Estimating a Time-to-Event Distribution from Right-Truncated Data in an Epidemic: a Review of Methods

By Shaun R. Seaman, Anne Presanis and Christopher Jackson

MRC Biostatistics Unit, Institute of Public Health, Forvie Site, Robinson Way, Cambridge, CB2 0SR.
Corresponding author: Shaun Seaman
shaun.seaman@mrc-bsu.cam.ac.uk

Abstract

Time-to-event data are right-truncated if only individuals who have experienced the event by a certain time can be included in the sample. For example, we may be interested in estimating the distribution of time from onset of disease symptoms to death and only have data on individuals who have died. This may be the case, for example, at the beginning of an epidemic. Right truncation causes the distribution of times to event in the sample to be biased towards shorter times compared to the population distribution, and appropriate statistical methods should be used to account for this bias. This article is a review of such methods, particularly in the context of an infectious disease epidemic, like coronavirus disease (CoVID-19). We consider methods for estimating the marginal time-to-event distribution, and compare their efficiencies. (Non-)identifiability of the distribution is an important issue with right-truncated data, particularly at the beginning of an epidemic, and this is discussed in detail. We also review methods for estimating the effects of covariates on the time to event. An illustration of the application of many of these methods is provided, using data on individuals who had died with CoVID-19 by 5th April 2020.

Keywords: CoVID-19, Cox regression, failure time, identifiability, relative efficiency, right-truncation, survival analysis.

1 Introduction

Data on time to an event are said to be right truncated if they come from a set of individuals who have been randomly sampled from a population using a sampling mechanism that selects only individuals who have experienced the event by a given time, called the truncation time. An example is data on the time from onset of CoVID-19 symptoms to death collected from a sample of individuals who all developed symptoms and died by 5th April 2020. Among individuals in the population whose symptoms began on, say, 20th March, only those whose time from symptom onset to death was less than 16 days could be included in the sample. For these people, the truncation time is 16 days. Likewise, among those whose onset was on 31st March, only those whose time to death was less than five days could be included. Their truncation time is five days. Among the sampled individuals, the proportion whose time to death is less than t days (e.g. $t = 14$ days) will be greater than the proportion in the population of people who (eventually) die with CoVID-19. So, a naive estimate of the distribution of time to death in this population will be biased. Moreover, the average time to death in sampled individuals whose symptom onset was on 31st March will be shorter than the average time in those whose onset was on 20th March, even if time to death is independent of onset time in the population.

Ideally, we would have a random sample of individuals who experience an initial event (e.g. onset of CoVID-19 symptoms) that places them at risk of a final event of interest (e.g. death with CoVID-19) and follow them to see how many experience the final event and the times from initial to final event. However, this is not always feasible. For example, in the case of CoVID-19 many people will have experienced symptoms but never reported them. Another example is HIV/AIDS, where many people will not have discovered they are infected with HIV (the initial event) until they were diagnosed with AIDS (the final event).

The analysis of right-truncated data requires statistical methods that account for the fact that each of the sampled individuals must have experienced the final event by their truncation time. The purpose of this article is to review such methods in the context of an infectious disease epidemic, like CoVID-19.

Much of the statistical methodology for right-truncated data was developed in the 1980's and early 1990's in the context of the HIV/AIDS epidemic (e.g. [1, 2]). However, some of it predated the 1980's. For example, some methods described in this article involve the idea of reversing the time axis, i.e. counting backwards in time from the final event to the intial event. This reversal has the effect of making right-truncated data left-truncated. The Lynden-Bell (1971)[3] estimator is the extension of the original Kaplan-Meier estimator to handle left truncation. Some of the other methods are based on Turnbull's (1976) general algorithm for estimating a distribution function under general patterns of censoring or truncation [4]. Recently, the emergence of CoVID-19 has highlighted the need for methods that handle right truncation. Data from early in the epidemic have been used to estimate the distributions of time from infection to symptom onset, symptom onset to hospitalisation, symptom onset to death, and hospitalisation to death. Although many researchers have accounted for the right truncation of the data (e.g. [5, 6, 7], some have not done so or it is unclear whether they have (e.g. [8, 9, 10]).

An important issue when estimating the distribution of time to event using right-truncated data is that of identifiability. When the maximum truncation time that can be observed in the sample is less than the maximum time to event in the population, the time-to-event distribution can only be estimated up to a constant of proportionality. In the article we shall pay particular attention to this issue.

The structure of the article is as follows. In Section 2 we look at methods that estimate the marginal distribution of time from onset to death in the population of people who (eventually) die from the infection. Some of these methods model both time of onset, and hence truncation time, and time from onset to death, whereas others model only the time from onset to death. Some methods are parametric; others, non-parametric. We review these methods and investigate their asymptotic relative efficiency (ARE) for estimating the mean time from onset to death. Section 3 looks at methods for estimating the effect of covariates on the distribution of time to death, and for testing independence between covariates and time to death. These include parametric and semi-parametric methods. An illustration of the application of some of these methods to CoVID-19 deaths is described in Section 4. Section 5 contains some practical recommendations and a brief discussion of more general truncation patterns and of censoring.

2 Estimating marginal distribution of time to event

Let X and Y denote the times of an individual's initial and final events, respectively, with $0 \leq X \leq Y$. These two events could be, respectively, onset of CoVID-19 symptoms and death from CoVID-19. Alternatively, they could be, for example, infection with CoVID-19 and hospitalisation, or HIV seroconversion and AIDS diagnosis. Let $T = Y - X$ denote the individual's time from initial to final event; we call this the 'delay'. We assume that the support of X includes zero and that X and T are either both continuous or both discrete variables; if they are discrete, we suppose, without loss of generality, that they take integer values and we interpret integrals as sums. Unless stated otherwise, we shall assume T is independent of X. Let $f_T^*(t)$ and $F_T^*(t)$ denote, respectively, the probability density (or mass) function of T and the distribution function of T.

We obtain an i.i.d. sample, $(x_1, t_1), \ldots, (x_n, t_n)$, from the probability distribution of (X, T) given $X + T \le \tau$ for some $\tau > 0$, i.e. from

$$f_{X,T}(x, t \mid X + T \le \tau) = \frac{f_X(x) f_T^*(t)\ I(x + t \le \tau)}{\int_0^\tau f_X(x')\ F_T^*(\tau - x')\ dx'.}, \tag{1}$$

where $f_X(x)$ denotes the conditional probability density (or mass) function of X given $X \le \tau$, and $I(.)$ denotes the indicator function. If X and T are discrete, we shall assume, again without loss of generality, that τ is an integer. We should like to estimate $F_T^*(t)$, but first we need to discuss whether this is possible.

2.1 Identifiability

The maximum truncation time in the sample is τ. If this is greater than the maximum delay in the population, then $F_T^*(\tau) = 1$ and $F_T^*(t)$ can be estimated from the data. If, on the other hand, the maximum truncation time is less than the maximum delay in the population, then $F_T^*(\tau) < 1$ and $F_T^*(t)$ is only (non-parametrically) identified up to a constant of proportionality. This is because the sampling mechanism does not allow us to observe values of T greater than τ, and so the data do not tell us what proportion, $1 - F_T^*(\tau)$, of individuals in the population have $T > \tau$. This lack of identifiability is manifest in equation (1): if $F_T^*(\tau) < 1$ and the probability density (or mass) function and distribution function of T are instead $f_T(t) = f_T^*(t)/F_T^*(\tau)$ and $F_T(t) = F_T^*(t)/F_T^*(\tau)$ $(0 < t \le \tau)$, then the joint distribution of X and T given $X + T \le \tau$ is

$$\frac{f_X(x)\ \{f_T^*(t)/F_T^*(\tau)\}\ I(x + t \le \tau)}{\int_0^\tau f_X(x')\ \{F_T^*(\tau - x')/F_T^*(\tau)\}\ dx'}, \tag{2}$$

which is still equal to the right-hand side of equation (1), because the two $F_T^*(\tau)$ terms in (2) cancel out.

Unlike $f_T^*(t)$ and $F_T^*(t)$, the functions $f_T(t)$ and $F_T(t)$ are identifiable from the data $(x_1, t_1), \ldots, (x_n, t_n)$. They are, respectively, the conditional probability density (or mass) function and conditional distribution function of T given $T \le \tau$.

In the absence of other information or assumptions, $F_T(t)$ (or equivalently $f_T(t)$) is all we can estimate. Obviously, if we know from other information that $F_T^*(\tau) = 1$, then estimating $F_T(t)$ is the same as estimating $F_T^*(t)$. Likewise, if we know that, for example, $F_T^*(\tau) = 0.8$, then $F_T^*(t) = F_T(t) \times 0.8$, and so we can estimate $F_T^*(t)$. Alternatively, if we assume a parametric model for $F_T^*(t)$ and fit this model to the data, then $F_T^*(\tau)$ (and hence $F_T^*(t)$ for all t) will be estimated, because $F_T^*(\tau)$ is a deterministic function of the model parameters. However, as we shall now illustrate, this estimate of $F_T^*(\tau)$ relies on extrapolation of the parametric model beyond the range of the data.

