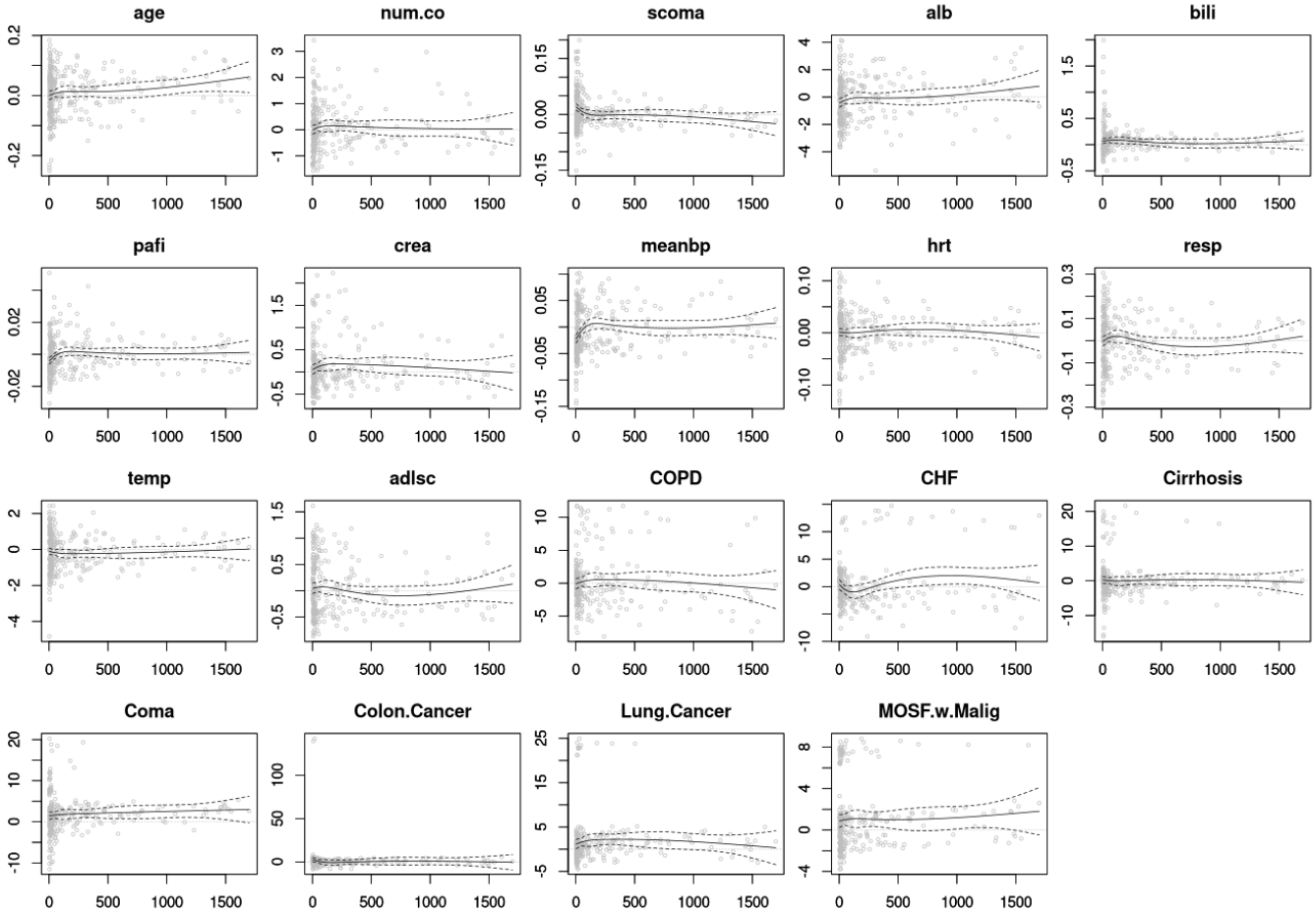
Supplementary Figure A.4: **Q–Q plots for all MLEs from the SUPPORT analysis**
Normal quantile-quantile plots for the maximum likelihood estimates for all parameters in the SUPPORT analysis. Results are shown from 100,000 multiple imputation draws.



Supplementary Figure A.5: **Q–Q plots for the posterior distribution of all parameters from the SUPPORT analysis**
Normal quantile-quantile plots for approximate posterior draws for all parameters in the SUPPORT analysis. Results are shown from 100,000 draws. Posterior samples are from the Approximate Bayesian pooling approach, assuming the posterior distribution under complete data is t-distributed.



Supplementary Figure A.6: **Martingale residuals for continuous covariates in the SUPPORT analysis**
Martingale residuals are shown for all covariates excluding categorical variables in the SUPPORT analysis. These are estimated on the singly-imputed dataset after imputing using the method recommended by the authors[73]. The plot approximates the appropriate functional form in the model - a linear trend indicates that including the untransformed covariate in the Cox model is appropriate.



Supplementary Figure A.7: **Schoenfeld residuals for the SUPPORT analysis**
Schoenfeld residuals are shown for all covariates in the SUPPORT analysis. These are estimated on the singly-imputed dataset after imputing using the method recommended by the authors[73]. A roughly constant line at 0 indicates the effect of the covariate is constant with time.

Additional discussion The martingale residuals (Supplementary Figure A.6) indicate that most parameters have an approximately linear effect in the Cox model. Some plots indicate higher order effects such as splines may be necessary, especially pafi, crea, meanbp, hrt, temp.

The Schoenfeld residuals (Supplementary Figure A.7) indicate most parameters have a time-constant effect on log-hazard, satisfying the proportional hazards assumption. This may not hold for some parameters, such as age, scoma, pafi, meanbp, lung cancer, and MOSF.w.Malig.

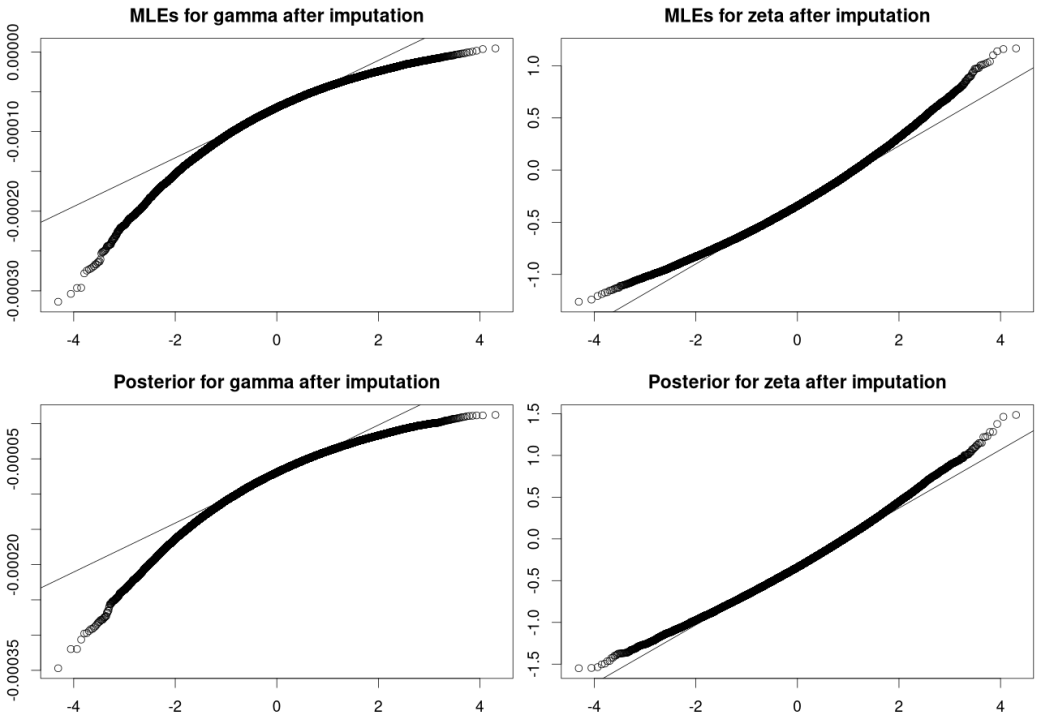
Our model could be improved upon by including spline effects and possibly extending to include time-varying effects. However we don't anticipate these changes will affect our results about pooling methods. Multiple imputation is still likely to improve upon single imputation, and Approximate Bayesian pooling is still likely to give very similar results to Rubin's rules. However we see no reason to expect the small differences between Rubin's rules and ABpool to disappear in a more complicated model.

A.3.3 COVID-19 vaccine efficacy

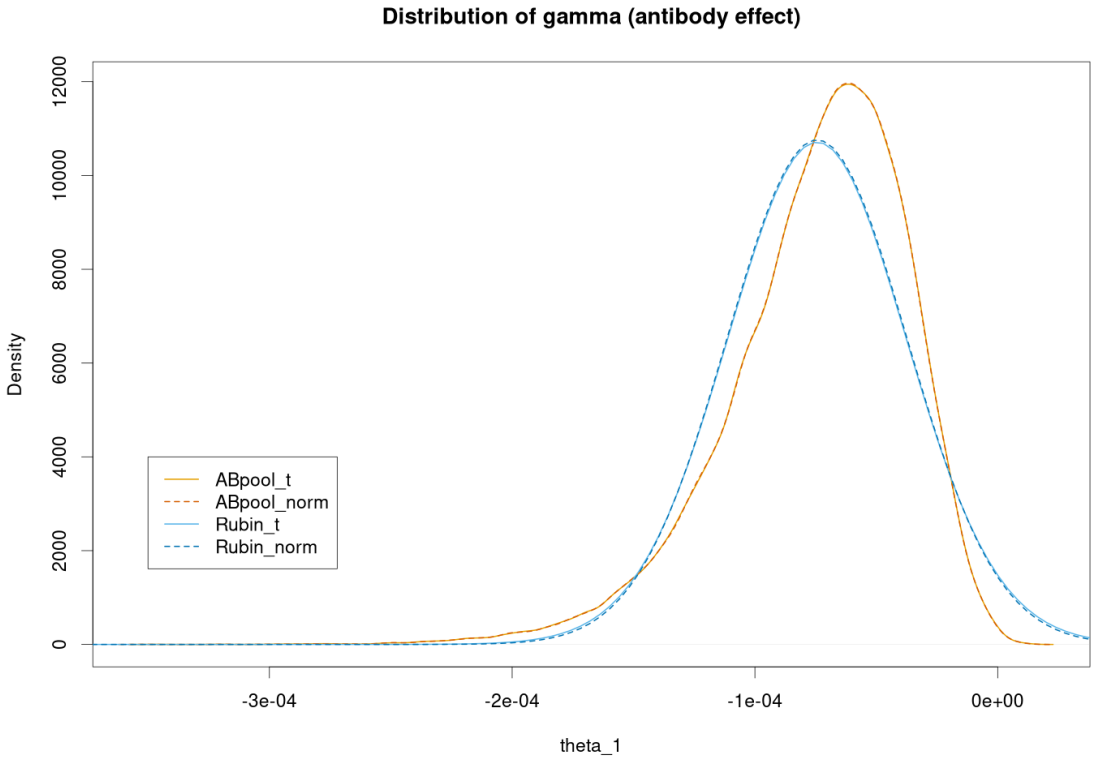
Supplementary Table A.3: **Credible intervals for antibody effect and vaccine efficacy at 10, 100, 1000 BAU/mL**

Outcome	Method	Mean	Median	2.5%	97.5%
Antibody effect	ABpool_t	-0.01153	-0.01072	-0.02494	-0.00274
	ABpool_norm	-0.01153	-0.01071	-0.02490	-0.00274
	Rubin_t	-0.01154	-0.01154	-0.02305	-0.00003
	Rubin_norm	-0.01154	-0.01154	-0.02281	-0.00027
VE(10)	ABpool_t	32.3	36.1	-20.9	63.1
	ABpool_norm	32.3	36.1	-20.7	63.0
	Rubin_t	35.4	35.4	-17.7	64.6
	Rubin_norm	35.4	35.4	-16.9	64.3
VE(100)	ABpool_t	76.2	76.1	64.3	88.7
	ABpool_norm	76.2	76.1	64.4	88.7
	Rubin_t	77.1	77.1	59.2	87.2
	Rubin_norm	77.1	77.1	59.5	87.1
VE(1000)	ABpool_t	99.6	100.0	97.3	100.0
	ABpool_norm	99.6	100.0	97.3	100.0
	Rubin_t	100.0	100.0	62.9	100.0
	Rubin_norm	100.0	100.0	70.5	100.0

Mean and median are the posterior mean and median respectively, as estimated by each method. Antibody effect is the effect on the log-hazard per increase of 1 BAU/mL. Vaccine efficacy (VE) is given in percentages.



Supplementary Figure A.8: **Q–Q plots for COVID-19 vaccine efficacy parameters** Normal quantile-quantile plots for the maximum likelihood estimates and posterior draws for γ and ζ, respectively the intercept and slope of the antibody effect on COVID-19 vaccine efficacy. Posterior samples are from the Approximate Bayesian pooling approach assuming a t-distributed complete-data posterior. 60,000 imputations are shown in each plot.



Supplementary Figure A.9: **The distribution of γ estimated by Rubin's rules, and approximate Bayesian pooling.**
The posterior distribution for the slope of the antibody effect γ. The orange and blue lines denote the posterior distribution estimated by approximate Bayesian pooling and Rubin's rules respectively. For approximate Bayesian pooling and Rubin's rules, the lighter shaded solid line denotes the method assuming a t-distribution given complete data, and the darker-shaded dashed line denotes the method assuming a Gaussian distribution given complete data.

Appendix B

Correlates of protection models and the need to account for antibody decay

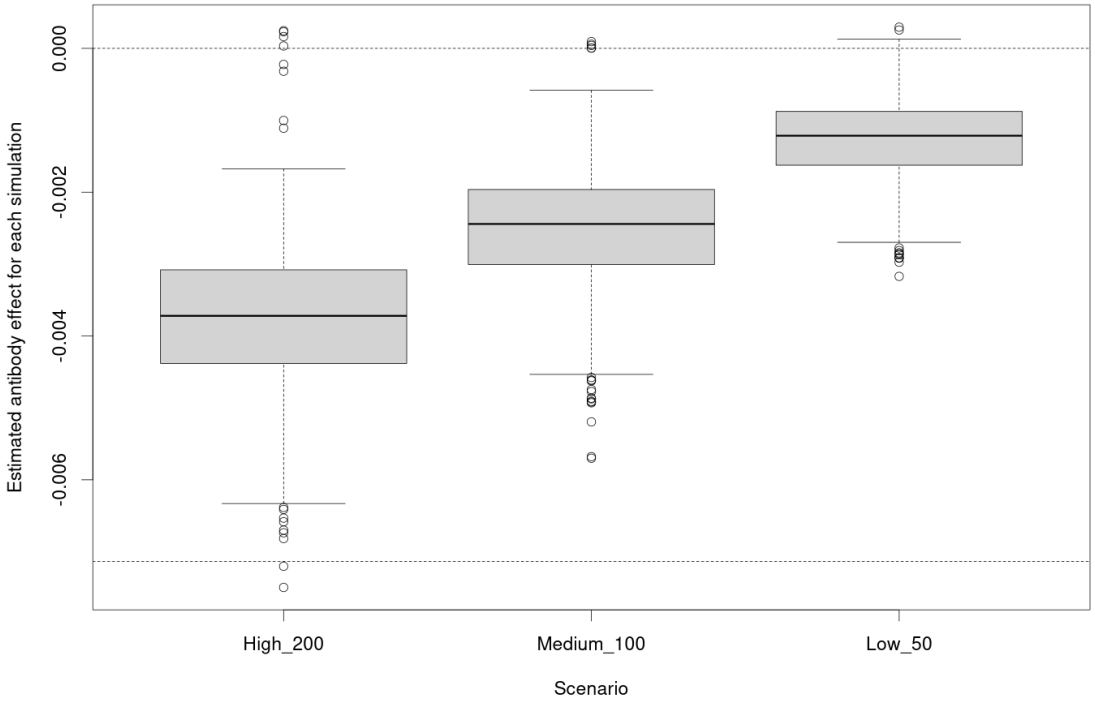
Supplementary Table B.1: **Rate of infection in the control group for COVID-19 vaccine studies**

Vaccine/Study	Rate (per 1,000 person-years)	Location in text	Calculation	Citation
ChAdOx1/COV002	149.2	Table 2		Voysey et al. (2021) [4]
ChAdOx1	215.4	Supplementary Results		Falsey et al. (2021) [135]
NVX-CoV2373/PREVENT	51.9	Main text - Efficacy		Dunkle et al. (2022) [136]
BNT162b2	72.9	Table 2	162/2.222	Polack et al. (2020) [137]
BNT162b2	141.6	Table 2	850/6.003	Thomas et al. (2021) [138]
Ad26.COV2.S/ENSEMBLE	113	Table 2	$351/3095.9 \times 1000$	Sadoff et al. (2021) [139]
mRNA-1273/COVE	56.5	Main text - Results		Baden et al. (2021) [140]

