

# Adjusting for time of infection or positive test when estimating the risk of a post-infection outcome in an epidemic

Statistical Methods in Medical Research

2022, Vol. 31(10) 1942–1958

© The Author(s) 2022



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/09622802221107105

[journals.sagepub.com/home/smm](https://journals.sagepub.com/home/smm)



Shaun R Seaman<sup>1</sup> , Tommy Nyberg<sup>1</sup> ,  
Christopher E Overton<sup>2,3,4</sup>, David J Pascall<sup>1,4</sup>, Anne M Presanis<sup>1</sup>  
and Daniela De Angelis<sup>1,4,5</sup>

## Abstract

When comparing the risk of a post-infection binary outcome, for example, hospitalisation, for two variants of an infectious pathogen, it is important to adjust for calendar time of infection. Typically, the infection time is unknown and positive test time used as a proxy for it. Positive test time may also be used when assessing how risk of the outcome changes over calendar time. We show that if time from infection to positive test is correlated with the outcome, the risk conditional on positive test time is a function of the trajectory of infection incidence. Hence, a risk ratio adjusted for positive test time can be quite different from the risk ratio adjusted for infection time. We propose a simple sensitivity analysis that indicates how risk ratios adjusted for positive test time and infection time may differ. This involves adjusting for a shifted positive test time, shifted to make the difference between it and infection time uncorrelated with the outcome. We illustrate this method by reanalysing published results on the relative risk of hospitalisation following infection with the Alpha versus pre-existing variants of SARS-CoV-2. Results indicate the relative risk adjusted for infection time may be lower than that adjusted for positive test time.

## Keywords

COVID-19, epidemic phase bias, selection bias

## 1 Introduction

Consider the problem of estimating the distribution of time between a first event (e.g. becoming infected with SARS-CoV-2) and a second event (e.g. testing positive for the virus) in the population of individuals who ultimately experience the second event. We shall call this time the ‘inter-event time’ or ‘delay’. Estimating this distribution may be complicated by the first event time not being observed and/or the available data being right-truncated on the second event time, due to only sampling individuals who experience the second event by a particular calendar time.

It has long been known that the inter-event times in the population of individuals who have experienced the second event before a given calendar time tend to be shorter than in the population of all individuals (who eventually experience the second event).<sup>1–3</sup> More precisely, the conditional probability that the inter-event time is less than  $t$  given that the second

<sup>1</sup>MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

<sup>2</sup>Department of Mathematics, University of Manchester, UK

<sup>3</sup>Clinical Data Science Unit, Manchester University NHS Foundation Trust, UK

<sup>4</sup>Joint Universities Pandemic and Epidemiological Research (JUNIPER) consortium, Cambridge, UK

<sup>5</sup>Statistics, Modelling and Economics Department, UKHSA, London, UK

### Corresponding author:

Shaun Seaman, MRC Biostatistics Unit, University of Cambridge, East Forvie Building, Forvie Site, Robinson Way, Cambridge, CB2 0SR, UK.

Email: [shaun.seaman@mrc-bsu.cam.ac.uk](mailto:shaun.seaman@mrc-bsu.cam.ac.uk)

event has occurred by a given calendar time is greater than the corresponding unconditional probability (unless all second events have occurred by that time).

It has also been noted that the conditional distribution of inter-event time given the actual calendar time of the second event depends on the marginal distribution of the first event time.<sup>4–6</sup> In particular, if the first event is generated by a Poisson process whose rate is increasing with calendar time, then the conditional distribution of the inter-event time given the calendar time of the second event is shifted towards zero compared to the unconditional distribution. On the other hand, if the rate is decreasing, the conditional distribution of the inter-event time is shifted away from zero compared to the unconditional distribution. This means that in the context of an infectious disease the time from infection (first event) to positive test (second event) in those who test positive at a given calendar time tends to be shorter than average when the incidence of infection is rising, and longer than average when the incidence is falling.

Now consider a third variable, which is measured at, or after, the time of the second event and is correlated with the inter-event time. Just as the distribution of inter-event time is affected by conditioning on the calendar time of the second event, so might the distribution of this variable. For example, an infected individual's viral load at time of positive test is a function of time since infection. Rydevik et al.<sup>5</sup> observed that this relation could be used to estimate an individual's infection time from that individual's viral load at the time of testing positive. Hay et al.<sup>6</sup> used this same idea to estimate the pattern of incidence of infection in the population from data on the distribution of viral load (measured as the cycle threshold) in a random sample of individuals who tested positive on a given day. If the mean viral load is high, this suggests most of the sampled individuals were infected recently, which is consistent with a rising incidence of infection. Conversely, if the mean viral load is low, this suggests less recent infection, and so a falling incidence. Hay et al.<sup>7</sup> investigated using such data to estimate simultaneously the pattern of incidence of infection and the dependence of the viral load on the time since infection. Similar work had previously been done in the field of HIV/AIDS (e.g. Authors in<sup>8–10</sup>).

In the present article, we consider the estimation of the distribution of a third variable where this variable is a binary outcome of interest. An association between this binary outcome and the inter-event time could arise due to factors that determine both. We take the first and second events to be infection and positive test, respectively, and the binary outcome to be hospitalisation within 14 days of the positive test, although what follows would apply to any other binary outcome, for example, death within 28 days of a positive test. Individuals with more severe infections may tend to experience symptom onset sooner after infection – and consequently be tested earlier – than average and also be more likely to become hospitalised. In this situation, the hospitalisation risk (i.e. the proportion ultimately hospitalised) in individuals who test positive before a particular calendar time would be higher than the risk in all individuals who eventually test positive. More importantly for this article, the hospitalisation risk in individuals who test positive *at* a particular calendar time will differ from the risk in all individuals who eventually test positive (unless the incidence of infection is constant over time). If the incidence of infection is rising, the former risk will be higher than the latter; if incidence is falling, it will be lower.

This dependence of the hospitalisation risk on the trajectory of incidence of infection is particularly relevant for any investigation of how the risk is changing over calendar time. Ideally, such an investigation might involve comparing the risks for individuals with different calendar times of infection. If, as is likely, infection times are unknown, it would be natural to instead compare the risks for individuals with different calendar times of positive test. The difficulty with interpreting this latter comparison is that, as noted above, even if the risk does not vary by calendar time of infection, it will depend on calendar time of positive test.

Another situation where one might condition on calendar time of positive test is when comparing the risks associated with two variants of a given pathogen, in this case SARS-CoV-2. Here, controlling for (i.e. conditioning on) time of infection would be important, because the 'exposure' (i.e. a binary variable for the variant) and the outcome (hospitalisation) may both depend on calendar time. The exposure would depend on calendar time if the ratio of the incidence rates of infection with the two variants varied over time. That would be the case if, for example, one variant emerged earlier but the other variant later became dominant. The hospitalisation outcome would depend on calendar time if measures designed to reduce the need for hospitalisation and/or policies on hospital admission changed over time. Failure to control for infection time when comparing the risks of hospitalisation for the two variants would mean comparing the risk in individuals infected with one variant, whose infection times may have been predominantly when pre-hospital treatments were less effective and/or hospital admission more encouraged, with the risk in individuals infected with the other variant, whose infection times were mostly when pre-hospital treatments were better or hospital admission more restricted. If infection times are unknown, it would be natural to control instead for the time of positive test as a proxy for infection time. The difficulty with this approach is that, even if the hospitalisation risk is the same for both variants and does not depend on the time of infection, once we condition on calendar time of positive

test a variant that has increasing incidence of infection will appear to have a higher risk than a variant that has a decreasing incidence.

Numerous studies have compared the risks of hospitalisation, intensive care unit admission and/or death in individuals infected with two variants of SARS-Cov-2 (either Alpha vs. pre-existing non-Alpha or Delta vs. Alpha), adjusting for calendar time of positive test, for example, the authors in.<sup>11–22</sup> In all these studies, the incidence of one variant has been rising while the incidence of the other has been falling or has been rising at a slower rate.

In this article, we describe in detail why and how the conditional risk of hospitalisation given time of positive test depends on the trajectory of incidence of infection, even when the conditional risk given time of infection does not. We also propose an easily implemented method that provides an indication of how much an estimate of the risk conditional on the positive test time might differ from the estimate one would have obtained if it had been possible to condition on the infection time. This method involves calculating the risk conditional on a *shifted* positive test time. For each individual who is not ultimately hospitalised, this shifted test time is the same as the actual positive test time, that is, there is no shift. However, for each individual who is ultimately hospitalised, the shifted time equals the actual positive test time plus the difference between the mean time from infection to positive test in individuals who do not become hospitalised and the mean time in individuals who do become hospitalised. This ensures that the shifted time from infection to positive test is uncorrelated with the hospitalisation outcome. Because this difference between mean times is unknown, our method requires the user to specify a range of plausible values for it.

The structure of the article is as follows. Section 2 defines our notation. Section 3 describes why and how the distribution of the delay conditional on the calendar time of positive test depends on the incidence of infection. Section 4 goes on to explain how this dependence affects the conditional risk of hospitalisation given calendar time of positive test. We introduce our proposed method in Section 5, and its performance is studied in Section 6. Practical application of the method is detailed in Section 7, and Section 8 illustrates its use on COVID-19 data from England. We conclude with a discussion in Section 9.

