Estimating distribution of length of stay in a multi-state model conditional on the pathway, with an application to patients hospitalised with Covid-19

Ruth H. Keogh[1,*], Karla Diaz-Ordaz[1], Nicholas P. Jewell[1], Malcolm G. Semple for the ISARIC4C Investigators[2,a], Liesbeth C. de Wreede[3], and Hein Putter[3]

[1]Department of Medical Statistics and Centre for Statistical Methodology, London School of Hygiene & Tropical Medicine, United Kingdom. ruth.keogh@lshtm.ac.uk
[2]NIHR Health Protection Research Unit, Institute of Infection, Veterinary and Ecological Sciences, Faculty of Health and Life Sciences, University of Liverpool, Liverpool, UK.
[3]Leiden University Medical Center, Leiden, Netherlands.
[a]The full list of ISARIC4C Investigators is available at: https://isaric4c.net/about/authors/
[*]Corresponding author: Ruth Keogh, ruth.keogh@lshtm.ac.uk

Abstract

Multi-state models are used to describe how individuals transition through different states over time. The distribution of the time spent in different states, referred to as 'length of stay', is often of interest. Methods for estimating expected length of stay in a given state are well established. The focus of this paper is on the distribution of the time spent in different states conditional on the complete pathway taken through the states, which we call 'conditional length of stay'. This work is motivated by questions about length of stay in hospital wards and intensive care units among patients hospitalised due to Covid-19. Conditional length of stay estimates are useful as a way of summarising individuals' transitions through the multi-state model, and also as inputs to mathematical models used in planning hospital capacity requirements. We describe non-parametric methods for estimating conditional length of stay distributions in a multi-state model in the presence of censoring, including conditional expected length of stay (CELOS). Methods are described for an illness-death model and then for the more complex motivating example. The methods are assessed using a simulation study and shown to give unbiased estimates of CELOS, whereas naive estimates of CELOS based on empirical averages are biased in the presence of censoring. The methods are applied to estimate conditional length of stay distributions for individuals hospitalised due to Covid-19 in the UK, using data on 42980 individuals hospitalised from March to July 2020 from the COVID19 Clinical Information Network.

Keywords: Multi-state model, length of stay, Covid-19, Illness-death model, State occupation.

1 Introduction

Multi-state models are used to describe how individuals transition through different states over time. The simplest multi-state model is the illness-death model, depicted in Figure 1A. Quantities of interest in multi-state modelling analyses include rates of transition from one state to another, the probability of being in a given state at a given time after entering another state, and the expected length of time spent in a given state. Analysis methods include non-parametric methods, including the Aalen-Johansen estimator, and methods that enable estimation of the impact of predictors on these quantities, including extensions to the Cox model, and fully-parametric methods. Andersen and Keiding (2002) and Putter et al. (2007) provide overviews of multi-state modelling methods, and details of the underlying theory are provided in the books by Andersen et al. (1993) and Aalen et al. (2008).

In this paper we consider descriptive analysis of multi-state systems, with a focus on estimating the distribution of the time spent in different states in a multi-state model, which is often referred to as 'length of stay', or 'state occupation time'. Beyersmann and Putter (2014) described non-parametric methods for estimating expected length of stay in multi-state models. Our interest is in the distribution of the time spent in different states *conditional on the complete pathway taken through the states*, which we refer to as *conditional length of stay*. In the illness-death model depicted in Figure 1A there are two possible complete pathways through the states: the pathway from state 1 to state 3, and the pathway from state 1 to state 2 to state 3. In the illness-death model therefore, conditional length of stay provides information about: (i) time spent in the healthy state among individuals who do not transition through the illness state (complete pathway: state 1 to state 3), (ii) time spent in the healthy state among individuals who do transition through the illness state (complete pathway: state 1 to state 2 to state 3), (iii) time spent in the illness state.

The concept of conditional length of stay involves conditioning on future events, which is rarely appropriate in analyses of times-to-event (Andersen and Keiding 2012). If our aim was to investigate causal effects of exposures on rates of transition between states, or other causal estimands, or if the aim was to develop a prognostic model, then conditioning on the patient's future pathway would not be appropriate for addressing the research question. Our consideration of conditional length of stay was motivated by questions about length of stay in hospital wards and intensive care units (ICU) among patients hospitalised due to Covid-19. Conditional length of stay estimates were of interest for two goals: (1) providing inputs to mathematical models which are used to inform resource requirements that are determined by patients' length of stay in different states; (2) providing a more comprehensive description of the multi-state system taking into account patient pathways, alongside unconditional length of stay estimates. The motivating example is described in more detail in Section 2.

Conditional length of stay has not, to our knowledge, been considered previously in the multi-state modelling literature. In this paper we describe non-parametric methods for estimating conditional length of stay distributions in a multi-state model, including the conditional expected length of stay in a given state (CELOS). These methods take into account that censoring can occur in every state. We also consider conditional length of stay distributions restricted to a particular time horizon, which are relevant when the full distribution of transition times is not observed in the data at hand due to limited follow-up. To describe the statistical methods we begin by focusing on an illness-death model (section 3). The methods are evaluated using a simulation study in section 4. In section 5 we extend the methods to the more complex multi-state model setting of the motivating example and apply them to estimate conditional length of stay in hospital and ICU for patients hospitalised with Covid-19 in the UK, using data from the ISARIC WHO CCP-UK COVID19 Clinical Information Network (CO-CIN) (Docherty et al. 2020). R code for implementing the methods is provided at https://github.com/ruthkeogh/lengthofstay.

2 Motivating example: patients hospitalised with Covid-19

The outbreak of Covid-19, caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was characterized as a pandemic by the World Health Organization on 11 March 2020 (WHO timelines 2020). According to UK government statistics (UK Government 2021), as of 3 April 2021 in the UK, 4,354,344 individuals had received a positive test for Covid-19, and a total of 458,868 hospitalisations and 126,955 deaths had been recorded (within 28 days of a positive Covid-19 test). Many patients require intensive care and, in the period up to 25 March 2021, 35,708 admissions to an intensive care unit (ICU) were recorded among patients in England, Wales and Northern Ireland with confirmed Covid-19 (ICNARC 2021).

Figure 2 illustrates a multi-state model for patients hospitalised with Covid-19 in the UK. The states are: 1. hospital ward; 2. intensive care unit (ICU); 3. hospital ward post-ICU; 4. Death in hospital; 5. Discharged from hospital. State 4 is an absorbing state. We also consider state 5 as an absorbing state - although patients can be discharged and readmitted, we did not consider this aspect. There are six possible complete pathways starting from state 1. Some individuals can start in state 2 (ICU), from which there are four possible complete pathways.

There were two main motivators for obtaining estimates of conditional length of stay in this study. The original motivator was a request to provide conditional length of stay estimates as inputs to mathematical models used in planning hospital capacity requirements. Molenberghs et al. (2020) discussed the importance of providing estimates of how long individuals require care in hospital and in ICU for planning hospital capacity requirements during the Covid-19 pandemic. Mathematical models are widely used to estimate hospital capacity requirements under different scenarios, for example varying the number of infected individuals and their age distribution. This is typically done using a simulation approach. One approach would be simulate how patients progress through the states of the multi-state model (Figure 2), using estimates of transition intensities. Expected lengths of stay in different states could then be estimated. However, this is computer intensive. Another approach, which is less computationally intensive, is to assign simulated patients at the time of hospital admission to one of the possible 'complete pathways' in the multi-state model with a given probability. This was the approach taken by Leclerc et al. 2021 from the London School of Hygiene & Tropical Medicine's Centre for Mathematical Modelling of Infectious Diseases group, for whom we provided estimates. They aimed to investigate how estimates of overall length of stay are influenced by the 'hospital bed pathways' taken by a patient, which may differ by region depending on the local patient population and local resource availability. It was concluded that national estimates of expected overall length of stay may not be appropriate for local forecasts of bed occupancy for COVID-19 (Leclerc et al. 2021).

A second motivator for this work was to show how we can provide descriptive information to the medical and scientific community and the general public about how long people hospitalised due to Covid-19 will be expected to spend receiving different levels of treatment in the hospital. Expected length of stay in hospital or ICU provides an overall summary, but conditional length of stay provides more detailed information that has also been of interest. Stays in the hospital ward (before a potential transfer to ICU) can end with death, discharge or a transfer to ICU. Conditional length of stay provides separate information on how long a patient requires in the hospital to recover and get discharged, and how long it takes for people in the hospital ward to become life-threatening ill and require intensive care. It also provides separate information on how long it takes for an individual admitted to ICU to recover, and how long a patient spends in ICU prior to death.

If all individuals in a given data set available for estimating length of stay had completed their stay, that is if their complete pathway was known, then expected lengths of stay and conditional expected lengths of stay in different states could be estimated empirically using observed averages. However, when the follow-up time of individuals is subject to censoring, empirical estimates based on the subset of individuals whose complete pathway is known will be biased. A number of authors have presented estimates of length of stay and conditional lengths of stay in different hospitalised states for Covid-19 patients (Vekaria 2020, Reig et al. 2020, Rees et al. 2020, Liu et al. 2020, Hazard et al. 2020). However, several have used empirical estimates (i.e. not accounting for censoring), and in other papers the approach taken was unclear. In this paper we show how traditional non-parametric multi-state modelling methods can be used to enable estimation of conditional lengths of stay. We discuss similarities and differences between our approach and that of other authors in Section 6.

3 Methods: illness-death model

3.1 Notation

We begin by considering the illness-death model depicted in Figure 1. The multi-state model is depicted in two different ways in Figures 1A and 1B. Figure 1A shows three states: 1. healthy state, 2. illness, 3. death. In Figure 1B the absorbing state of death is divided into two components: $3^{(1)}$ - death directly from the healthy state, $3^{(2)}$ - death from the illness state. These are two representations of the same model. In Figure 1B there is only one arrow going into any given state, in contrast with Figure 1A where there are two arrows going into state 3. Below it will be shown how the representation in Figure 1B is helpful for estimating conditional length of stay, and subsequent notation will refer to the model representation in Figure 1B.

Using standard notation for multi-state models we let $X(t)$ denote the state occupied at time t after entering state 1. We let $P_{1k}(s,t) = \Pr(X(t) = k|X(s) = 1)$ denote the probability of being in state k ($k = 1, 2, 3^{(1)}, 3^{(2)}$) at time t conditional on having been in state 1 at time s. The intensities of transitions from state 1 to state k ($k = 2, 3^{(1)}$) at time t are denoted $\lambda_{1k}(t)$. For transitions out of state 2 we assume a clock-reset (i.e. semi-Markov) approach and let $X^{(2)}(t)$ denote the state occupied at time t after entering state 2. We define the transition probability $P_{2k}(s,t) = \Pr(X^{(2)}(t) = k|X^{(2)}(s) = 2)$ as the probability of being in state ($k = 2, 3^{(2)}$) at time t after entering state 2, having been in state 2 at time s after entering state 2. The transition intensity from state 2 to state $3^{(2)}$ at time t after entering state 2 is denoted $\lambda^{(2)}_{23^{(2)}}(t)$. In the motivating example, a clock-reset approach for the ICU and hospital-post-ICU states was considered most reasonable.

There are two possible complete pathways through the multi-state system: $1 \to 3^{(1)}$, $1 \to 2 \to 3^{(2)}$. We may also allow people to start in state 2, and the only possible pathway for those people is $2 \to 3^{(2)}$. Let $P_{k|p}(t)$ denote the probability that the time spent in state k is $\geq t$, conditional on the complete pathway being p. We are interested in the distribution of time spent in state 1 conditional on the complete pathway being $1 \to 3^{(1)}$ or $1 \to 2 \to 3^{(2)}$, defined by the probabilities $P_{1|13^{(1)}}(t)$ and $P_{1|123^{(2)}}(t)$ respectively. We are also interested in the distribution of time spent in state 2 conditional on the complete pathway being equivalently $1 \to 2 \to 3^{(2)}$, defined by the probabilities $P_{2|123^{(2)}}(t)$. For those people who start in state 2 we are interested in $P_{2|23^{(2)}}(t)$. For the purposes of describing the methods, we assume that $P_{2|123^{(2)}}(t) = P_{2|23^{(2)}}(t)$, meaning that the distribution of time spent in state 2 (conditional on entering state 2) does not depend on whether the person started in state 1 or state 2. This assumption could be relaxed by estimating $P_{2|123^{(2)}}(t)$ and $P_{2|23^{(2)}}(t)$ separately. Below we consider estimation of $P_{1|13^{(1)}}(t)$, $P_{1|123^{(2)}}(t)$, $P_{2|123^{(2)}}(t)$, and $P_{2|23^{(2)}}(t)$.