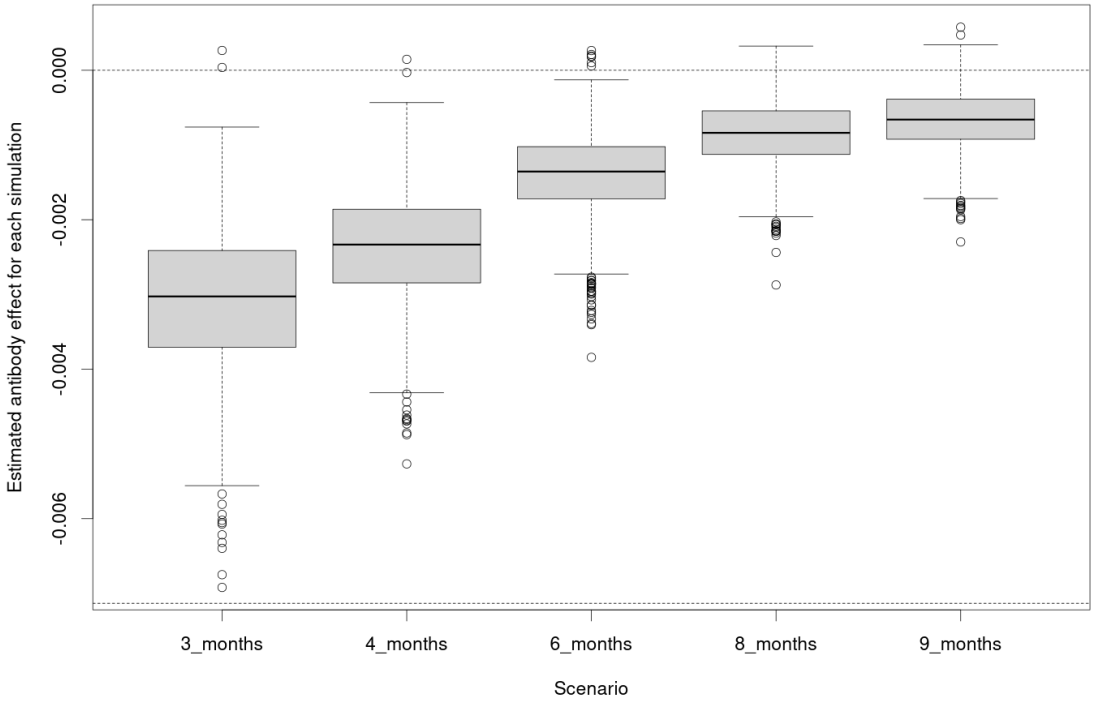
Supplementary Table B.2: **Additional information from simulation study**

	Endemic regime			Epidemic regime				
	200	100	50	3 months	4 months	6 months	8 months	9 months
VE (%)	61.5	51.5	37.1	56.5	49.2	35.1	23.4	19.0
Total events	301.9	301	300.5	242.1	265.7	291.1	298.8	299.9
Events in vaccine arm	83.6	98.0	115.6	73.1	89.2	114.2	129.2	133.7
Events in control arm	218.3	202.9	184.9	168.9	176.5	176.9	169.6	166.1
Study end time (d)	118.5	195.2	330.1	365.0	361.8	319.4	311.4	326.2
Mean event time in vaccine arm (d)	52.3	97.6	177.9	74.2	101.1	157.1	212.6	240.5
Mean follow-up time in vaccine arm (d)	88.3	164.1	297.4	331.2	327.7	286.5	279.9	295.0
Mean follow-up time in control arm (d)	86.8	161.9	294.4	325.4	322.9	284.5	279.2	294.5
Maximum coverage for gamma (%)	91.7	90.9	91.1	92.3	91.1	91.9	92.2	91.2

The mean values of various additional information about the simulated data, averaged across the 1000 simulations per scenario. The column names for the endemic regime give the baseline hazard in the control group per 1000 person-years. The column names for the epidemic regime give the timing of the peak infection rate. VE is estimated as 1-RR, with RR being the ratio of events in the vaccinated group to events in the control group. The maximum coverage for gamma shows the largest value of coverage achieved at any value of gamma; the full coverage curve is shown in Figures 3.1 & 3.2

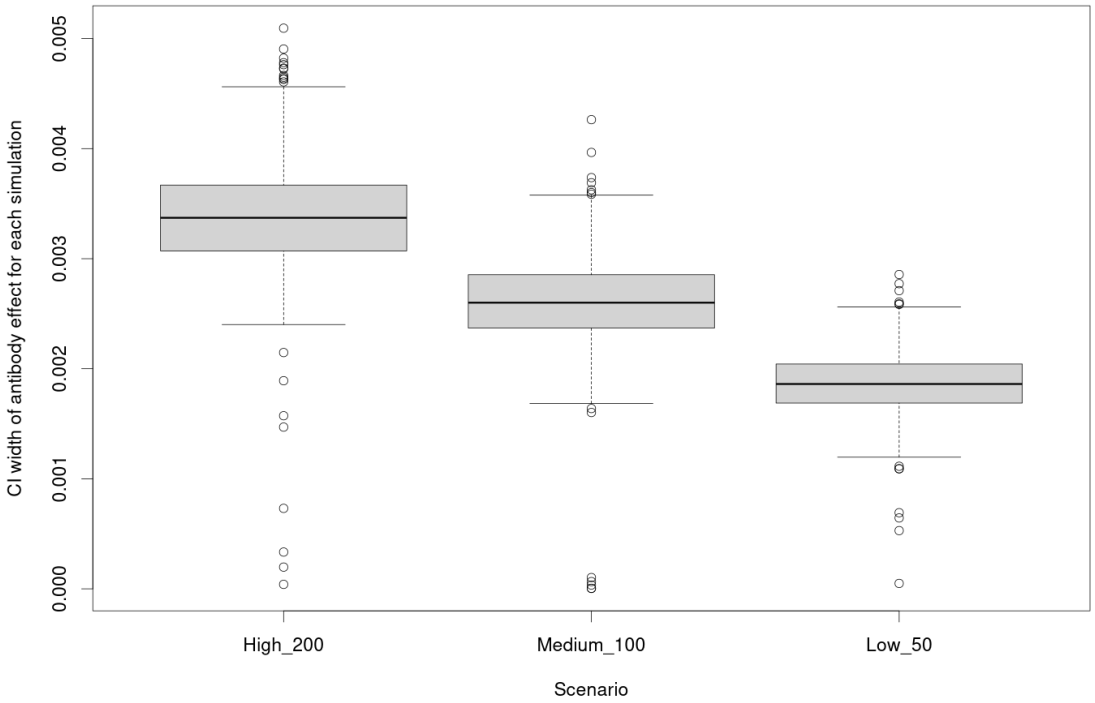


Supplementary Figure B.1: **Boxplot of the estimates for the antibody effect parameter γ for each simulation in the endemic regime**
A boxplot of the estimates $\hat{\gamma}$ across the 1000 simulated datasets in the high, medium and low constant hazard scenario. $\hat{\gamma}$ gives the estimated effect of antibody levels at time 0 on subsequent log hazard for infection. The baseline hazard is equivalent to 200, 100 and 50 cases per 1000 person-years in the high, medium and low case rate scenarios respectively. The dotted lines show the true values of γ in the extreme cases of no antibody waning ($\gamma = \gamma_0 = -0.007$) and all antibody levels waning to 0 ($\gamma = 0$). Note the boxplot only shows the range of estimates $\hat{\gamma}$, and does not display the uncertainty on the estimates.

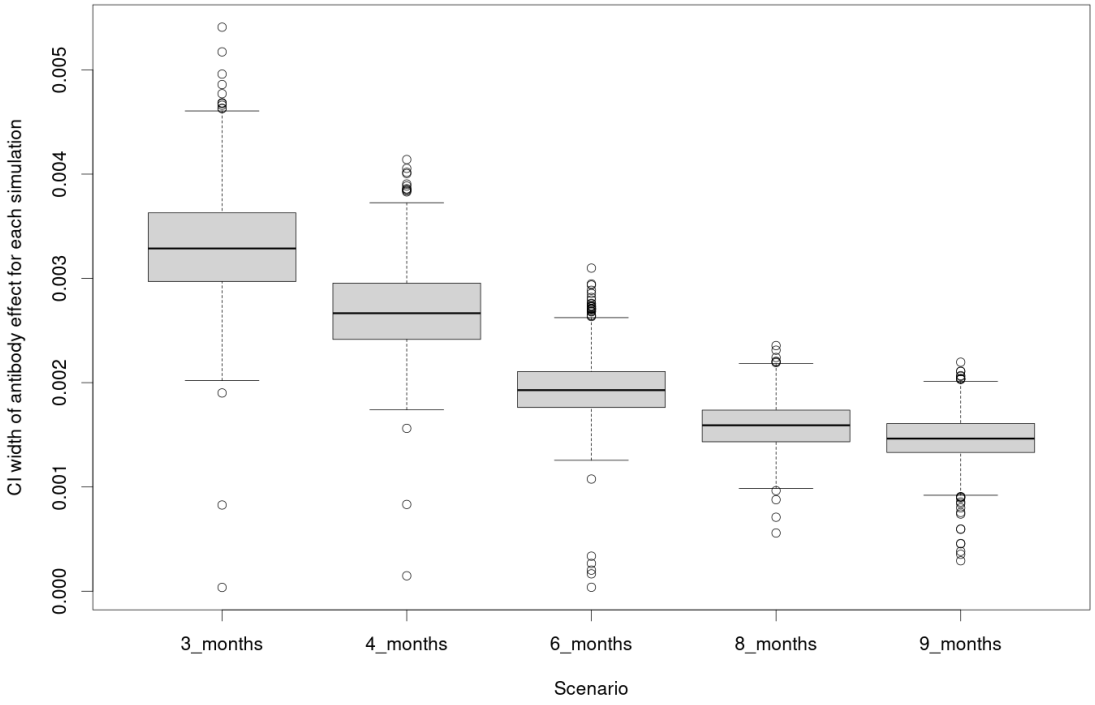


Supplementary Figure B.2: **Boxplot of the estimates for the antibody effect parameter γ for each simulation in the epidemic regime**

A boxplot of the estimates $\hat{\gamma}$ across the 1000 simulated datasets for different timings of peak infection rates in the epidemic regimes. $\hat{\gamma}$ gives the estimated effect of antibody levels at time 0 on subsequent log hazard for infection. The baseline hazard peaks at 3, 4, 6, 8 and 9 months in the respective scenarios. The dotted lines show the true values of γ in the extreme cases of no antibody waning ($\gamma = \gamma_0 = -0.007$) and all antibody levels waning to 0 ($\gamma = 0$). Note the boxplot only shows the range of estimates $\hat{\gamma}$, and does not display the uncertainty on the estimates.



Supplementary Figure B.3: **Boxplot of the CI width for the antibody effect parameter γ for each simulation in the endemic regime**
A boxplot of the confidence interval (CI) width for γ across the 1000 simulated datasets in the high, medium and low constant hazard scenario. γ gives the effect of antibody levels at time 0 on subsequent log hazard for infection. The baseline hazard is equivalent to 200, 100 and 50 cases per 1000 person-years in the high, medium and low case rate scenarios respectively. Note the boxplot shows the width of the confidence intervals, not values of confidence bounds.



Supplementary Figure B.4: **Boxplot of the CI width for the antibody effect parameter γ for each simulation in the epidemic regime**
A boxplot of the confidence interval (CI) width for γ across the 1000 simulated datasets for different timings of peak infection rates in the epidemic regimes. γ gives the effect of antibody levels at time 0 on subsequent log hazard for infection. The baseline hazard peaks at 3, 4, 6, 8 and 9 months respectively. Note the boxplot shows the width of the confidence intervals, not values of confidence bounds.

Appendix C

Joint modelling of COVID-19 correlates of protection, antibody decay and efficacy waning after vaccination

C.1 Supplementary Methods

C.1.1 Model

Longitudinal component

See the Methods section in the main text for full definitions of parameters and observed data. Assume there are N individuals in the study, of whom n received two doses of the ChAdOx1 nCoV-19 vaccine (the vaccine arm), and $N - n$ received two doses of a MenACWY control vaccine (the control arm). For individual $i = n + 1, \ldots, N$ in the control arm, we assume their true antibody levels are zero, $A_i(t) \equiv 0$. For individual $i = 1, \ldots, n$ in the vaccine arm, we assume the jth antibody observation $\log Y_i(t_{ij})$ is given by

$$\log Y_i(t_{ij}) = \log A_i(t_{ij}) + \epsilon_{ij} \tag{C.1}$$

where random error $\epsilon_{ij} \sim t_\nu(0, \sigma^2)$ follows the Student-t distribution with ν degrees of freedom, location 0 and scale σ^2. We used the Student-t distribution as QQ-plots showed better fit than using Gaussian error, and it is more robust to outliers than a Gaussian error distribution [36]. We chose $\nu = 4$ by examining QQ-plots of the residuals from an earlier model using Gaussian random error. We considered different integer values of ν, and chose $\nu = 4$ as the QQ-plot was closest to a straight line.
The formulation of equations (4.2) and (4.3) can be alternatively written as

$$\log A_i(t) = \alpha_0 + x_{Li}^\top \beta_{L0} + \tau_0 \eta_{0i} - t \exp\left(\alpha_1 + x_{Li}^\top \beta_{L1} + \tau_1 \left(\rho \eta_{0i} + \sqrt{1-\rho^2}\eta_{1i}\right)\right) \quad \text{(C.2)}$$

where η_{0i}, η_{1i} are independent standard normal random variables, $i = 1, \ldots, n$. This formulation is used in the code implementing the model.

Time-to-infection component

For a given infection outcome (any COVID-19 infection or primary symptomatic COVID-19 infection), let δ_i be the event indicator for individual i, equal to 1 if they suffered (any/symptomatic) infection, and 0 otherwise. Let
$\mathcal{R}_i = \{j : S_j + T_j \geq S_i + T_i > S_j \ \& \ G_j = G_i\}$ be the set of individuals at risk at calendar time $S_i + T_i$, who are in the same study site as individual i. Then the partial likelihood is given by

$$L_P(\gamma, \zeta, \beta_D \,|\, T, \delta, X) = \prod_{i=1}^{N} \frac{\delta_i \exp\left(Z_i \left(\gamma A_i(T_i - 7) + \zeta\right) + x_{Di}^\top \beta_D\right)}{\sum_{j \in \mathcal{R}_i} \exp\left(Z_j \left(\gamma A_j(T_i + S_i - S_j - 7) + \zeta\right) + x_{Dj}^\top \beta_D\right)} \quad \text{(C.3)}$$

The Cox model maximises the partial likelihood with respect to parameters γ, ζ, β_D.

C.1.2 Bayesian implementation of longitudinal component

Accounting for informative missingness

Missing data occurred in the antibody measurements for two reasons. Firstly, longitudinal antibody observations were censored, that is, excluded if they occur after the infection event or the end of the at-risk period. Secondly, blood samples were tested for antibody levels retrospectively by case-cohort sampling, meaning antibody data is

missing for a large proportion of non-cases. Hence the missingness of antibody data depends by design on the total time at risk and the final infection status. Because antibody levels affect the risk of infection, these time-to-event variables are informative of the antibody levels. In order to account for this informatively missing-at-random antibody data, we also included infection data in the longitudinal covariates. Namely, we included the indicator for any COVID-19 infection, the indicator for primary symptomatic COVID-19 infection, and the Nelson–Aalen estimate for cumulative hazard of any COVID-19 NAAT+ test in the vaccine arm. This approach was first suggested for the Cox model by White and Royston (2009) [64]. Moreno-Bentacur et al. (2018) used a similar approach in a joint model, suggesting to also include the interaction between other covariates and the Nelson–Aalen estimator, although they note the interaction term did not improve the estimates [46]. The inclusion of both infection indicators accounts for the censoring at any infection event. This can further be seen as a multiple imputation approach to fitting a joint model with two competing risks of non-symptomatic COVID-19 infection and symptomatic COVID-19 infection.

Covariate effects

The multiplicative effect due to an increase of 1 in the jth covariate x_{Lj} on the antibody level at PB28 and the half-life of antibody level decay is given by $\exp(\beta_{L0j})$ and $\exp(-\beta_{L1j})$ respectively. This can be derived from Equation (C.2).

Standardisation

We standardised the data before estimating the posterior, in order to improve mixing in the Hamiltonian Monte Carlo algorithm. We recentered the time since second vaccination variable at 28 days post second dose, and divided by the standard deviation of the antibody observation times. We standardised the covariates and observed antibody levels by subtracting the mean and dividing by the standard deviation of each variable. We then fit the model on the standardised data, before transforming the parameters back to the original scale.

Priors

We fit the following independent priors to the parameters, where θ' indicates the parameter θ transformed to the standardised scale:

$$\alpha'_0,\ \alpha'_1 \sim \text{Normal}(0, 2^2) \tag{C.4}$$

$$\beta'_{L0}, \beta'_{L1} \sim \text{Normal}(0, 2^2 I_{p_L}) \tag{C.5}$$

$$\tau'_0,\ \tau'_1,\ \sigma' \sim \text{Student-t}(\text{d.f.} = 2, \text{mean} = 0, \text{scale} = 1) \tag{C.6}$$

$$\rho' \sim \text{Uniform}(-1, 1) \tag{C.7}$$

where I_{p_L} denotes the identity matrix with p_L rows. These priors, combined with equations (4.2)-(4.3) induce a prior on $A_i(t)$.