## 2 Notation

We shall consider the population to be everyone who is at risk of infection from some calendar time zero. Time can be measured discretely or continuously. Suppose for now that all infections result in positive tests. In Section 9, we shall discuss the consequences of violation of this assumption. If an individual has two or more separate episodes of infection, we only consider the first episode.

For each individual in the population, let  $I$  denote the (calendar) time of infection, and let  $L$  be the delay (' $L$ ' for 'lag') between infection and positive test. Now,  $T = I + L$  is the time of positive test. Let  $H$  equal 1 if the individual is hospitalised within 14 days of positive test, and 0 otherwise. In Section 7, we shall use  $V$  to denote a binary indicator of which of two variants has infected an individual, and use  $U$  to denote a vector of variables that are fixed from the time of infection, for example, age and ethnicity. There we shall be interested in the odds ratio of hospitalisation associated with  $V$  adjusted for  $U$  and  $I$ .

## 3 Delay distribution conditional on test time

In this section and Section 4, we shall assume, for simplicity, that  $L$  is independent of  $I$ . Using Bayes' Rule,  $f_L(l | T = t)$ , the conditional probability distribution function of the delay given the positive test occurs at time  $t$ , can be shown to be related to the unconditional probability distribution function  $f_L(l)$  by

$$f_L(l | T = t) = \frac{f_{T,L}(t, l)}{f_T(t)} = \frac{f_{I,L}(t - l, l)}{f_T(t)} = f_L(l) \times \frac{f_I(t - l)}{f_T(t)} \quad (1)$$

If the incidence of infection is increasing over the period prior to time  $t$ ,  $f_I(t - l)$  will be a decreasing function of  $l$ . So, for any  $l_1 < l_2$ , we have  $f_I(t - l_1) > f_I(t - l_2)$  and so equation (1) implies

$$\frac{f_L(l_1 | T = t)}{f_L(l_1)} = \frac{f_I(t - l_1)}{f_T(t)} > \frac{f_I(t - l_2)}{f_T(t)} = \frac{f_L(l_2 | T = t)}{f_L(l_2)}. \quad (2)$$

From inequality (2), we have

$$\frac{f_L(l_1 | T = t)}{f_L(l_2 | T = t)} > \frac{f_L(l_1)}{f_L(l_2)}.$$

That is, conditioning on  $T = t$  shifts probability mass from larger values of  $L$  to smaller values. So, if we look only at those individuals whose positive test time is  $t$ , then small delays will be over-represented and long delays under-represented. This is not surprising, since an individual with positive test time  $t$  had a short delay if he was infected recently and a long delay if he was infected long ago, and there are more individuals infected recently than individuals infected long ago.

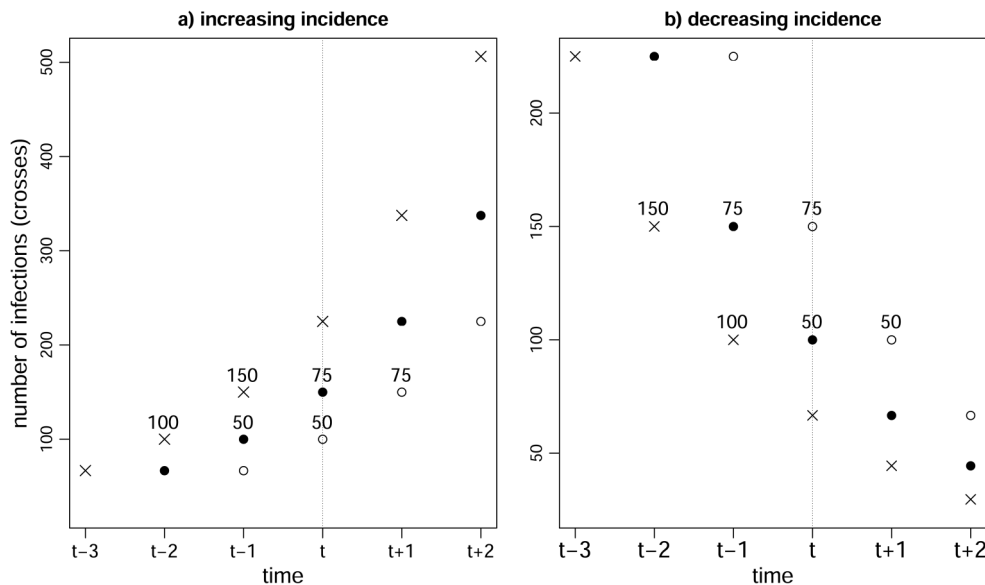
Conversely, if the incidence of infection is decreasing,  $f_I(t - l)$  will be a increasing function of  $l$ , and so

$$\frac{f_L(l_1 | T = t)}{f_L(l_2 | T = t)} < \frac{f_L(l_1)}{f_L(l_2)}.$$

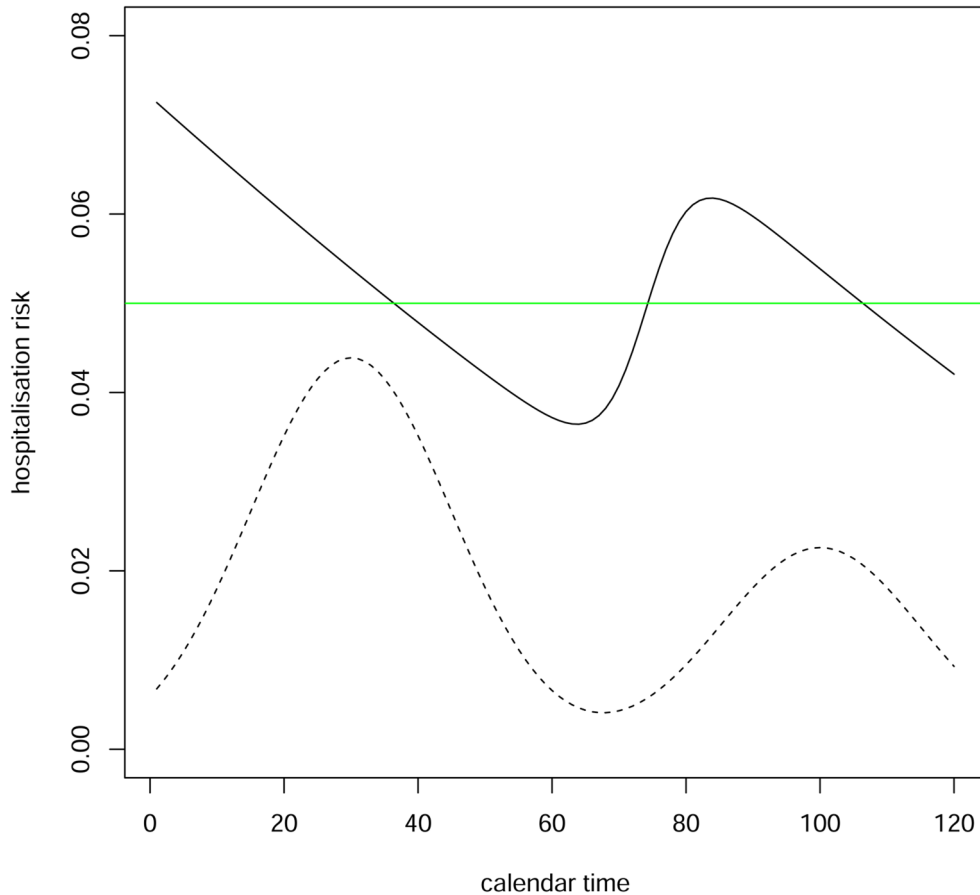
That is, conditioning on  $T = t$  shifts probability mass from smaller values of  $L$  to larger values: long delays are over-represented and short delays under-represented. In this situation, there are fewer individuals infected recently than individuals infected long ago.

*Example 1:* Suppose half of infected individuals test positive on the day after they are infected and the other half test positive two days after they are infected. That is,  $P(L = 1) = P(L = 2) = 0.5$ . Further, suppose that 100 individuals are infected on day  $t - 2$  and 150 individuals are infected on day  $t - 1$  (so incidence is increasing). Then 125 individuals will test positive on day  $t$  and, of these, 50 were infected on day  $t - 2$  and 75 were infected on day  $t - 1$ . So, the proportion of these 125 individuals whose delay was one day is  $75/125 = 0.6 > 0.5$ . Conversely, suppose that 150 individuals are infected on day  $t - 2$  and 100 individuals are infected on day  $t - 1$  (incidence is decreasing). Then 125 individuals will again test positive on day  $t$ , but the proportion of these whose delay is one day is only  $50/125 = 0.4 < 0.5$ . Figure 1 illustrates this example.