2.2 Parametric modelling of joint distribution of X and T

2.2.1 A joint-conditional likelihood

We can estimate $F_T^*(t)$ by parameterising the distributions of X and T in terms of distinct parameters λ and θ respectively, and maximising the likelihood function corresponding to equation (1), viz.

$$L_1 = L_1(\lambda, \theta) = \prod_{i=1}^{n} \left\{ f_X(x_i; \lambda) f_T^*(t_i; \theta) \Big/ \int_0^\tau f_X(x'; \lambda) F_T^*(\tau - x'; \theta)\, dx' \right\} \tag{3}$$

For example, if we assume $f_X(x; \lambda) \propto \exp(\lambda x)$ and $T \sim \text{Gamma}(\theta_1, \theta_2)$, then

$$L_1 = \frac{1}{\left(\int_0^\tau \exp(\lambda x') \int_0^{\tau - x'} t'^{(\theta_1 - 1)} \exp(-\theta_2 t')\, dt'\, dx' \right)^n} \prod_{i=1}^{n} \exp(\lambda x_i) t_i^{\theta_1 - 1} \exp(-\theta_2 t_i).$$

We might call L_1 the ‘joint-conditional’ likelihood, because it is based on the joint distribution of X and T and is conditional on the final event occurring by time τ. Notice that equation (3) can be equivalently written as

$$L_1 = L_1(\lambda, \theta) = \prod_{i=1}^{n} \left\{ f_X(x_i; \lambda) f_T(t_i; \theta) \Big/ \int_0^\tau f_X(x'; \lambda) F_T(\tau - x'; \theta)\, dx' \right\}. \tag{4}$$

This highlights that L_1 depends on the distribution of T only through $F_T(t)$, its conditional distribution given $T \leq \tau$. In the example given immediately above, this distribution is

$$F_T(t) = \frac{\int_0^t t'^{(\theta_1-1)} \exp(-\theta_2 t')\ dt'}{\int_0^\tau t'^{(\theta_1-1)} \exp(-\theta_2 t')\ dt'} \qquad (0 < t \leq \tau). \tag{5}$$

Any distribution $F_T^*(t)$ for which equation (5) describes $F_T(t) = F_T^*(t)/F_T^*(\tau)$ would yield the same likelihood L_1. Thus, the data do not distinguish between $T \sim \text{Gamma}(\theta_1, \theta_2)$ and, for example, $F_T^*(t) = F_T(t)$. In the former case, $F^*(\tau) = \theta_2^{\theta_1} \int_0^\tau t'^{(\theta_1-1)} \exp(-\theta_2 t')\ dt'/\Gamma(\theta_1)$; in the latter case, $F^*(\tau) = 1$. Another possibility is that $F_T^*(t) = F_T(t) \times 0.001$, in which case $F^*(\tau) = 0.001$. In short, $F^*(\tau)$ could theoretically be anywhere between 0 and 1 and only $F_T(t)$ is non-parametrically identified.

In practice, we might believe that the parametric model accurately describes the whole of the delay distribution, or we might have additional information that makes us confident that $F_T^*(\tau) = 1$, or the data themselves might suggest that $F_T^*(\tau)$ equals 1 or is close to 1. The latter might be the case if there were a reasonably large number of sampled individuals with truncation times close to τ and all or almost all of these individuals had delays that were far less than τ. The absence of any longer delays among these people who could have been sampled even if their delays had been longer might lead one to conclude that longer delays are rare. However, caution is warranted, because there remains a possibility that the distribution of delays is bimodal, with a fraction of the population having much longer delays than the rest. For example, this might be the case for CoVID-19 if some very seriously ill patients are kept alive on ventilators for a long period of time.

We should be careful not to be misled by a good fit of the parametric model to the data. We can assess the fit on this parametric model only over the range $t \in [0, \tau]$. In all three examples of distributions that give rise to equation (5), this fit is perfect. In particular, if the proportion, $\theta_2^{\theta_1} \int_0^\tau t^{(\theta_1-1)} \exp(-\theta_2 t)\ dt/\Gamma(\theta_1)$, of the probability mass of the $\text{Gamma}(\theta_1, \theta_2)$ distribution that lies in the interval

$[0, \tau]$ is far from 1, then we are only judging the fit of the model in the early part of the distribution. Note that while it is common in time-to-event studies (without right truncation) for administrative censoring to prevent assessment of the fit of a parametric model in the tail of the distribution, this censoring prevents $F_T^*(t)$ from being (non-parametrically) identified only for times t after the censoring time.

2.2.2 An equivalent likelihood

So far, we have assumed the data $(x_1, t_1), \ldots, (x_n, t_n)$ represent a sample from the population and that the sample size n is fixed. An alternative framework involves assuming that initial events are generated from a (non-homogeneous) Poisson process with rate $h(t)$ at time t, the delays T are independent of the initial event times, and we observe all N individuals for whom $X + T \leq \tau$ [2]. Now N is a random variable with $N \sim \text{Poisson}(\alpha C)$, where $\alpha = \int_0^\tau h(t)\, dt$ and $C = C(\lambda, \theta) = \int_0^\tau f_X(x'; \lambda) F_T(\tau - x'; \theta)\, dx'$. Also, conditional on $N = n$, $(X_1, T_1), \ldots, (X_n, T_n)$ are i.i.d. with distribution given by equation (1) with $f_X(x) = h(x)/\alpha$. Hence the joint distribution of $(N, X_1, T_1, \ldots, X_N, T_N)$ is

$$\begin{aligned} f_{N,X_1,T_1,\ldots,X_N,T_N}(n, x_1, t_1, \ldots, x_N, t_N) &= \frac{\alpha^n \exp(-\alpha C)}{n!} \\ &\quad \times \prod_{i=1}^{n} f_X(x_i) f_T(t_i)\; I(x_i + t_i \leq \tau), \end{aligned}$$

which gives rise to the likelihood function

$$L_2 = L_2(\alpha, \lambda, \theta) = \alpha^n \exp\{-\alpha C(\lambda, \theta)\} \prod_{i=1}^{n} f_X(x_i; \lambda) f_T(t_i; \theta). \tag{6}$$

Kalbfleisch and Lawless (1989)[2] show that the maximum likelihood (ML) estimates and observed and estimated Fisher Information for (λ, θ) are the same whether they are obtained from L_1 or L_2. Thus, θ (and λ) can be estimated from whichever of these likelihoods is most computationally convenient, even if the assumption that the initial events are generated from a Poisson process does not hold (as might be the case, for example, if the initial events occur in clusters).

2.2.3 Applications

Kalbfleisch and Lawless (1989)[2] illustrate the use of L_2. They estimate the distribution of time T from infection to onset of AIDS in blood transfusion patients. A common assumption is that at the beginning of an epidemic, cases arise from a Poisson process with rate $h(t) = \lambda_0 \exp(\lambda t)$. This means that $\alpha = \lambda_0\{\exp(\lambda\tau) - 1\}/\lambda$ and $f_X(x; \lambda) = \lambda \exp(\lambda x)/\{\exp(\lambda\tau) - 1\}$. Kalbfleisch and Lawless make this assumption and assume that T has a Weibull distribution. They obtain the ML estimate of (λ, θ) using a Fisher scoring algorithm. Another example of the use of L_2 is given by Cox and Medley (1989)[11], who estimate the distribution of the time T taken for an AIDS diagnosis to be reported to the Communicable Disease Surveillance Centre. They allow the rate of AIDS diagnoses to be increasing sub-exponentially, by using $h(t; \lambda) = \lambda_0 \exp(\lambda_1 t + \lambda_2 t^2)$, and test the null hypothesis that $\lambda_2 = 0$. They consider several parametric models for the distribution of the reporting delay T.

Salje et al. (2020)[7] use equation (1), the basis of the likelihood L_1, to estimate $f_T(t)$, but do not use ML. In their application, T is the time from hospitalisation with CoVID-19 to death. They assume $f_X(x)$, the distribution of hospitalisation times in the population, is known and model $f^*_T(t)$ as a mixture of an exponential distribution and a log normal distribution. They estimate the parameters of this mixture distribution by finding the values that minimise the sum of squared differences between the distribution of $f_T(t \mid X + T \leq \tau)$ implied by equation (1) and the observed distribution of delays T in the sample.

2.3 Modelling the distribution of T

2.3.1 Parametric modelling

A third likelihood function that can be used to estimate θ comes from factorising $f_{X,T}(x, t \mid X + T \leq \tau)$ as $f_X(x \mid X + T \leq \tau) \times f^*_T(t \mid T \leq \tau - x)$ and using only the second factor. This yields the likelihood

$$L_3 = L_3(\theta) = \prod_{i=1}^{n} \frac{f_T^*(t_i;\ \theta)}{F_T^*(t_i;\ \theta)} = \prod_{i=1}^{n} \frac{f_T(t_i;\ \theta)}{F_T(\tau - x;\ \theta)}$$

For example, if we assume $T \sim \text{Gamma}(\theta_1, \theta_2)$, then

$$L_3 = \prod_{i=1}^{n} \frac{t_i^{\theta_1 - 1} \exp(-\theta_2 t_i)}{\int_0^{\tau - x_i} t'^{(\theta_1 - 1)} \exp(-\theta_2 t')\, dt'}. \tag{7}$$

We might refer to L_3 as the 'conditional-on-initial (event time)' likelihood. The issues regarding the identifiability of $F_T^*(t)$ that were discussed in Section 2.2 continue to apply when L_3, rather than L_1, is used.

2.3.2 Non-parametric modelling

L_3 is also the basis of a non-parametric estimator, $\hat{F}_T^{(\text{NP})}(t)$, of $F_T(t)$. This estimator is obtained by applying the familiar Kaplan-Meier estimator in reverse-time, i.e. to $\tau - T$. By reversing time, right truncation of T (i.e. T must be $\leq \tau - X$) becomes left truncation of $\tau - T$ (i.e. $\tau - T$ must be $\geq X$). Left truncation (or 'late entry') is easily handled by the original Kaplan-Meier estimator (or more accurately, the Lynden-Bell (1971)[3] estimator, which extends the Kaplan-Meier estimator to handle left truncation). For simplicity, we shall now describe the estimator $\hat{F}_T^{(\text{NP})}(t)$ for discrete T; Lagakos et al. (1988)[12] describe the corresponding estimator for continuous T.