Model fitting

We implement the model using Hamiltonian Monte Carlo [141] in Stan [6]. We ran 4 chains of 20000 iterations each, discarding the first 5000 iterations as warm-up and combining all the remaining iterations from each chain. We checked for convergence by examining trace plots and checking all values of the potential scale reduction factor were less than 1.01.

C.1.3 Approximate Bayesian inference procedure

We estimate the joint posterior distribution by multiple imputation.
The joint posterior distribution for the antibody trajectories $\boldsymbol{A}(t) = (A_1(t), \ldots, A_n(t))^\top$, longitudinal parameters $\theta_L = (\alpha_0, \alpha_1, \beta_{L0}, \beta_{L1}, \tau_0, \tau_1, \rho, \sigma\ )$ and time-to-event parameters $\theta_D = (\gamma, \zeta, \beta_D)^\top$, given data $D_{obs} = (Y, T, \delta, X_L, X_D)$ is given by

$$\pi\left(\boldsymbol{A}(t), \theta_D, \theta_L \,|\, D_{obs}\right) = \pi\left(\theta_D \,|\, \boldsymbol{A}(t), D_{obs}\right)\ \pi\left(\boldsymbol{A}(t) \,|\, \theta_L, D_{obs}\right)\ \pi\left(\theta_L \,|\, D_{obs}\right) \tag{C.8}$$

as the time-to-event parameters θ_D are independent of longitudinal parameters θ_L given antibody trajectories $\boldsymbol{A}(t)$.
Hence to sample from the joint posterior distribution, we may first draw a sample of antibody trajectories $\boldsymbol{A}(t)^{(k)} \sim \pi\left(\boldsymbol{A}(t) \,|\, \theta_L, D_{obs}\right)\ \pi\left(\theta_L \,|\, D_{obs}\right)$, then draw a sample of

time-to-event parameters $\theta_D^{(k)} \sim \pi\left(\theta_D \mid \boldsymbol{A}(t), D_{obs}\right)$, given the sample of antibody trajectories. This is equivalent to treating the antibody intercepts and slopes a_0, a_1 as missing data and imputing from an imputation model. The antibody decay model approximates the distribution of antibody trajectories given the antibody and time-to-event data $\pi\left(\boldsymbol{A}(t) \mid Y, T, \delta, X_L, X_D\right)$. Note Section C.1.2 means our samples are drawn from an approximation of the appropriate posterior, which conditions on the time-to-event data T, δ.

We first impute the latent true antibody trajectories from the Bayesian posterior defined in Section C.1.2. We then run the time-to-event model to estimate the distribution of the time-to-event parameters given the imputation.

For $k = 1, \ldots, K$, we impute the latent values of the true antibody level using draws of the random intercept $a_{0i}^{(k)}$ and random slope $a_{1i}^{(k)}$, $i = 1, \ldots, n$ from the posterior of the longitudinal model. A set of imputed intercepts and slopes $a_0^{(k)}$, $a_1^{(k)}$ defines the true antibody trajectories $A_i^{(k)}(t)$, $i = 1, \ldots, n$ in Eq. (4.4).

We then run a Cox proportional hazards model with time-varying covariate $A_i^{(k)}(t-7)$ and hazard given by Eq. (4.4). The model gives us the maximum likelihood estimate (MLE) $\hat{\theta}_I^{(k)} = \left(\hat{\gamma}^{(k)}, \hat{\zeta}^{(k)}\right)^\top$ and corresponding covariance matrix $\widehat{\mathrm{Var}}_{\theta_D}^{(k)}$ for the time-to-infection parameters, given the kth imputation of antibody trajectories.

Next we sample from the asymptotic distribution of the Cox maximum likelihood estimators $\hat{\theta}_I^{(k)}$, such that $\theta_D^{(k)} \sim N\left(\hat{\theta}_I^{(k)}, \widehat{\mathrm{Var}}_{\theta_D}^{(k)}\right)$. This approximates a Bayesian posterior with diffuse improper priors [65], namely the distribution of $\pi\left(\theta_D^{(k)} \mid \boldsymbol{A}(t)^{(k)}, Y, T, \delta, X_L, X_D\right)$ in Eq. (C.8). Then our samples $\theta_D^{(k)}, \boldsymbol{A}^{(k)}(t)$ are approximate draws from the joint posterior $\pi(\theta_D, \boldsymbol{A}(t) \mid Y, T, \delta, X_L, X_D)$. Posterior quantities of interest can be calculated from these samples and summarised by posterior medians, means, and 95% credible intervals by calculating quantiles.

This procedure is equivalent to imputing missing antibody intercepts and slopes, and then performing Approximate Bayesian pooling assuming a Gaussian complete-data posterior (Chapter 2). Note due to the large sample size, the degrees of freedom from a t-distributed complete-data posterior is large, so a Gaussian assumption is reasonable.

C.1.4 Calculation of posterior quantities

Antibody corresponding to a given vaccine efficacy

The antibody level A required for a given vaccine efficacy VE is then given by the inverse of Eq. (4.5)

$$\mathrm{A}(\mathrm{VE}) = \left(\log(1-\mathrm{VE}) - \zeta\right)/\gamma \tag{C.9}$$

When this is negative, we report antibody of 0, as negative antibodies are not possible.

Mean vaccine efficacy over time

The mean vaccine efficacy at time t is then given by

$$\mathrm{VE}(t) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{VE}(\mathrm{A}_i(t-7)) = 1 - \frac{1}{n}\sum_{i=1}^{n}\exp(\gamma \mathrm{A}_i(t-7)+\zeta) \tag{C.10}$$

Note this is describing a counterfactual scenario where individuals are exposed to COVID-19 at time t and not exposed prior. In real-life populations, individuals who test positive are protected from future infection (for a period) and drop out from the sum. This will be more likely for individuals who are higher risk for COVID-19, including namely unvaccinated individuals. Hence the observed vaccine efficacy in a population will be shrunk towards a null effect compared to the counterfactual vaccine efficacy we report here. This is sometimes known as the differential depletion of susceptibles bias [120].

Quantiles of vaccine efficacy

Let $\mathbf{A}(t)$ be the vector with ith entry $\mathrm{A}_i(t)$. Let $Q(\boldsymbol{x}, q)$ be the qth quantile of the vector $\boldsymbol{x}$. Then the qth quantile of vaccine efficacy at time t, VE_q is then given by

$$\mathrm{VE}_q(t) = Q\left(\mathrm{VE}(\mathbf{A}(t-7)), q\right) = 1-\exp(\gamma Q(\mathbf{A}(t-7), q)+\zeta) = \mathrm{VE}(Q(\mathbf{A}(t-7), q)) \tag{C.11}$$

That is, the qth quantile of vaccine efficacy at time t is equal to the vaccine efficacy given at the qth quantile of antibody levels at time $t-7$. As in Section C.1.4 above, this calculates a counterfactual vaccine efficacy, in which individuals are not exposed to the virus until time t. In an observed population individuals who have the lowest antibody

response will be infected earlier on average, and drop out of the at-risk group.

Calculating covariate effects on vaccine efficacy

To calculate covariate effects on vaccine efficacy, we predicted efficacy in new individuals with given covariate values. As mentioned in the main text, the event indicators and Nelson–Aalen (N–A) estimate of cumulative hazard for a new individual are unknown, as they depend on the infection outcome. We first built an imputation model to sequentially impute the missing event indicators (infection, symptomatic) and N–A estimate, which appear in the longitudinal antibody model, given the known covariates.

Imputing the missing time-to-event data We imputed the missing time-to-event data sequentially. We ran regression models on the observed dataset, then drew predictions from them for the new individuals to impute the missing values.
We first ran a logistic model to predict the probability of any COVID-19 infection using all longitudinal covariates x_L except the event indicators and Nelson–Aalen estimate (logmI). We then ran a logistic model to predict the probability of primary symptoms given infection occurred, using all longitudinal covariates x_L except N–A estimate (logmS). We then ran a linear model to predict the N–A estimate using all longitudinal covariates x_L including infection and primary symptomatic status (linmNA).
For a new individual, we imputed the infection indicator to be 1 with probability equal to their predicted probability of infection from the logmI model above, and 0 otherwise. If the imputed infection indicator is 1, we next imputed the symptomatic infection indicator to be 1 with probability equal to their predicted probability of symptomatic infection given they were infected, predicted from the logmS model above, and 0 otherwise. Given the imputed event indicators, we imputed the N–A estimate to be the predicted value from the linear model linmNA, plus a randomly sampled residual from linmNA. If the resulting imputated value is less than 0, we replace it with 0, as the N–A estimate is non-negative by definition. Together, these imputations are an approximate draw from the distribution of the event indicators and N–A estimate given the longitudinal covariates.

Predicted vaccine efficacy for a new individual We imputed these missing covariates n times. For each set of imputed covariates, we drew from the predictive distribution of antibody trajectories for a new individual with the given covariate values, $\mathrm{A}_j^{\mathrm{pred}}(t), j = 1, \ldots, n$. We then averaged these to calculate the mean risk of infection given the known covariate values

$$\mathrm{VE}(t) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{VE}(\mathrm{A}_j^{\mathrm{pred}}(t)) = 1 - \frac{1}{n}\sum_{j=1}^{n} \exp(\gamma \mathrm{A}_j^{\mathrm{pred}}(t) + \zeta) \tag{C.12}$$

Predicted relative vaccine efficacy against Omicron BA.4/5 infection

To estimate the relative efficacy against Omicron BA.4/5 infection over time compared with vaccine non-responders, we considered three scenarios. We assumed the estimated mean, 95% upper and lower CI bounds for the relationship between anti-spike IgG antibody levels and relative efficacy against Omicron BA.4/5 infection from Wei et al. (2023) [82] to be the true relationship in each scenario respectively. We extracted the estimate and 95% CI from Wei et al. (Supplementary Fig. 1 from Wei et al.), at every 500 BAU/mL (0, 500, 1000, 1500, ..., 8000), and at every 10% relative efficacy (0%, 10%, 20%, ..., 80%). We then fitted a LOESS curve to each of the three curves (95% lower confidence bound, estimate, 95% upper confidence bound), with a span of 0.3. The span was chosen such that the curve visually fit the data well, with no signs of overfitting. We then predicted the relative efficacy against Omicron infection RE(A) at a given antibody level A from the fitted LOESS curve. The mean relative efficacy against Omicron in the trial is then given by

$$\mathrm{RE}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{RE}(\mathrm{A}_i(t-7)) \tag{C.13}$$

C.2 Supplementary Results

C.2.1 Predicted efficacy over time against the Omicron BA.4/5 variant

We extrapolated our results to predict what the observed relative efficacy over time in our study might have been, were the trial conducted during the Omicron period. We predicted efficacy relative to vaccine non-responders, defined below. Wei et al. (2023)

estimated the relationship between anti-spike IgG antibody levels and relative efficacy against Omicron BA.4/5 variant infection after booster vaccination following two initial doses of COVID-19 vaccine [82]. They calculated efficacy relative to vaccinated individuals who produced antibody levels of 16 BAU/mL, deemed a non-response. We applied their estimated relationship between anti-spike IgG antibody levels and relative efficacy against the Omicron BA.4/5 variant to the predicted antibody levels over time from this study. We considered three scenarios: assuming their reported mean, upper- and lower 95% confidence limit for the relationship between anti-spike IgG and relative efficacy against Omicron BA.4/5 to be the true relationship. We refer to these as the mean efficacy, high efficacy and low efficacy scenarios respectively. This analysis assumes the relationship between antibody levels and efficacy after two doses of ChAdOx1 nCoV-19 is the same as the relationship after three vaccine doses (of possibly different vaccines). This assumption is unlikely to hold, which may bias our analysis (see Supplementary Discussion). In the mean efficacy scenario, we estimated relative efficacy at day 35, 97 and 189 to be 18.5% (95% CrI: 17.8, 19.2), 11.0% (10.5, 11.6) and 4.9% (4.4, 5.5) respectively (Supplementary Fig. C.11). In the high and low efficacy scenarios respectively, we estimate relative efficacy of 23.4% (22.6, 24.2) and 13.6% (12.9, 14.2) at day 35, 14.8% (14.1, 15.4) and 7.4% (7.0, 7.9) at day 97 and 7.1% (6.5, 7.8) and 3.1% (2.7, 3.5) at day 189. The credible intervals do not reflect the uncertainty in the relationship between antibody levels and relative efficacy. This indicates the antibodies induced by two doses of ChAdOx1 nCoV-19 might only induce minimal protection against Omicron BA.4/5 of at most 24.2% at day 35, waning to at most 7.8% at day 189, relative to a vaccinated individual with no antibodies (see Discussion).

C.2.2 Model fit

To demonstrate how the model works, Supplementary Fig. C.12 shows the estimated anti-spike IgG antibody trajectories over time for 12 different randomly chosen individuals We randomly chose 12 individuals, 3 each with 0, 1, 2 and 3 antibody observations available respectively, and plotted their estimated true anti-spike IgG levels over time. The plot demonstrates that individuals with no antibody observations are predicted to have a typical antibody trajectory, with a large uncertainty. Individuals

with more observations have less uncertainty around their estimated antibody levels over time.

C.2.3 Computation time

The longitudinal model was run in Stan with 4 chains, in parallel on a SLURM compute cluster. This took 2h 23m (9h 33m total walltime), and required 52GB of RAM to generate 15000 iterations per chain, after 5000 iterations of warmup. The Cox model was run in parallel using the snowfall package in R, on the imputed datasets from each iteration from the longitudinal model. This was run on a SLURM compute cluster separately for each outcome (symptomatic, all infections). Each outcome was run using 24 parallel nodes, taking 2h 37m (2 days 12h, 6m total walltime) and 2h 43m (2 days 12h, 35m total walltime) for the symptomatic and all infection analyses respectively. The Cox analysis required 41GB and 38GB of RAM respectively. The results from the Cox and longitudinal models were combined to obtain the vaccine efficacy results reported in the paper in a third step using R and the snowfall package. This was run on 24 parallel nodes on a SLURM compute cluster, taking 1h 28m (1 day 11h 10m total walltime), and 1h 22m (1 day 8h 50m total walltime) for the symptomatic and all infection outcomes respectively. The final analysis required 49GB and 50GB of RAM respectively.

C.2.4 Calculated parameters

Posterior summaries of the estimated parameters are reported as follows. The intercept, slope and related variance and correlation parameters are reported in Supplementary Table C.4. The multiplicative covariate effects on 28 days post second dose and half-life are reported in Fig. 4.4 and Supplementary Tables C.5 & C.6. These are given by $\exp(\beta_0)$ and $\exp(-\beta_1)$ respectively, and the natural logarithm can be applied to calculate β_0 and β_1. For the time-to-event model, the covariate effects on hazard (including antibody and vaccination) are reported in Supplementary Fig. C.7, and Supplementary Table C.7.

C.3 Supplementary Discussion

Our data was collected in a period when the Alpha variant was dominant, and considered individuals without prior infection who received two doses of vaccine. In contrast, the Omicron variant is currently dominant, and many individuals have received three or more doses of vaccine and been previously infected. While some findings such as effects of dose interval on vaccine efficacy will likely still apply, our results are therefore not directly applicable to current and future populations. To aid with this, we extrapolated our results to examine what the efficacy over time since second dose might have been in the Omicron BA.4/5 period. We used the estimated relationship between efficacy and antibody levels reported by Wei et al. (2023) [82] to estimate the level of Omicron BA.4/5 protection induced by antibody trajectories observed in this trial. This analysis should be interpreted with caution for three reasons: firstly, Wei et al., and thus also our analysis, use vaccine non-responders (vaccinated individuals whose antibody level was 16 BAU/mL) as the comparison in their calculation of relative efficacy instead of unvaccinated controls. It is not clear therefore how our analysis, or that of Wei et al., might relate to protection against Omicron BA.4/5 compared with unvaccinated individuals. Secondly, the relationship reported by Wei et al. was calculated from observational data, which may introduce bias into our analysis. Thirdly, our analysis assumed that the relationship between antibody levels and efficacy was the same after two doses in our study, as in the Wei et al. study. Their study combined individuals vaccinated with two, three or four vaccine doses (83.9% had three doses), including individuals who received the ChAdOx1 nCoV-19 and BNT162b2 as their first two doses, and different booster dose vaccines. The relationship between antibody levels and efficacy may differ across these groups, which may introduce some bias into our analysis. We suggest that protection from two doses of vaccine is likely to be no greater than after three doses, at the same level of antibodies. Therefore any such bias is likely to result in our reported estimates being an over-estimate of the effectiveness of two doses against Omicron BA.4/5, compared with vaccine non-responders. These estimates are all calculated for individuals with no prior infection, which is rare today. Levels of vaccine-induced protection may be different for individuals with a prior infection.