We assume that data are available on a cohort of individuals and we let $\mathcal{T}_1 = \{t_1, \ldots, t_{J_1}\}$ denote the set of ordered observed times of transition out of state 1 (to state 2 or to state $3^{(1)}$). Similarly, $\mathcal{T}_2 = \{t_1^{(2)}, \ldots, t_{J_2}^{(2)}\}$ denotes the set of ordered observed times of transition from state 2 to state $3^{(2)}$.

3.2 Conditional distribution of time spent in state 1

By using the illness-death model in the format as depicted in Figure 1B we can express the probabilities $P_{1|p}(t)$ in terms of the multi-state transition probabilities $P_{1k}(s,t)$. First, $P_{1|13^{(1)}}(t)$ can be written

$$\begin{aligned} P_{1|13^{(1)}}(t) =& \Pr(X(t) = 1 | X(\infty) = 3^{(1)}) \\ =& \frac{\Pr(X(\infty) = 3^{(1)} | X(t) = 1) \Pr(X(t) = 1)}{\Pr(X(\infty) = 3^{(1)})} \\ =& \frac{P_{13^{(1)}}(t,\infty) P_{11}(0,t)}{P_{13^{(1)}}(0,\infty)} \end{aligned} \tag{1}$$

Similarly, we can write

$$\begin{aligned} P_{1|123^{(2)}}(t) =& \Pr(X(t) = 1 | X(\infty) = 3^{(2)}) \\ =& \frac{\Pr(X(\infty) = 3^{(2)} | X(t) = 1) \Pr(X(t) = 1)}{\Pr(X(\infty) = 3^{(2)})} \\ =& \frac{P_{13^{(2)}}(t,\infty) P_{11}(0,t)}{P_{13^{(2)}}(0,\infty)} \end{aligned} \tag{2}$$

Using established results for multi-state models (Aalen et al. 2008, Ch.3) we can write the transition probabilities $P_{11}(s,t)$, $P_{13^{(1)}}(s,t)$ and $P_{13^{(2)}}(s,t)$ as functions of the transition intensities as follows:

$$\begin{aligned} P_{11}(s,t) =& \Pr(X(t) = 1 | X(s) = 1) \\ =& e^{-\int_s^t (\lambda_{12}(x) + \lambda_{13^{(1)}}(x)) dx} \end{aligned} \tag{3}$$

$$\begin{aligned} P_{13^{(1)}}(s,t) =& \Pr(X(t) = 3^{(1)} | X(s) = 1) \\ =& \int_s^t P_{11}(s,u^-) P_{13^{(1)}}(u^-,u) du \\ =& \int_s^t e^{-\int_s^{u^-} (\lambda_{12}(x) + \lambda_{13^{(1)}}(x)) dx} \lambda_{13^{(1)}}(u) du \end{aligned} \tag{4}$$

$$\begin{aligned} P_{13^{(2)}}(s,t) =& \Pr(X(t) = 3^{(2)} | X(s) = 1) \\ =& \int_s^t \int_0^{t-u} P_{11}(s,u^-) P_{12}(u^-,u) P_{22}^{(2)}(0,v^-) P_{23^{(2)}}^{(2)}(v^-,v) dv du \\ =& \int_s^t \int_0^{t-u} e^{-\int_s^{u^-} (\lambda_{12}(x) + \lambda_{13^{(1)}}(x)) dx} \lambda_{12}(u) e^{-\int_0^{v^-} \lambda_{23^{(2)}}^{(2)}(x) dx} \lambda_{23^{(2)}}^{(2)}(v) dv du \end{aligned} \tag{5}$$

The transition intensities $\lambda_{1k}(t)$ ($k = 2, 3^{(1)}, 3^{(2)}$) can be estimated non-parametrically using $\hat{\lambda}_{1k}(t) = d_{1k}(t)/n_1(t)$, where $d_{1k}(t)$ denotes the number of transitions from state 1 to state k at time t, and $n_1(t)$ denotes the number at risk of transitioning to state 1 from state k at time t, i.e. the number of individuals observed to be in state 1 just before time t. Note that $\hat{\lambda}_{13^{(1)}}(t_j) = 0$ for times $t_j \in \mathcal{T}_1$ that are times of transition from state 1 to state 2 but not times of transition from state 1 to state $3^{(1)}$, and similarly $\hat{\lambda}_{12}(t_j) = 0$ for times $t_j \in \mathcal{T}_1$ that are times of transition from state 1 to state $3^{(1)}$ but not times of transition from state 1 to state 2.

Suppose first that the full distribution of transition times out of state 1 and state 2 is observed in the data. Note that this does not preclude the presence of censoring. In Section 3.4 we discuss estimation of $P_{k|p}(t)$ when the full distribution of transition times is not observed. The probabilities in (3), (4), and (5) can be estimated using

$$\widehat{P}_{11}(s,t) = \prod_{s<t_j\leq t} \left(1 - \hat{\lambda}_{12}(t_j) - \hat{\lambda}_{13^{(1)}}(t_j)\right) \tag{6}$$

$$\widehat{P}_{13^{(1)}}(s,t) = \sum_{s<t_j\leq t} \hat{\lambda}_{13^{(1)}}(t_j) \prod_{s<u<t_j} \left(1 - \hat{\lambda}_{12}(u) - \hat{\lambda}_{13^{(1)}}(u)\right). \tag{7}$$

$$\begin{aligned}\widehat{P}_{13^{(2)}}(s,t) = \sum_{s<t_j\leq t} \sum_{0<t_j^{(2)}<t-t_j} &\left(\prod_{s<u<t_j} \left(1 - \hat{\lambda}_{12}(u) - \hat{\lambda}_{13^{(1)}}(u)\right)\right) \hat{\lambda}_{12}(t_j) \\ &\times \left(\prod_{0<v<t_j^{(2)}} \left(1 - \hat{\lambda}_{23^{(2)}}(v)\right)\right) \hat{\lambda}^{(2)}_{23^{(2)}}(t_j^{(2)})\end{aligned} \tag{8}$$

It follows from the above that $P_{1|13^{(1)}}(t)$ (equation (1)) can be estimated using

$$\widehat{P}_{1|13^{(1)}}(t) = \frac{\sum_{t_j>t} \hat{\lambda}_{13^{(1)}}(t_j) \prod_{u<t_j} \left(1 - \hat{\lambda}_{12}(u) - \hat{\lambda}_{13^{(1)}}(u)\right)}{\sum_{t_j\in\mathcal{T}_1} \hat{\lambda}_{13^{(1)}}(t_j) \prod_{u<t_j} \left(1 - \hat{\lambda}_{12}(u) - \hat{\lambda}_{13^{(1)}}(u)\right)} \tag{9}$$

and $P_{1|123^{(2)}}(t)$ (equation (2)) can be estimated using

$$\widehat{P}_{1|123^{(2)}}(t) = \frac{\sum_{t_j>t} \hat{\lambda}_{12}(t_j) \prod_{u<t_j} \left(1 - \hat{\lambda}_{12}(u) - \hat{\lambda}_{13^{(1)}}(u)\right)}{\sum_{t_j\in\mathcal{T}_1} \hat{\lambda}_{12}(t_j) \prod_{u<t_j} \left(1 - \hat{\lambda}_{12}(u) - \hat{\lambda}_{13^{(1)}}(u)\right)} \tag{10}$$

3.3 Conditional distribution of time spent in state 2

The probability of being in state 2 for time t or longer (conditional on reaching state 2) conditional on the pathway being $1 \to 2 \to 3^{(2)}$ or $2 \to 3^{(2)}$ can be written

$$\begin{aligned}P_{2|123^{(2)}}(t) =& \Pr(X^{(2)}(t) = 2 | X^{(2)}(\infty) = 3^{(2)}) \\ =& \frac{\Pr(X^{(2)}(\infty) = 3^{(2)} | X^{(2)}(t) = 2) \Pr(X^{(2)}(t) = 2)}{\Pr(X^{(2)}(\infty) = 3^{(2)})} \\ =& \frac{P_{23^{(2)}}(t,\infty) P_{22}(0,t)}{P_{23^{(2)}}(0,\infty)}\end{aligned} \tag{11}$$

where $P_{23^{(2)}}(0,\infty) = 1$ and $P_{23^{(2)}}(t,\infty) = 1$. The transition probabilities $P_{22}(s,t)$ can be written

$$P_{22}(s,t) = e^{-\int_s^t \lambda_{23^{(2)}}(x)dx}. \tag{12}$$

If the full distribution of transition times is observed, these probabilities can be estimated for any s and t using

$$\widehat{P}_{22}(s,t) = \prod_{s<t_j^{(2)}\leq t} \left(1 - \hat{\lambda}^{(2)}_{23^{(2)}}(t_j^{(2)})\right) \tag{13}$$

Therefore $P_{2|123^{(2)}}(t)$ can be estimated using

$$\widehat{P}_{2|123^{(2)}}(t) = \prod_{0<t_j^{(2)}\leq t} \left(1 - \hat{\lambda}^{(2)}_{23^{(2)}}(t_j^{(2)})\right) \tag{14}$$

This is simply the Kaplan-Meier estimate, because once a person reaches state 2 there is only one subsequent state to which they can transition. The transition intensity $\lambda^{(2)}_{23^{(2)}}(t)$ can be estimated by $\hat{\lambda}^{(2)}_{23^{(2)}}(t) = d_{2k}(t)/n_2(t)$, where $d_{2k}(t)$ denotes the number of transitions from state 2 to state $3^{(2)}$ at time t after entering state 2, and $n_2(t)$ denotes the number at risk of transitioning to state 2 from state $3^{(2)}$ at time t after entering state 2.

3.4 Estimation when the full distribution of transition times is not observed

Above we assumed for estimation that the full distributions of transition times out of state 1 and state 2 were observed in the data. Suppose instead that there is censoring in the observed data in such a way that the full distributions of transition times are not observed. This means that the last observed time of censoring or transition out of a given state (state 1 or state 2) will be a censoring time rather than a transition time. In this case it is not possible to estimate the probabilities $P_{1|13^{(1)}}(t)$ and $P_{1|123^{(2)}}(t)$. We note that this problem does not arise if the data are only subject to uninformative censoring prior to the last transition time, but rather is specific to 'late' censoring which results in the full distribution of transition times not being observed. In this situation, we can consider instead $P^{\tau}_{1|13^{(1)}}(t)$ - the probability of spending time t or longer in state 1 conditional on transitioning to state $3^{(1)}$ *before time* τ, and $P^{\tau}_{1|123^{(2)}}(t)$ - the probability of spending time t or longer in state 1 conditional on transitioning to state 2 before time τ (because subsequent transition to state $3^{(2)}$ is then inevitable). The probabilities $P^{\tau}_{1|13^{(1)}}(t)$ and $P^{\tau}_{1|123^{(2)}}(t)$ can be estimated for times $\tau \leq t^*_{J_1}$, where $t^*_{J_1}$ denotes the latest observed follow-up time in state 1 (including both transition times and censoring times). To estimate $P^{\tau}_{1|13^{(1)}}(t)$ and $P^{\tau}_{1|123^{(2)}}(t)$, the results in equations (9) and (10) can be applied, with the sums in the denominators changed from $\sum_{t_j\in\mathcal{T}_1}$ to $\sum_{t_j\leq\tau}$.