*Example 2:* Verity et al.<sup>23</sup> (see also Seaman et al.<sup>3</sup>) showed that if infections are generated by a Poisson process with rate at time  $t$  proportional to  $\exp(\lambda t)$  for some  $\lambda$ , and the delay  $L$  has a gamma distribution with shape  $\alpha$  and rate  $\beta$ , then the conditional distribution of  $L$  given  $T = t$  is gamma with shape  $\alpha$  and rate  $\beta + \lambda$ . If  $\lambda > 0$ , then the incidence is rising (exponentially) and the conditional mean delay,  $\alpha/(\beta + \lambda)$ , is less than the unconditional mean  $\alpha/\beta$ . If instead  $\lambda < 0$ , then the



**Figure 1.** Illustration of Example 1. Crosses show numbers of infections. Black and white circles represent cases with delays of 1 and 2 days, respectively. In left-hand graph, infection incidence is increasing. A total of 100 individuals are infected on day  $t - 2$ , of whom half (i.e. 50) test positive with a delay of 2 days on day  $t$ . In addition, 150 individuals are infected on day  $t - 1$ , of whom half (i.e. 75) test positive with a delay of 1 day on day  $t$ . So,  $75/(50 + 75) = 60\%$  of the cases who test positive on day  $t$  have a delay of 1 day. In right-hand graph, incidence is decreasing. A total of 150 individuals are infected on day  $t - 2$ , of whom 75 test positive with a delay of 2 days on day  $t$ . In addition, 100 individuals are infected on day  $t - 1$ , of whom 50 test positive with a delay of 1 day on day  $t$ . So,  $50/(50 + 75) = 40\%$  of the cases who test positive on day  $t$  have a delay of 1 day.



**Figure 2.** Hospitalisation risk conditional on positive test time (solid black line) when risk conditional on infection time is 0.05 (green line). Incidence of infection is shown (dotted line). Time from infection to positive test is assumed to have a gamma distribution with mean 4 and variance 8 for the ultimately hospitalised individuals and a gamma distribution with mean 7 and variance 14 for the ultimately non-hospitalised individuals.

incidence is falling and the conditional mean delay is greater than the unconditional mean. If  $\lambda = 0$ , the two gamma distributions are the same.

*Example 3:* Figure 2 shows how the hospitalisation risk conditional on positive test time varies according to positive test time in a scenario where the incidence of infection first rises then falls, then rises and falls again. Here, the hospitalisation risk conditional on infection time is 5% irrespective of the infection time, and the mean time from infection to positive test is shorter in individuals who are ultimately hospitalised than in those who are not.

#### 4 Hospitalisation risk conditional on test time

As we have seen, conditioning on the positive test time changes the distribution of the delay in circumstances where the delay is independent of the time of infection. If hospitalisation is more common in individuals with shorter delays than in those with longer delays, that is,  $P(H = 1 | I = t, L = l)$  is a decreasing function of  $l$ , then conditioning on the positive test time might be expected also to change the probability of hospitalisation. The effect of conditioning on  $T = t$  will depend on whether the incidence of infection is rising or falling. If it is rising, we might expect the proportion of hospitalisations to be increased, because short delays are over-represented. Conversely, if the incidence is falling, we might expect the proportion of hospitalisations to be decreased. We now confirm mathematically that this is true when the risk of hospitalisation either does not depend on the time of infection or changes little over the course of all but the longest delays.

Suppose that almost all delays are at most  $l^*$  for some constant  $l^*$  (i.e.  $P(L > l^*) \approx 0$ ). Also, assume that  $P(H = 1 | I = t - l, L = l) \approx P(H = 1 | I = t, L = l)$  for all  $0 < l \leq l^*$ . The risk of hospitalisation conditional on time of infection can then be written as

$$\begin{aligned} P(H = 1 | I = t) &= \int_0^{\infty} P(H = 1 | I = t, L = l) f_L(l) dl \\ &\approx \int_0^{l^*} P(H = 1 | I = t, L = l) f_L(l) dl. \end{aligned} \quad (3)$$

Likewise, the risk of hospitalisation conditional on positive test time is

$$\begin{aligned} P(H = 1 | T = t) &= \int_0^{\infty} P(H = 1 | I = t - l, L = l) f_L(l | T = t) dl \\ &\approx \int_0^{l^*} P(H = 1 | I = t - l, L = l) f_L(l | T = t) dl. \\ &\approx \int_0^{l^*} P(H = 1 | I = t, L = l) f_L(l | T = t) dl. \end{aligned} \quad (4)$$

Expressions (3) and (4) are both weighted averages of  $P(H = 1 | I = t, L = l)$ . In expression (3) the weighting function is  $f_L(l)$ ; in (4) it is  $f_L(l | T = t)$ . If the incidence of infection is increasing over the period  $[t - l^*, t]$ , then, as explained in Section 3, conditioning on  $T = t$  shifts probability mass from larger values of  $L$  to smaller values. Hence, the weighted average in expression (4) gives more weight to small values of  $l$  (and less weight to large values of  $l$ ) than does the weighted average in expression (3). This, combined with our assumption that  $P(H = 1 | I = t, L = l)$  is a decreasing function of  $l$ , implies that (4) is greater than (3). That is,  $P(H = 1 | T = t) > P(H = 1 | I = t)$ . On the other hand, if the incidence of infection is decreasing over the period  $[t - l^*, t]$ , then (as explained in Section 3) conditioning on  $T = t$  shifts probability mass from smaller values of  $L$  to larger values, with the result that  $P(H = 1 | T = t) < P(H = 1 | I = t)$ .

*Example 1 continued:* Suppose  $P(H = 1 | I = t, L = 1) = 0.05$  and  $P(H = 1 | I = t, L = 2) = 0.01$ . Then  $P(H = 1 | I = t) = (0.05 + 0.01)/2 = 0.03$ . If the incidence of infection is increasing, then of the 125 individuals who test positive on day  $t$ , the expected number who are hospitalised is  $50 \times 0.01 + 75 \times 0.05 = 4.25$ , corresponding to a proportion of  $4.25/125 = 0.034$  (which is  $>0.03$ ). If, on the other hand, the incidence of infection is decreasing, then of the 125 individuals who test positive on day  $t$ , the expected number who are hospitalised is  $75 \times 0.01 + 50 \times 0.05 = 3.25$ , corresponding to a proportion of  $0.026$  (which is  $<0.03$ ). The ratio of these two proportions is  $0.034/0.026 = 1.31$ , and so the hospitalisation risk conditional on time of positive test would differ by 31% between a period of epidemic growth and a period of epidemic decline.

## 5 Hospitalisation risk conditional on infection time plus random delay

Now suppose we had a different proxy of infection time such that the difference between this proxy and the actual infection time were (unlike the difference between positive test time and infection time) not associated with the hospitalisation outcome. If we conditioned on this proxy, we might achieve the goal of approximately adjusting for time of infection without creating a measure of hospitalisation risk that depends on the trajectory of the infection incidence. We now describe such a proxy.

Suppose, hypothetically, that each individual who becomes infected at time  $I = i$  and tests positive at time  $T = t$  is randomly assigned a time variable  $T^0$  sampled from the conditional distribution of  $T$  given  $I = i$  and  $H = 0$ , that is, the distribution of positive test time in those who are infected at the same time and who are not ultimately hospitalised. This time  $T^0$  will be our proxy of infection time. By construction,  $T^0 - I$  is not associated with the outcome  $H$ .

We cannot actually carry out this assignment in practice, because we do not observe  $I$ . However, under the following working assumption, we shall still be able to estimate  $P(H = 1 | T^0 = t)$ .

Assumption 1

$$f_T(t | I, H = 1) = f_T(t + c | I, H = 0),$$

where  $c$  is some known constant.

Assumption 1 means that the distribution of time from infection to positive test (conditional on time of infection) in ultimately hospitalised individuals equals the corresponding distribution in ultimately non-hospitalised individuals shifted by  $c$  days. So, the mean time from infection to test in the ultimately hospitalised is  $c$  days less than the mean in the ultimately non-hospitalised. We might expect that  $c > 0$ .

If Assumption 1 holds, then (see the Appendix for proof)

$$P(H = 1 | T^0 = t) = P(H = 1 | T + cH = t). \tag{5}$$

It follows from equation (5) that  $P(H = 1 | T^0 = t)$  can be consistently estimated simply by creating the new variable  $T^* = T + cH$  (which equals  $T$  for ultimately non-hospitalised cases and  $T + c$  for ultimately hospitalised cases) and calculating the proportion who have  $H = 1$  among those sampled individuals with  $T^* = t$  (or, if time is continuous,  $T^* \approx t$ ).

The hospitalisation risk conditional on  $T^0$ , i.e.  $P(H = 1 | T^0 = t)$ , has the desirable property that  $P(H = 1 | T^0 = t) = P(H = 1 | I = t)$  if  $P(H = 1 | I = t)$  does not depend on  $t$ . More generally, if  $P(H = 1 | I = t) \approx P(H = 1 | I = t + l)$  for all  $0 < l \leq l^*$ , then  $P(H = 1 | T^0 = t) \approx P(H = 1 | I = t)$ . That is, when the risk conditional on  $I$  is constant or changes only slowly over time, conditioning on  $T^0$  yields almost the same risk as conditioning on  $I$ .