We estimated low efficacy after two doses of ChAdOx1 nCoV-19 vaccine against the Omicron BA.4/5 variant, of 18.5% at 35 days, 7.6% at 140 days and 4.1% at 209 days, relative to vaccine non-responders (Supplementary Fig. C.11). Supplementary Table C.9 compares our methods and results to previous studies of vaccine efficacy against the Omicron variant. Andrews et al. (2022b) used a test-negative design to estimate VE against Omicron [98] and is vulnerable to the aforementioned biases from observational studies. Andrews et al. reported a much higher vaccine efficacy against Omicron at 2-4 weeks since second dose, but report a small negative vaccine efficacy at 25 weeks since second dose, which may be due to bias. While the results from our study are not directly comparable with theirs, it is however interesting that they reported higher initial VE against Omicron than our estimated relative efficacy, with both studies waning to negligible protection after 25 weeks. This might indicate that vaccine non-responders are initially protected against infection relative to control individuals, but this protection wanes by 25 weeks.

C.4 Supplementary Tables

Supplementary Table C.1: **Baseline characteristics of vaccinated (ChAdOx1) and control arms**

	ChAdOx1 (n=4605)	Control (n=4423)
Age (years)		
Median (IQR)	45 (34, 56)	45 (34, 55)
18-55 years	3435 (74.6%)	3423 (77.4%)
55-69 years	568 (12.3%)	503 (11.4%)
70+ years	602 (13.1%)	497 (11.2%)
Sex		
Male	1934 (42.0%)	1765 (39.9%)
Female	2671 (58.0%)	2658 (60.1%)
Ethnicity		
White	4235 (92.0%)	4092 (92.5%)
Other	370 (8.0%)	331 (7.5%)
Comorbidities		
None	3475 (75.5%)	3348 (75.7%)
Comorbidity	1130 (24.5%)	1075 (24.3%)
BMI (kg/m^2)		
Median (IQR)	25.5 (23, 28.8)	25.4 (22.9, 28.9)
BMI <30 kg/m^2	3704 (80.4%)	3531 (79.8%)
BMI ≥30 kg/m^2	901 (19.6%)	892 (20.2%)
Healthcare worker status		
Non-healthcare worker	1759 (38.2%)	1550 (35.0%)
Healthcare worker (<1 patient with COVID-19 per day)	1973 (42.8%)	2008 (45.4%)
Healthcare worker (≥1 patient with COVID-19 per day)	873 (19.0%)	865 (19.6%)
Interval between first and second dose		
Median (IQR) days	74 (42, 89)	76 (49, 89)
<6 weeks	1722 (37.4%)	1673 (37.8%)
6-8 weeks	1215 (26.4%)	1294 (29.3%)
9-11 weeks	555 (12.1%)	497 (11.2%)
≥12 weeks	1113 (24.2%)	959 (21.7%)
First dose strength		
Standard Dose	2979 (64.7%)	2913 (65.9%)
Low Dose	1626 (35.3%)	1510 (34.1%)
NAAT+ case*		
Negative	4398 (95.5%)	4046 (91.5%)
NAAT+ case	207 (4.5%)	377 (8.5%)
Primary symptomatic NAAT+ case†		
Non-case	4534 (98.5%)	4206 (95.1%)
NAAT+ primary symptomatic case	71 (1.5%)	217 (4.9%)
Time from day 28 post second dose until infection		
Median (IQR) days	64 (38.5, 98)	59 (35, 91)
Total follow up time at risk		
Median (IQR) days	102 (77, 128)	97 (73, 122)

Values are counts (percentages) unless stated otherwise. *NAAT+ case denotes nucleic acid amplification test positive participants. †Primary symptomatic NAAT+ denotes nucleic acid amplification test positive participants with at least one qualifying symptom.

Supplementary Table C.2: **Observed anti-spike IgG antibody measurements by visit, arm (ChAdOx1/Control) and case**

		ChAdOx1				Control
		Non-case n=4398 (95.5%)	Case n=207 (4.5%)	Symptomatic case n=71 (1.5%)	Total n=4605	Total n=4423
PB28	Number of antibody measurements available	1155 (26.3%)	173 (83.6%)	59 (83.1%)	1328 (28.8%)	852 (19.3%)
	Median (IQR) anti-spike IgG value BAU/mL	219 (119, 383)	197 (103, 322)	165 (100, 276)	214 (115, 370)	0.3 (<LLOQ, 0.5)
	Number of antibody measurements <LLOQ	0 (0%)	0 (0%)	0 (0%)	0 (0%)	357 (41.9%)
PB90	Number of antibody measurements available	519 (11.8%)	64 (30.9%)	23 (32.4%)	583 (12.7%)	62 (1.4%)
	Median (IQR) anti-spike IgG value BAU/mL	115 (66, 206)	93 (47, 168)	78 (62, 178)	112 (64, 199)	0.3 (<LLOQ, 0.6)
	Number of antibody measurements <LLOQ	0 (0%)	0 (0%)	0 (0%)	0 (0%)	21 (33.9%)
PB182	Number of antibody measurements available	58 (1.3%)	1 (0.5%)	1 (1.4%)	59 (1.3%)	9 (0.2%)
	Median (IQR) anti-spike IgG value BAU/mL	65 (45, 117)	30 (NA)	30 (NA)	64 (44, 117)	0.2 (<LLOQ, 0.3)
	Number of antibody measurements <LLOQ	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4 (44.4%)
Total	Number of antibody measurements available	1732 (39.4%)	238 (115%)	83 (116.9%)	1970 (42.8%)	923 (20.9%)
	Median (IQR) anti-spike IgG value BAU/mL	175 (92, 320)	164 (82, 280)	137 (69, 243)	172 (91, 312)	0.3 (<LLOQ, 0.5)
	Number of antibody measurements <LLOQ	0 (0%)	0 (0%)	0 (0%)	0 (0%)	382 (41.4%)

Only ChAdOx1 vaccinated participants are split by case (Non-case, Case, Symptomatic case). The number of antibody measurements available, median and interquartile range, and number of measurements below the lower limit of quantification (LLOQ) 0.21 BAU/mL is shown. Antibody measurements are given in BAU/mL to the nearest whole number for the ChAdOx1 arm, and to 1 decimal place for the Control arm. Outliers are not included in this table (see Supplementary Table C.3).

Supplementary Table C.3: **The 8 outlying antibody values from ChAdOx1 vaccinated participants which were removed from the analysis**

Case	Days since second dose	Visit	Anti-spike IgG value (BAU/mL)
Non-case	40	PB28	<LLOQ
Non-case	34	PB28	0.6
Non-case	27	PB28	0.3
Non-case	89	PB90	<LLOQ
Non-case	103	PB90	4363.8
Symptomatic case	95	PB90	3.3
Non-case	91	PB90	0.7
Non-case	91	PB90	<LLOQ

Values less than the lower limit of quantification 0.21 BAU/mL are written as < LLOQ. All except the 5th value were removed due to being extremely low, which may lead to issues in the model fit, encouraging the variance of random intercepts and slopes to be too large (See Supplementary Fig. C.2)

Supplementary Table C.4: **Posterior summary of parameters in the longitudinal antibody model, for antibody measurements converted to BAU/mL**

Parameter	Mean	Median	0.025 quantile	0.975 quantile
α_0	5.761	5.762	5.513	6.006
α_1	−4.770	−4.770	−5.032	−4.506
σ_e	0.206	0.206	0.184	0.226
τ_0	0.756	0.756	0.722	0.791
τ_1	0.162	0.170	0.014	0.287
ρ	−0.134	−0.103	−0.767	0.269

Time $t = 0$ is set to be 28 days after the second dose in the model implementation. A reference individual in the model is one with age 18-55, male, white, no comorbidity, $\geq$ 12 weeks interval between first and second dose, BMI $< 30\text{kg/m}^2$, not a healthcare worker, standard first dose, negative COVID-19 and 0 Nelson–Aalen estimate.

Supplementary Table C.5: **The multiplicative covariate effects on anti-spike IgG level 28 days post second dose**

	Posterior estimates			
Multiplicative effect on anti-spike IgG level 28 days post second dose	Median	Mean	2.5% quantile	97.5% quantile
Age (56-69 vs 18-55 years)	0.86	0.86	0.71	1.03
Age (≥70 vs 18-55 years)	0.78	0.78	0.63	0.96
Sex (Female vs Male)	1.08	1.08	0.99	1.19
Ethnicity (Non-white vs White)	1.18	1.18	1.00	1.39
Comorbidity (Comorbidity vs None)	1.00	1.00	0.90	1.11
BMI (≥30 vs <30 kg/m^2)	1.00	1.00	0.90	1.12
Healthcare worker (HCW) status (HCW facing <1 COVID-19 patient per day vs Not a HCW)	0.84	0.84	0.73	0.95
Healthcare worker (HCW) status (HCW worker facing ≥1 COVID-19 patient per day vs Not a HCW)	0.90	0.90	0.77	1.06
Interval between first and second dose (9-11 weeks vs ≥12 weeks)	0.83	0.83	0.74	0.93
Interval between first and second dose (6-8 weeks vs ≥12 weeks)	0.70	0.70	0.61	0.81
Interval between first and second dose (<6 weeks vs ≥12 weeks)	0.51	0.52	0.43	0.61
First dose (Low dose vs Standard dose)	1.11	1.11	0.99	1.24
COVID-19 outcome (Positive vs Negative)	0.85	0.86	0.71	1.01
COVID-19 outcome (Primary symptomatic vs Negative)	0.76	0.76	0.61	0.95
Cumulative hazard (Nelson–Aalen estimator for cumulative hazard)	0.97	0.97	0.92	1.03

A forest plot is given in Fig. 4.4

Supplementary Table C.6: **The multiplicative covariate effects on half-life of the decay in anti-spike IgG levels**

	Posterior estimates			
Multiplicative effect on anti-spike IgG level half-life	Median	Mean	2.5% quantile	97.5% quantile
Age (56-69 vs 18-55 years)	1.22	1.25	0.89	1.74
Age (≥70 vs 18-55 years)	0.86	0.86	0.62	1.15
Sex (Female vs Male)	0.90	0.90	0.81	0.99
Ethnicity (Non-white vs White)	1.09	1.09	0.90	1.34
Comorbidity (Comorbidity vs None)	0.92	0.92	0.82	1.04
BMI (≥30 vs <30 kg/m^2)	0.84	0.84	0.74	0.94
Healthcare worker (HCW) status (HCW facing <1 COVID-19 patient per day vs Not a HCW)	0.96	0.96	0.83	1.11
Healthcare worker (HCW) status (HCW worker facing ≥1 COVID-19 patient per day vs Not a HCW)	0.88	0.88	0.76	1.03
Interval between first and second dose (9-11 weeks vs ≥12 weeks)	0.99	0.99	0.88	1.11
Interval between first and second dose (6-8 weeks vs ≥12 weeks)	0.97	0.97	0.84	1.12
Interval between first and second dose (<6 weeks vs ≥12 weeks)	0.96	0.98	0.75	1.27
First dose (Low dose vs Standard dose)	1.09	1.09	0.97	1.23
COVID-19 outcome (Positive vs Negative)	0.86	0.86	0.72	1.04
COVID-19 outcome (Primary symptomatic vs Negative)	0.78	0.78	0.64	0.97
Cumulative hazard (Nelson–Aalen estimator for cumulative hazard)	1.01	1.01	0.96	1.07

A forest plot is given in Fig. 4.4

Supplementary Table C.7: **The multiplicative covariate effects on hazard against symptomatic COVID-19 infection and any COVID-19 infection**

Covariate multiplicative effect	Posterior estimates			
Multiplicative effect on hazard for symptomatic COVID-19 infection	Median	Mean	2.5% quantile	97.5% quantile
Antibody level (Effect due to increase of 100 BAU/mL)	0.34	0.32	0.08	0.76
Vaccination direct effect (ChAdOx1 nCoV-19 vs control)	0.71	0.73	0.38	1.53
Age (56-69 vs 18-55 years)	0.50	0.50	0.29	0.89
Age (≥70 vs 18-55 years)	0.16	0.16	0.06	0.42
Sex (Female vs Male)	0.88	0.88	0.68	1.13
Ethnicity (Non-white vs White)	1.02	1.02	0.67	1.56
Comorbidity (Comorbidity vs None)	0.98	0.98	0.73	1.32
BMI (≥30 vs <30)	1.39	1.39	1.06	1.82
Healthcare worker (HCW) status (HCW facing <1 COVID-19 patient per day vs Not a HCW)	1.42	1.42	1.00	2.02
Healthcare worker (HCW) status (HCW worker facing ≥1 COVID-19 patient per day vs Not a HCW)	2.20	2.20	1.49	3.24
Multiplicative effect on hazard for any COVID-19 infection	Median	Mean	2.5% quantile	97.5% quantile
Antibody level (Effect due to increase of 100 BAU/mL)	0.49	0.48	0.27	0.75
Vaccination direct effect (ChAdOx1 nCoV-19 vs control)	0.92	0.93	0.64	1.39
Age (56-69 vs 18-55 years)	0.66	0.66	0.46	0.93
Age (≥70 vs 18-55 years)	0.53	0.53	0.36	0.80
Sex (Female vs Male)	0.93	0.93	0.78	1.11
Ethnicity (Non-white vs White)	0.86	0.86	0.62	1.19
Comorbidity (Comorbidity vs None)	0.95	0.95	0.77	1.16
BMI (≥30 vs <30)	1.13	1.13	0.92	1.38
Healthcare worker (HCW) status (HCW facing <1 COVID-19 patient per day vs Not a HCW)	1.18	1.18	0.93	1.50
Healthcare worker (HCW) status (HCW worker facing ≥1 COVID-19 patient per day vs Not a HCW)	1.65	1.65	1.26	2.17

A forest plot is given in Supplementary Fig. C.7

Supplementary Table C.8: **Anti-spike IgG antibody levels required to give different levels of vaccine efficacy against any COVID-19 infection or symptomatic COVID-19**

	Required anti-spike IgG antibody level (BAU/mL)	
Vaccine efficacy (%)	Symptomatic COVID-19	Any COVID-19 infection
0	0 (0, 18)	0 (0, 27)
5	0 (0, 20)	0 (0, 31)
10	0 (0, 23)	3 (0, 36)
15	0 (0, 26)	11 (0, 41)
20	0 (0, 28)	20 (0, 47)
25	0 (0, 32)	29 (0, 54)
30	2 (0, 35)	38 (0, 62)
35	8 (0, 39)	48 (0, 72)
40	16 (0, 44)	59 (19, 85)
45	24 (0, 50)	71 (41, 102)
50	32 (0, 57)	84 (59, 125)
55	42 (0, 67)	99 (73, 154)
60	52 (0, 82)	116 (85, 191)
65	63 (24, 104)	135 (97, 233)
67	69 (37, 116)	143 (102, 253)
70	78 (51, 139)	156 (109, 285)
75	96 (65, 192)	182 (124, 346)
80	117 (76, 266)	214 (141, 423)
85	144 (88, 365)	254 (163, 522)
90	182 (105, 507)	311 (194, 662)
95	247 (133, 757)	408 (247, 900)
98	311 (161, 1002)	506 (300, 1140)
99	397 (197, 1331)	635 (369, 1458)

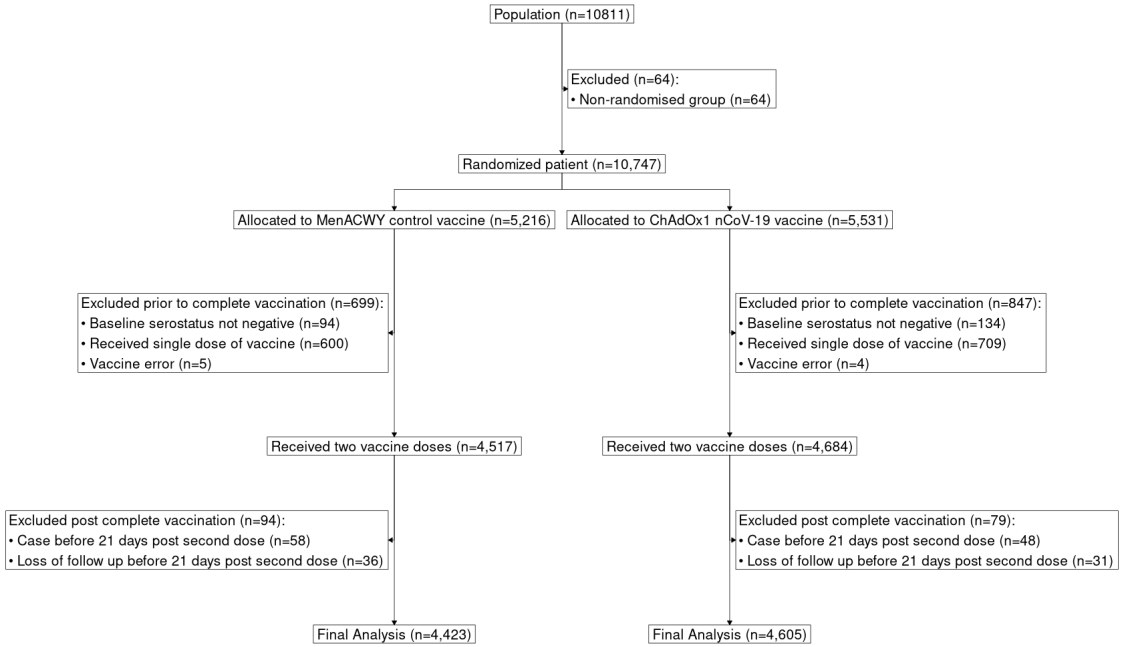
Posterior medians (and 95% credible intervals) are reported. When the required antibody level is predicted to be negative, 0 BAU/mL is reported, as negative antibody levels are not possible.