For time spent in state 2, $P_{2|123^{(2)}}(t)$ can be estimated for any $t \leq t^*_{J_2}$, where $t^*_{J_2}$ denotes the latest observed follow-up time in state 2 (including both transition times and censoring times). We may also be interested in $P^{\tau}_{2|123^{(2)}}(t)$, which we define at the probability of spending time t or longer in state 2 conditional on transitioning to state $3^{(2)}$ before time τ, which can be written $P^{\tau}_{2|123^{(2)}}(t) = \frac{P_{23^{(2)}}(t,\tau)P_{22}(0,t)}{P_{23^{(2)}}(0,\tau)}$, and estimated using

$$\widehat{P}^{\tau}_{2|123^{(2)}}(t) = \frac{\sum_{t<t_j^{(2)}\leq\tau} \hat{\lambda}^{(2)}_{23^{(2)}}(t_j^{(2)}) \prod_{u<t_j^{(2)}} \left(1 - \hat{\lambda}^{(2)}_{23^{(2)}}(u)\right)}{\sum_{0<t_j^{(2)}\leq\tau} \hat{\lambda}^{(2)}_{23^{(2)}}(t_j^{(2)}) \prod_{u<t_j^{(2)}} \left(1 - \hat{\lambda}^{(2)}_{23^{(2)}}(u)\right)} \tag{15}$$

3.5 Conditional expected length of stay (CELOS)

Above we focused on the distribution of conditional lengths of stay. The expected time spent in a given state conditional on the pathway is one way of summarising the distribution. We refer to this as *conditional expected length of stay* (CELOS) and let $\text{CELOS}_{k|p}$ denote the expected length of stay in state k conditional on the complete pathway being p. The (unconditional) expected length of stay in state k can be written in terms of the state occupation probability: $E_k = \int_0^\infty \Pr(X(t) = k)dt$ (Beyersmann and Putter (2014)). It follows that $\text{CELOS}_{k|p}$ can be written

$$\text{CELOS}_{k|p} = \int_0^\infty P_{k|p}(t)dt \tag{16}$$

The conditional expected length of stay in state 1 among those who do not transition to state 2, denoted $\text{CELOS}_{1|13^{(1)}}$, can therefore be estimated using

$$\widehat{\text{CELOS}}_{1|13^{(1)}} = \sum_{t_j \in \mathcal{T}_1} (t_j - t_{j-1}) \times \widehat{P}_{1|13^{(1)}}(t_{j-1}) \tag{17}$$

where $t_0 = 0$ and $P_{1|13^{(1)}}(t_0) = 1$. $\text{CELOS}_{1|13^{(1)}}$ can equivalently be estimated using $\widehat{\text{CELOS}}_{1|13^{(1)}} = \sum_{t_j \in \mathcal{T}_1} t_j \times (\widehat{P}_{1|13^{(1)}}(t_{j+1}) - \widehat{P}_{1|13^{(1)}}(t_j))$. The expression in (17) is similar to that used by Beyersmann and Putter (2014) for restricted expected length of stay. Similarly, $\widehat{\text{CELOS}}_{1|123^{(2)}} = \sum_{t_j \in \mathcal{T}_1}(t_j - t_{j-1}) \times \widehat{P}_{1|123^{(2)}}(t_{j-1})$ and $\widehat{\text{CELOS}}_{2|23^{(2)}} = \sum_{t_j \in \mathcal{T}_2}(t_j - t_{j-1}) \times \widehat{P}_{2|23^{(2)}}(t_{j-1})$.

In studies where there is censoring such that the full distribution of transition times is not observed, we discussed above that the conditional probabilities $P_{1|13^{(1)}}(t)$ and $P_{1|123^{(2)}}(t)$ cannot be estimated, and $P_{2|123^{(2)}}(t)$ can only be estimated for times t up to the latest observed transition time. Beyersmann and Putter (2014) discussed *restricted* expected length of stay in the multi-state modelling context, defined as $E_k^\tau = \int_0^\tau \Pr(X(t) = k)dt$, which is the expected time spent in state k up to time τ. This is an extension to the multi-state setting of restricted mean survival time (RMST), proposed by Irwin (1949) (see also Royston and Parmar (2013) for example), which is the mean survival up to a particular time horizon.

We define *restricted* conditional expected length of stay (RCELOS) as the expected length of stay in a given state up to time τ conditional on the pathway taken up to time τ:

$$\text{RCELOS}^\tau_{k|p} = \int_0^\tau P_{k|p}(t)dt. \tag{18}$$

$\text{RCELOS}^\tau_{1|13^{(1)}}$ and $\text{RCELOS}^\tau_{1|123^{(2)}}$ can be estimated using

$$\widehat{\text{RCELOS}}^\tau_{1|13^{(1)}} = \sum_{t_j \in \mathcal{T}_1, t_j \le \tau} (t_j - t_{j-1}) \times \widehat{P}^\tau_{1|13^{(1)}}(t_{j-1})$$

and

$$\widehat{\text{RCELOS}}^\tau_{1|123^{(2)}} = \sum_{t_j \in \mathcal{T}_1, t_j \le \tau} (t_j - t_{j-1}) \times \widehat{P}^\tau_{1|123^{(2)}}(t_{j-1}).$$

$\text{RCELOS}^\tau_{2|123^{(2)}}$ is the same as the restricted (unconditional) length of stay in state 2 and is estimated using $\widehat{\text{RCELOS}}^\tau_{2|123^{(2)}} = \sum_{t_j^{(2)} \in \mathcal{T}_2, t_j^{(2)} \le \tau}(t_j^{(2)} - t_{j-1}^{(2)}) \times \widehat{P}_{2|123^{(2)}}(t_{j-1}^{(2)})$. We may also be interested in

$$\widehat{\text{RCELOS}}^{\tau *}_{2|123^{(2)}} = \sum_{t_j^{(2)} \in \mathcal{T}_2, t_j^{(2)} \le \tau} (t_j^{(2)} - t_{j-1}^{(2)}) \times \widehat{P}^\tau_{2|123^{(2)}}(t_{j-1}^{(2)})$$

which estimates the expected length of stay in state 2 conditional on transitioning to state $3^{(2)}$ before time τ after entering state 2.

3.6 Software

The conditional state occupation probabilities $P_{k|p}(t)$ and $\text{CELOS}_{k|p}$, and the restricted equivalents $P^{\tau}_{k|p}(t)$ and $\text{RCELOS}^{\tau}_{k|p}$ can be estimated 'manually' by obtaining estimates of the transition intensities $\lambda_{1k}(t)$ ($k = 2, 3^{(1)}$) and $\lambda^{(2)}_{23^{(2)}}(t)$, and applying the formulae given above. In the illness-death setting that we have considered so far, it is also possible to make use of some of the features of the `mstate` package in R (De Wreede et al. 2011, Putter et al. 2020), notably the `probtrans` function which can provide an estimate of the probability of having entered state 2. However, the `probtrans` function does not currently allow a clock-reset approach, which we assume here, which means that it cannot be used without modification beyond the illness-death setting.

4 Simulation study

We conducted a simulation study with the primary aims of checking the results in Section 3 and of demonstrating the bias if a naive analysis is used, in which empirical probabilities and means are calculated from the data ignoring censoring. The simulation also aims to illustrate some of the considerations needed when estimating restricted length of stay. R code is provided at https://github.com/ruthkeogh/lengthofstay, enabling the simulation results to be replicated.

4.1 Simulating data

Data were generated from the multi-state model depicted in Figure 1 for $N = 1000$ individuals. We consider three scenarios. In scenario (1) transition times were generated from exponential distributions using constant transition intensities $\lambda_{12} = 0.005$, $\lambda_{13^{(1)}} = 0.1$, $\lambda^{(2)}_{23^{(2)}} = 0.3$. In the motivating example transition times are recorded in terms of dates, resulting in ties. To mimic this discrete time setting of the motivating example, all times were rounded up to the next whole number in this scenario. In scenario (2) transition times were generated from Weibull hazard models of the form $\lambda(t) = \kappa\gamma t^{\kappa-1}$ for each transition, where κ is the shape parameter and γ is the rate parameter. For $\lambda_{12}(t)$, $\lambda_{13^{(1)}}(t)$, and $\lambda^{(2)}_{23^{(2)}}(t)$ we used $(\kappa = 0.75, \gamma = 0.05)$, $(\kappa = 0.75, \gamma = 0.1)$, and $(\kappa = 1.25, \gamma = 0.3)$ respectively. In practice, there is likely to be heterogeneity of transition intensities between individuals. We therefore considered a scenario (3) in which we incorporated individual frailties. This was done using Weibull transition hazards as in scenario (2), and individual frailties generated from a log-normal distribution with mean 0 and variance 1 and independently across transitions.

In all three scenarios censoring times were generated from an exponential model with hazard λ_0. We consider situations with no censoring ($\lambda_0 = 0$) and with substantial censoring ($\lambda_0 = 0.2$) designed to result in the full distribution of transition times not being observed. In the situation with censoring, the choice of λ_0 resulted in an average of 53% of individuals having their transition out of state 1 censored in scenario (1), 67% in scenario (2), and 60% in scenario (3).

There are 6 scenarios in total: scenarios (1), (2) and (3), each with and without censoring. We generated 1000 simulated data sets under each scenario.

4.2 Estimands

The estimands of interest were the CELOS ($\text{CELOS}_{1|13^{(1)}}$, $\text{CELOS}_{1|123^{(2)}}$, $\text{CELOS}_{2|123^{(2)}}$) and the RCELOS ($\text{RCELOS}^{\tau}_{1|13^{(1)}}$, $\text{RCELOS}^{\tau}_{1|123^{(2)}}$, $\text{RCELOS}^{\tau}_{2|123^{(2)}}$, $\text{RCELOS}^{\tau^*}_{2|123^{(2)}}$) for a time horizon of $\tau = 5$. We note that the RCELOS with a large τ correspond to the CELOS. For the RCELOS we present results for a time horizon of $\tau = 5$ because the maximum observed times spent in states 1 and 2 in the simulated data sets was typically greater than 5 in all scenarios, meaning that we expect to be able to obtain unbiased estimate of the RCELOS with $\tau = 5$ in situations with and without censoring. In practice, the time horizon may be selected as the maximum observed transition or censoring time in each state.

For scenario (1), where transition times are integers, we also obtained estimates of the probabilities $P_{1|13^{(1)}}(t)$, $P_{1|123^{(2)}}(t)$ and $P_{2|123^{(2)}}(t)$ (corresponding to the CELOS) and $P^{\tau}_{1|13^{(1)}}(t)$, $P^{\tau}_{1|123^{(2)}}(t)$ and $P^{\tau}_{2|123^{(2)}}(t)$ for $\tau = 5$ (corresponding to the RCELOS).

4.3 Methods and true values

We applied the multi-state analysis methods described in Section 3. We also calculated the empirical ("naive") estimates in each simulated data set. For example, the naive estimate of $\text{CELOS}_{1|13^{(1)}}$ was calculated as the mean observed time of entering state $3^{(1)}$ in those who transition to that state, excluding individuals who were censored. The naive estimate of $\text{RCELOS}^{\tau}_{1|13^{(1)}}$ was calculated as the mean observed time of entering state $3^{(1)}$ in those who transition to that state and who do so before time τ, excluding individuals who were censored. The naive estimates of $P_{1|13^{(1)}}(t)$ and $P^{\tau}_{1|13^{(1)}}(t)$ were calculated as the proportion of individuals who transitioned to state $3^{(1)}$ whose time of transition to $3^{(1)}$ was $\geq t$ (and $T_3 \leq \tau$ for $P^{\tau}_{1|13^{(1)}}(t)$), excluding individuals who were censored.

In scenarios without censoring we expect the estimates of the CELOS to be (asymptotically) unbiased using both the naive approach and using our formulae. In scenarios with censoring the CELOS cannot always be estimated. Given the quite substantial censoring generated in the censoring scenarios, we expect the estimates of the CELOS to be biased both under the naive approach and using our formulae.

The true values of the estimands were approximated by simulating a data set of one million individuals for scenarios (1), (2) and (3) *without censoring* and calculating the empirical values, as in the naive approach.