Let $D_j = \sum_{i=1}^{n} I(T_i = j)$ be the number of delays observed to equal j $(j = 0, \ldots, \tau)$, and let $M_j = \sum_{i=1}^{n} I(T_i \leq j \leq \tau - X_i)$ be the number of delays observed to be at most j in those individuals whose truncation time $\tau - X$ is such that a delay of j would have been observed. Let $\hat{G}_j = D_j / M_j$, which is a consistent estimator of $P(T = j \mid T \leq j)$, the hazard in reverse time (the usual hazard in forward time is $P(T = j \mid T \geq j)$). Since $P(T \leq t \mid T \leq \tau) = P(T \leq t \mid T \leq t + 1) \times P(T \leq t + 1 \mid T \leq t + 2) \times \ldots \times P(T \leq \tau - 1 \mid T \leq \tau)$,

$$\hat{F}_T^{(\text{NP})}(t) = \prod_{j=t+1}^{\tau} (1 - \hat{G}_j) \tag{8}$$

is a consistent estimator of $F_T(t)$. $\hat{F}_T^{(\mathrm{NP})}(t)$ is asymptotically normally distributed and its variance can be consistently estimated using the following adaptation of the Greenwood formula for the variance of the Kaplan-Meier estimator [12]: $\widehat{\mathrm{Var}}\{\hat{F}_T^{(\mathrm{NP})}(t)\} = \{\hat{F}_T^{(\mathrm{NP})}(t)\}^2 \sum_{j=t+1}^{\tau} D_j/\{M_j(M_j - D_j)\}$. 95% confidence limits for $\hat{F}_T^{(\mathrm{NP})}(t)$ can be calculated as $\exp\big(-\exp\big[\hat{K}(t) \pm 1.96 \times \sqrt{\widehat{\mathrm{Var}}\{\hat{K}(t)\}}\big]\big)$, where $\hat{K}(t) = \log\{-\log \hat{F}_T^{(\mathrm{NP})}(t)\}$ and $\widehat{\mathrm{Var}}\{\hat{K}(t)\} = \{\log \hat{F}_T^{(\mathrm{NP})}(t)\}^{-2} \sum_{j=t+1}^{\tau} D_j/\{M_j(M_j - D_j)\}$. This guarantees that the confidence interval lie within the interval $[0, 1]$.

An alternative way to calculate the same estimator $\hat{F}_T^{(\mathrm{NP})}(t)$ is to define $Y_{x,t}$ (for $x = 0, \ldots, \tau$; $t = 0, \ldots, \tau - x$) as the number of sampled individuals with $X = x$ and $T = t$ and fit the model $Y_{xt} \sim \mathrm{Poisson}\{\exp(\lambda_x + \theta_t)\}$. Now $\hat{F}_T^{(\mathrm{NP})}(t) = \sum_{j=0}^{t} \exp(\hat{\theta}_j)/\sum_{j=0}^{\tau} \exp(\hat{\theta}_j)$, where $\hat{\theta}_j$ is the ML estimate of θ_j [1]. Obtaining an estimate of the variance of $\hat{F}_T^{(\mathrm{NP})}(t)$ this way is, however, more difficult.

The non-parametric estimator $\hat{F}_T^{(\mathrm{NP})}(t)$ can be used to check the fit of a parametric model for T. When doing this, it is important to compare $\hat{F}_T^{(\mathrm{NP})}(t)$ with the parametric estimate of $F_T(t)$, rather than of $F_T^*(t)$. For example, if we maximise the likelihood L_3 given by equation (7), we obtain both an estimate of $F_T^*(t) = P(T \leq t)$, the distribution function of a gamma distribution, and an estimate of $F_T(t) = F_T^*(t)/F_T^*(\tau) = P(T \leq t \mid T \leq \tau)$, the distribution function of a truncated gamma distribution. $\hat{F}_T^{(\mathrm{NP})}(t)$ should be compared with the latter.

It is important to notice that only the M_τ individuals with $X = 0$, and hence a truncation time of τ, contribute to the calculation of $\hat{G}_\tau$. Therefore, if M_τ is small, $\hat{G}_\tau$ will have large variance, which causes $\hat{F}_T^{(\mathrm{NP})}(t)$ also to have large variance not just for $t = \tau$ but for all values of t. In this case, it is advisable to choose a value $\tau^* < \tau$ such that M_{τ^*} is reasonably large and replace the estimator $\hat{F}_T^{(\mathrm{NP})}(t)$ of $F_T(t)$ by $\hat{F}_T^{(\mathrm{NP}.\tau^*)}(t) = \prod_{j=t+1}^{\tau^*}(1 - \hat{G}_j)$, which is an estimate of

$P(T \leq t \mid T \leq \tau^*)$.

2.3.3 Relative efficiency of likelihoods L_1 and L_3

Unlike L_1 (or equivalently L_2), L_3 does not require a model $f_X(x; \lambda)$ to be specified for the distribution of the initial event times. This eliminates the risk that such a model may be misspecified. However, it has the disadvantage that some of the information in the data is being discarded, which makes L_3 less efficient than L_1, especially when τ is small. Brookmeyer and Gail (1988)[13] found that the ML estimator of θ based on L_3 could be considerably less asymptotically efficient that the estimator based on L_3 when the density of X is known. Jewell (1990)[14] and Kalbfleisch and Lawless (1989)[2] also comment that the loss of efficiency from using L_3 can be considerable. In Section 2.5 we carry out a more extensive study of relative efficiency, comparing: i) L_1 treating the distribution of X as known; ii) L_1 treating this distribution as unknown; iii) L_3; and iv) the likelihood L_4 described in Section 2.4.

A simple example illustrates the information that L_1 uses but L_3 does not. Suppose X and T are discrete, with X equal to either 0 or 1 and T equal to either 0 or 1, and $\tau = 1$. We observe ten individuals with $(X, T) = (0, 0)$, ten individuals with $(X, T) = (0, 1)$, and no individuals with $(X, T) = (1, 0)$. If we use L_3, we would estimate $f_T(0) = f_T(1) = 0.5$. Now suppose that we know that $f_X(0) = f_X(1) = 0.5$. Unlike L_3, L_1 uses this information. Since we have only observed individuals with $X = 0$, it seems likely that there are quite a few individuals with $X = 1$ whom we have not observed. Since we have not observed them, they must all have $T = 1$. This suggests that $f_T(1) > 0.5$.

In Section 2.2, we considered the use of L_1 only with parametric models. If X and T are discrete variables (possibly formed by discretising continuous variables), the distribution of X can instead be modelled non-parametrically.

Writing $\lambda = (\lambda_0, \ldots, \lambda_\tau)$, with $\lambda_x = f_X(x)$, L_1 then becomes

$$L_1^{(\mathrm{NP})} = \frac{1}{\left(\sum_{x=0}^{\tau} \lambda_x \, F_T(\tau - x')\right)^n} \prod_{i=1}^{n} \lambda_{x_i} f_T(t_i; \theta). \tag{9}$$

Here, $F_T(t;\theta)$ can be a non-parametric or parametric model. Kalbfleisch and Lawless (1989)[2] show that the ML estimate of θ obtained from $L_1^{(\mathrm{NP})}$ is identical to that obtained from L_3. This is true whether T is modelled parametrically or non-parametrically and regardless of how finely time (if continuous) is discretised. This shows that when L_1 is more efficient than L_3, this greater efficiency comes from the modelling assumptions L_1 makes about the marginal distribution of X.

2.4 Modelling the distribution of X given Y

Verity et al. (2020)[5] proposed a fourth likelihood function, which arises from factorising $f_{X,T}(x, t \mid X + T \leq \tau)$ as $f_Y(y \mid X + T \leq \tau) \times f_X(x \mid Y = y)$, where $y = x + t$, and using only the second factor. This factor can be written as

$$f_X(x \mid Y = y) = \frac{f_Y(y \mid X = x; \theta) \times f_X(x; \lambda)}{\int_0^y f_Y(y \mid X = x'; \theta) \times f_X(x'; \lambda) \, dx'} \tag{10}$$

They assume that the initial events follow a Poisson process with rate $h(t) = \lambda_0 \exp(\lambda t)$ and that $T \sim \mathrm{Gamma}(\theta_1, \theta_2)$. Equation (10) then becomes

$$\begin{aligned} f_X(x \mid Y = y) &= \frac{(y-x)^{\theta_1 - 1} \exp\{-\theta_2(y-x)\} \times \exp(\lambda x) \times I(x \leq y)}{\int_0^y (y - x')^{\theta_1 - 1} \exp\{-\theta_2(y - x')\} \times \exp(\lambda x') \, dx'} \\ &= \frac{t^{\theta_1 - 1} \exp(-\theta_2 t) \times \exp(-\lambda t) \times I(t \leq y)}{\int_0^y t'^{(\theta_1 - 1)} \exp(-\theta_2 t') \times \exp(-\lambda t') \, dt'}. \\ &= \frac{t^{\theta_1 - 1} \exp\{-(\theta_2 + \lambda) t\}}{\int_0^y t'^{(\theta_1 - 1)} \exp\{-(\theta_2 + \lambda) t'\} \, dt'}, \end{aligned}$$

which is the density of a truncated gamma distribution with shape θ_1, rate $\theta_2 + \lambda$ and truncated to $[0, y]$. This gives the likelihood

$$L_4 = L_4(\lambda, \theta) = \prod_{i=1}^{n} \frac{t_i^{\theta_1 - 1} \exp\{-(\theta_2 + \lambda) t_i\}}{\int_0^{y_i} t'^{(\theta_1 - 1)} \exp\{-(\theta_2 + \lambda) t'\} \, dt'} \tag{11}$$

We might refer to L_4 as the 'conditional-on-final (event time)' likelihood. It is evident from equation (11) that only θ_1 and $\theta_2 + \lambda$ are identified. Practical use of L_4 therefore requires that λ be known.