Supplementary Table C.9: **Comparison of methods and results with previous studies considering waning vaccine efficacy against the Delta, Gamma and Omicron variants**

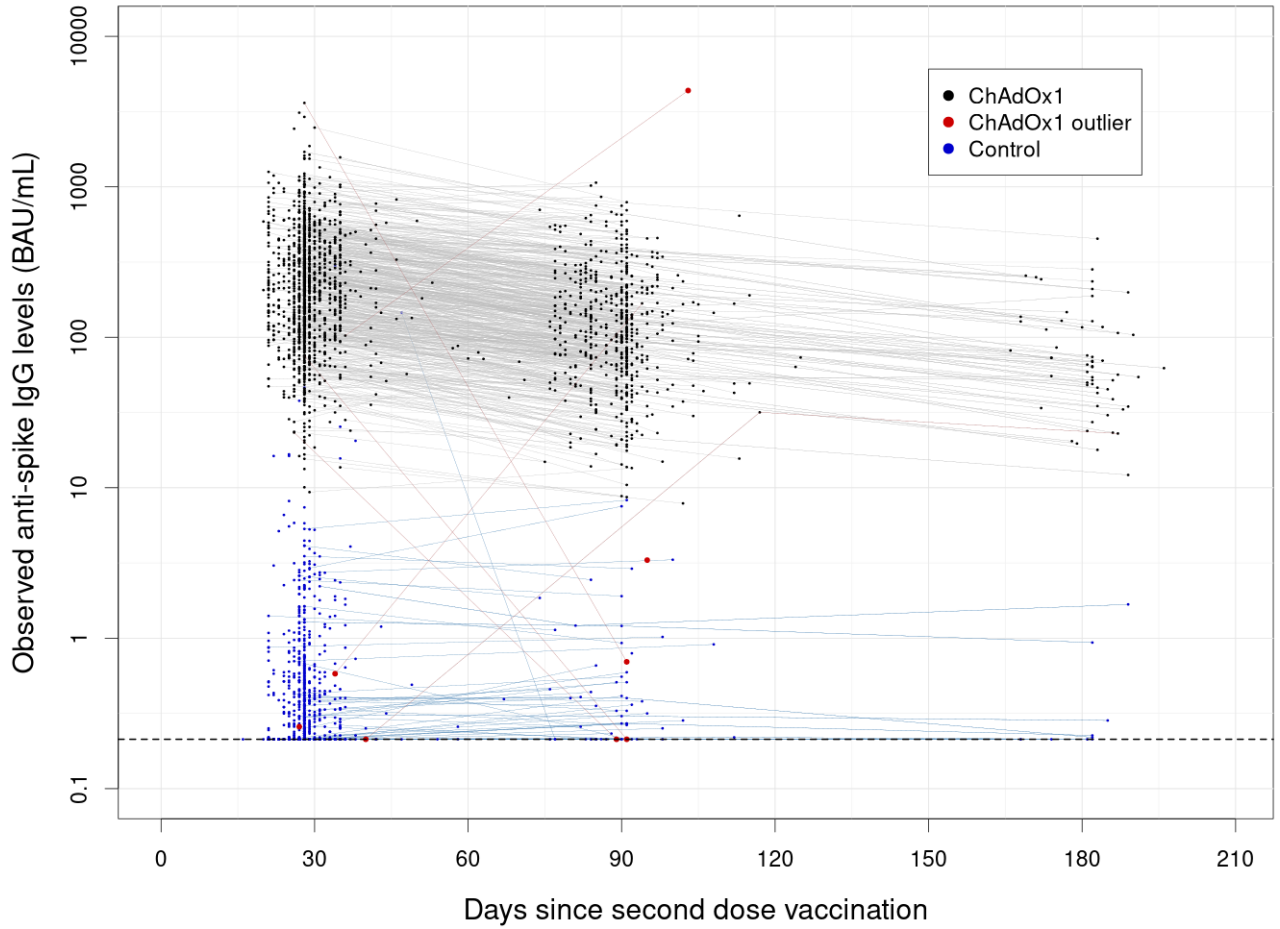
Paper	Primary course vaccine	Booster vaccine	Dominant variant	Two-dose /booster /prior infection	Study design	Randomised vaccine /control assignment	Baseline comparator	Outcome	Location of result in paper	Group	Weeks since dose	Vaccine efficacy (%)
Katikireddi et al. (2022) [95]	ChAdOx1†		Delta Gamma	Two dose	Cohort* Test-negative	x	Unvaccinated	Severe disease* Symptomatic	Table 3	Scotland (Delta) Brazil (Gamma)	4-5 10-11 18-19 4-5 10-11 18-19	67.3 (65.3, 69.1) 59.3 (57.2, 61.4) 44.6 (41.5, 47.6) 68.4 (67.8, 68.9) 63.2 (62.2, 64.2) 57.7 (55.4, 60.0)
Menni et al. (2022) [96]	ChAdOx1† BNT162b2* mRNA-1273*	BNT162b2* mRNA-1273*	Delta	Two dose Booster dose*	Cohort	x	Unvaccinated	Any infection (mostly symptomatic)	(Figure 1A) Supp. Table 1	All ages	4 9 13 17 22 26	83.1 (82.2, 84) 79.3 (78.6, 80.0) 76.7 (76.0, 77.5) 75.8 (75.0, 76.4) 75.7 (74.9, 76.4) 75.2 (74.3, 76.1)
Horne et al. (2022) [97]	ChAdOx1† BNT162b2*		Delta	Two dose	Cohort	x	Unvaccinated	Any infection	Supp. Table 7	Age 40-64 Age 65+	3-6 11-14 23-26 3-6 11-14 23-26	22 (19, 25) −25 (−30, −20) −86 (−79, −93) 57 (42, 68) 30 (20, 38) −23 (−36, −11)
Andrews et al. (2022b) [98]	ChAdOx1† BNT162b2*	BNT162b2 mRNA-1273 ChAdOx1†*	Omicron Delta*	Two dose Booster dose	Test-negative	x	Unvaccinated	Symptomatic	Table 3	Two dose BNT162b2 booster mRNA-1273 booster	2-4 5-9 15-19 25+ 2-4 5-9 ≥10 2-4 5-9	48.9 (39.2, 57.1) 33.7 (25.0, 41.5) 17.8 (13.4, 21.9) −2.7 (−4.2, −1.2) 62.4 (61.8, 63.0) 52.9 (52.1, 53.7) 39.6 (38.0, 41.1) 70.1 (69.5, 70.7) 60.9 (59.7, 62.1)
Hogan et al. (2023) [142]‡	ChAdOx1† BNT162b2* mRNA-1273*	BNT162b2* mRNA-1273* ChAdOx1†*	Omicron Delta*	Two dose Booster dose*	Modelling	x	Unvaccinated	Mild disease	Table 2	All recipients	13 26	12 (11.5, 12.4) 6.1 (5.8, 6.4)§
Our Omicron analysis¶	ChAdOx1†		Omicron BA.4/5	Two dose	Modelling	x	Vaccine non-responders	Any infection	Supp. Fig. C.11	All recipients (Scenario: Mean efficacy)	5 10 20 26 29.9	18.5 (17.8, 19.2) 13.8 (13.3, 14.5) 7.6 (7.1, 8.2) 5.3 (4.8, 5.8) 4.1 (3.7, 4.7)

We include studies of two doses of the ChAdOx1 nCoV-19 vaccine, and a study of efficacy after a booster dose, against the Delta, Gamma, and Omicron COVID-19 variants. Results are given with 95% confidence intervals or 95% credible intervals, depending on the paper. Results are presented for different age groups, where appropriate. *We do not report these results here, see the original paper. †ChAdOx1 refers to the ChAdOx1 nCoV-19 vaccine in this table. ‡Hogan et al. (2023) [142] fit a decay model to the results from Andrews et al. (2022b) [98]. The model does not appear to fit the data well - see Fig. 2 (top panel, mild disease) from their paper. §These results at 26 weeks were extrapolated from the model for timepoints beyond the data. ¶Our analysis predicts efficacy against Omicron by applying the relationship between anti-spike IgG levels and relative efficacy reported by Wei et al. [82] to the antibody trajectories predicted by our model.

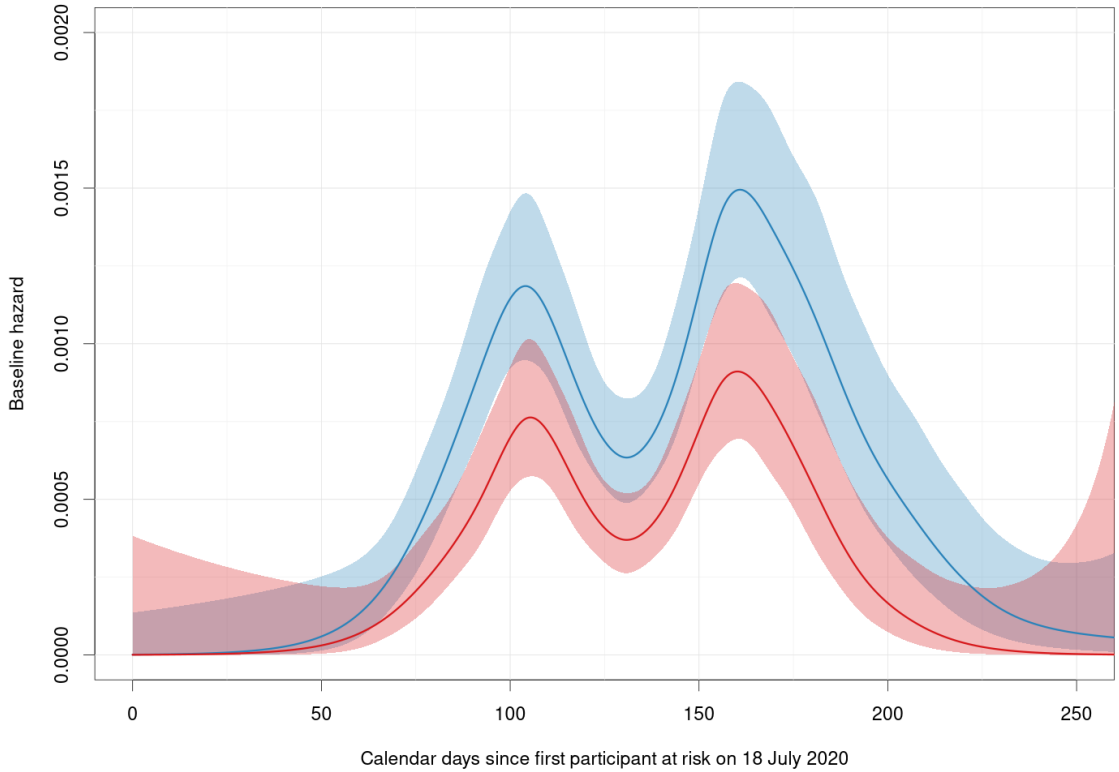
C.5 Supplementary Figures



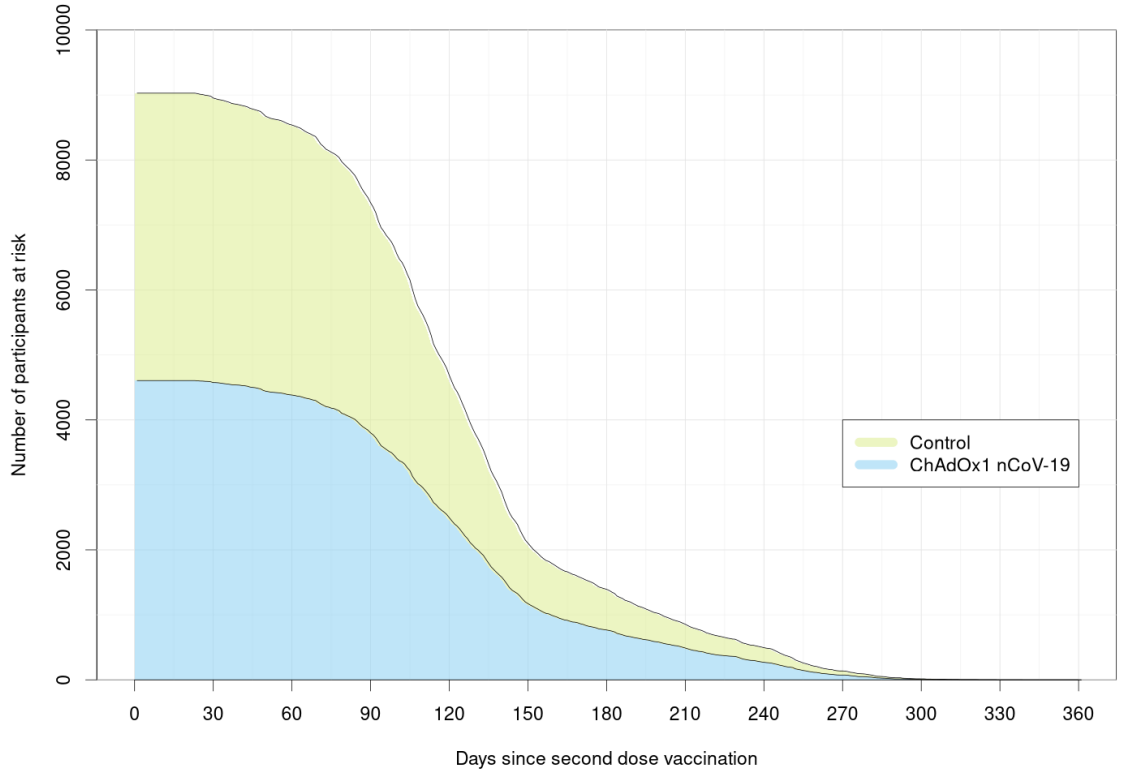
Supplementary Figure C.1: **Flow chart showing inclusion of COV002 participants in analysis cohort.**



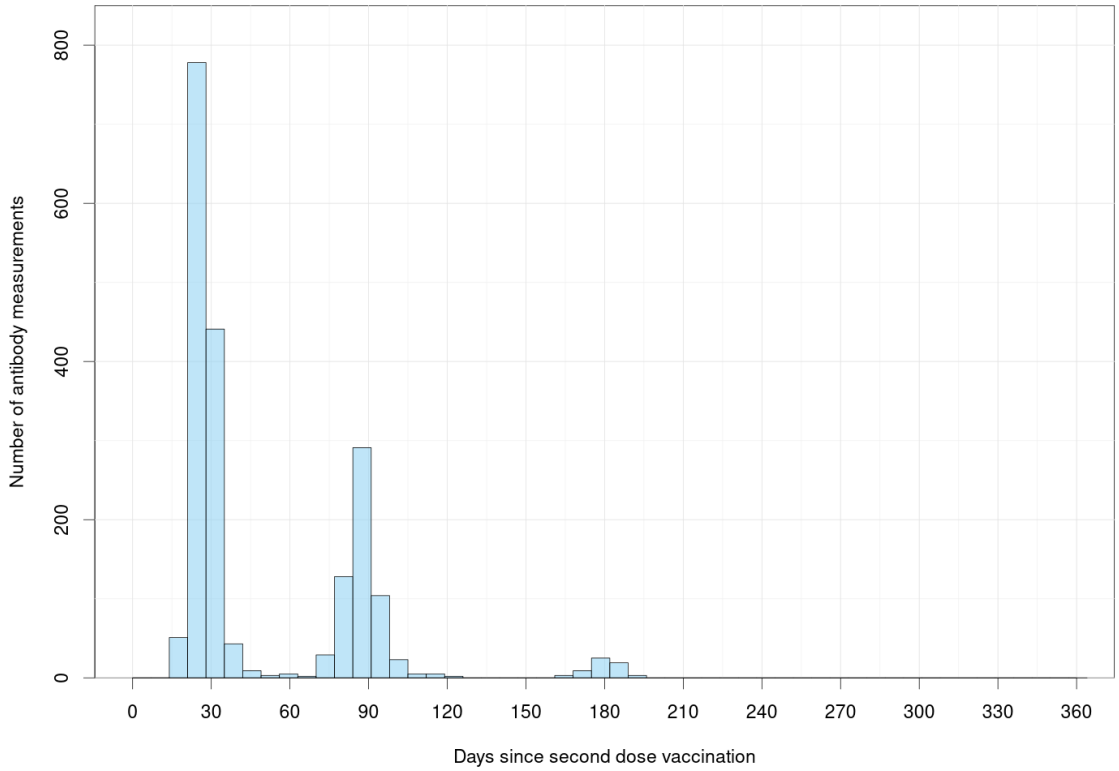
Supplementary Figure C.2: **Observed anti-spike IgG antibody levels over time since vaccination, including outliers and measurements from control participants.** Each point represents an observation, and consecutive observations from the same individual are connected by a line. Black points denote included observations from vaccine arm participants. Red points denote observations excluded as outliers from vaccine arm participants. Blue points denote observations from control arm participants, which were not included in our model. When one of the points which a line connects is an outlier, a red line is used. Blue lines connect observations from control participants. A dotted horizontal line is drawn at the lower limit of quantification (LLOQ) 33 AU/mL (0.21 BAU/mL). Observations below the LLOQ are plotted at the LLOQ.