For each estimand, we present the mean estimate across the 1000 simulated data sets and the empirical standard deviation. We also present the bias using the mean difference between the 1000 estimates and the true value, and corresponding Monte-Carlo standard error, which is calculated as the empirical standard deviation of the estimates divided by $\sqrt{1000}$ (the square-root of the number of simulated data sets). In scenario (1), averages of probability estimates at a given time t are obtained only from those simulated data sets in which t was an observed transition time.

4.4 Results

Simulation results for the CELOS and RCELOS estimates for Scenarios (1), (2), and (3) are summarised in Tables 1-3.

When there is no censoring, the naive estimates of the CELOS and RCELOS are identical to those obtained from the multi-state analysis, as we would expect. The estimates are (approximately) unbiased, with very small bias in some values (according to the MCE) being attributed to the finite sample size.

When there is censoring the CELOS estimates are biased both using the naive approach and the multi-state analysis. Again, this is what we expect to see. The censoring induced by the data generating mechanisms results in the latest observed transition or censoring time typically being a censoring time. The bias from the multi-state analysis does not arise because there is a problem with the method, but because the conditional mean cannot be estimated when the full distribution of transition times in not observed, highlighting that restricted estimates are required in this situation. We note that the bias is smaller from the multi-state analysis than from the naive analysis, but it is still substantial in all three scenarios. The bias is in the direction of under-estimating the conditional expected length of stay. We chose a high hazard for censoring in this simulation. The bias due to ignoring censoring will clearly depend strongly on the extent and distribution of censoring. In the motivating example shown later, the amount of censoring is much lower.

Estimates of the RCELOS obtained using the multi-state analysis are (approximately) unbiased in all three scenarios, including when there is censoring. The naive estimates are unbiased only when there is no censoring. When there is censoring the naive analysis results in estimates that are biased downwards, i.e. under-estimating the RCELOS.

Supplementary Figures S1-S4 show plots of the estimated distribution of time spent in different states conditional on the pathway taken in scenario (1), for situations without censoring and with censoring. These demonstrate clearly how bias arises in the naive approach when there is censoring, with small values of t being over-represented relative to large values of t due to incomplete follow-up, resulting in an underestimate of the CELOS and RCELOS.

5 Application to hospitalisation for Covid-19

5.1 Data

The International Severe Acute Respiratory and emerging Infections Consortium WHO Clinical Characterisation Protocol UK (ISARIC WHO CCP-UK) study was established in the wake of the influenza A H1N1 pandemic (2009) and the emergence of Middle East respiratory syndrome coronavirus (2012). Further details about ISARIC WHO CCP-UK can be found at https://isaric4c.net. A key component of the ISARIC WHO CCP-UK study is the COVID19 Clinical Information Network (CO-CIN), which has collected clinical care data in near-real time from 208 hospitals in England, Scotland, and Wales on patients admitted to hospital since January 2020. Data were collected by clinical research nurses and administrators from clinical notes and entered into an online database. The clinical features of patients in this cohort have been described previously (Docherty et al. 2020).

We used CO-CIN data on individuals with proven or a high likelihood of infection with SARS-CoV-2 leading to COVID-19 disease with hospital admission from 10 March to 19 July 2020 (130 days). Information recorded includes patient characteristics, level of care (ward based, high dependency unit, or intensive care unit), complications, and dates of entering the following states: admission to hospital ward, admission to ICU (defined as high dependency unit or intensive care unit), stepping down from ICU to the general ward, death in hospital, and discharge. We include patients who had been admitted for a separate condition but had tested positive for SARS-CoV-2 during their hospital stay. A small proportion of individuals whose age or sex was not recorded were excluded.

The majority of individuals start in the hospital ward state, and the remainder start in the ICU admission state. The “discharge” state included individuals recorded with the outcomes “discharged alive” or “palliative discharge”. Individuals with the outcomes “hospitalized” or “transfer to other facility” were assumed alive and still in hospital or ICU at their outcome date. Some individuals have no outcome recorded because they were still within their care episode at the date of data extraction. These individuals were censored at the last date at which they had any information recorded in the data. When more than one event/transition was recorded on the same date for a given individual, we assumed the events occurred in quick succession and modified the data. For example if an individual was recorded as having been admitted to ICU on the same date as hospital admission, and then recorded as dying on the same date, the time of ICU admission was considered to be 0.25 days and the time of death 0.5 days.

5.2 Methods

Figure 2A illustrates the multi-state model for the more complex motivating example, in which there are 5 states. For patients starting in state 1 (hospital ward) there are 6 possible pathways. In the data, some individuals are observed to be admitted directly to ICU and therefore start in state 2. Therefore, we are also interested in the three possible pathways than a patient can follow if they start in state 2. The probabilities $P_{k|p}(t)$ for this setting are summarised in Table 4. In Figure 2B the two absorbing states of discharge (state 4) and death (state 5) depicted in Figure 2A are each divided into three states. State 4 is divided into states $4^{(1)}$ for people who are discharged from the hospital ward, state $4^{(2)}$ for people who are discharged from ICU, and $4^{(3)}$ for people who are discharged from the ward after ICU. Similarly state 5 is divided into states $5^{(1)}$, $5^{(2)}$, $5^{(3)}$, depending on the state from which an individual transitions to the death state.

The methods outlined for the simpler illness-death model can be extended to this more complex multi-state model and details are provided in the Supplementary Materials.

5.3 Results

The data contained the records of a total of 74722 individuals. After restricting to those with a proven or a high likelihood of infection with SARS-CoV-2 and admitted to hospital between 10 March and 19 July 2020 there remain 43256 individuals for analysis. We excluded 270 individuals with missing data on age or sex. The sample used for the analysis contains 42980 individuals, including 24776 males (58%) and 18204 females (42%). Table 5 summarises the numbers of observed transitions between states. The majority of individuals start in the hospital ward state (39571, 92%), with the remainder starting in ICU. A total of 7816 (18%) of individuals entered the ICU state (including those who start in that state), of whom the majority (89%) went back to the hospital ward after ICU, prior to death or discharge. There were 12058 deaths (28%) and 24456 (57%) individuals were discharged, with the remaining 15% of patients being censored.

We began by summarising how patients transition through the multi-state model using plots of state occupation probabilities, estimated non-parametrically. Figure S1 shows the resulting estimated state occupation probabilities. These show that the majority of transitions out of the hospital ward (pre ICU) have occurred by around 40 days. There are longer tails on the state occupation estimates after entering the ICU state. After entering the hospital ward after being in ICU, the plot shows that individuals who then die tend to do so quickly and the majority of deaths and discharges occurred within 10 days. The maximum time of transition out of state 1 (hospital ward pre-ICU) was 103 days, the maximum time of transition out of state 2 (ICU) was 107 days and the the maximum time of transition out of state 3 (hospital ward after ICU) was 89 days.

The (unconditional) expected lengths of stay in the hospital ward, in ICU and in the hospital ward after entering ICU were estimated using the methods of Beyersmann and Putter (2014), using the `ELOS` function from the `mstate` package in R. For individuals admitted go the hospital ward, the expected length of stays are: 8.99 days (95% CI 8.87, 9.11) in the hospital ward, 12.36 days (11.99, 12.77) in ICU, and 9.44 days (8.65, 10.20) in the hospital ward after ICU. For individuals admitted directly to ICU, the expected length of stays are: 14.36 days (13.79, 14.89) in ICU, and 9.26 days (8.37, 10.12) in the hospital ward after ICU.

We applied the methods described in section 5.2 to estimate the conditional length of stay distributions (Table 4) and corresponding CELOS. Preliminary investigations indicate that the length of follow-up available in this data set captures almost the full distribution of time spent in each state, and therefore permits estimation of the CELOS (as opposed to RCELOS). For comparison, we also calculated the naive estimates of the CELOS, which exclude the 15% of patients who were censored. Bootstrapping (percentile method) was used to estimate 95% confidence intervals (CI) for the CELOS estimates.

CELOS estimates are shown in Table 6 and the corresponding full conditional distributions in Figures 4 and 5. We focus on the results obtained for individuals who started their stay in the hospital ward, as opposed to in ICU. Individuals who were discharged at the end of their stay tend to spend longer in any given state (1, 2 or 3) compared with patients who die at the end of their stay. Among patients who did not go to the ICU, the expected time spent in hospital was 8.07 days in those who died at the end of their stay and 10.23 days in those who were discharged. Figure 4 (first panel) shows the long tail on the distributions. Time spent in the hospital ward (pre-ICU) was much shorter in those who transition to ICU, being just over 4 days. Figure 4 (first panel) shows a large drop off in the curves after 1 day for the curves corresponding to pathways through ICU. Because we have assumed a clock-reset approach, the time spent in hospital conditional on going to ICU does not depend on the states entered after ICU.

Patients who went to ICU followed by the hospital ward were estimated to spend an average of 12.38 days in ICU. Time spent in ICU was slightly shorter in those who did not subsequently return to the hospital ward (CELOS 7.71 days for those die in ICU and 9.76 days for those who are discharged directly from ICU), but these estimates are based on small numbers and the confidence intervals are wide. In those who go to ICU and then return to the hospital ward, the time spent in the hospital ward after ICU tended to be very short in patients who died (CELOS 1.03 days), suggesting that some individuals are returned to the ward from ICU when it is known that they are close to death. The expected time spent in the hospital ward after ICU was 10.77 days in those who were subsequently discharged. Figure 4 (third panel) shows a very large drop off in the distribution after 1 day for individuals who die. The distribution of time spent in states 2 and 3 was similar for patients who started in state 1 and patients who started in state 2 (i.e. were admitted directly to ICU).

The estimates of conditional length of stay using the naive analysis (excluding censored observations) tend to underestimate the true values (Table 6), which we expect from the simulation results and from theory.

6 Discussion

We have presented methods for estimating distributions of length of stay in a multi-state model conditional on the pathway taken through the states in the model. We also showed how the conditional length of stay distribution can be summarised in terms of a conditional expected length of stay (CELOS) or restricted CELOS (RCELOS), which is appropriate when there is censoring such that the last observed time in the state of interest is a censoring time rather than a transition time. The methods are non-parametric and do not rely on distributional assumptions. We described the methods for the widely used illness-death multi-state model and also provided details of the extension to the more complex multi-state model relevant for transitions of hospitalised Covid-19 patients. We assumed a clock-reset approach in which the transition intensities in a given state depend on time since entering that state, but not on previous states visited or duration spent in previous states. Extensions to our approach could relax this assumption, for example by specifying

Cox models for the transition intensities and including previous state and time spent in previous states as covariates.

The methods were assessed using a simulation study based on an illness-death model. The results show that in situations with censoring such that the full distribution of transition times is not observed, the naive estimates of the conditional length of stay distributions are biased, giving under estimates of the RELOS due to small transition times being over-represented in the data and higher transition times not being observed. The proposed multi-state approach gives approximately unbiased estimates. The results highlight that care should be taken when interpreting expected length of stay results when there is censoring and in finite samples - in these situations the restricted conditional length of stay (RCELOS) (up to a chosen time horizon τ) is an appropriate summary measure. We have also provided example R code for creating a simulated data set and for implementing the methods.

Alongside describing new methods, we applied the methods to estimate conditional length of stay in different states in patients hospitalised with Covid-19 in the UK using data on 42980 patients. Results were presented in terms of distributions and conditional expected length of stay in the hospital ward, in ICU, and in the hospital ward after ICU. The CELOS in the hospital ward in patients not admitted to ICU was 9.58 days, CELOS in ICU (among those admitted to ICU) was 12.38 days (in those who stepped down to the hospital ward after ICU, which was the majority), and the CELOS in the hospital ward after ICU (in those who entered that state) was 6.88 days, though this differed considerably between patients who subsequently died and those who were discharged.