In practice, it is unlikely that we shall know the true value of  $c$ . However, one may be able to specify a range of plausible values for it and then investigate how sensitive the estimate of  $P(H = 1 | T^0)$  is to this value. We shall illustrate this approach in Section 8.

Assumption 1 states that the distribution of delay in hospitalised individuals equals the distribution of delay in non-hospitalised individuals shifted by some number ( $c$ ) of days. This assumption may well be false. In particular, it implies that the minimum delay in non-hospitalised cases cannot be less than  $c$ . So, in Section 6, we shall investigate the extent to which  $P(H = 1 | T^*)$  differs from  $P(H = 1 | T^0)$  when one delay distribution is not a shifted version of the other but we set  $c$  to be equal to  $E(L | I, H = 1) - E(L | I, H = 0)$ , that is, the difference between the mean of the two delay distributions.

### 6 Investigation of proposed method

Suppose the incidence of infection at time  $t$  is proportional to  $\exp(\lambda t)$ . If  $\lambda > 0$ , then  $d = \log(2)/\lambda$  is the doubling time; if  $\lambda < 0$ , then  $-d$  is the halving time. Suppose that  $P(H = 1 | I = t)$ , the risk of hospitalisation for an individual who is infected at time  $t$  is  $r = 0.05$  and does not depend on the time of infection. Assume that the delay in the non-hospitalised cases has a gamma distribution with mean 7 and variance 14 (i.e. shape  $\alpha = 3.5$  and rate  $\beta = 0.5$ ). We consider two scenarios for the distribution of delay in the hospitalised cases. In Scenario 1, it is a gamma distribution with mean 4 and variance 8 (shape  $\alpha_1 = 2$  and rate  $\beta = 0.5$ ). In Scenario 2, it is an equal mixture of a gamma distribution with mean 7 and variance 14 and a gamma distribution with mean 1 and variance 2 (shape  $\alpha_2 = 0.5$  and rate  $\beta = 0.5$ ). So, in both scenarios, the difference between the mean delay in the hospitalised and non-hospitalised is three days. Figure 3 shows, for each scenario, the distributions of delay in the non-hospitalised (black) and hospitalised (red) cases. The dotted line shows the distribution in the non-hospitalised cases shifted by three days.

The number of non-hospitalised individuals who test positive at time  $t$  is proportional to

$$\begin{aligned} (1 - r) \int_{-\infty}^t \exp(\lambda u) \frac{\beta^\alpha}{\Gamma(\alpha)} (t - u)^{\alpha-1} \exp\{-\beta(t - u)\} du \\ = (1 - r) \exp(\lambda t) \frac{\beta^\alpha}{(\beta + \lambda)^\alpha}. \end{aligned}$$

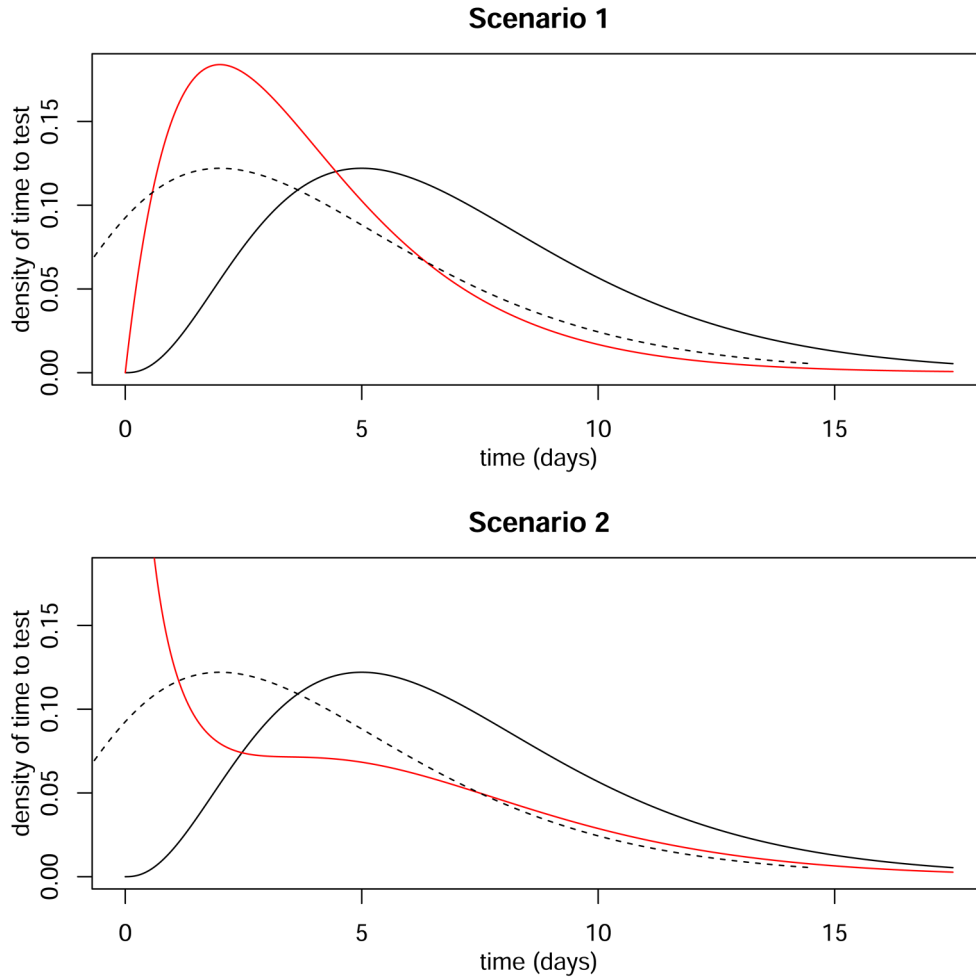
Similarly, the number of hospitalised individuals who test positive at time  $t$  is proportional to

$$r \exp(\lambda t) \frac{\beta^{\alpha_1}}{(\beta + \lambda)^{\alpha_1}}$$

in Scenario 1 and is

$$r \exp(\lambda t) \times \frac{1}{2} \left\{ \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + \frac{\beta^{\alpha_2}}{(\beta + \lambda)^{\alpha_2}} \right\}$$

in Scenario 2.



**Figure 3.** Distributions of time from infection to positive test. Solid black line is distribution for ultimately non-hospitalised individuals. Dotted line is same distribution shifted by three days. Red line is distribution for ultimately hospitalised individuals.

So, the risk when we condition on time of positive test  $T$  is

$$P(H = 1 | T = t) = \frac{r \frac{\beta^{\alpha_1}}{(\beta + \lambda)^{\alpha_1}}}{(1 - r) \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + r \frac{\beta^{\alpha_1}}{(\beta + \lambda)^{\alpha_1}}} \tag{6}$$

in Scenario 1 and is

$$\begin{aligned} P(H = 1 | T = t) &= \frac{\frac{r}{2} \left\{ \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + \frac{\beta^{\alpha_2}}{(\beta + \lambda)^{\alpha_2}} \right\}}{(1 - r) \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + \frac{r}{2} \left\{ \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + \frac{\beta^{\alpha_2}}{(\beta + \lambda)^{\alpha_2}} \right\}} \\ &= \frac{r \left\{ \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + \frac{\beta^{\alpha_2}}{(\beta + \lambda)^{\alpha_2}} \right\}}{(2 - r) \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + r \frac{\beta^{\alpha_2}}{(\beta + \lambda)^{\alpha_2}}} \end{aligned} \tag{7}$$

in Scenario 2.



If we condition on  $T^* = T + 3H$ , the risk is

$$\begin{aligned}
 P(H = 1 \mid T^* = t) &= \frac{r \exp \{\lambda(t - 3)\} \frac{\beta^{\alpha_1}}{(\beta + \lambda)^{\alpha_1}}}{(1 - r) \exp(\lambda t) \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + r \exp \{\lambda(t - 3)\} \frac{\beta^{\alpha_1}}{(\beta + \lambda)^{\alpha_1}}} \\
 &= \frac{r \frac{\beta^{\alpha_1}}{(\beta + \lambda)^{\alpha_1}}}{(1 - r) \exp(3\lambda) \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + r \frac{\beta^{\alpha_1}}{(\beta + \lambda)^{\alpha_1}}}
 \end{aligned} \tag{8}$$

in Scenario 1, and is

$$\begin{aligned}
 P(H = 1 \mid T^* = t) &= \frac{\frac{r}{2} \exp \{\lambda(t - 3)\} \left\{ \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + \frac{\beta^{\alpha_2}}{(\beta + \lambda)^{\alpha_2}} \right\}}{(1 - r) \exp(\lambda t) \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + \frac{r}{2} \exp \{\lambda(t - 3)\} \left\{ \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + \frac{\beta^{\alpha_2}}{(\beta + \lambda)^{\alpha_2}} \right\}} \\
 &= \frac{r \left\{ \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + \frac{\beta^{\alpha_2}}{(\beta + \lambda)^{\alpha_2}} \right\}}{\{(2 - 2r) \exp(3\lambda) + r\} \frac{\beta^\alpha}{(\beta + \lambda)^\alpha} + r \frac{\beta^{\alpha_2}}{(\beta + \lambda)^{\alpha_2}}}
 \end{aligned} \tag{9}$$

in Scenario 2.