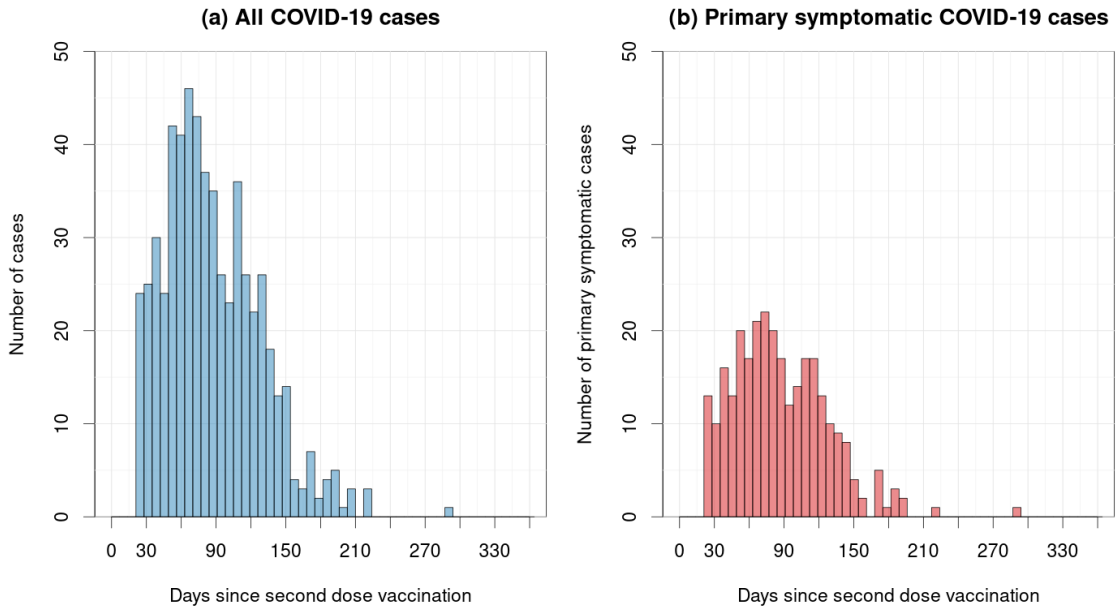
Supplementary Figure C.3: **The estimated baseline hazard of infection in the trial by calendar days.** The (lower) red line and shaded region give the estimate and 95% confidence interval for baseline hazard of primary symptomatic COVID-19 infection. The (upper) blue line and shaded region give the estimate and 95% confidence interval for baseline hazard of any COVID-19 infection. The shaded regions overlap. Both were estimated from the control group only, using penalised restricted cubic splines in the R package survPen [143]. The baseline hazard is the hazard rate for an unvaccinated individual with baseline covariates. Few individuals had yet entered the study prior to 50 days since the start, and few remained in the study beyond 250 days since the start, hence the wide confidence intervals in those regions.



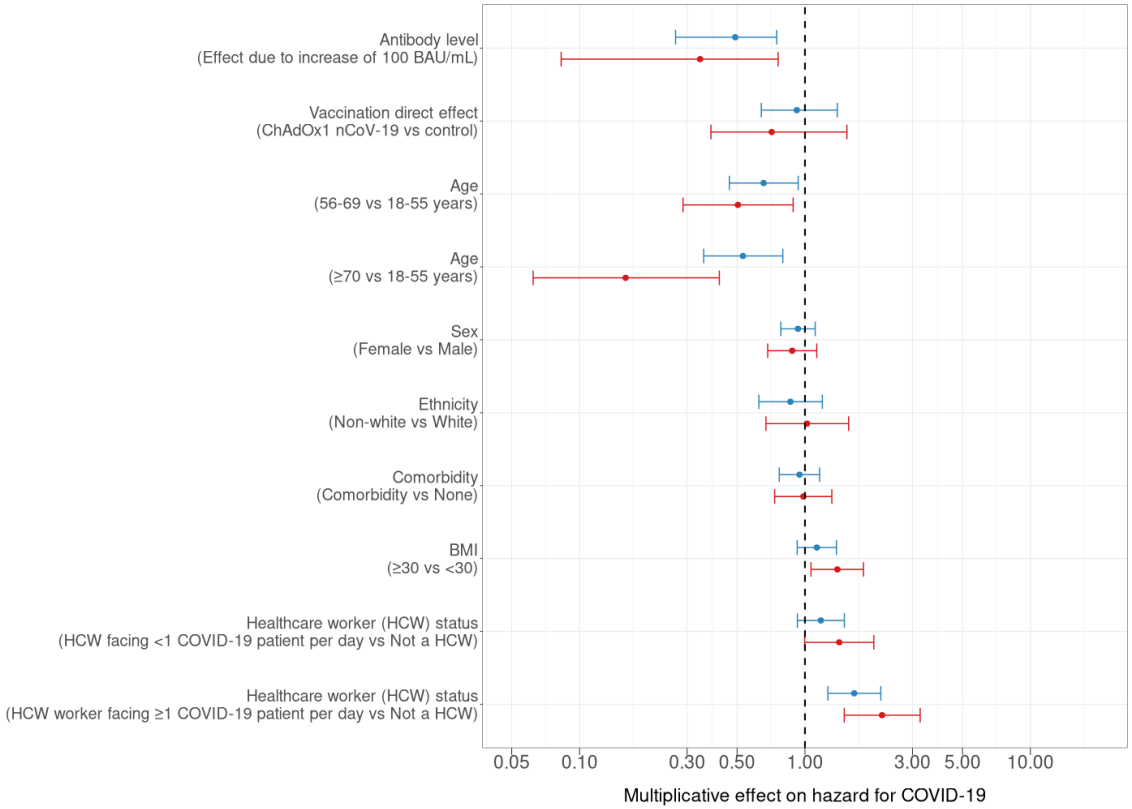
Supplementary Figure C.4: **The number of individuals at risk over time since second dose in the vaccine and control arms.** The upper green area denotes the number of ChAdOx1 nCoV-19 (AZD1222) vaccinated participants at risk over time since second dose. The lower blue area denotes the number of control participants at risk over time since second dose. As such, the upper black line denotes the total number of participants at risk over time since second dose, and the lower black line the number of control participants at risk over time. Participants are no longer considered at-risk after leaving the study, returning a positive COVID-19 test, and receiving a subsequent vaccination, among other reasons.



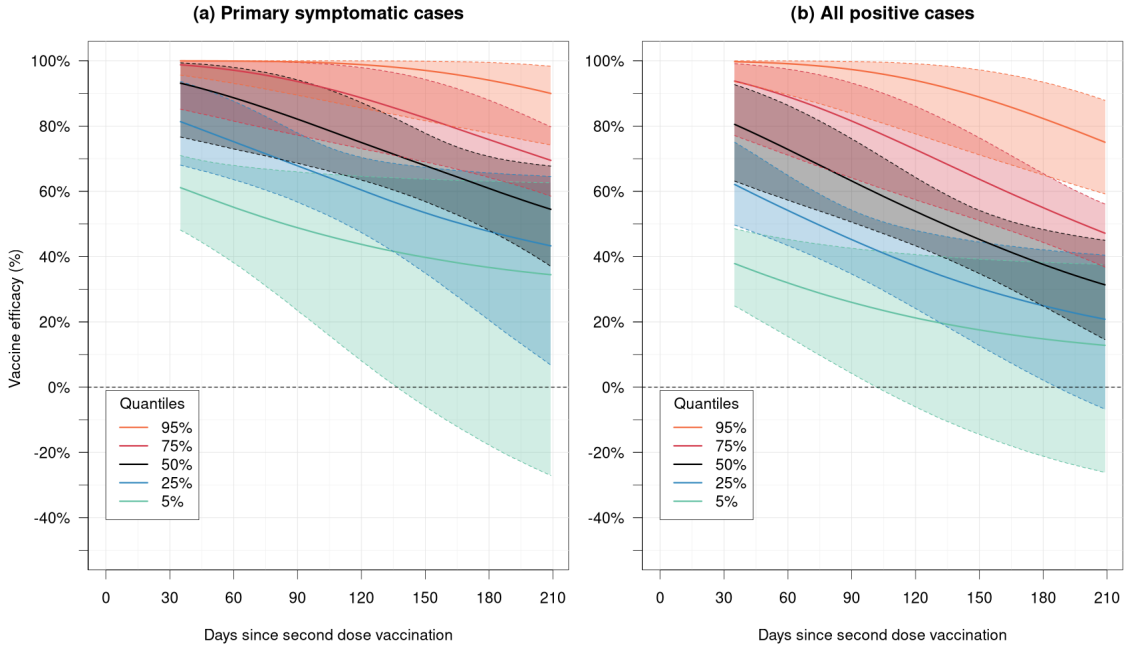
Supplementary Figure C.5: **Histogram of the timings of anti-spike IgG samples since second dose vaccination.** Only anti-spike IgG samples from ChAdOx1 nCoV-19 (AZD1222) vaccinated participants are included. Samples from individuals more than 7 days prior to entering the at-risk period or less than 7 days before leaving the at-risk period are excluded. Hence, samples less than 7 days before returning a positive COVID-19 test, or after returning a positive test, are excluded.



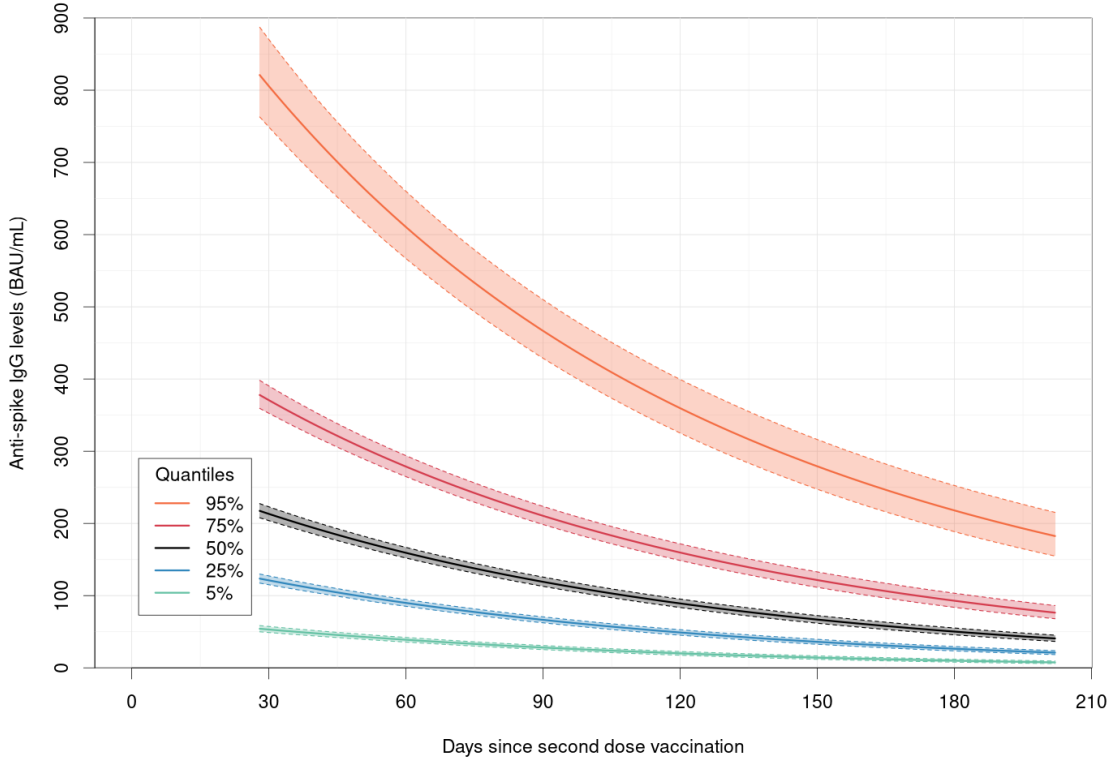
Supplementary Figure C.6: **Histogram of the onset time since second dose of (a) any COVID-19 infection, (b) primary symptomatic COVID-19 infection.** (a) The blue histogram denotes the timings of the onset of any COVID-19 infection. (b) The red histogram denotes the timings of the onset of primary symptomatic COVID-19 infection.



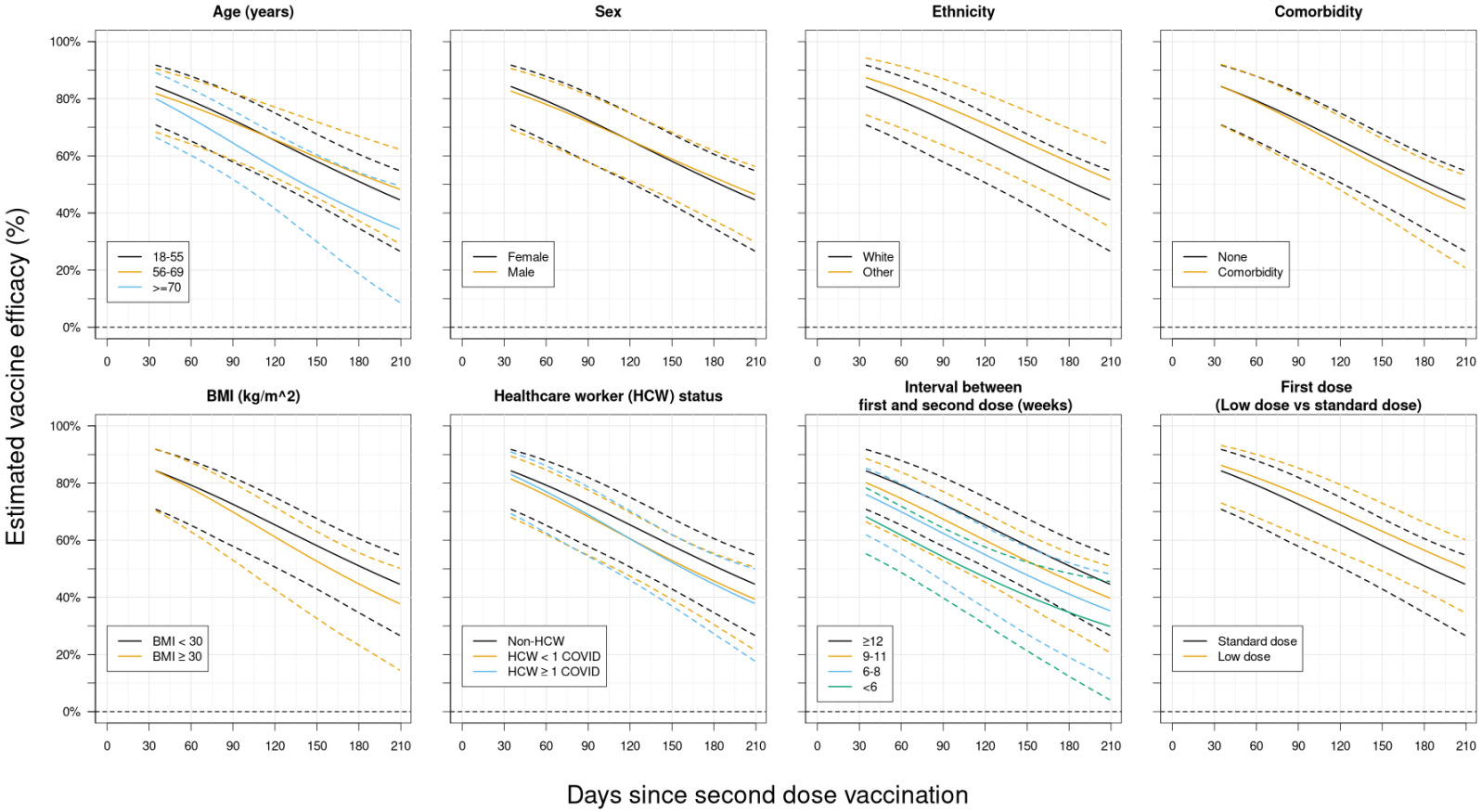
Supplementary Figure C.7: **Covariate effects on hazard of COVID-19 infection.** The multiplicative effects due to different covariates on hazard of symptomatic COVID-19 infection and any COVID-19 infection. Points denote posterior medians and lines denote 95% credible intervals. Red denotes hazard against symptomatic infection and blue denotes hazard against any COVID-19 infection. These results are also given in Supplementary Table C.7



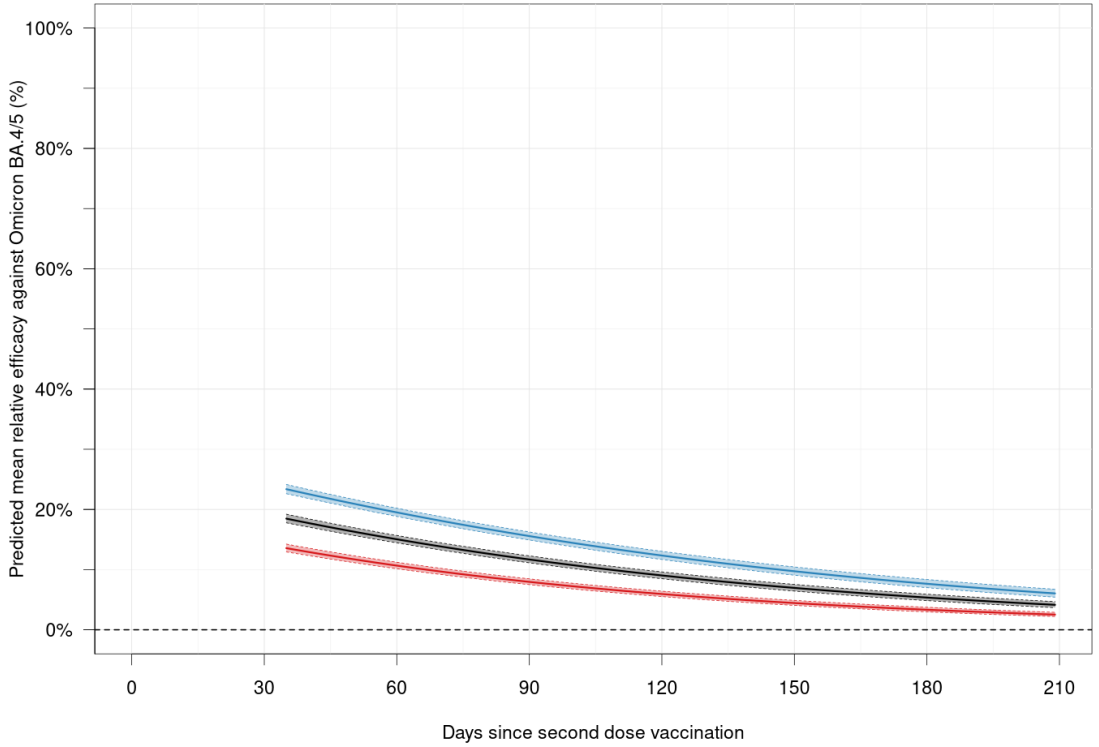
Supplementary Figure C.8: **Quantiles of vaccine efficacy as a function of time since vaccination.** The curves show posterior medians and shaded regions 95% credible intervals. The 95%, 75%, 50%, 25% and 5% quantiles are plotted. (a) shows VE against symptomatic COVID-19 and (b) against all COVID-19 infection.