Conditional length of stay in a state of a multi-state model involves conditioning on what happens to an individual in the future, which is usually best avoided in time-to-event analyses (Andersen and Keiding 2012). However, our estimands were carefully defined as conditional on the pathway, and we have shown that they enable a nuanced description of the multi-state system, as well as providing inputs that can be used in mathematical models. A different aim in a multi-state model could be to provide information about the risk of certain transitions occurring for an individual given their characteristics, or to estimate how certain covariates are associated with rates of transition. In that case conditioning on the pathway taken, or on any other future information, would be in appropriate for the question at hand. In the Covid-19 literature, multi-state modelling methods have been used by a number of authors to investigate time spent in different states in the context of patients hospitalised with Covid-19, and both unconditional and conditional lengths of stay have been estimated. Vekaria (2020) estimated conditional lengths of stays using data on 6208 Covid-19 patients in the UK observed in the COVID-19 Hospitalisation in England Surveillance System (CHESS) from March to May 2020. They took a parametric modelling approach and fitted Weibull models for each transition in a multi-state model, which was combined with a simulation procedure to obtain conditional length of stay estimates. Their estimates are in line with ours. They estimated a mean of 4 days spent in hospital prior to ICU admission (our estimate: 4.23 days). In those who did not go to ICU the expected time to death was 8.8 days (our estimate: 8.07 days) and the expected time to discharge 11.3 days (our estimate: 10.23 days). Among individuals who stepped down to the hospital ward after ICU, the expected time to discharge was 6.2 days (our estimate: 10.77 days). The expected time from ICU admission to death was 17.4 days (we did not obtain an equivalent estimate). They stated that they did not observe any individuals who stepped down from ICU to the hospital ward and then died. We observed individuals who transitioned from ICU to the hospital ward, however our results showed that a high proportion of these individuals died a short time after returning to the ward, suggesting that it may be appropriate to class some of these deaths as deaths in ICU. Data on the reason for a patient going to the Ward after ICU would facilitate this. There may have been different ways of recording death after ICU admission in the CHESS and CO-CIN data sets.

Reig et al.(2020) performed multi-state modelling using data on 213 patients admitted to a German hospital (February-May 2020). They considered the following states: regular ward, ICU (without mechanical ventilation), mechanical ventilation, extracorporeal membrane oxygenation (ECMO), death and discharge. In those admitted to the regular ward, the expected length of stay in the regular ward was 13.6 days, and expected length of stay in ICU was 0.8 days - this appears not be be conditional on actually going to ICU and so has a different interpretation than our estimates. In patients admitted directly to ICU the expected length of stay in ICU was 5.6 days. Hazard et al. (2020) used non-parametric multi-state modelling analysis to estimate restricted expected length of stay in ventilated and non-ventilated among Covid-19 patients admitted to ICU using data from two small published data sets from the US (n=24) and the US, Europe and Japan (n=53). The estimated total length of stay in ICU up to 28 days was 15.05 days (95% CI 9.29-21.66) in the larger study, which involved patients treated with remdesivir.

Rees et al. (2020) conducted a systematic review of estimated length of stay in Covid-19 patients based on studies published up to 12 April 2020. They identified 52 studies, most of which were from China. In studies from China the median length of stay in hospital was 14 days (interquartile range 10-19 days), and in studies outside of China the median length of stay in hospital was 5 days (interquartile range 3-9 days). Median length of stay in ICU was 8 days in studies from China, and 7 days outside of China. We estimated the full distribution of length of stay in different states and the means. For use in planning capacity requirements, means are more appropriate than medians as summary measures. Rees et al. (2020) noted that patients discharged alive tended to have longer length of stay compared with those who died, which we also found. In a study of trajectories among patients hospitalised with Covid-19 in France, Boelle et al. (2020) found that the median time to death in those who went to ICU was 20 days, and the median time to discharge from ICU was 17 days. In those who did not go to ICU, the median time to death was 9 days, and median time to discharge was also 9 days. They used parametric modelling methods, though it was not entirely clear how they estimated the length of stay. In a study from Australia, Liu et al. (2020) found that the median time spent in hospital was 9 days and the median time spent in ICU was 6 days; their results appear to be based on patients with death or discharge observed.

The methods described in this paper are non-parametric and do not incorporate covariates. The methods could be applied to subsets of patients defined by characteristics such as age group and sex. In further work it is of interest to extend the methods to incorporate several covariates simultaneously. This could be done, for example, by using semi-parametric Cox models for the transition intensities, and it should be straightforward to implement this using the `mstate` package in R. It would also be of interest to investigate extensions of the work of Klinten Grand and Putter (2016) who used pseudo-observations to construct regression models for expected length of stay in multi-state models, which enables estimation of associations between covariates and length of stay to be quantified.

Figure 1: Illness-death multistate model

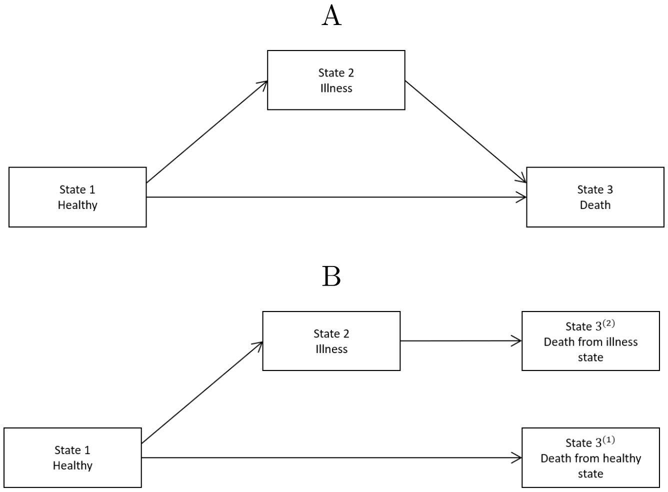


Figure 2: Multi-state model for patients hospitalised due to Covid-19

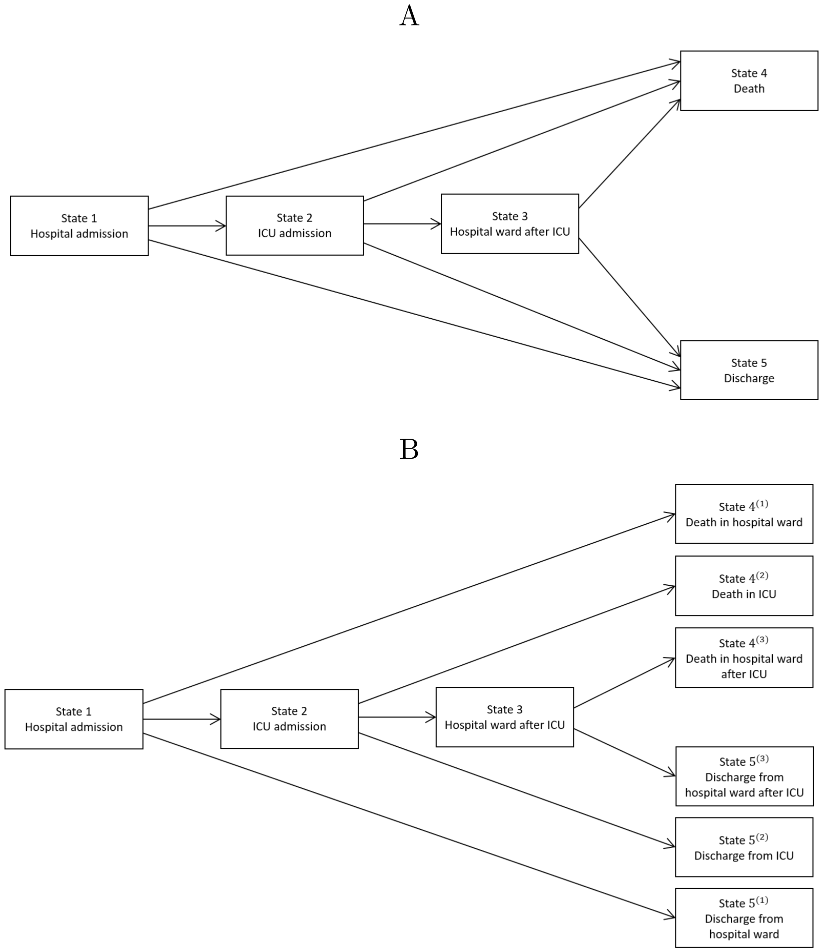


Table 1: Simulation results for scenario (1) (exponential data generating model). CELOS and RCELOS estimates and corresponding bias, obtained using the naive analysis (ignoring censored observations) and the multi-state analysis, for scenarios with and without censoring. Est: mean of estimates from 1000 simulated data sets. SD: empirical standard deviation. MCE: Monte-Carlo standard error.

	Without censoring		With censoring	
	Est (SD)	Bias (MCE)	Est (SD)	Bias (MCE)
Conditional expected length of stay (CELOS)				
Naive analysis				
$\text{CELOS}_{1\|13^{(1)}}$	7.172 (0.253)	-0.000 (0.008)	3.390 (0.168)	-3.782 (0.005)
$\text{CELOS}_{1\|123^{(2)}}$	7.172 (0.356)	-0.016 (0.011)	3.394 (0.299)	-3.794 (0.009)
$\text{CELOS}_{2\|123^{(2)}}$	3.853 (0.177)	0.007 (0.006)	2.544 (0.214)	-1.303 (0.007)
Multi-state analysis				
$\text{CELOS}_{1\|13^{(1)}}$	7.172 (0.253)	-0.000 (0.008)	6.447 (0.912)	-0.725 (0.029)
$\text{CELOS}_{1\|123^{(2)}}$	7.172 (0.356)	-0.016 (0.011)	6.342 (1.252)	-0.847 (0.040)
$\text{CELOS}_{2\|123^{(2)}}$	3.853 (0.177)	0.007 (0.006)	3.575 (0.434)	-0.272 (0.014)
Restricted conditional expected length of stay (RCELOS) with $\tau = 5$				
Naive analysis				
$\text{RCELOS}^{\tau}_{1\|13^{(1)}}$	2.704 (0.077)	0.003 (0.002)	2.337 (0.084)	-0.365 (0.003)
$\text{RCELOS}^{\tau}_{1\|123^{(2)}}$	2.705 (0.103)	0.001 (0.003)	2.341 (0.152)	-0.364 (0.005)
$\text{RCELOS}^{\tau}_{2\|123^{(2)}}$	2.428 (0.080)	0.011 (0.003)	2.098 (0.132)	-0.319 (0.004)
$\text{RCELOS}^{\tau^*}_{2\|123^{(2)}}$	2.999 (0.084)	0.007 (0.003)	2.336 (0.150)	-0.656 (0.005)
Multi-state analysis				
$\text{RCELOS}^{\tau}_{1\|13^{(1)}}$	2.704 (0.077)	0.003 (0.002)	2.705 (0.095)	0.004 (0.003)
$\text{RCELOS}^{\tau}_{1\|123^{(2)}}$	2.705 (0.103)	0.001 (0.003)	2.704 (0.126)	-0.001 (0.004)
$\text{RCELOS}^{\tau}_{2\|123^{(2)}}$	2.428 (0.080)	0.011 (0.003)	2.425 (0.153)	0.008 (0.005)
$\text{RCELOS}^{\tau^*}_{2\|123^{(2)}}$	2.999 (0.084)	0.007 (0.003)	2.998 (0.154)	0.005 (0.005)

Table 2: Simulation results for scenario (2) (Weibull data generating model). CELOS and RCELOS estimates and corresponding bias, obtained using the naive analysis (ignoring censored observations) and the multi-state analysis, for scenarios with and without censoring. Est: mean of estimates from 1000 simulated data sets. SD: empirical standard deviation. MCE: Monte-Carlo standard error.