Table 1 shows the results of applying equations (6) to (9) in Scenarios 1 and 2, when  $d = 4$  and when  $d = 10$ .

It also shows the results when  $d = -4$  or  $d = -10$ , meaning that the incidence is falling with a halving time is 4 or 10 days. We see that the risks conditional on  $T$  are indeed different from the risk conditional on time of infection  $I$ , that is,  $r = 0.05$ . The proposed method produces conditional risks  $P(H = 1 \mid T^*)$  that are close to  $r$ .

Finally, Table 2 shows the risk ratios that Table 1 implies when comparing two variants both of which have the same risk  $r = 0.05$ , but one of which has a doubling time of 4 (respectively, 10) days and the other has a halving time of 4 (respectively, 10) days. The risk ratio conditional on time of infection is  $r/r = 1$ . The risk ratios conditional on  $T$  vary from 1.5 to 2.8. When we instead condition on  $T^*$ , the conditional risk ratios vary from 0.95 to 1.04, that is, they are much closer to 1.

### 7 Practical implementation when comparing risks of two variants

In practice, when comparing the risks of hospitalisation (or other post-infection outcome, like death) associated with two variants, researchers may adjust not only for positive test time but also for variables like age and ethnicity. In this section, we detail how researchers can apply our proposed sensitivity analysis in this situation. We shall begin by assuming that the researchers are using logistic regression (as did, e.g. the authors in<sup>15,16,22</sup>), and then go on to address the slightly more complicated situation where there is administrative censoring of the outcome and the researchers are using Cox regression or parametric survival regression (as did, e.g. the authors in<sup>14,17-21</sup>).

**Table 1.** Risks when adjusted for  $T$  and  $T^*$ . If  $d = 4$  or 10, the doubling time is 4 or 10 days. If  $d = -4$  or  $-10$ , the halving time is 4 or 10 days.

Scenario	$d$	$T$	$T^*$
1	4	0.0760	0.0466
1	10	0.0601	0.0494
1	-4	0.0270	0.0447
1	-10	0.0404	0.0493
2	4	0.0830	0.0511
2	10	0.0612	0.0503
2	-4	0.0326	0.0536
2	-10	0.0414	0.0504

**Table 2.** Risk ratios when adjusted for  $T$  and  $T^*$ . One variant has doubling time  $d = 4$  (or  $d = 10$ ) days and the other has halving time  $d = 4$  (or  $d = 10$ ) days.

Scenario	$d$	$T$	$T^*$
1	4	2.810	1.044
1	10	1.489	1.003
2	4	2.551	0.954
2	10	1.480	0.997

Let  $V$  be a binary indicator for variant and  $U$  denote a vector of other variables that are fixed at the time of infection, for example, age and ethnicity. Suppose the researchers have specified a particular logistic regression model, with  $H$ , the indicator of hospitalisation, as the outcome and with  $V$ ,  $U$  and  $T$  (or functions of  $T$ , e.g. calendar week of positive test) as the covariates, and suppose they have fitted this model to all the individuals whose positive test time  $T$  lies in some chosen interval of calendar time  $[\tau_1, \tau_2]$ . The coefficient of  $V$  in this model represents the log odds of hospitalisation for variant adjusted for  $U$  and positive test time  $T$ .

To apply our method, assume that  $f_T(t | I, X, H = 1) = f_T(t + c | I, X, H = 0)$ , where  $X = (V, U)$ . This is like Assumption 1 but conditional on  $X$ , and it implies

$$P(H = 1 | T^0 = t, X) = P(H = 1 | T + cH = t, X) \quad (10)$$

(see the Appendix for proof). Choose a range of plausible values for  $c$ . In Section 8, we illustrate how one might choose this range. Then, for any given value in this range (e.g.  $c = 1$  day or  $c = 2$  days), calculate  $T^* = T + cH$  for each individual who has tested positive. Identify the set of individuals whose  $T^*$  lies in the interval  $[\tau_1, \tau_2]$ . Note that this set differs from the set whose  $T$  lies in  $[\tau_1, \tau_2]$ , because (assuming that  $c > 0$ ) it excludes hospitalised individuals with  $\tau_2 - c < T \leq \tau_2$  and includes hospitalised individuals with  $\tau_1 - c \leq T < \tau_1$ .<sup>a</sup> Fit the same logistic regression model used earlier to this set of individuals, replacing the covariate  $T$  (or functions of  $T$ ) in the model with a covariate  $T^*$  (or functions of  $T^*$ ). The coefficient of  $V$  in this model is the log odds of hospitalisation for variant adjusted for  $U$  and  $T^*$ . Equation (10) implies that this is also the log odds adjusted for  $U$  and  $T^0$  when the true value of  $c$  is used. The Supplemental Materials contain a simulation study that demonstrates the use of this method.

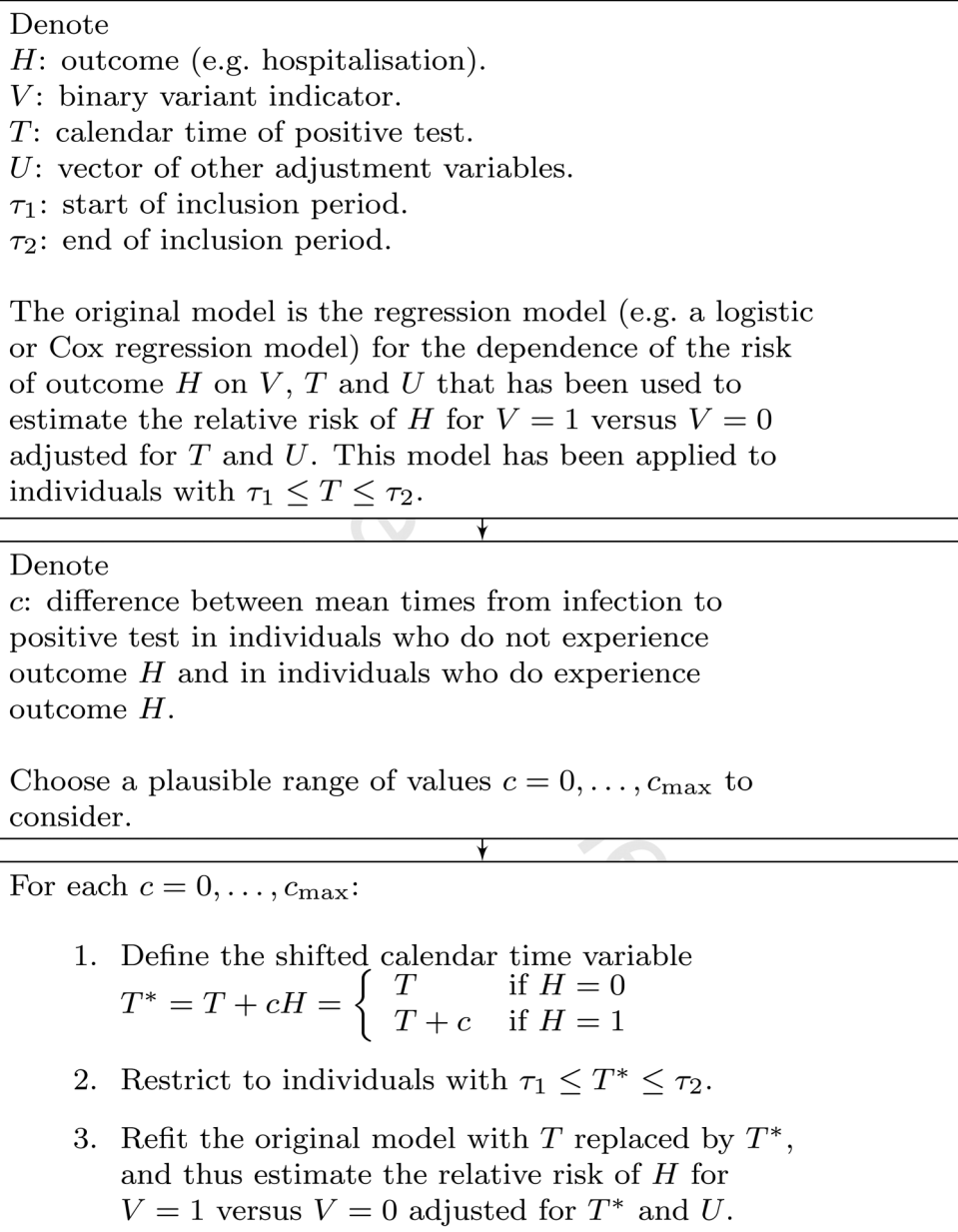
So far, we have assumed that the binary outcome  $H$  is observed for everyone who tests positive and that logistic regression has been used. In practice, there may be administrative censoring. This would occur if some of the sampled individuals test positive less than 14 days before time  $\tau_2$ , had not yet been hospitalised by time  $\tau_2$ , and no data were available on hospitalisations after time  $\tau_2$ . In this situation, researchers may use the *time* from positive test to hospitalisation as the outcome (rather than the binary indicator of hospitalisation within 14 days), right-censoring this time at 14 days, and fit a Cox regression model (or a parametric survival model) of this time-to-event outcome on  $V$ ,  $U$  and  $T$  to all the individuals whose  $T$  lies in the interval  $[\tau_1, \tau_2]$ . When the binary outcome is rare and is fully observed, the hazard ratios (HRs) estimated from this Cox regression would be approximately equal to the odds ratio estimates from logistic regression.<sup>24</sup> Our proposed sensitivity analysis involves fitting exactly the same Cox regression (or parametric survival) model that the researchers have used, with exactly the same time-to-event outcome. The only difference is that: (1) the covariate  $T$  in this model is replaced by  $T^*$ ; and (2) this model is fitted to all the individuals whose  $T^*$  lies in  $[\tau_1, \tau_2]$ . A difficulty arises because  $T^*$  is calculated from  $T$  and  $H$ , the last of which is unobserved for the censored individuals. To avoid compromising the simplicity of the proposed method, we suggest assuming that individuals whose hospitalisation status is unknown due to this administrative censoring have  $H = 0$  for the purpose of calculating  $T^*$  (and so setting  $T^* = T$ ). Provided that hospitalisation within 14 days is uncommon, this assumption will be true for the great majority of censored individuals.