Verity et al. (2020)[5], analysing data on 24 CoVID-19 deaths that occurred in China very early in the epidemic, assumed $\lambda = 0.14$ per day and estimated that the mean time from onset of symptoms to death was 19 (95% CI 16–50) days. They did not explain why they used L_4, rather than L_1. In view of the small sample, it was probably impractical to estimate both λ and θ. However, L_1 could have been used instead, also with λ fixed at 0.14. The appeal of L_4 may have been ease of use: it is just the likelihood of a truncated gamma distribution. Another advantage of L_4 relative to L_1, which may have been relevant, is that the former is a valid likelihood no matter how individuals are sampled, provided that the sampling probabilities depend only on Y. Validity of L_1 as a likelihood requires a simple random sample of individuals with $X + T \leq \tau$.

Equation (11) is derived from the assumptions that $f_X(x) \propto \exp(\lambda t)$ and T has a gamma distribution. The 'conditional-on-final' likelihood could also be derived from other assumed distributions for X and T, but would not have the form of a truncated gamma distribution.

2.5 Study of asymptotic relative efficiency of estimators

We carried out a study of the AREs of the ML estimators of the expected delay, $E(T)$, based on i) L_1 with known λ ('L_1^{kwn}'), ii) L_3, and iii) L_4 (with known λ), all relative to the ML estimator based on iv) L_1 with unknown λ ('L_1^{est}'). We assumed $f_x(x) \propto \exp(\lambda x)$ and $T \sim \text{Gamma}(\theta_1, \theta_2)$, and considered multiple scenarios defined by different combinations of values of λ, θ_1, θ_2 and τ. For the distribution of X, we used $\lambda = 0$, 0.035, 0.07, 0.14 and 0.28. For the distribution of T, we used $\theta_1 = 1$, 2, 5 and 10, and set $\theta_2 = \theta_1/19$, so that T has mean $E(T) = 19$. The mean of 19 (days) was chosen, because it was the estimate calculated by Verity et al. (2020) early in the CoVID-19 pandemic. τ was varied from 10 to 60. Note that the AREs are invariant to the choice of $E(T) = 19$, in the sense that they do not change if λ and θ_2 are both multiplied by some constant and τ is divided by the same constant (keeping θ_1 unchanged).

To calculate an ARE, we first calculated the asymptotic variance of each of the four ML estimators of $\theta = (\theta_1, \theta_2)$. Then we obtained the corresponding asymptotic variance of each of the four ML estimators of $E(T)$ using the Delta Method. The ratio of the asymptotic variances of two estimators of $E(T)$ is their ARE. The formulae used for these calculations are given in the Supplemental Materials.

Figure 1 shows the results. Each row corresponds to a different value of λ; each column, to different value of θ_1. The x-axis of each graph represents τ and the y-axis represents the ARE. The ARE of L_1^{kwn} (relative to L_1^{est}) varies from slightly over 1.0 to about 1.4. It increases with λ; it also increases with τ, at least when $\tau \leq 30$. The ARE of L_3 (relative to L_1^{est}) varies from 0.67 to almost 1. It decreases with increasing λ or θ_1, and mostly increases with τ. In particular, it is close to 1 when X is uniformly distributed ($\lambda = 0$) and T is exponentially distributed ($\theta_1 = 1$), and is 0.67 when $\lambda = 0.28$, $\theta_1 = 10$ and $\tau = 10$. When X is uniformly distributed ($\lambda = 0$), L_4 has exactly the same efficiency as L_3 (see Supplemental Material for proof). However, as λ increases, L_4 becomes relatively more efficient, and approaches the efficiency of L_1^{kwn}, especially for larger values of τ.

We also calculated the AREs for the ML estimators of the median of T. These were almost identical to the AREs for the expectation, $E(T)$ (see Supplemental Material).

3 Estimating and testing covariate effects

Let Z be a covariate or vector of covariates. We assume, unless stated otherwise, that Z is independent of X. We may be interested in testing whether T is independent of Z and/or estimating the effect of Z on T.

3.1 Parametric models

A parametric model $f_T^*(t \mid Z = z; \beta)$ can be specified for the distribution of T given Z. For example, T might be assumed to have a gamma distribution with $\log E(T \mid Z) = \beta_0 + \beta_1 Z$ and shape parameter β_2. Then $\beta = (\beta_0, \beta_1, \beta_2)$ can be estimated by maximising L_1, L_2 or L_3. Just as in the case with no covariates (Section 2.2.2), the ML estimate and Fisher information for β obtained from L_1 and L_2 are identical [2]. If Z is a function of X, this same method can still be applied using either L_1 or L_3. A likelihood ratio test or Wald test can be used for the null hypothesis that one or more elements of the vector β equal zero.

Kalbfleisch and Lawless (1989)[2] give an example of using L_2 with a Weibull regression model for the effect of age Z on time T from HIV infection to AIDS diagnosis.

3.2 Semi-parametric models

Brookmeyer and Liao (1990)[15] propose a generalisation of the discrete-time estimator $\hat{F}_T^{(\mathrm{NP})}(t)$ to estimate covariate effects. Fit the τ binomial regression models $g\{P(T = j \mid T \leq j \leq \tau - X, Z = z) = \beta_{0j} + \beta_1 z$ $(j = 1, \ldots, \tau)$ simultaneously, where g is a specified link function. Let $\hat{\beta}_{0j}$ and $\hat{\beta}_1$ denote the resulting estimates. Then calculate $\hat{F}_T(t \mid Z = z) = \prod_{j=t+1}^{\tau}\{1 - g^{-1}(\hat{\beta}_{0j} + \hat{\beta}_1 z)\}$. In the absence of covariates, this is equivalent to $\hat{F}_T^{(\mathrm{NP})}(t)$. Brookmeyer and Liao recommend using the complementary log log link, $g(p) = \log\{-\log(1-p)\}$, because the model then implies $F_T(t \mid Z = z) = \{F_T(t \mid Z = 0)\}^{\exp(\beta_1 z)}$, which provides an interpretation of β_1. The null hypothesis that β_1 (or a subvector of a vector β_1) equals zero can be tested using a likelihood ratio test or Wald test. This method can also be used when Z is a function of X.

Kalbfleisch and Lawless (1991)[16] extend the method of Brookmeyer and Liao (1990) to continuous time and derive score tests of the null hypothesis that

$\beta_1 = 0$. Following a similar approach, Lagakos et al. (1988)[12] had earlier described a log rank test for testing independence of T and a binary covariate Z.

The Poisson regression approach to calculating $\hat{F}_T^{(\mathrm{NP})}(t)$, described in Section 2.3.2, is extended by Brookmeyer and Damiano (1989)[1] to perform a likelihood ratio test of the global null hypothesis that $\beta_1 = 0$. This is done by including interaction terms in the Poisson model. This approach is less useful, however, for testing whether a set of covariates is conditionally independent of T given another set of covariates or for estimating covariate effects [15].

3.3 Proportional hazards models

A common assumption in both parametric and semi-parametric models for a time-to-event outcome is that hazards are proportional. Brookmeyer and Liao's (1990)[15] semi-parametric model, described in Section 3.2, assumes proportional hazards, but in reverse time. This differs from the usual proportional hazards assumption, which is in forward time and states that hazard ratios β_1 are constant over time and $1 - F_T^*(t \mid Z = z) = \{1 - F_T^*(t \mid Z = 0)\}^{\exp(\beta_1 z)}$. Brookmeyer and Liao's model implies that when $\beta_1 z > 0$ (respectively, $\beta_1 z < 0$), the (forward-time) hazard ratio comparing $Z = z$ to $Z = 0$ is initially greater (less) than one and decreases (increases) monotonically over time, becoming equal to one at time $\tau - 1$.

Finkelstein et al. (1993)[17] show that the (forward-time) proportional hazards assumption suffices to identify $F_T^*(\tau)$, provided that the hazard ratios of the covariates in the model do not all equal zero. When the hazard ratios all equal zero, $F_T^*(\tau)$ is not identified, just as in the non-parametric case with no covariates. Finkelstein et al. describe how to fit the semi-parametric proportional hazards model by ML. Provided that the hazard ratios of the covariates do not all equal zero, this provides estimates of the hazard ratios and $F_T^*(\tau)$. Unfortunately, the identification of $F_T^*(\tau)$ relies entirely on the proportional hazard assumption. If this does not hold, the estimate of $F_T^*(\tau)$ can be heavily biased. Moreover, if there is only one covariate Z in the model, its hazard ratio is estimated very imprecisely. Finkelstein et al. discourage the use of their method for estimating $F_T^*(\tau)$ or the hazard ratio of a single covariate. When there are multiple covariates with non-zero hazard ratios in the model (and possibly additional covariates with zero hazard ratio), Finkelstein et al. find that these hazard ratios can be estimated more precisely. However, it is unclear how big might be the effect of a small violation of the proportional hazards assumption on the bias of these estimates when $F_T(\tau) < 1$. Alioum and Commenges (1996)[18] suggest that when there is only one covariate, its hazard ratio could be estimated twice, once with $F_T^*(\tau)$ fixed at its lowest value considered plausible, and once at its highest plausible value. However, the resulting range of hazard ratios in their example is very wide.

Perhaps the main use of Finkelstein et al.'s method is for hypothesis testing with multiple covariates. Brookmeyer and Liao (1990), Lagakos et al. (1988) and Brookmeyer and Damiano (1989) described simpler hypothesis testing methods. However, if one wants to test whether one set of covariates is independent of T given another set of covariate, then Lagakos et al.'s log rank test cannot be used, and although the binomial regression model of Brookmeyer and Liao or the Poisson model of Brookmeyer and Damiano could be used, the parameters in these models do not have interpretations as standard hazard ratios, whereas those in Finkelstein et al.'s model do.