	Without censoring		With censoring	
	Est (SD)	Bias (MCE)	Est (SD)	Bias (MCE)
Conditional expected length of stay (CELOS)				
Naive analysis				
$\text{CELOS}_{1\|13^{(1)}}$	14.947 (0.759)	0.012 (0.024)	2.708 (0.223)	-12.227 (0.007)
$\text{CELOS}_{1\|123^{(2)}}$	14.927 (1.097)	-0.013 (0.035)	2.683 (0.399)	-12.257 (0.013)
$\text{CELOS}_{2\|123^{(2)}}$	2.444 (0.109)	-0.004 (0.003)	1.847 (0.184)	-0.600 (0.006)
Multi-state analysis				
$\text{CELOS}_{1\|13}$	14.947 (0.759)	0.012 (0.024)	7.950 (2.596)	-6.985 (0.082)
$\text{CELOS}_{1\|123^{(2)}}$	14.927 (1.097)	-0.013 (0.035)	7.430 (3.187)	-7.510 (0.101)
$\text{CELOS}_{2\|123^{(2)}}$	2.444 (0.109)	-0.004 (0.003)	2.388 (0.252)	-0.060 (0.008)
Restricted conditional expected length of stay (RCELOS) with $\tau = 5$				
Naive analysis				
$\text{RCELOS}^{\tau}_{1\|13^{(1)}}$	1.933 (0.090)	-0.001 (0.003)	1.533 (0.097)	-0.401 (0.003)
$\text{RCELOS}^{\tau}_{1\|123^{(2)}}$	1.918 (0.129)	-0.014 (0.004)	1.527 (0.177)	-0.405 (0.006)
$\text{RCELOS}^{\tau}_{2\|123^{(2)}}$	1.940 (0.074)	0.001 (0.002)	1.640 (0.147)	-0.300 (0.005)
$\text{RCELOS}^{\tau^*}_{2\|123^{(2)}}$	2.265 (0.084)	-0.002 (0.003)	1.790 (0.164)	-0.478 (0.005)
Multi-state analysis				
$\text{RCELOS}^{\tau}_{1\|13^{(1)}}$	1.933 (0.090)	-0.001 (0.003)	1.927 (0.119)	-0.006 (0.004)
$\text{RCELOS}^{\tau}_{1\|123^{(2)}}$	1.918 (0.129)	-0.014 (0.004)	1.920 (0.178)	-0.012 (0.006)
$\text{RCELOS}^{\tau}_{2\|123^{(2)}}$	1.940 (0.074)	0.001 (0.002)	1.943 (0.165)	0.004 (0.005)
$\text{RCELOS}^{\tau^*}_{2\|123^{(2)}}$	2.260 (0.084)	-0.008 (0.003)	2.237 (0.174)	-0.031 (0.006)

Table 3: Simulation results for scenario (3) (Weibull data generating model with individual frailty). CELOS and RCELOS estimates and corresponding bias, obtained using the naive analysis (ignoring censored observations) and the multi-state analysis, for scenarios with and without censoring. Est: mean of estimates from 1000 simulated data sets. SD: empirical standard deviation. MCE: Monte-Carlo standard error.

	Without censoring		With censoring	
	Est (SD)	Bias (MCE)	Est (SD)	Bias (MCE)
Conditional expected length of stay (CELOS)				
Naive analysis				
$\text{CELOS}_{1\mid 13^{(1)}}$	18.053 (1.872)	-0.064 (0.059)	2.101 (0.166)	-16.016 (0.005)
$\text{CELOS}_{1\mid 123^{(2)}}$	20.893 (2.703)	-0.073 (0.085)	2.237 (0.299)	-18.729 (0.009)
$\text{CELOS}_{2\mid 123^{(2)}}$	3.347 (0.254)	0.000 (0.008)	1.717 (0.192)	-1.629 (0.006)
Multi-state analysis				
$\text{CELOS}_{1\mid 13^{(1)}}$	18.053 (1.872)	-0.064 (0.059)	6.066 (2.324)	-12.051 (0.073)
$\text{CELOS}_{1\mid 123^{(2)}}$	20.893 (2.703)	-0.073 (0.085)	6.378 (2.922)	-14.587 (0.092)
$\text{CELOS}_{2\mid 123^{(2)}}$	3.347 (0.254)	0.000 (0.008)	2.701 (0.501)	-0.646 (0.016)
Restricted conditional expected length of stay (RCELOS) with $\tau = 5$				
Naive analysis				
$\text{RCELOS}^{\tau}_{1\mid 13^{(1)}}$	1.676 (0.083)	-0.006 (0.003)	1.316 (0.083)	-0.366 (0.003)
$\text{RCELOS}^{\tau}_{1\mid 123^{(2)}}$	1.743 (0.106)	-0.007 (0.003)	1.378 (0.147)	-0.372 (0.005)
$\text{RCELOS}^{\tau}_{2\mid 123^{(2)}}$	1.673 (0.071)	0.001 (0.002)	1.378 (0.125)	-0.294 (0.004)
$\text{RCELOS}^{\tau^*}_{2\mid 123^{(2)}}$	2.310 (0.089)	0.001 (0.003)	1.588 (0.142)	-0.721 (0.004)
Multi-state analysis				
$\text{RCELOS}^{\tau}_{1\mid 13^{(1)}}$	1.676 (0.083)	-0.006 (0.003)	1.679 (0.108)	-0.003 (0.003)
$\text{RCELOS}^{\tau}_{1\mid 123^{(2)}}$	1.743 (0.106)	-0.007 (0.003)	1.751 (0.151)	0.002 (0.005)
$\text{RCELOS}^{\tau}_{2\mid 123^{(2)}}$	1.673 (0.071)	0.001 (0.002)	1.675 (0.150)	0.003 (0.005)
$\text{RCELOS}^{\tau^*}_{2\mid 123^{(2)}}$	2.299 (0.090)	-0.010 (0.003)	2.254 (0.167)	-0.056 (0.005)

Table 4: Summary of possible pathways for the multistate model in Figure 2, and notation for distribution of time spent in a given state k conditional on a given pathway p, $P_{k|p}(t)$. The last column shows the observed number of individuals following each pathway in the CO-CIN data. State 4 denotes death and state 5 denotes discharge.

	State 1	State 2	State 3	
Pathway	(Hospital ward)	(ICU)	(Ward after ICU)	Number observed to follow each pathway
Starting in state 1 (Hospital ward)				
$1 \rightarrow 4$	$P_{1|14^{(1)}}(t)$	-	-	9209
$1 \rightarrow 5$	$P_{1|15^{(1)}}(t)$	-	-	20766
$1 \rightarrow 2 \rightarrow 4$	$P_{1|124^{(2)}}(t)$	$P_{2|124^{(2)}}(t)$	-	89
$1 \rightarrow 2 \rightarrow 5$	$P_{1|125^{(2)}}(t)$	$P_{2|125^{(2)}}(t)$		94
$1 \rightarrow 2 \rightarrow 3 \rightarrow 4$	$P_{1|1234^{(3)}}(t)$	$P_{2|123^{(3)}4}(t)$	$P_{3|1234^{(3)}}(t)$	1561
$1 \rightarrow 2 \rightarrow 3 \rightarrow 5$	$P_{1|1235^{(3)}}(t)$	$P_{2|1235^{(3)}}(t)$	$P_{3|1235^{(3)}}(t)$	2042
Starting in state 2 (ICU)				
$2 \rightarrow 4$	-	$P_{2|24^{(2)}}(t)$		98
$2 \rightarrow 5$	-	$P_{2|25^{(2)}}(t)$		134
$2 \rightarrow 3 \rightarrow 4$	-	$P_{2|234^{(3)}}(t)$	$P_{3|234^{(3)}}(t)$	1101
$2 \rightarrow 3 \rightarrow 5$	-	$P_{2|235^{(3)}}(t)$	$P_{3|235^{(3)}}(t)$	1420

Table 5: Number (%) of transitions between states and censorings among 42980 patients in the CO-CIN data.

	To state					
	State 1	State 2	State 3	State 4	State 5	
From state	Hospital ward	ICU	Ward after ICU	Death	Discharge	Censored
State 1: Hospital ward	-	4407 (11%)	-	9209 (23%)	20766 (52%)	5189 (13%)
State 2: ICU	-	-	6940 (89%)	187 (2%)	228 (3%)	461 (6%)
State 3: Ward after ICU	-	-	-	2662 (38%)	3462 (50%)	816 (12%)
Total starting in this state	39571 (92%)	3409 (8%)	-	-	-	-
Total entering this state	39571 (92%)	7816 (18%)	6940 (16%)	12058 (28%)	24456 (57%)	-

Table 6: Conditional expected length of stay (CELOS) in states 1 (hospital ward), 2 (ICU) and 3 (ward after ICU) for Covid-19 hospitalised patients using the CO-CIN data: Naive estimates (excluding censored observations) and estimates obtained using the multi-state analysis. State 4 denotes death and state 5 denotes discharge.

Pathway	State 1	State 2	State 3
	Hospital ward	ICU	Ward after ICU
Naive analysis			
Starting in state 1 (Hospital ward)			
$1 \rightarrow 4^{(1)}$	7.14 (6.98, 7.29)	-	-
$1 \rightarrow 5^{(1)}$	8.59 (8.47, 8.72)	-	-
$1 \rightarrow (4^{(1)}$ or $5^{(1)})$	8.14 (8.05, 8.24)	-	-
$1 \rightarrow 2 \rightarrow 4^{(2)}$	4.33 (3.37, 5.49)	7.47 (5.88, 9.18)	-
$1 \rightarrow 2 \rightarrow 5^{(2)}$	4.38 (3.55, 5.32)	7.73 (5.84, 10.15)	-
$1 \rightarrow 2 \rightarrow (4^{(2)}$ or $5^{(2)})$	4.35 (3.71, 5.05)	7.61 (6.29, 9.13)	-
$1 \rightarrow 2 \rightarrow 3 \rightarrow 4^{(3)}$	4.12 (3.86, 4.41)	10.23 (9.79, 10.67)	0.88 (0.74, 1.03)
$1 \rightarrow 2 \rightarrow 3 \rightarrow 5^{(3)}$	3.13 (2.95, 3.31)	11.90 (11.38, 12.43)	8.41 (8.05, 8.80)
$1 \rightarrow 2 \rightarrow 3 \rightarrow (4^{(3)}$ or $5^{(3)})$	3.56 (3.41, 3.71)	11.18 (10.84, 11.56)	5.15 (4.91, 5.43)
Starting in state 2 (ICU)			
$2 \rightarrow 4^{(2)}$	-	7.27(6.06, 8.48)	-
$2 \rightarrow 5^{(2)}$	-	12.01 (9.85, 14.15)	-
$2 \rightarrow (4^{(2)}$ or $5^{(2)})$	-	10.00 (8.62, 11.33)	-
$2 \rightarrow 3 \rightarrow 4^{(3)}$	-	10.91 (10.37, 11.46)	0.91 (0.72, 1.12)
$2 \rightarrow 3 \rightarrow 5^{(3)}$	-	13.23 (12.60, 13.87)	8.37 (7.92, 8.79)
$2 \rightarrow 3 \rightarrow (4^{(3)}$ or $5^{(3)})$	-	12.22 (11.78, 12.64)	5.11 (4.83, 5.39)
Multi-state analysis			
Starting in state 1 (Hospital ward)			
$1 \rightarrow 4^{(1)}$	8.07 (7.83,8.33)	-	-
$1 \rightarrow 5^{(1)}$	10.23 (10.05,10.44)	-	-
$1 \rightarrow (4^{(1)}$ or $5^{(1)})$	9.58 (9.45,9.73)	-	-
$1 \rightarrow 2 \rightarrow 4^{(2)}$	4.23 (3.86,4.82)	7.71 (5.90,9.59)	-
$1 \rightarrow 2 \rightarrow 5^{(2)}$	4.23 (3.86,4.82)	9.76 (5.90,15.61)	-
$1 \rightarrow 2 \rightarrow (4^{(2)}$ or $5^{(2)})$	4.23 (3.86,4.82)	8.77 (6.48,11.84)	-
$1 \rightarrow 2 \rightarrow 3 \rightarrow 4^{(3)}$	4.23 (3.86,4.82)	12.38 (11.98,12.79)	1.03 (0.82,1.24)
$1 \rightarrow 2 \rightarrow 3 \rightarrow 5^{(3)}$	4.23 (3.86,4.82)	12.38 (11.98,12.79)	10.77 (9.24,13.18)
$1 \rightarrow 2 \rightarrow 3 \rightarrow (4^{(3)}$ or $5^{(3)})$	4.23 (3.86,4.82)	12.38 (11.98,12.79)	6.88 (5.92,8.45)
Starting in state 2 (ICU)			
$2 \rightarrow 4^{(2)}$	-	7.45 (6.10,8.76)	-
$2 \rightarrow 5^{(2)}$	-	13.38 (10.75,16.45)	-
$2 \rightarrow (4^{(2)}$ or $5^{(2)})$	-	10.92 (9.31,12.79)	-
$2 \rightarrow 3 \rightarrow 4^{(3)}$	-	14.63 (13.70,15.19)	1.13 (0.87,1.39)
$2 \rightarrow 3 \rightarrow 5^{(3)}$	-	14.63 (13.70,15.19)	11.07 (9.45,13.09)
$2 \rightarrow 3 \rightarrow (4^{(3)}$ or $5^{(3)})$	-	14.63 (13.70,15.19)	7.13 (6.13,8.41)

Figure 3: Estimated state occupation probabilities up to 100 days after entering each state, applied to Covid-19 hospitalised patients using the CO-CIN data. For individuals admitted to the hospital ward. See Supplementary Figure S1 for corresponding plots for individuals admitted directly to ICU.