Figure 4 summarises this procedure for carrying out the sensitivity analysis.

## 8 Application to COVID-19 hospitalisation and mortality data

### 8.1 Hazard ratio conditional on $T$ and $T^0$

Based on record linkage of routine healthcare data in England, Nyberg et al. (2021) recently reported a higher risk of hospital admission and mortality in COVID-19 cases infected with the Alpha variant (B.1.1.7) than in cases infected with pre-existing non-Alpha variants.<sup>18</sup> The study included data on 839,278 cases who had their first positive tests between 23 November 2020 and 31 January 2021 and whose positive test sample had been assessed for S gene target failure (SGTF). SGTF was used as a proxy test for the Alpha variant, and had positive and negative predictive values >90%



**Figure 4.** Summary of implementation of proposed sensitivity analysis.

during the study period.<sup>25</sup> As determined by SGTF, 592,409 cases had Alpha and 246,869 cases had non-Alpha variants. Using stratified Cox regression, Nyberg et al. estimated that, after adjustment for date of positive test and other covariates, the HR of hospital admission within 14 days of positive test was 1.52 (95% CI 1.47–1.57) for Alpha versus non-Alpha cases. Full details on the dataset, including inclusion criteria, variant classification, outcome definitions, and adjustment methods, are available in the recent publication.<sup>18</sup>

The Alpha variant was first detected in England in November 2020. During the study period the prevalence of the Alpha variant among newly test-positive cases, as determined by SGTF, increased from 16% in the week commencing 23 November 2020 to 94% in the week commencing 25 January 2021, and the prevalence of the pre-existing variants decreased accordingly.<sup>25</sup> During this period, the number of cases of the Alpha variant had a considerably higher growth rate than the number of cases of non-Alpha variants.<sup>26</sup> So, the aforementioned HR of 1.52 conditional on the positive test time  $T$  would be expected to be greater than the corresponding HR conditional on the infection time  $I$ .

We reanalysed the study dataset to estimate the HRs conditional on  $T^0$  for different choices of  $c$ . Column 2 of Table 3 shows the adjusted HR (and associated 95% CI) of hospital admission for Alpha versus non-Alpha from stratified Cox

**Table 3.** Hazard ratios for the two outcomes hospital admission and death (and 95% confidence intervals) conditional on  $T^*$  for COVID-19 cases with Alpha compared to non-Alpha variants, as determined based on S gene target failure. The assumed difference in mean number of days from infection to first positive test between individuals without the outcome and with the outcome is  $c$ . HRs were estimated using stratification for calendar week of positive test, age group, sex, ethnicity, index of multiple deprivation quintile, and region of residence (Public Health England Centres); and including strata-specific linear terms for exact date of positive test, exact age, and index of multiple deprivation rank; see<sup>18</sup>.

$c$	Hospitalisation outcome		Death outcome	
	All COVID cases HR (95% CI)	Symptomatic cases HR (95% CI)	All COVID cases HR (95% CI)	Symptomatic cases HR (95% CI)
0	1.52 (1.47–1.57)	1.49 (1.44–1.54)	1.59 (1.44–1.74)	1.42 (1.29–1.56)
1	1.41 (1.36–1.45)	1.37 (1.33–1.42)	1.43 (1.30–1.57)	1.32 (1.20–1.45)
2	1.31 (1.27–1.35)	1.28 (1.24–1.32)	1.33 (1.21–1.46)	1.23 (1.12–1.35)
3	1.21 (1.17–1.25)	1.19 (1.15–1.22)	1.23 (1.12–1.35)	1.14 (1.04–1.25)
4	1.13 (1.09–1.16)	1.10 (1.07–1.14)	1.15 (1.04–1.26)	1.07 (0.97–1.17)
5	1.04 (1.01–1.07)	1.02 (0.99–1.05)	1.06 (0.97–1.17)	1.00 (0.91–1.10)

regression models fitted after adding  $c$  days to the date of first positive test for those cases who were ultimately hospitalised. The HR when  $c = 0$  is that in the original publication.<sup>18</sup> As expected, there was a trend of decreasing HR with increasing  $c$ . However, the HR remained greater than one and statistically significant when  $c$  was between 1 and 5 days.

## 8.2 Plausible upper bound for $c$

We now consider what might be a plausible upper bound for  $c$ . The cases used in the analysis of Nyberg et al. all tested positive through the pillar-two national SARS-CoV-2 testing programme. Pillar-two is a broad mass community testing programme of individuals seeking testing due to symptoms or contact tracing efforts, and 85.5% of the included cases reported COVID-19 symptoms at the time of their positive test (88.5% in hospitalised cases and 85.4% in non-hospitalised cases). The mean time from symptom onset to positive test was 2.65 days (standard deviation 2.13) in the symptomatic hospitalised cases, and 2.51 days (standard deviation 2.12) in the symptomatic non-hospitalised cases. These percentages and means were similar in both Alpha and non-Alpha cases (data not shown). These figures indicate no obvious association between hospitalisation and time from symptom onset to test. Thus, we might reasonably assume that the mean time is 2.5 days in both hospitalised and non-hospitalised cases.

It has been estimated that the mean incubation time (i.e. time from infection to symptom onset) in cases who eventually experience symptoms is 5.74 days.<sup>27</sup> Since most of the cases in the data set were not (ultimately) hospitalised, this can be regarded as the mean incubation time in non-hospitalised cases who eventually experience symptoms. We assume that the mean incubation time in hospitalised cases (all of whom must have been symptomatic or pre-symptomatic at time of positive test) is less than or equal to the mean time in non-hospitalised cases who eventually have symptoms, but is not less than two days.

We shall assume that the mean time from infection to positive test in the 14.6% of (ultimately) non-hospitalised cases who were asymptomatic at time of positive test is not more than 12 days. This choice of 12 days is somewhat arbitrary, but does not seem unreasonable given that some of these cases will have been pre-symptomatic at time of positive test. It now follows that the mean time from infection to positive test in the non-hospitalised cases is not more than 8.8 days (i.e. mean incubation 5.74 days plus mean time from symptoms to test 2.5 days for the 85.4% of symptomatic cases, and 12 days for the 14.6% of asymptomatic cases). Assuming that the mean time from infection to positive test in the hospitalised cases is at least 4.5 days (i.e. mean incubation 2 days plus mean time from symptoms to test 2.5 days), it follows that a plausible upper bound for  $c$  is 4.3 days (i.e.  $8.8 - 4.5$ ).

## 8.3 Analysis of symptomatic cases

The analysis described above includes a proportion (14.5%) of cases who were asymptomatic prior to their positive test. We now consider a modified analysis that focuses on symptomatic cases. It is the same as the analysis above, except that we now exclude cases who were both asymptomatic prior to positive test and not (ultimately) hospitalised.

A total of 4.3% (36,233/839,278) of the cases were hospitalised. After excluding the 117,494 non-hospitalised cases who were asymptomatic at time of positive test, 5.0% (36,233/721,784) of the remaining cases were hospitalised.

Column 3 of Table 3 shows the results from analysing this subgroup consisting of symptomatic and/or hospitalised pillar-two cases (henceforth referred to as ‘symptomatic cases’). Consistent with the results using the full dataset, the HR point estimates were  $\geq 1.10$  when  $c \leq 4$  but was lower when  $c = 5$ .

Using the same logic as described above, the mean time from infection to positive test in the symptomatic non-hospitalised cases is not more than 8.2 days (i.e. mean incubation 5.74 days plus mean time from symptoms to test 2.5 days). Assuming again that the mean time from infection to positive test in the hospitalised cases is at least 4.5 days, a plausible upper bound for  $c$  is 3.7 days (i.e.  $8.2 - 4.5$ ).