Finkelstein et al.'s method involves estimating the baseline hazard. Vakulenko-Lagun et al. (2019)[19] propose a simpler Cox regression approach with inverse probability weighting. This uses a modification of the Cox partial likelihood, which does not depend on the baseline hazard. This method involves weighting the observed individuals so that they represent both themselves and those individuals who were not observed because of the right truncation. The method requires either that $F_T^*(\tau) = 1$ or that $F_T^*(\tau \mid Z = 0)$ is known. If $F_T^*(\tau \mid Z = 0)$ is unknown, a sensitivity analysis can be performed, analysing the data using a range of plausible values of $F_T^*(\tau \mid Z = 0)$. This method can be applied using the R package *coxrt*. We are not aware of software being available for implementing Finkelstein et al.'s method.

4 Application to CoVID-19

The first case of CoVID-19 in the UK was reported on the 30th January 2020. Public Health England (PHE) receives reports every day of CoVID deaths from National Health Service England, the Demographics Batch Service (DBS) and Health Protection Teams (HPT). DBS and HPT also report date of symptom onset, when available, for these deceased individuals. Here we illustrate some of the methods discussed above using data available early in the epidemic, specifically at 9th April. We estimate the distribution of time ('delay') from symptom onset to death and investigate the effects of sex and age on this distribution. This is intended only as a simple illustration of methods; results should be interpreted with caution.

To allow for reporting delays, we exclude deaths occurring between 6th and 9th April; around 80% of deaths are reported within four days [20]. Of the remaining 7415 deaths, the symptom onset date was known for 316 (4.3%). Of these 316, we excluded 12 because of missing sex or age. The remaining 304 constitute the sample we shall use. 180 were male and 124 female; 25 were aged under 65, 33 aged 65–74, 100 aged 75–84, and 146 were aged over 85. Figure 2 shows the distribution of onset times X. The distribution is skewed, with most onsets being in the second half of March. This reflects exponential growth in the early phase of the epidemic. The earliest observed onset date was 1st February (time zero); the two individuals with onset on that day have the maximum truncation time of

$\tau = 64$ days. Only 13 other individuals have onsets before 2nd March (time 30), and so truncation times $\tau - X$ greater than 34 days; most truncation times are less than 20 days.

The mean delay in the sample is 7.1 days (range: 0–52 days). As only those who die by 5th April can be included in the sample (right truncation), the mean in the population could be much greater. Using the R package *flexsurv* [21], we estimated the distribution of delays in the population by fitting two parametric models: a gamma distribution and a log normal distribution. Each was fitted in four ways: by maximising L_1 with unknown λ ('L_1^{est}'), L_1 with known λ ('L_1^{kwn}'), L_3 and L_4. For L_1^{kwn} and L_4, we assumed $\lambda = 0.14$, the estimate calculated early in the epidemic by Verity et al. (2020)[5]. Figure 3 shows the estimates of the survival distribution (i.e. $1 - F_T^*(t)$). These are quite diverse. For example, estimated survival at 30 days varies from 0.21 to 0.88. However, as expected, these estimates are all greater than the proportion (0.02) of the sample who have delays greater than 30 days. The estimates from the log normal model are systematically greater than those from the gamma model, and all estimates have wide associated confidence intervals. For a given model, the estimates from L_3 and L_1^{est} are similar to one another, and those from L_1^{kwn} and L_4 are almost identical to one another. This is perhaps not surprising, given our findings in Section 2.5 for the gamma distribution. There we found that: 1) when $\lambda \geq 0.07$, L_3 and L_1^{est} have similar asymptotic efficiency, unless τ is small compared to the mean delay $E(T)$; and 2) L_3 and L_1^{est} have similar asymptotic efficiencies when $\lambda < 0.14$ and the shape parameter of the gamma distribution equals 1. For the CoVID data, the estimates of $E(T)$ from the gamma model varied from 19 for L_3 or L_1^{est} to 36 for L_1^{kwn} or L_4, and the estimates of the shape varied from 1.16 to 1.24.

For both the gamma and log normal models, the estimates of λ from L_1^{est} were 0.11 (95% CI 0.09–0.12). The difference between this estimate and the assumed value of $\lambda = 0.14$ used by L_1^{kwn} and L_4 may explain why the estimates of survival from L_1^{est} and L_3 are lower than those from L_1^{kwn} and L_4. Compared to $\lambda = 0.11$, $\lambda = 0.14$ implies a later average onset time, $E(X)$, in the population, and hence a higher proportion of people in that population that has delays too long to be sampled ($T > \tau - X$). This implied greater extent of right truncation implies a greater difference between the mean delay in the sample and the mean delay in the population.

We used the non-parametric estimate of survival conditional on $T \leq \tau^*$, i.e. $1 - \hat{F}_T^{(\text{NP})}(t)$, to assess the fit of the parametric models. To avoid having wide confidence intervals for $\hat{F}_T^{(\text{NP})}(t)$, we used $\tau^* = 31$. This ensures that at least $M_{\tau^*} = 20$ individuals had truncation times of τ^* days. Figure 4 compares $1 - \hat{F}_T^{(\text{NP})}(t)$ with the corresponding parametric estimates of survival conditional on $T \leq \tau^*$, i.e. with the estimates of $1 - F_T(t)/F_T(\tau^*)$. The fit of the models using L_1^{est} and L_3 is reasonable during the first 15 days, but less good thereafter. The fit when L_1^{kwn} or L_4 is used is considerably worse, presumably because the data do not support the choice of $\lambda = 0.14$. Note that the difference between the *conditional* (on $T \leq \tau^*$) survivor curves estimated from the gamma model and the corresponding estimates from the log normal model is much less obvious than the differences between the corresponding *unconditional* survivor functions (shown in Figure 3). This illustrates the point made in Sections 2.1 and 2.2.1 that two models can give similar estimates of $F_T(t)$ and yet very different estimates of $F_T^*(t)$.

Next we fitted the gamma and log normal models with sex and age group (0–64, 65–74, 75–84 and ≥ 85) as covariates, again using the R package *flexsurv*. This was done using the likelihood L_3; *flexsurv* does not currently allow L_1 or L_4 to be used with covariates. The gamma (respectively, log normal) model assumes that the log rate (respectively, mean of the log delay) is a linear function of the covariates. Delays were estimated to be longer for males than females and for younger than for older people. Both effects were borderline-significant in the gamma model. Neither was significant in the log normal model, although age was found to be significant when a trend test was used (see Supplemental Materials).

If the gamma or log normal model is misspecified, tests of covariate effects may not be valid. Brookmeyer and Liao's method allow tests that do not depend on parametric assumptions. Using this method, we found again that delays are longer for males and younger people. Neither effect reached statistical significance, although age was significant when a trend test was used (see Supplemental Materials).

Finally, we used Vakulenko-Lagun's et al.'s Cox regression method. This requires that either $F_T^*(\tau) = 1$ or we specify a value for $F_T^*(t \mid Z = 0)$. Figure 3 suggests it is unlikely that $F_T^*(\tau)$ is close to 1. Using the R package *coxrt*, we fitted the model that included the covariate sex, varying $F_T^*(t \mid Z = 0)$ over the full range from 1.0 to 0.1, where here $Z = 0$ means male. The method uses inverse probability weighting to account for the right truncation, with the weights being functions of a Kaplan-Meier estimate of the truncation time distribution. As explained in the Supplemental Materials, this Kaplan-Meier estimate could not be calculated for our data set, until we excluded the 12 individuals with onset times before 1st March. This makes 1st March the new time zero, and so τ now equals 36. Figure 5 shows how the estimated log hazard ratio associated with being female changes as $P(T \geq \tau \mid Z = 0) = 1 - F_T^*(\tau \mid Z = 0)$ changes. Females are estimated to have a greater hazard than males (and hence shorter mean delay) and the estimated log hazard ratio increases as $P(T \geq \tau \mid Z = 0)$ increases. However, the confidence intervals, calculated using 1000 bootstrap samples, indicate that this effect is not significant, at least not until $P(T \geq \tau \mid Z = 0) = 0.1$. There were convergence problems: the percentage of bootstrap samples for which convergence was not achieved was 0.0% when $P(T \geq \tau \mid Z = 0)$ is 0.2 or less, 0.3% when it is 0.3, 1.4% when it is 0.5, 3.1% when it is 0.7, and 7.9% when it is

0.9. This may make the estimated confidence intervals unreliable for the largest values of $P(T \geq \tau \mid Z = 0)$. We also tried to fit the model with age group, both as a four-level unordered categorical variable and an ordinal categorical variable with linear effect, but the fitting algorithm did not converge. We could, however, fit the model with age as a binary variable, although again with some convergence problems in the bootstrap samples (see Supplemental Materials).

5 Discussion

We have considered maximum likelihood estimation of the marginal distribution of the delay T, using four likelihoods. Likelihoods L_1 and L_2 are based on the joint distribution of T and the time of the initial event X; L_3 on the distribution of T given X; and L_4 on the distribution of T given the time of the final event $Y = X + T$. Estimates from L_1 and L_2 are identical. L_3 has the advantage of not requiring a model for $f_X(x)$ but the disadvantage of yielding the least efficient estimates. L_4 requires $f_X(x)$ to be known exactly. When $f_X(x)$ is known, L_1 is more efficient than L_4. L_1 also has the advantage over L_4 that it can be used when $f_X(x)$ is a function of unknown parameters. However, L_4 has the advantage that, unlike L_1 and L_3, it yields valid estimates even when the sampling probabilities depend on the actual values of Y, rather than only on whether $Y \leq \tau$.

In our study of asymptotic efficiency, we found the ARE of L_3 relative to L_1 varied between 0.67 and 1 when $f_X(x)$ was unknown. Because L_3 does not use information on $f_X(x)$, these AREs became more marked when $f_X(x)$ was known, varying between 0.58 and 0.92. AREs of 0.67 and 0.58 correspond to reductions in sample size of 18% and 24%, respectively. These AREs were calculated assuming a gamma distribution for the delay and exponential growth over time in the number of initial events. In the early phase of an epidemic, the assumption of exponential growth may be tenable, but it is unlikely to hold later on. More research on the AREs when X and T have other distributions is warranted, as well as on finite-sample relative efficiencies.