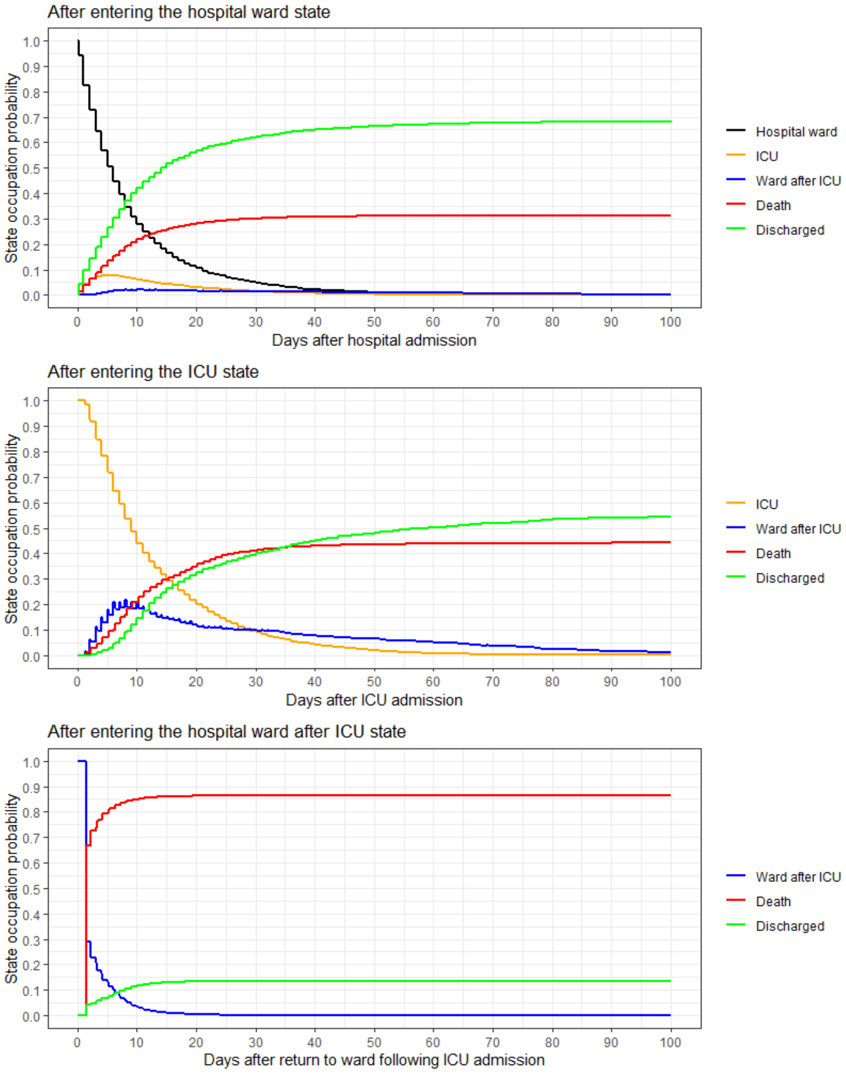


Figure 4: Summary of the distribution of time spent in the hospital ward (pre-ICU), in ICU, and in the hospital ward after ICU conditional on the pathway taken, for patients who are admitted to the hospital ward. The plots how the probability that the time spent in state k is $\geq t$ days conditional on the pathway p: $P_{k|p}(t)$.

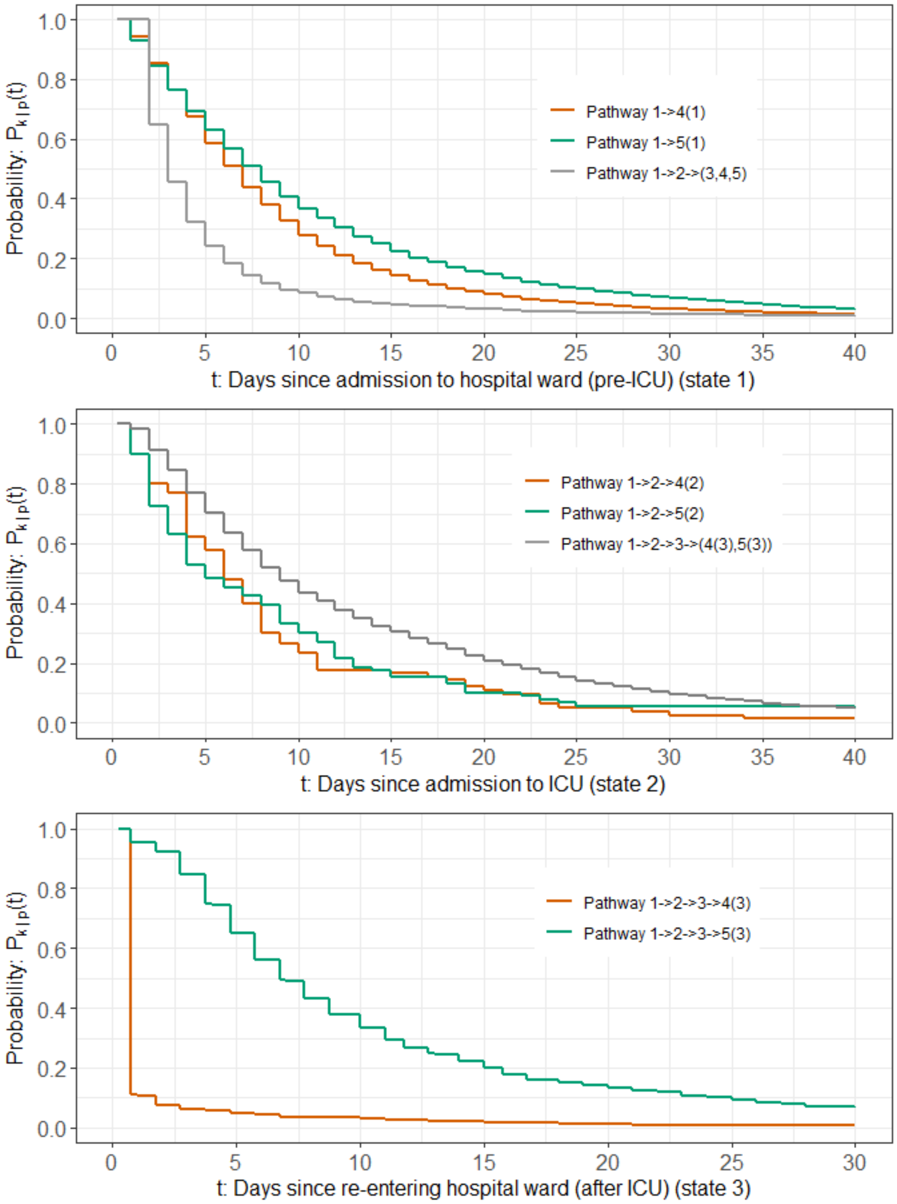
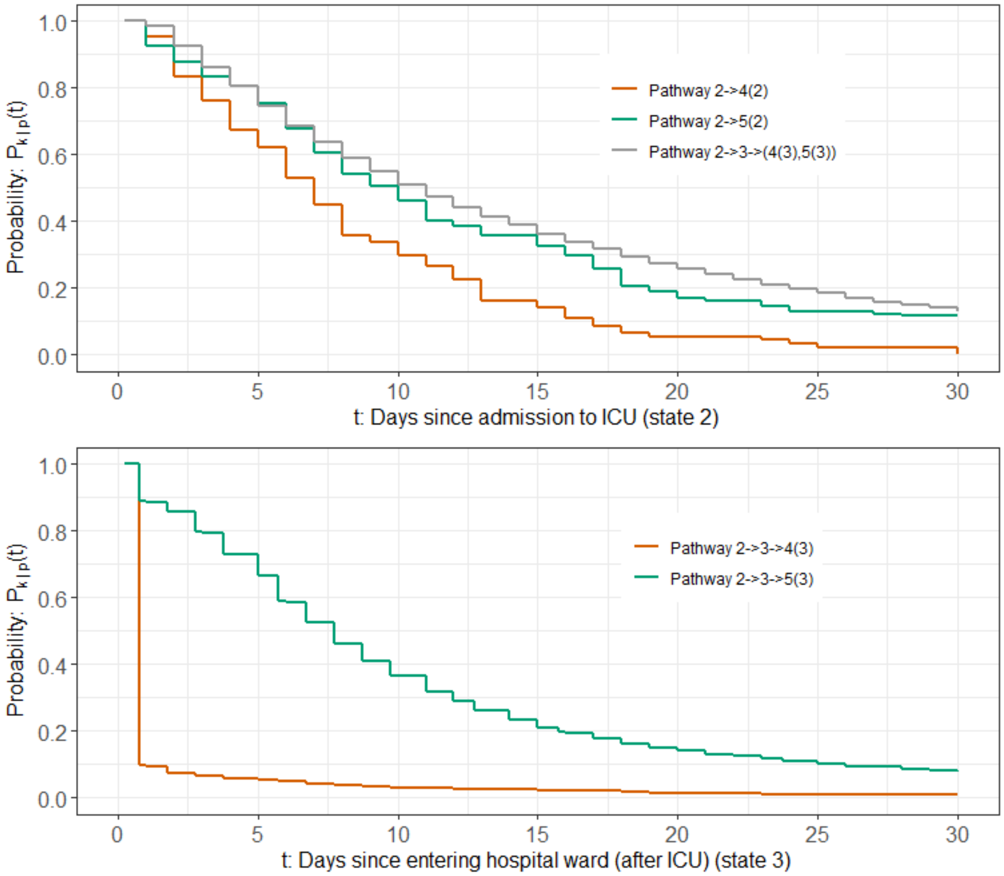


Figure 5: Summary of the distribution of time spent in ICU, and in the hospital ward after ICU conditional on the pathway taken, for patients who are admitted directly to ICU. The plots how the probability that the time spent in state k is $\geq t$ days conditional on the pathway p: $P_{k|p}(t)$. Estimated distribution of time spent in the ICU and hospital ward after ICU conditional on the pathway taken.



Acknowledgements

We thank Quentin Leclerc, Gwen Knight, and Nicholas Davies (Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine) for posing the scientific questions that motivated this work.

This work uses data provided by patients and collected by the NHS as part of their care and support. We are grateful to the frontline NHS clinical and research staff and volunteer medical students who collected the data in challenging circumstances; and the generosity of the participants and their families for their individual contributions in these difficult times.

Funding

RHK is funded by A UK Research and Innovation Future Leaders Fellowship (MR/S017968/1). KDO is supported by a Royal Society-Wellcome Trust Sir Henry Dale Fellowship (218554/Z/19/Z).