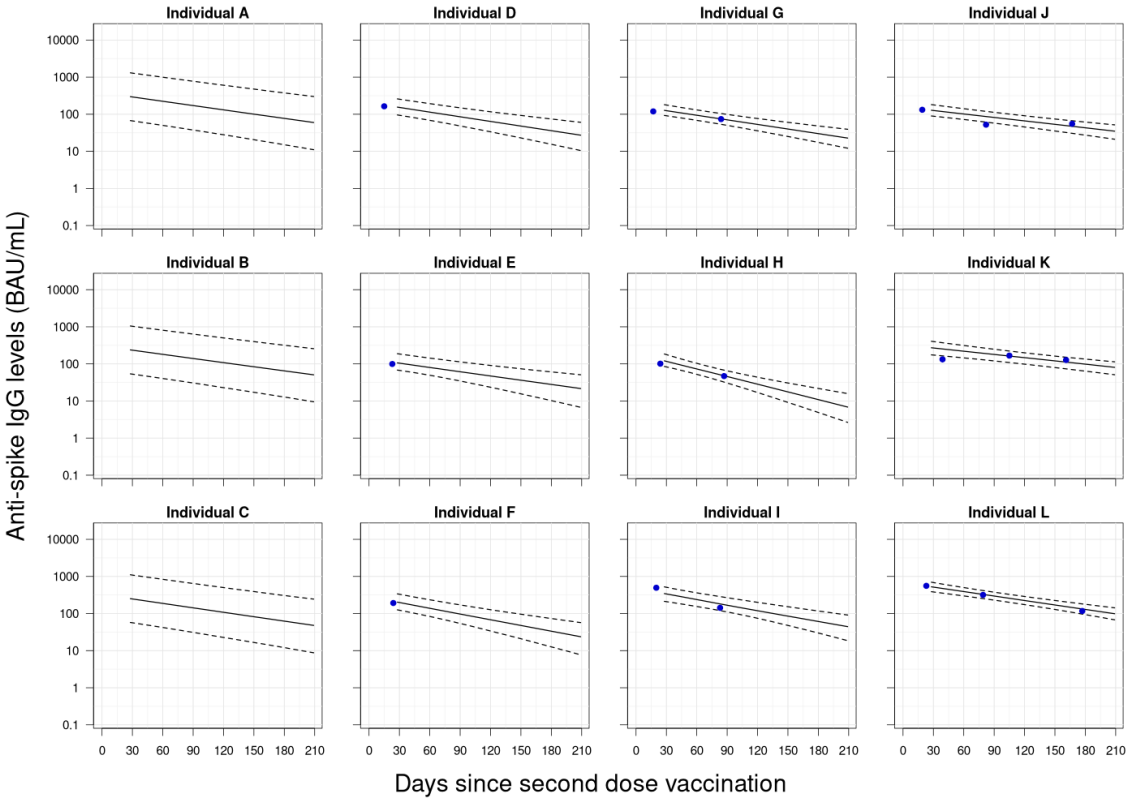
Supplementary Figure C.9: **Quantiles of anti-spike IgG antibody levels as a function of time since vaccination.** The curves show posterior medians and shaded regions 95% credible intervals. The 95%, 75%, 50%, 25% and 5% quantiles are plotted.



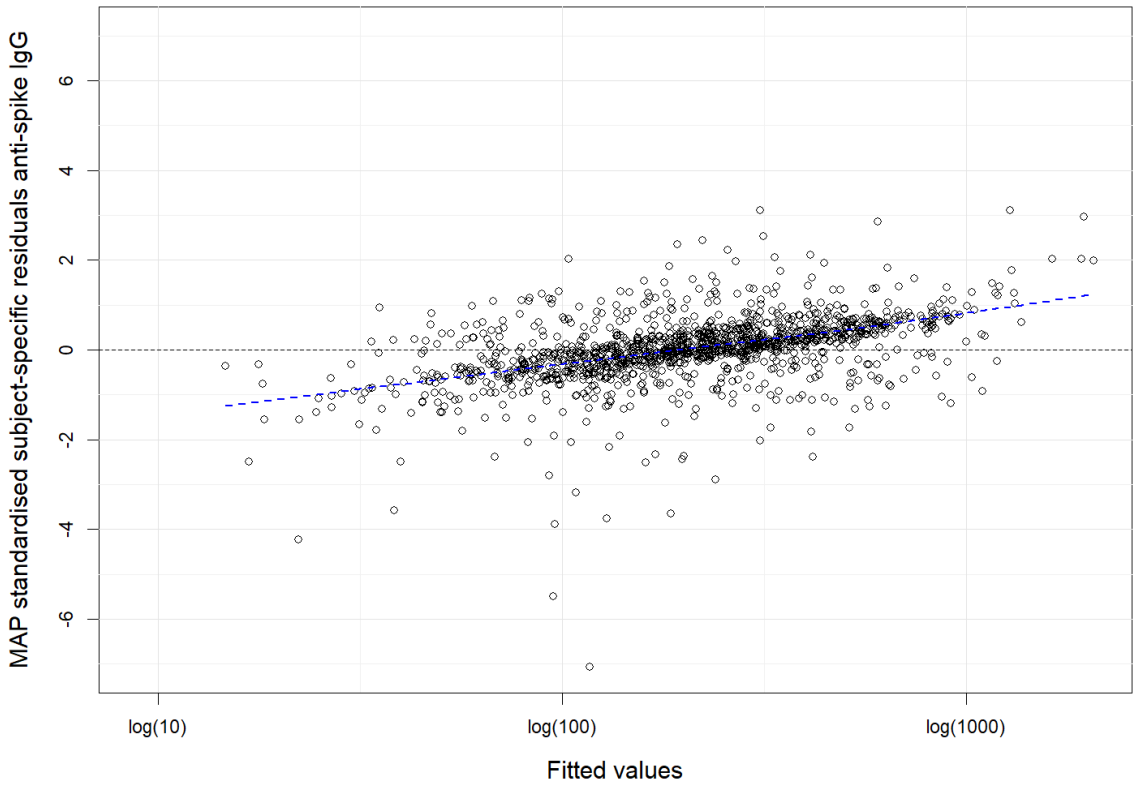
Supplementary Figure C.10: **Vaccine efficacy against any COVID-19 infection plotted against time since vaccination for different covariate combinations.** The black lines represent an individual with "baseline" covariates: age 18-55, female, white ethnicity, no comorbidities, BMI<30 kg/m^2, >12 week interval between first and second dose, non-HCW, and receiving a standard dose as their first dose. All other lines represent an individual with one of the above covariates changed from these baseline values. The solid lines show posterior medians and dashed lines 95% credible intervals. "HCW < 1 COVID" refers to a healthcare worker facing <1 COVID-19 patient per day, "HCW ≥ 1 COVID" refers to a healthcare worker facing ≥1 COVID-19 patient per day.



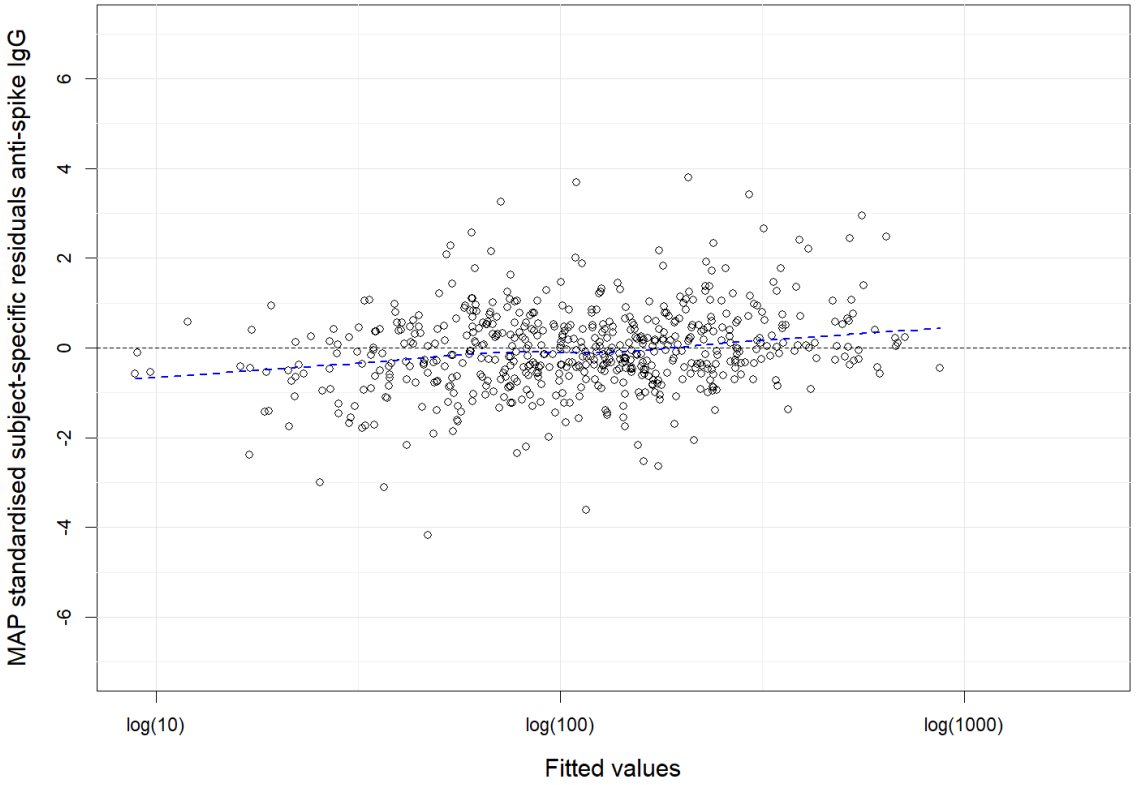
Supplementary Figure C.11: **Predicted relative efficacy against SARS-CoV-2 Omicron BA.4/5 variant infection as a function of time since vaccination.** The black, blue and red lines represent estimates calculated using the reported posterior mean, and upper and lower 95% credible bounds for the relationship between relative efficacy and anti-spike IgG antibody levels from Wei et al. (2023). The solid lines show posterior medians and dashed lines 95% credible intervals, for the mean relative efficacy in the study.



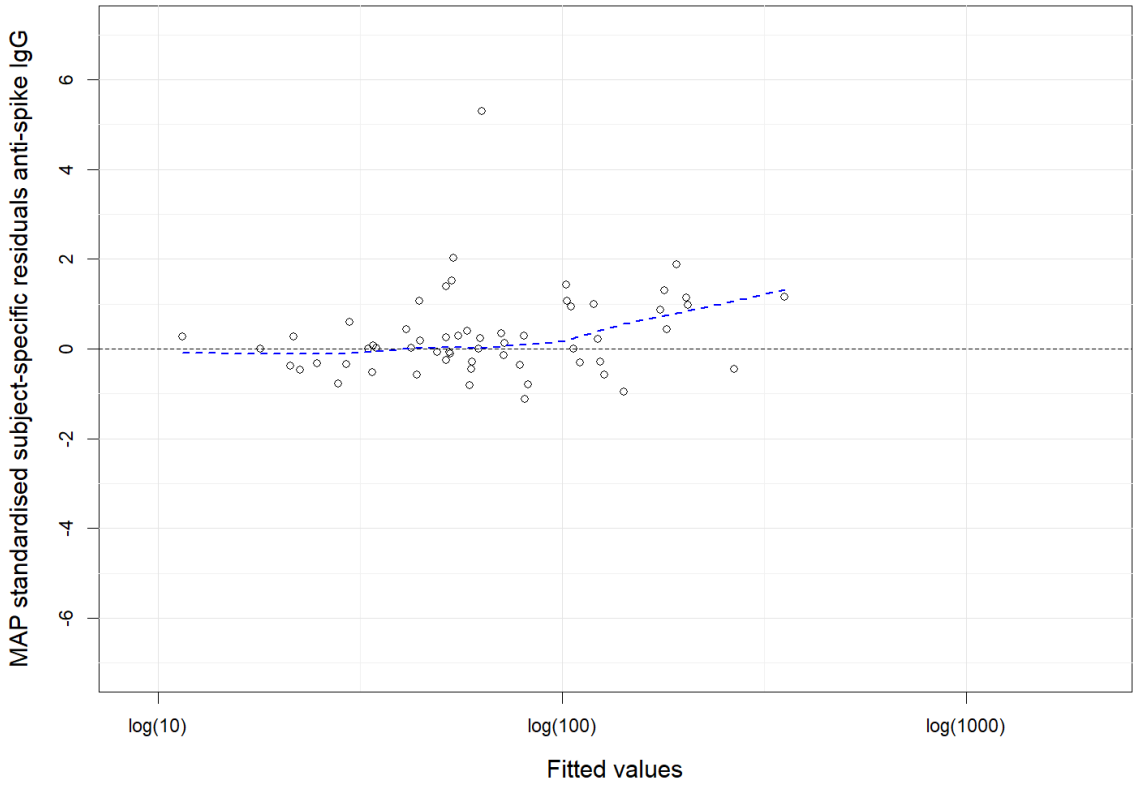
Supplementary Figure C.12: **Estimated true anti-spike IgG levels for 12 randomly chosen individuals.** We randomly chose 3 individuals with 0, 1, 2 and 3 antibody observations available respectively, and plotted their estimated true anti-spike IgG levels over time. The solid line represents the posterior median, and the dashed lines represent a 95% credible interval. The blue points show the observed antibody levels. To be clear, the 3 plots on each row are 3 separate individuals (not the same individual with data consecutively being added).