The non-parametric estimator, $\hat{F}_T^{(\mathrm{NP})}(t)$, of the delay distribution has the attraction of not requiring distributional assumptions. It does, however, only estimate the distribution of T conditional on $T \leq \tau$; the unconditional distribution of T is estimable only by using parametric assumptions. One use of $\hat{F}_T^{(\mathrm{NP})}(t)$ is to assess the fit of parametric models over the range $t \in [0, \tau]$. However, the confidence intervals associated with $\hat{F}_T^{(\mathrm{NP})}(t)$ may be very wide at the beginning of an epidemic, when the numbers of sampled individuals with large truncation times $\tau - X$ may be small.

To estimate the effects of covariates on the delay, and to test whether these effects are non-zero, L_1 or L_3 can be used with parametric models. Brookmeyer and Liao's (1990) semi-parametric model can also be used, particularly for the purpose of testing whether the effect of a single covariate is zero. The semi-parametric Cox regression method of Vakulenko-Lagun et al. (2019) allows hazard ratios of covariates to be estimated under a standard proportional hazards assumption. However, it does assume that the covariates are independent of the truncation time $\tau - X$, and hence of X. Moreover, it requires that $F_T^*(\tau)$ be equal to one or that an interval can be specified within which $F_T^*(\tau)$ is believed to lie. As this interval becomes wider, the uncertainty in the hazard ratios increases. We also had some convergence problems when fitting these models to the CoVID-19 data (Section 4).

The R package *flexsurv* can be used to fit parametric models using L_1, L_3 and L_4, and also to calculate $\hat{F}_T^{(\mathrm{NP})}(t)$. Brookmeyer and Liao's (1990) method can be applied using any software for fitting generalised linear models. The R package *coxrt* applies the method of Vakulenko-Lagun et al. (2019). We have focussed on

ML estimation, but also Bayesian analyses can be carried out using the likelihoods L_1, L_2, L_3 and L_4. Indeed, the analysis of Verity et al. (2020) was Bayesian, using L_4 with informative priors.

In addition to being right-truncated, Y may be censored. This is easily handled in parametric models by replacing $f_T^*(t_i)$ in L_1 and L_3 by $F_T^*(t_i^U) - F_T^*(t_i^L)$, where $[t_i^L, t_i^U]$ is the interval within which individual i's delay is known to lie. If individual i is left-censored, $t_i^L = 0$; if right-censored, $t_i^U = \infty$. The estimator $\hat{F}_T^{(\mathrm{NP})}(t)$ is easily extended to allow left censoring of Y [22]. The non-parametric estimator of $F_T(t)$ under general censoring of both X and Y and right truncation of Y is described by Sun (1995) [23]. Alioum and Commenges (1996) generalise Finkelstein et al.'s (1993) method to allow interval censoring of Y. Double truncation (i.e. simultaneous left and right truncation) of Y is addressed by, among others, [4, 24, 25, 18, 23, 26, 27]. The R package *DTDA* [28] can be used to calculate the non-parametric estimator of $F_T(t)$ when Y is double truncated (it can also be used when Y is only right-truncated, but calculating $\hat{F}_T^{(\mathrm{NP})}(t)$ using equation (8) is faster). Brookmeyer and Gail (1988)[13] showed that when Y is double-truncated and the distributions of X and T are modelled parametrically, the efficiency gain from using L_1 rather than L_3 to estimate $f_T(t)$ could be considerably greater than the efficiency gains we showed in Section 2.5. In the Supplemental Materials, we extend our study of ARE to double-truncated data and replicate Brookmeyer and Gail's finding.

$F_T(t)$ and $F_T^*(t)$ describe the distribution of T in the population of individuals who will (eventually) experience the final event. They do not describe what proportion of the population will never experience the final event. An alternative sampling mechanism randomly samples individuals who experience any one of a number of mutually exclusive types of final event by time τ. For example, one might have a random sample from the population of individuals who develop CoVID-19 symptoms and go on to die or recover by time τ. This situation of competing risks and right truncation is discussed by Hudgens et al. (2001)[29] and de Una-Alvarez (2020)[30], who describe how to estimate cumulative incidence functions. These functions describe what proportion of individuals who die or recover by time τ will die by time t and what proportion will recover by time t ($t \leq \tau$).

In addition to the problems of right-truncation, censoring and competing risks, other issues can complicate the estimation of a time-to-event distribution early in an epidemic. Overton et al.[31] describes some of these, which include the possibility that some individuals who experience the final event leave the country before being detected, changes over time in the definition of the final event or in the format of the data being collected, and delays in reporting the final event.

Acknowledgements

This work was funded by UKRI Medical Research Council (Seaman: MC_UU_00002/10 and Presanis, Jackson: MC_UU_00002/11) and via a grant from the MRC UKRI / DHSC NIHR COVID-19 rapid response call (Presanis, grant ref: MC_PC_19074), and was supported by the NIHR Cambridge BRC. We acknowledge the support of the Public Health England (PHE) Epidemiology and Modelling Cells in providing and discussing the data used. The views expressed are those of the authors and not necessarily those of PHE, the NHS, the NIHR or the Department of Health and Social Care.

Conflicts of interest

The Authors declare that there is no conflict of interest.

Data Availability

Data used in the analyses of this paper are available upon the signing of a data-sharing agreement with Public Health England.

References

[1] Brookmeyer R and Damiano A. Statistical methods for short-term projections of AIDS incidence. *Statistics in Medicine* 1989; **8**: 23–34.

[2] Kalbfleisch JD and Lawless JF. Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association* 1989; **84**: 360–372.

[3] Lynden-Bell D. A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices of the Royal Astronomical Society* 1971; **155**: 95–118.

[4] Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* 1976; **38**: 290–295.

[5] Verity R, Okell LC, Dorigatti I, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infectious Disease* 2020; **20**: 669–677.

[6] Pellis L, Scarabel F, Stage HB, et al. Challenges in control of COVID-19: short doubling times and long delay to effect of interventions. 2020; MedRxiv, https://doi.org/10.1101/2020.04.12.20059972 (11/06/2020).

[7] Salje H, Kiem CT, Lefrancq N, et al. Estimating the burden of SARS-CoV-2 in France. *Science* 2020; **369**: 208–211.

[8] Backer JA, Klinkenberg D, and Wallinga J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance* 2020; **25**: 1–6.

[9] Lauer SA, Grantz KH, Qifang B, et al. The incubation period of Coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine* 2020; **172**: 577–582.

[10] Wu JT, Leung K, Bushman M, et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nature Medicine* 2020; **26**: 506–510.

[11] Cox DR and Medley GF. A process of events with notification delay and the forecasting of AIDS. *Philosophical Transactions of the Royal Society of London*, 1989; **325**: 135–145.

[12] Lagakos SW, Barraj LM, and De Gruttola V. Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika* 1988; **75**: 515–523.

[13] Brookmeyer R and Gail MH. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association* 1988; **83**: 301–308.

[14] Jewell NP. Some statistical issues in studies of the epidemiology of AIDS. *Statistics in Medicine* 1990; **9**: 1387–1416.

[15] Brookmeyer R and Liao J. The analysis of delays in disease reporting: Methods and results for the Acquired Immunodeficiency Syndrome. *American Journal of Epidemiology* 1990; **132**: 355–365.

[16] Kalbfleisch JD and Lawless JF. Regression models for right truncated data with applications to AIDS incubation times and reporting lag. *Statistica Sinica* 1991; **1**: 19–32.

[17] Finkelstein DM, Moore DF, and Schoenfeld DA. A proportional hazards model for truncated AIDS data. *Biometrics* 1993; **49**: 731–740.

[18] Alioum A and Commenges D. A proportional hazards model for arbitrarily censored and truncated data. *Biometrics* 1996; **52**: 512–524.

[19] Vakulenko-Lagun B, Mandel M, and Betensky RA. Inverse probability weighting methods for Cox regression with right-truncated data. *Biometrics* 2019; **76**: 484–495.

[20] Seaman SR, Samartsidis P, Kall M, and DeAngelis D. Nowcasting CoVID-19 deaths in England by age and region. 2020; MedRxiv, https://doi.org/10.1101/2020.09.15.20194209 (16/09/2020).

[21] Jackson CH. flexsurv: a platform for parametric survival modeling in R. *Journal of Statistical Software* 2016; **70**: 1–33.

[22] Cui J. Nonparametric estimation of a delay distribution based on left-censored and right-truncated data. *Biometrics* 1999; **55**: 345–349.

[23] Sun J. Estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. *Biometrics* 1995; **51**: 1096–1104.

[24] Frydman H. A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *Journal of the Royal Statistical Society, Series B* 1994; **56**: 71–74.

[25] Efron B and Petrosian V. Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association* 1999; **94**: 824–834.

[26] Mandel M, Álvarez de Uña J, Simon DK, and Betensky RA. Inverse probability weighted Cox regression for doubly truncated data. *Biometrics* 2018; **74**: 481–487.

[27] Rennert L and Xie SX. Cox regression model with doubly truncated data. *Biometrics* 2018; **74**: 725–733.

[28] Moreira C, Álvarez de Uña J, and Crujeiras RM. DTDA: An R package to analyze randomly truncated data. *Journal of Statistical Software* 2010; **37**: 1–20.

[29] Hudgens MG, Satten GA, and Longini IM. Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics* 2001; **57**: 74–80.

[30] Álvarez de Uña J. Nonparametric estimation of the cumulative incidences of competing risks under double truncation. *Biometrical Journal* 2020; **62**: 852–867.

[31] Overton CE, Stage HB, Ahmad S, et al. Using statistics and mathematical modelling to understand infectious disease outbreaks: COVID-19 as an example. *Infectious Disease Modelling* 2020; **5**: 409–441.