## 8.4 Mortality

Using the same stratified Cox regression approach, Nyberg et al. (2021) also considered mortality as a secondary outcome, and reported an adjusted HR of death within 28 days of 1.59 (95% CI 1.44–1.74) for Alpha versus non-Alpha. Columns 4 and 5 of Table 3 show the original ( $c = 0$ ) and adjusted HR of this outcome, after adding  $c$  days to the date of first positive test for those cases who ultimately died within 28 days of testing positive. There was a trend towards lower HR estimates with increasing  $c$ . The HRs were statistically significantly greater than 1 for  $c \leq 4$  when using all cases, and for  $c \leq 3$  when restricting to symptomatic cases, but non-significant for higher  $c$ .

## 9 Discussion

In this article, we have highlighted the difference between the risk of a binary post-infection outcome (which, in this paper, is hospitalisation) conditional on the time of infection  $I$  and the corresponding risk conditional on the time of positive test  $T$ , and noted that the latter is a function of the trajectory of incidence of infection over calendar time. One way to interpret this difference is as a bias: if the goal is to estimate the risk conditional on infection time and if this risk differs from the risk conditional on the positive test time, then an (asymptotically) unbiased estimator of the latter risk will be an (asymptotically) biased estimator of the former risk. One might call this ‘epidemic phase bias’, since its direction and magnitude depend on whether the incidence of infection is falling or rising, and how quickly. This ‘bias’ may affect the results from a number of studies, for example, the authors in.<sup>11–22</sup> We have proposed a simple, easily implemented sensitivity analysis. This involves a third risk: that conditional on  $T^0$ , the infection time plus a random time that is independent of the outcome. This third risk equals the risk conditional on infection time when the latter does not change over time (i.e. as a function of infection time), and is approximately equal to it when the risk conditional on infection time changes slowly over time. More generally, the two risks differ, but both have the advantage of not depending on the trajectory of incidence of infection.

As with other sensitivity analysis approaches, for example, for addressing unmeasured confounding<sup>28</sup> and missing data,<sup>29,30</sup> ours does not yield a single estimate of the risk. It does, however, provide an indication of how sensitive the estimated risk is to the epidemic phase. If the incidence of infection is constant over calendar time, the estimated risk will not change as  $c$  is varied; if incidence is changing rapidly, the estimate will be very sensitive to the choice of  $c$ . When a risk ratio comparing two variants is of interest, sensitivity will be least when the variant-specific incidences of infection are both following the same trajectory (constant, increasing at the same exponential rate, or decreasing at the same rate), and will be greatest when one incidence is increasing rapidly and the other is decreasing rapidly.

The proposed method is likely to be most useful when a range of plausible values can be specified for  $c$ , the difference between the mean time from infection to positive test in the cases who experience the outcome and the corresponding mean time in those who do not experience the event. For our re-analysis of the healthcare data from England (Section 8) we suggested that the difference between these mean times is unlikely to be greater than 4.3 days when all cases are included. We found that the estimated hazard ratio (Alpha vs. non-Alpha variant) of hospital admission within 14 days declined from 1.52 (when  $c = 0$ ) to 1.41, 1.31, 1.21 or 1.13, respectively, when  $c = 1, 2, 3$  or 4, although all HR estimates were statistically significantly greater than 1. When we restricted the analysis to symptomatic cases, we argued that a plausible upper bound for  $c$  is 3.7 days. The hazard ratio using these cases declined from 1.49 (when  $c = 0$ ) to 1.37, 1.28, 1.19 or 1.10, respectively, when  $c = 1, 2, 3$  or 4, and was still significant. The results were similar for the mortality outcome. Hence, although this sensitivity analysis does not refute the reported association between the Alpha variant and risk of hospital admission or mortality, the range of plausible values of  $c$  includes the possibility that differences in epidemic phase might have caused overestimation to some extent of the strength of association.

The true value of  $c$  is uncertain, however, and to establish the precise impact of the epidemic phase ‘bias’, further research is needed to estimate the difference, between ultimately hospitalised and non-hospitalised (or ultimately deceased and non-deceased) cases, in their mean time from infection to onset of symptoms, and their mean time from symptom onset to positive test. Of these, the difference in time from infection to symptom onset (i.e. incubation time) is the more uncertain.

Estimates are available (e.g. Rai et al.<sup>27</sup>) for the mean incubation time in the population of all cases who experience symptoms. Since hospitalisation is an uncommon outcome, this mean should approximately equal the mean incubation time in ultimately non-hospitalised cases. Harder to estimate is the mean incubation time in cases who are ultimately hospitalised. For some cases it may be possible to establish retrospectively that they almost certainly became infected on a particular day, perhaps because they are known to have been exposed to an infected individual on that day and unlikely to have been exposed otherwise during the period around that day. If a sufficient number of such cases were identified, it ought to be possible to use them to estimate the mean incubation time in ultimately hospitalised and non-hospitalised cases separately, and hence the difference between these means. However, the numbers of cases required to do this with precision could be large, because hospitalisation is uncommon. There might also be a concern that these cases with known infection times were unrepresentative of all cases.

It would be straightforward in theory to allow  $c$  to depend on observed variables  $X$ . Simply assume that  $f_T(t | I, X, H = 1) = f_T\{t + c(X) | I, X, H = 0\}$ , for some function  $c(x)$  of  $x$ , and redefine  $T^*$  slightly as  $T^* = T + c(X)H$ . One might also want to allow  $c$  to depend on the unknown infection time  $I$ . For example, if the mean time from infection to positive test in the non-hospitalised cases is getting smaller over the study period, then it is quite likely that the difference between the mean time in non-hospitalised cases and the mean time in hospitalised cases is also getting smaller. If it were necessary to allow  $c$  to depend on  $I$  over the study period, then a crude but practical way of doing this would be to specify  $c$  as a function of  $T$  and calculate  $T^*$  as  $T^* = T + c(T)$ .

We have focused on an observed binary outcome  $H$ , but also briefly addressed right-censoring of this outcome. We proposed that  $T^*$  be calculated as though the censored individuals did not experience the outcome (i.e.  $H = 0$ ). This is a reasonable approximation when the outcome is rare and the proportion of censored individuals is small. For more common outcomes or when the extent of censoring is larger, it would be preferable to use a more sophisticated approach. More research is needed on this, but one possibility may be the following. Fit the Cox model to the original data. Estimate the baseline hazard. Use this estimated baseline hazard and the estimated hazard ratios from the Cox model to calculate the probability  $p_i$  that a censored individual  $i$  has  $H = 1$ . Then create two copies of each censored individual  $i$ : one with  $H = 1$ ,  $T^* = T + c$  and weight  $p_i$ ; and one with  $H = 0$ ,  $T^* = T$  and weight  $1 - p_i$ . An obvious drawback of this method is that  $p_i$  would be calculated from a model that implicitly assumed  $c = 0$ . A more refined version might begin by calculating  $T^*$  as though the censored individuals all had  $H = 0$ , then using the resulting fitted Cox model and estimated baseline hazard to calculate  $p_i$ .

We have assumed that all infections result in a positive test. This is obviously not true in reality. However, this issue affects all studies of risks of post-infection outcomes in samples of individuals who have tested positive, and is not specific to this article. There is not a problem if those individuals who test positive are representative of all infected individuals. Otherwise, the estimated risks must be interpreted as risks conditional on eventually testing positive.

We have focused on a setting where most delays are measured in days or a small number of weeks. Here, researchers may view time of positive test as a good proxy for time of infection. In a setting where the mean delay is much larger or where there is a long tail in the delay distribution, both the actual positive test time  $T$  and its shifted counterpart  $T^*$  will typically be very different from the infection time  $I$ . In that setting, although adjustment for  $T^*$  can still remove the dependence of the risk on the trajectory of incidence to which the risk adjusted for  $T$  is subject, neither of these adjusted risks should be seen as an approximation of the risk conditional on  $I$ .

Finally, if additional information is available on the incidence of infection with each variant over time, it may be possible to estimate the hospitalisation risk without using data on positive test times. This could be done using deconvolution techniques, such as those developed in the 1980s and 1990s for back-calculation in the context of the HIV/AIDS epidemic. There the purpose was to estimate the distribution of HIV infection times from the observed distribution of AIDS onset times and an assumed-known distribution of time from infection to AIDS onset. For example, Rosenberg and Gail<sup>31</sup> described how to do this using software for Poisson regression with identity link function. In the context of the present article, the purpose would be to estimate, for each variant, the distribution of time from infection to hospitalisation from the observed distribution of hospitalisation times and an assumed-known distribution of infection times. It may be possible to do this by applying, for example, an adaptation of the Poisson regression method of Rosenberg and Gail with an additional offset term for the total number of infections observed so far.

## Acknowledgements

We acknowledge the support of the UK Health Security Agency (UKHSA) Epidemiology and Modelling Cells in providing and discussing the data used. The authors received no financial support for the research, authorship and/or publication of this article.