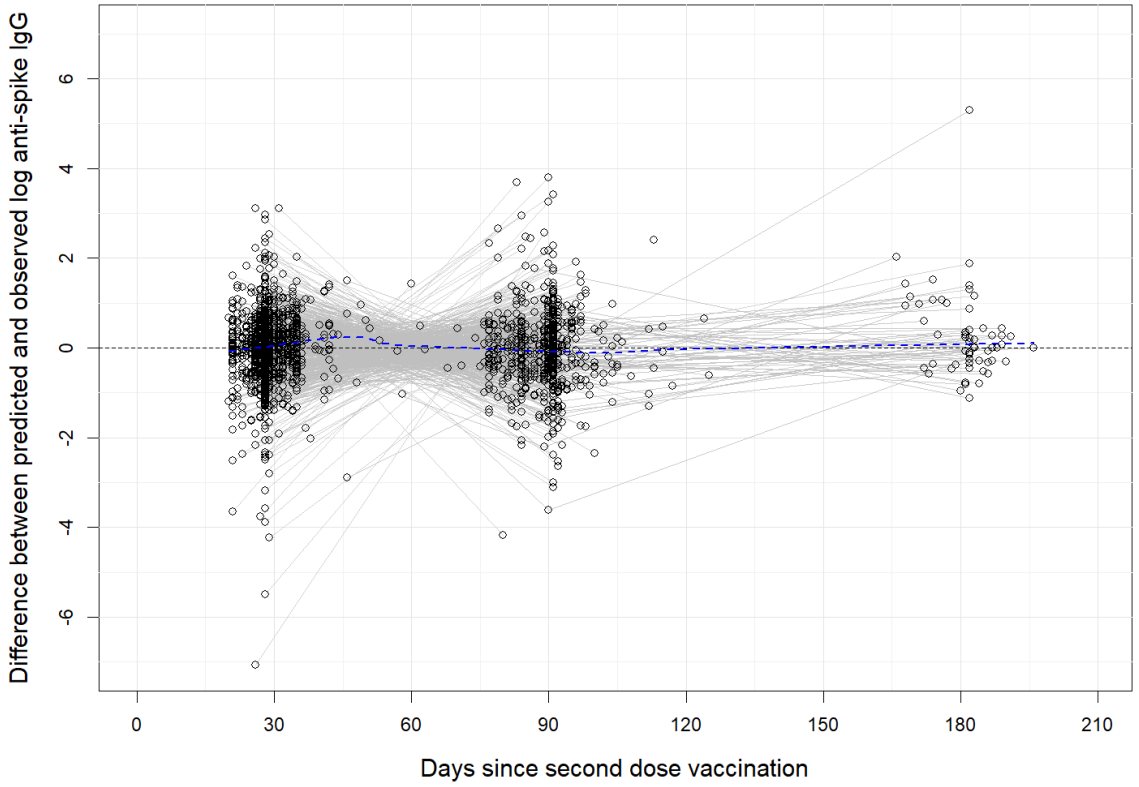
Supplementary Figure C.13: **MAP standardised subject-specific residuals for the PB28 visit, plotted against posterior predicted mean** The y-axis value gives the MAP standardised subject-specific residuals, defined in section 4.3.2. These are plotted against the mean among replicated data drawn from the posterior predictive for an observation for the same individual at the time of their PB28 visit. The blue dotted line shows a loess curve fitted to the points. The plot shows a strong positive linear trend, indicating the posterior mean pools strongly between the individual observed values and the sample mean of the observed values. This is unsurprising, since for most individuals with a PB28 observation, this is their only observation available in the study, and the model contains a random intercept and random slope.



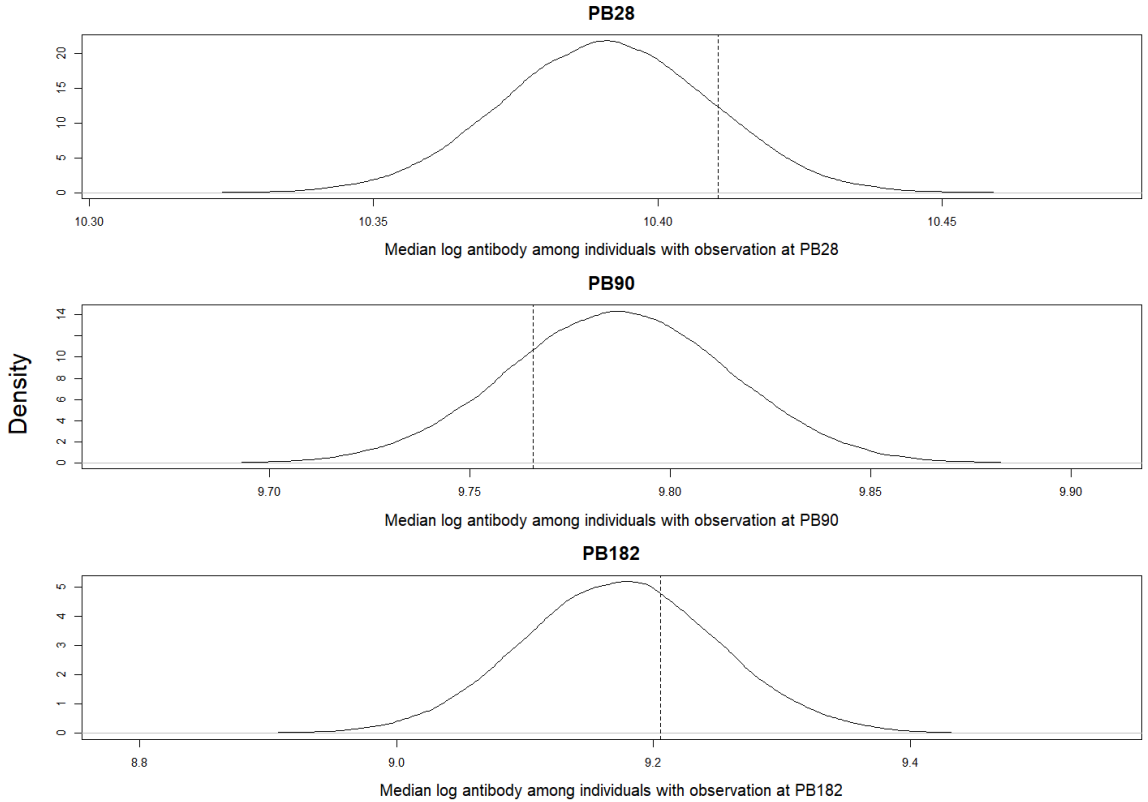
Supplementary Figure C.14: **MAP standardised subject-specific residuals for the PB90 visit, plotted against posterior predicted mean** The y-axis value gives the MAP standardised subject-specific residuals, defined in section 4.3.2. These are plotted against the mean among replicated data drawn from the posterior predictive for an observation for the same individual at the time of their PB28 visit. The blue dotted line shows a loess curve fitted to the points. The plot shows a weak positive linear trend, indicating the posterior mean pools weakly between the individual observed values and the sample mean of the observed values. This is unsurprising, since for most individuals with a PB90 observation, two observations are available in the study (PB28 and PB90), and the model contains a random intercept and random slope.



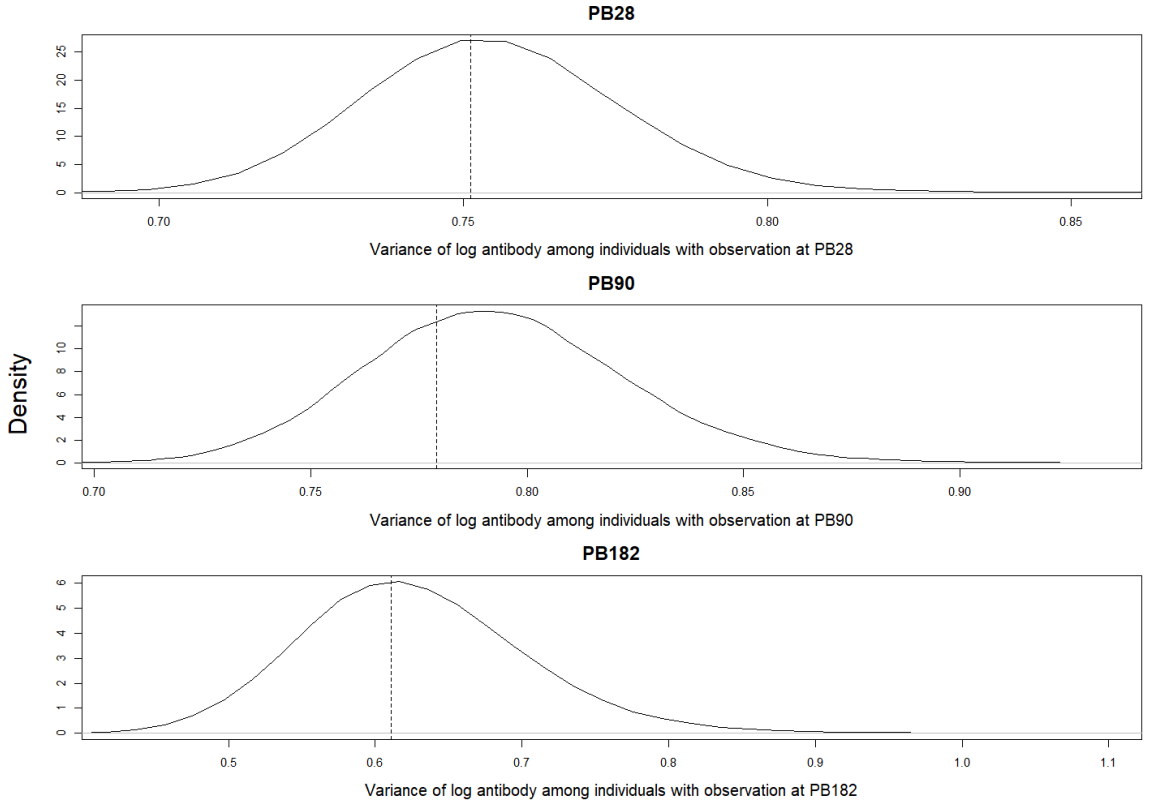
Supplementary Figure C.15: **MAP standardised subject-specific residuals for the PB182 visit, plotted against posterior predicted mean** The y-axis value gives the MAP standardised subject-specific residuals, defined in section 4.3.2. These are plotted against the mean among replicated data drawn from the posterior predictive for an observation for the same individual at the time of their PB28 visit. The blue dotted line shows a loess curve fitted to the points. The plot shows a slight positive trend, indicating the posterior mean may pool weakly between the individual observed values and the sample mean of the observed values. This is unsurprising, since for most individuals with a PB90 observation, three observations are available in the study (PB28, PB90, PB182), and the model contains a random intercept and random slope. This is sparse data to fit two random effects, so observing some pooling is unsurprising.



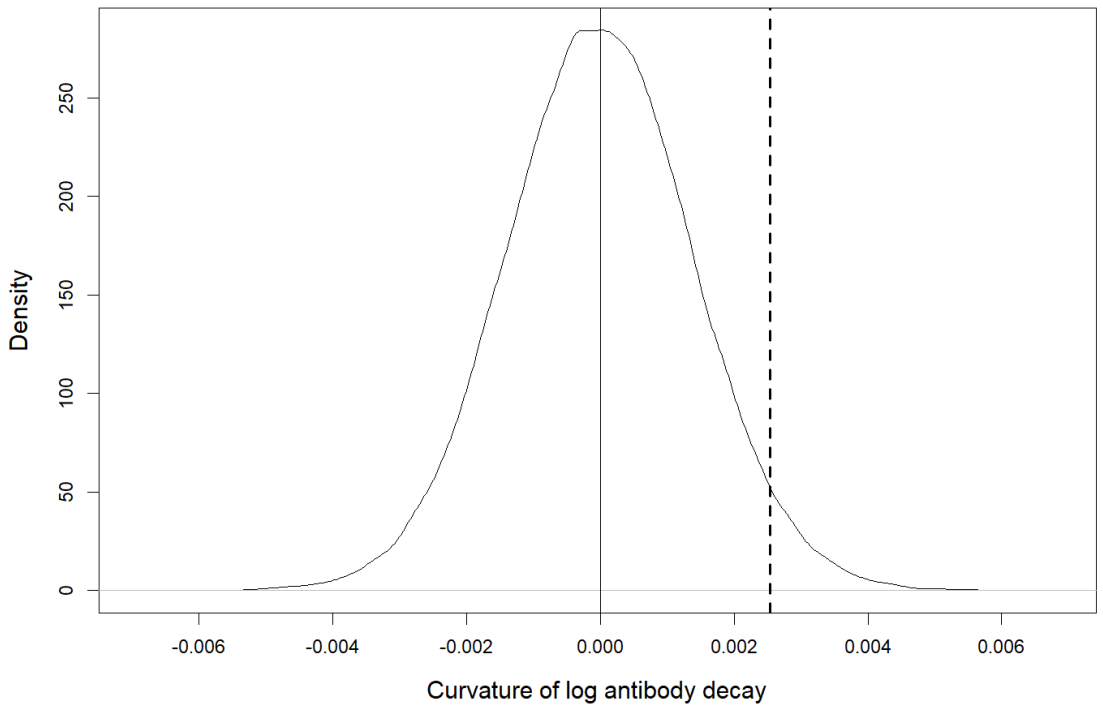
Supplementary Figure C.16: **MAP standardised subject-specific residuals against days since second dose vaccination** The y-axis value gives the MAP standardised subject-specific residuals, defined in section 4.3.2. These are plotted against days since vaccination. The grey lines join two consecutive points for the same individual. The blue dotted line shows a loess curve fitted to the points. The residuals are centered around 0 over time, indicating the model is giving an adequate fit to the data.



Supplementary Figure C.17: **Posterior predictive distribution of the median antibody level at each visit, compared with the observed value** The curve shows the density of the posterior predictive distribution of the median antibody level among individuals observed at each visit. The vertical dotted line shows the observed median antibody level at each visit. The lines each lie towards the bulk of the distribution, indicating a reasonable fit. However, the observed median at PB28 is lower than the posterior mean, and the median at PB90 higher, indicating the model may be overestimating the antibody levels around 28 days after vaccination, and underestimating the levels around 90 days after vaccination. The posterior predictive probability of the median being larger than the observed value at PB28 is 0.15, and smaller than the observed value at PB90 is 0.22.

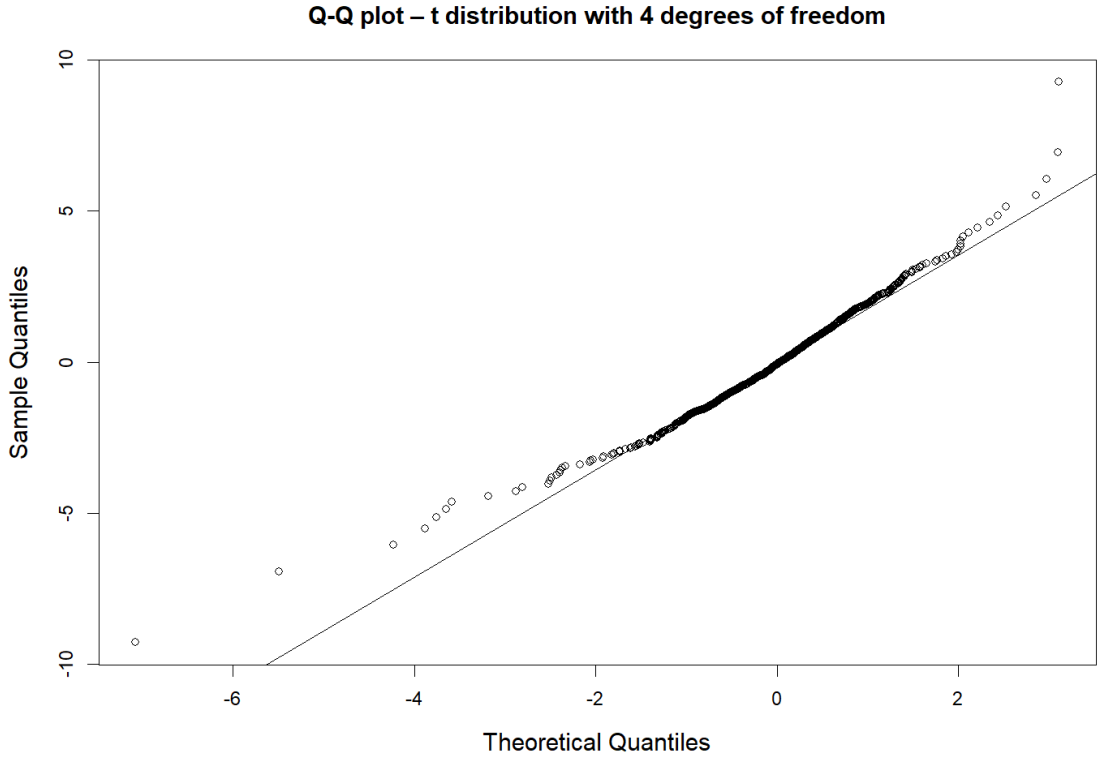


Supplementary Figure C.18: **Posterior predictive distribution for the variance of the antibody level at each visit, compared with the observed value** The curve shows the density of the posterior predictive distribution for the variance of the antibody level among individuals observed at each visit. The vertical dotted line shows the observed variance of the antibody level at each visit. The lines each lie close to the peak of the distribution, indicating a good fit.



Supplementary Figure C.19: **Posterior predictive distribution for the mean antibody curvature between visits, compared with the observed value**

The curve shows the density of the posterior predictive distribution for the mean log antibody curvature among individuals with observations at all three visits. Curvature is defined in section 4.3.2. The vertical dotted line lies towards the tail of the distribution, indicating the model is not capturing the observed curvature in the data well. The model assumes log antibodies decay linearly, whereas the observed data appear to indicate the decay may be faster at first, before slowing down. The posterior predictive probability of the curvature being larger than the observed value at PB28 is 0.035.



Supplementary Figure C.20: **QQ-plot comparing the MAP standardised subject-specific residuals to a t-distribution with 4 degrees of freedom** The plot compares the approximate observed error distribution in our model to the assumed error distribution. If the two distributions are the same, the points will go through the straight line. The bulk of the distribution lies close to the straight line, indicating the assumption for the error distribution is reasonable.