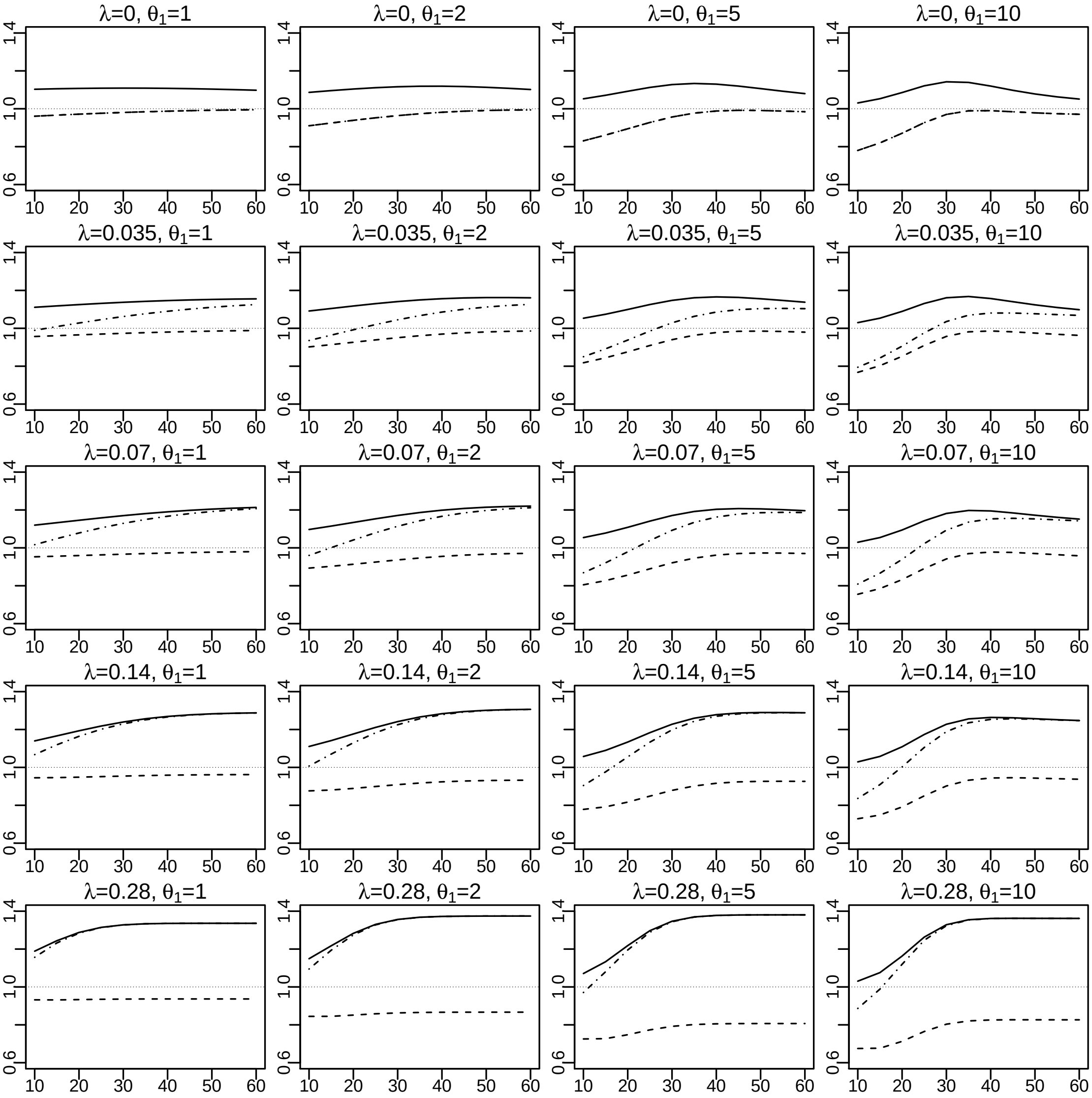

Figure 1: For each of five values of λ and of four values of θ_1, graph show the asymptotic relative efficiency (ARE) of the estimator of $E(T)$ based on i) L_1^{kwn} (solid line), ii) L_3 (broken line) and iii) L_4 (dot-dash line) all compared to iv) L_1^{est}. The x-axis of each graph is τ and the y-axis is the ARE.