## Data availability

This analysis was based on routine healthcare data, which cannot be made available to others by the study authors. Requests to access these non-publicly available data are handled by the Public Health England Office for Data Release (<https://www.gov.uk/government/publications/accessing-public-health-england-data/about-the-phe-odr-and-accessing-data>).


## Declaration of conflicting interests


The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

## Funding

This work was funded by UKRI Medical Research Council: core Unit funding (Seaman: MC UU 00002/10; De Angelis, Presanis: MC UU 00002/11), JUNIPER consortium (Overton, Pascall: MR/V038613/1), MRC UKRI / DHSC NIHR COVID-19 rapid response call (Presanis, De Angelis, Nyberg: MC PC 19074); by the Wellcome Trust and Royal Society (Overton, 202562/Z/16/Z); by the NIHR Health Protection Unit in Behavioural Science and Evaluation (De Angelis); and was supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed are those of the authors and not necessarily those of UKHSA, the NHS, the NIHR or the Department of Health and Social Care.

## ORCID iDs

Shaun R Seaman  <https://orcid.org/0000-0003-3726-5937>

Tommy Nyberg  <https://orcid.org/0000-0002-9436-0626>

## Supplemental material

Supplemental material for this article is available online.

## Note

- a. Alternatively, if data on individuals with  $T < \tau_1$  are unavailable, one can instead identify the set of individuals whose  $T^*$  lies in the interval  $[\tau_1 + c, \tau_2]$ .

## References

1. Brookmeyer R and Damiano A. Statistical methods for short-term projections of AIDS incidence. *Stat Med* 1989; **8**: 23–34.
2. Kalbfleisch J and Lawless J. Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. *J Am Stat Assoc* 1989; **84**: 360–372.
3. Seaman S, Jackson C and Presanis A. Estimating a Time-to-Event Distribution from Right-Truncated Data in an Epidemic: a Review of Methods. *Statistical Methods in Medical Research*. 2021; <https://doi-org.ezp.lib.cam.ac.uk/10.1177/09622802211023955>.
4. Keyfitz N and Caswell H. 5. *In: Applied Mathematical Demography*. 3rd ed. New York: Springer, 2005.
5. Rydevik G, Innocent G, Marion G et al. Using combined diagnostic test results to hindcast trends of infection from cross-sectional data. *PLOS Comput Biol* 2015; **12**: 1–19.
6. Hay J, Kennedy-Shaffer L, Kanjilal S et al. Estimating epidemiologic dynamics from cross-sectional viral load distributions. *Science* 2021; **373**: eabh0635.
7. Hay J, Kennedy-Shaffer L and Mina M. Viral loads observed under competing strain dynamics. *MedRxiv* 2021. DOI:10.1101/2021.07.27.21261224.
8. Muñoz A, Wang MC, Bass S et al. Acquired immunodeficiency syndrome (AIDS)-free time after human immunodeficiency virus type 1 (HIV-1) seroconversion in homosexual men. *Am J Epidemiol* 1989; **130**: 530–539.
9. Dubin N, Berman S, Marmor M et al. Estimation of time since infection using longitudinal disease-marker data. *Stat Med* 1994; **13**: 231–244.
10. Sommen C, Commenges D, Le Vu S, et al. Estimation of the distribution of infection times using longitudinal serological markers of HIV: implications for the Estimation of HIV incidence. *Biometrics* 2011; **67**: 467–475.
11. Bager P, Wohlfahrt J, Fonager J et al. Risk of hospitalisation associated with infection with SARS-CoV-2 lineage B.1.1.7 in Denmark: an observational cohort study. *Lancet Infect Dis* 2021; **21**: 1507–1517.
12. Bager P, Wohlfahrt J, Rasmussen M, et al. Hospitalisation associated with SARS-CoV-2 Delta variant in Denmark. *Lancet Infectious Diseases* 2021; **21**: 1351.
13. Challen R, Brooks-Pollock E, Read J, et al. Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *BMJ* 2021; **372**: n579.
14. Davies N, Jarvis C, CMMID COVID-19 Working Group, Edmunds W et al. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature* 2021; **593**: 270–274.
15. Fisman DN and Tuite AR. Evaluation of the relative virulence of novel SARS-CoV-2 variants: a retrospective cohort study in Ontario, Canada. *Can Med Assoc J* 2021; **193**: E1619–E1625.

16. Funk T, Pharris A, Spiteri G et al. Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. *Eurosurveillance* 2021; **26**: pii=2100348.
17. Grint D, Wing K, Williamson E et al. Case fatality risk of the SARS-CoV-2 variant of concern B.1.1.7 in England. *Eurosurveillance* 2021; **26**: pii=2100256.
18. Nyberg T, Twohig K, Harris R et al. Risk of hospital admission for patients with SARS-CoV-2 variant B.1.1.7: cohort analysis. *BMJ* 2021; **373**: n1412.
19. Patone M, Thomas K, Hatch R et al. Mortality and critical care unit admission associated with the SARS-CoV-2 lineage B.1.1.7 in England: an observational cohort study. *Lancet Infect Dis* 2021; **11**: 1518–1528.
20. Sheikh A, McMenamin J, Taylor B et al. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. *Lancet* 2021; **397**: 2461–2462.
21. Twohig K, Nyberg T, Zaidi A et al. Hospital admission and emergency care attendance risk for SARS-CoV-2 Delta (B.1.617.2) compared with Alpha (B.1.1.7) variants of concern: a cohort study. *Lancet Infect Dis* 2022; **22**: 35–42.
22. Whittaker R, Kristofferson A, Seppälä E et al. Trajectories of hospitalisation for patients infected with SARS-CoV-2 variant B.1.1.7 in Norway, December 2020–April 2021. *J Infect* 2021; **83**: e14–e17.
23. Verity R, Okell L, Dorigatti I et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* 2020; **20**: 669–677.
24. Annesi I, Moreau T and Lellouch J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Stat Med* 1989; **8**: 1515–1521.
25. Public Health England. Investigation of SARS-CoV-2 variants of concern in England: Technical briefing 5. Public Health England; 2021. <https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201>, accessed on 27/07/2021.
26. Davies N, Abbott S, Barnard R et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* 2021; **372**: eabg3055.
27. Rai B, Shukla A and Dwivedi L. Incubation period for COVID-19: a systematic review and meta-analysis. *J Public Health (Bangkok)* 2021. <https://doi-org.ezp.lib.cam.ac.uk/10.1007/s10389-021-01478-1>.
28. VanderWeele T and Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med* 2017; **167**: 268–274.
29. Dufouil C, Brayne C and Clayton D. Analysis of longitudinal studies with death and drop-out: a case study. *Stat Med* 2004; **23**: 2215–2226.
30. Rotnitzky A, Robins J and Scharfstein D. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J Am Stat Assoc* 1998; **93**: 1321–1339.
31. Rosenberg P and Gail M. Backcalculation of flexible linear models of the human immunodeficiency virus infection curve. *J R Stat Soc Ser C (Applied Statistics)* 1991; **40**: 269–282.

## Appendix: Proof of equations (5) and (10)

Since equation (5) is a special case of equation (10), it suffices to prove that the latter is implied by the equation  $f_T(t | I, X, H = 1) = f_T(t + c | I, X, H = 0)$ .

The probability distribution function of the observed data  $(X, T, H)$  at  $(x, t, h)$  is

$$f_{X,T,H}(x, t, h) = \int f_I(i) f_{H|I}(h | i) f_{X|I,H}(x | i, h) f_{T|I,X,H}(t | i, x, h) di.$$

By definition,  $T^0$  is sampled from

$$f_{T|I,X,H}(t | i, x, 0)$$

independently of  $T$ . So, the joint probability distribution function of  $(X, T^0, H)$  evaluated at  $(x, t, h)$  is

$$f_{X,T^0,H}(x, t, h) = \int f_I(i) f_{H|I}(h | i) f_{X|I,H}(x | i, h) f_{T|I,X,H}(t | i, x, 0) di. \quad (11)$$

Let  $T^* = T + cH$ . Then the joint probability distribution function of  $(X, T^*, H)$  evaluated at  $(i, t, h)$  is

$$f_{X,T^*,H}(x, t, h) = \int f_I(i) f_{H|I}(h | i) f_{X|I,H}(x | i, h) f_{T|I,X,H}(t - hc | i, x, h) di. \quad (12)$$

Assumption 1 can be written as

$$f_{T|I,X,H}(t - c | i, x, 1) = f_{T|I,X,H}(t | i, x, 0).$$



So, equation (12) becomes

$$f_{X,T^*,H}(x, t, h) = \int f_I(i) f_{H|I}(h | i) f_{X|I,H}(x | i, h) f_{T|I,X,H}(t | i, x, 0) di. \quad (13)$$

The right-hand sides of lines (13) and (11) are the same. So,  $(X, T^*, H)$  and  $(X, T^0, H)$  have the same joint distribution. Therefore, the conditional distribution of  $H$  given  $X$  and  $T^0$  is the same as the conditional distribution of  $H$  given  $X$  and  $T^*$ .