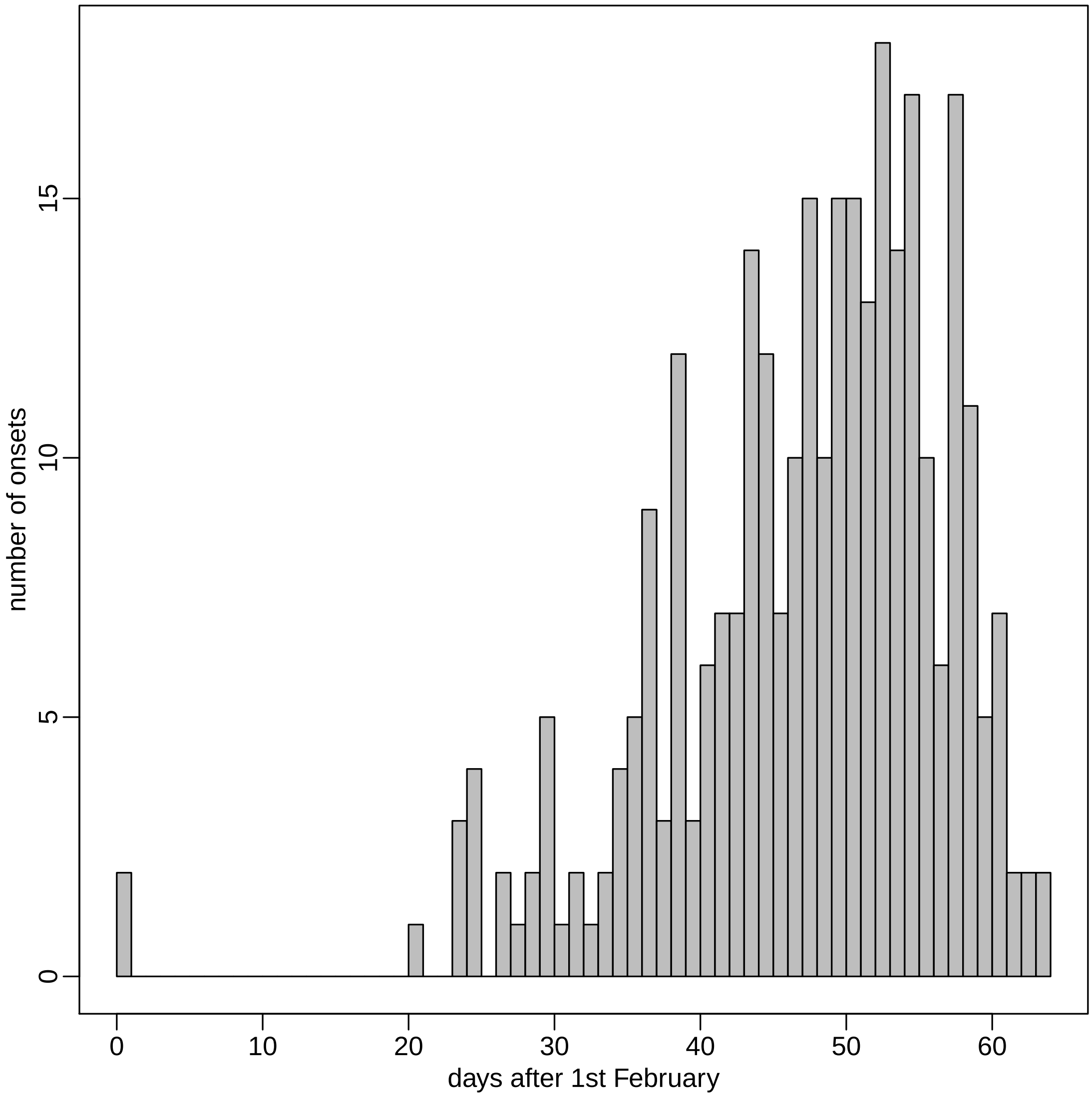

Figure 2: Distribution of symptom onset times in the sample of 304 individuals

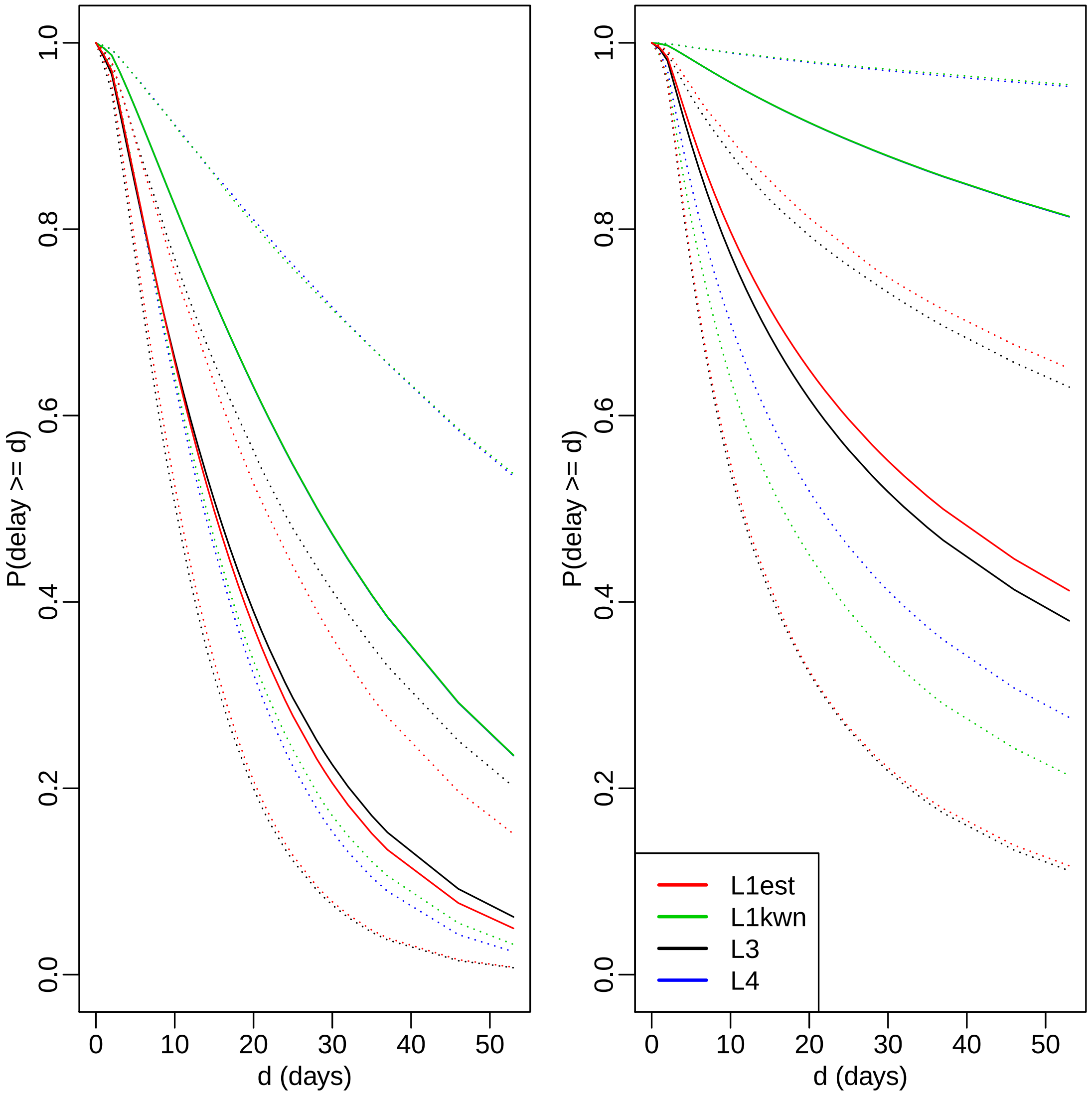

Figure 3: Estimated survival curves from the gamma model (left) and log normal model (right), obtained using likelihoods L_1^{est}, L_1^{kwn}, L_3 and L_4. Dotted lines represent 95% confidence intervals. (Estimates using L_1^{kwn} and L_4 are so close they may be hard to distinguish.)

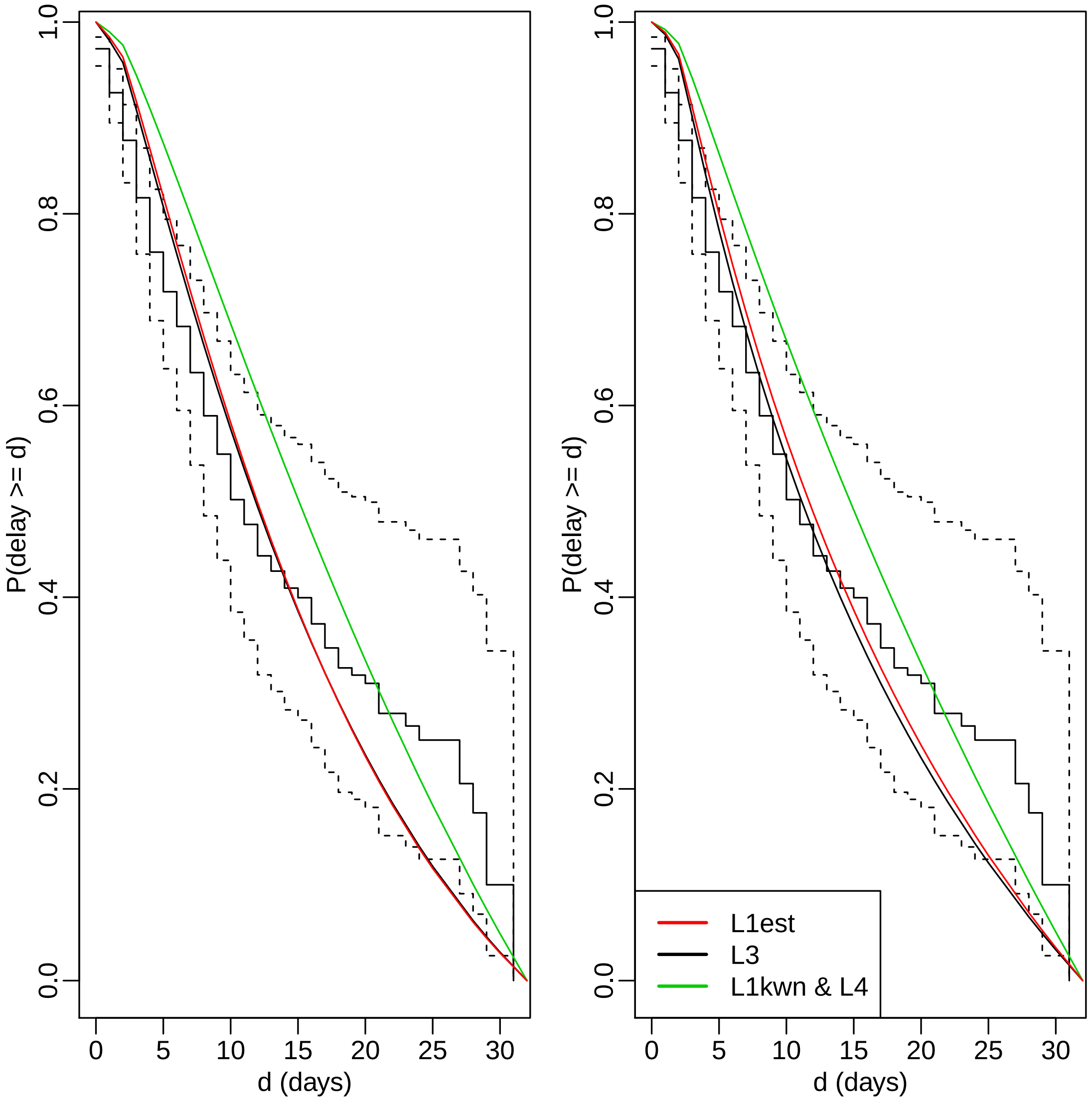

Figure 4: Comparison of non-parametric estimate (step function) of survival conditional on delay being less than 31 days with corresponding estimates from the gamma model (left) and log normal model (right). Dotted lines represent 95% confidence intervals for the non-parametric estimate. (Estimates using L_1^{kwn} and L_4 are so close that they are shown by a single line, and estimates using L_1^{est} and L_3 are so close that they may be difficult to distinguish.)

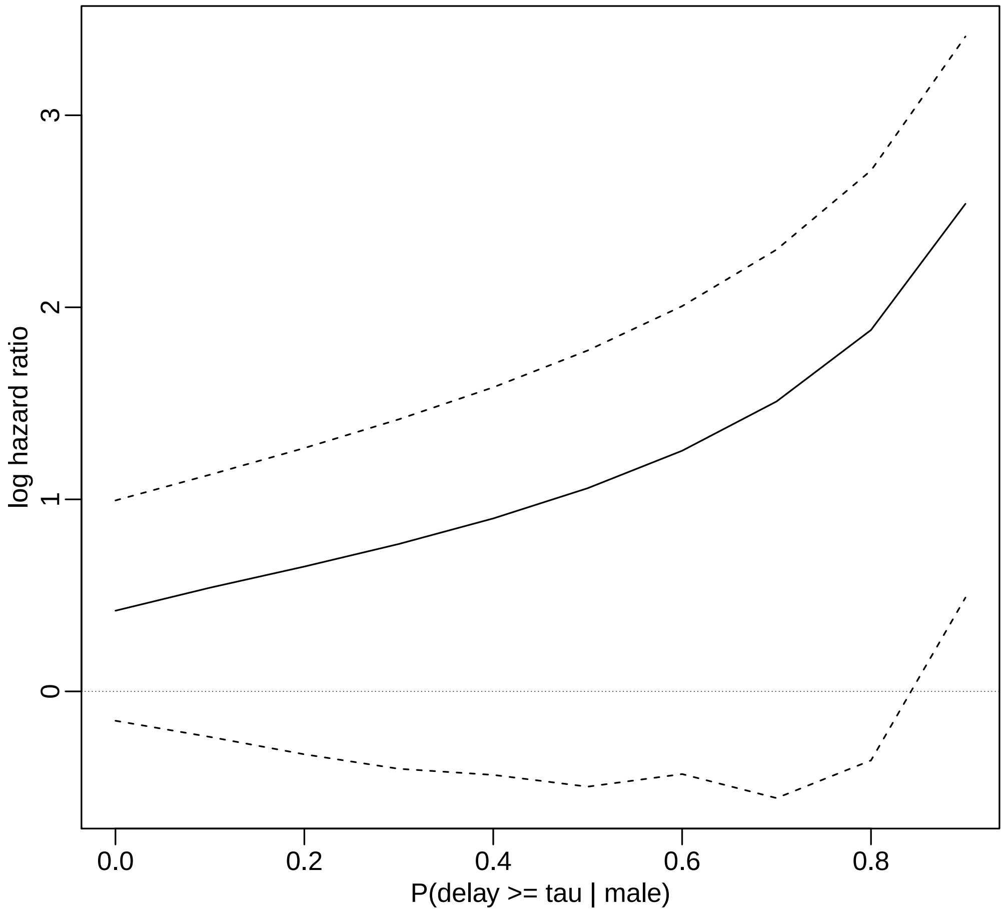

Figure 5: Estimate of log hazard ratio associated with sex=female as a function of $P(T \geq \tau \mid \text{sex=male})$. Dotted lines indicate 95% confidence limits calculated by bootstrap; these may be unreliable when $P(T \geq \tau \mid \text{sex=male})$ is large (see text).