ISARIC CCP UK is supported by grants from: the National Institute for Health Research (NIHR; award CO-CIN-01), the Medical Research Council (MRC; grant MC_PC_19059), and by the NIHR Health Protection Research Unit (HPRU) in Emerging and Zoonotic Infections at University of Liverpool in partnership with Public Health England (PHE), in collaboration with Liverpool School of Tropical Medicine and the University of Oxford (award 200907), NIHR HPRU in Respiratory Infections at Imperial College London with PHE (award 200927), and NIHR Clinical Research Network for providing infrastructure support for this research. The views expressed are those of the authors and not necessarily those of the NIHR, MRC, or PHE.

Conflict of interest

The authors declare that they have no conflict of interest.

Availability of data and material

The CO-CIN data was collated by ISARIC4C Investigators. The study protocol is available at https://isaric4c.net/protocols. ISARIC4C welcomes applications for data and material access through the Independent Data and Material Access Committee (https://isaric4c.net).

Code availability

R code for implementing the methods and the simulation study is provided at https://github.com/ruthkeogh/length enabling the simulation results to be replicated.

Ethics approval

Ethical approval for ISARIC CCP UK was given by the South Central-Oxford C Research Ethics Committee in England (reference 13/SC/0149), and by the Scotland A Research Ethics Committee (reference 20/SS/0028). The study was registered at https://www.isrctn.com/ISRCTN66726260.

References

[1] Aalen PK, Borgan Ø, Gjessing HK. Survival and Event History Analysis: A process point of view. Springer: 2008.

[2] Andersen PK, Borgan Ø, Gill RD, Keiding N. Statistical Models Based on Counting Processes. Springer: Berlin, 1993.

[3] Andersen PK, Keiding N. Multi-state models for event history analysis. Statistical Methods in Medical Research 2002; 11: 91-115.

[4] Andersen PK, Keiding N. Interpretability and importance of functionals in competing risks and multistate models. Statistics in Medicine 2012; 31: 1074-1088.

[5] Beyersmann J, Putter H. A note on computing average state occupation times. Demographic Research 2014; 30: 1681-1696.

[6] Boelle P-Y, Delory T, Maynadier X, et al. Trajectories of Hospitalization in COVID-19 Patients: An Observational Study in France. J. Clin. Med. 2020, 9(10), 3148; https://doi.org/10.3390/jcm9103148

[7] De Wreede L, Fiocco M, Putter H. mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. Journal of Statistical Software 2011; 38: 7.

[8] Docherty AB, Harrison EM, Green CA, et al. Features of 20,133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. BMJ 2020;369:m1985. doi: https://doi.org/10.1136/bmj.m1985

[9] Hazard D, Kaier K, von Cube M, et al. Joint analysis of duration of ventilation, length of intensive care, and mortality of COVID-19 patients: a multistate approach. BMC Medical Research Methodology 2020; 20: 206.

[10] Intensive Care National Audit and Research Centre (ICNARC). ICNARC report on COVID-19 in critical care: England, Wales and Northern Ireland 26 March 2021. 2021. https://www.icnarc.org/Our-Audit/Audits/Cmp/Reports. Accessed 9 April 2021.

[11] Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. Journal of Hygiene 1949; 47: 188–189.

[12] Klinten Grand M, Putter H. Regression models for expected length of stay. Statistics in Medicine 2016; 35: 1178-1192.

[13] Leclerc QJ, Fuller NM, Keogh RH, et al. Importance of patient bed pathways and length of stay differences in predicting COVID-19 hospital bed occupancy in England. BMC Health Services Research 21, 566 (2021). https://doi.org/10.1186/s12913-021-06509-x.

[14] Liu B, Spokes P, Alfaro-Ramirez M, Ward K, Kaldor J. Hospital outcomes after a COVID-19 diagnosis from January to May 2020 in New South Wales Australia. Commun Dis Intell (2018) 2020; 44 (https://doi.org/10.33321/cdi.2020.44.97) Epub 24/12/2020

[15] Molenberghs G, Buyse M, Abrams S, et al. Infectious diseases epidemiology, quantitative methodology, and clinical research in the midst of the COVID-19 pandemic: Perspective from a European country. Contemporary Clinical Trials 99 (2020) 106189

[16] Putter H, Fiocco M, Geskus R. Tutorial in biostatistics: Competing risks and multi-state models. Statistics in Medicine 2007; 26: 2389-2430.

[17] Putter H, de Wreede L, Fiocco M, Geskus R. Package 'mstate'. R package. 2020. https://cran.r-project.org/web/packages/mstate/index.html

[18] Rees EM, Nightingale ES, Jafari Y, et al. COVID-19 length of hospital stay: a systematic review and data synthesis. BMC Medicine 2020; 18: 270.

[19] Rieg S, von Cube M, Kalbhenn J, et al. COVID-19 in-hospital mortality and mode of death in a dynamic and non-restricted tertiary care model in Germany. PLoS One. 2020; 15(11): e0242127.

[20] Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Medical Research Methodology volume 13, Article number: 152 (2013)

[21] UK Government. Coronavirus (COVID-19) in the UK. https://coronavirus.data.gov.uk/. Accessed 9 April 2021.

[22] Vekaria B, Overton C, Wisniowski A, et al. Hospital length of stay for COVID-19 patients: Data-driven methods for forward planning.2020 Hospital length of stay for COVID-19 patients: Data-driven methods for forward planning. https://www.researchsquare.com/article/rs-56855/latest.pdf. https://github.com/thomasallanhouse/covid19-los/blob/master/manuscript.pdf (2020)

[23] World Health Organisation. Timeline of WHO's response to COVID-19. 2020. https://www.who.int/news-room/detail/29-06-2020-covidtimeline. Accessed 30 June 2020.

Supplementary Material

S1 Application to hospitalisation for Covid-19

In this section we extend the methods for estimating conditional length of stay distributions to the more complex multi-state model illustrated in Figure 2 in the main text.

We assume a clock reset approach, with the clock being reset to 0 after a person enters state 2 or state 3. We let $X(t)$, $X^{(2)}(t)$, and $X^{(3)}(t)$ denote the state occupied at time t after entering states 1 (hospital), 2 (ICU), and 3 (Ward after ICU) respectively. We let $P_{1k}(s,t) = \Pr(X(t) = k|X(s) = 1)$ denote the probability of being in state k ($k = 1, 2, 3, 4^{(21}, 4^{(2)}, 4^{(3)}, 5^{(1)}, 5^{(2)}, 5^{(3)}$) at time t conditional on having been in state 1 at time s after entering state 1. Similarly, $P_{2k}(s,t) = \Pr(X^{(2)}(t) = k|X^{(2)}(s) = 2)$ denotes the probability of being in state k ($k = 2, 3, 4^{(2)}, 4^{(3)}, 5^{(2)}, 5^{(3)}$) at time t after entering state 2, having been in state 2 at time s after entering state 2. Similarly, $P_{3k}(s,t) = \Pr(X^{(3)}(t) = k|X^{(3)}(s) = 2)$ denotes the probability of being in state k ($k = 3, 4^{(3)}, 5^{(3)}$) at time t after entering state 3, having been in state 3 at time s after entering state 3.

Transition intensities from state 1 to state k are denoted $\lambda_{1k}(t)$. Transition intensities from state 2 to state k at time t after entering state 2 are denoted $\lambda^{(2)}_{2k}(t)$, and transition intensities from state 3 to state k at time t after entering state 3 are denoted $\lambda^{(3)}_{3k}(t)$.

There are six possible complete pathways through the multi-state system: $1 \to 4^{(1)}$, $1 \to 5^{(1)}$, $1 \to 2 \to 4^{(2)}$, , $1 \to 2 \to 5^{(2)}$, $1 \to 2 \to 3 \to 4^{(3)}$, $1 \to 2 \to 3 \to 5^{(3)}$. As in the illness-death example from the main text, we let $P_{k|p}(t)$ denote the probability that the time spent in state k is $\geq t$, conditional on the complete pathway being p. Interest lies in the distribution of time spent in states 1, 2, and 3, conditional on the complete pathway.

S1.1 Conditional distribution of time spent in state 1

The conditional probabilities $P_{1|p}(t)$ can be written in terms of conditional probabilities involving $X(t)$. For example,

$$\begin{aligned} P_{1|14^{(1)}}(t) =& \Pr(X(t) = 1|X(\infty) = 4^{(1)}) \\ =& \frac{\Pr(X(\infty) = 4^{(1)}|X(t) = 1)\Pr(X(t) = 1)}{\Pr(X(\infty) = 4^{(1)})} \\ =& \frac{P_{14^{(1)}}(t,\infty)P_{11}(0,t)}{P_{14^{(1)}}(0,\infty)} \end{aligned} \tag{S1}$$

Similar expressions hold for $P_{1|15^{(1)}}(t)$, $P_{1|124^{(2)}}(t)$, $P_{1|125^{(2)}}(t)$, $P_{1|1234^{(3)}}(t)$, $P_{1|1235^{(3)}}(t)$, and they involve the probabilities $P_{11}(s,t)$, $P_{14^{(1)}}(s,t)$, $P_{14^{(2)}}(s,t)$, $P_{14^{(3)}}(s,t)$, $P_{15^{(1)}}(s,t)$, $P_{15^{(2)}}(s,t)$, $P_{15^{(3)}}(s,t)$. We have the following expressions for $P_{11}(s,t)$, $P_{14^{(1)}}(s,t)$, $P_{14^{(2)}}(s,t)$, $P_{14^{(3)}}(s,t)$:

$$P_{11}(s,t) =P(X(t) = 1|X(s) = 1) = e^{-\int_s^t (\lambda_{12}(x)+\lambda_{14^{(1)}}(x)+\lambda_{15^{(1)}}(x))dx} \tag{S2}$$

$$P_{14^{(1)}}(s,t) = P(X(t) = 4^{(1)}|X(s) = 1) = \int_s^t \lambda_{14^{(1)}}(u)e^{-\int_s^{u^-} (\lambda_{12}(x)+\lambda_{14^{(1)}}(x)+\lambda_{15^{(1)}}(x))dx}du \tag{S3}$$

$$
\begin{aligned}
P_{14^{(2)}}(s,t) &= P(X(t)=4^{(2)}|X(s)=1) \\
&= \int_s^t \int_0^{t-u} \left\{ \lambda_{12}(u) e^{-\int_s^{u^-} (\lambda_{12}(x)+\lambda_{14^{(1)}}(x)+\lambda_{15^{(1)}}(x))dx} \right\} \left\{ \lambda^{(2)}_{24^{(2)}}(v) e^{-\int_0^{v^-} (\lambda^{(2)}_{23}(x)+\lambda^{(2)}_{24^{(2)}}(x)+\lambda^{(2)}_{25^{(2)}}(x))dx} \right\} dvdu
\end{aligned} \tag{S4}
$$

$$
\begin{aligned}
P_{14^{(3)}}(s,t) &= P(X(t)=4^{(3)}|X(s)=1) \\
&= \int_s^t \int_0^{t-u} \int_0^{t-u-v} \left\{ \lambda_{12}(u) e^{-\int_s^{u^-} (\lambda_{12}(x)+\lambda_{14^{(1)}}(x)+\lambda_{15^{(1)}}(x))dx} \right\} \left\{ \lambda^{(2)}_{23}(v) e^{-\int_0^{v^-} (\lambda^{(2)}_{23}(x)+\lambda^{(2)}_{24^{(2)}}(x)+\lambda^{(2)}_{25^{(2)}}(x))dx} \right\} \\
&\quad \left\{ \lambda^{(2)}_{34^{(3)}}(w) e^{-\int_0^{w^-} (\lambda^{(3)}_{34^{(3)}}(x)+\lambda^{(3)}_{35^{(3)}}(x))dx} \right\} dwdvdu
\end{aligned} \tag{S5}
$$

Similar expressions can be written for $P_{15^{(1)}}(s,t)$, $P_{15^{(2)}}(s,t)$, $P_{15^{(3)}}(s,t)$.

The transition intensities $\lambda_{1k}(t)$, $\lambda^{(2)}_{2k}(t)$, $\lambda^{(3)}_{3k}(t)$ can be estimated non-parametrically. We let $\mathcal{T}_1 = \{t_1, \ldots, t_{J1}\}$ denote the set of ordered observed times of transition out of state 1, $\mathcal{T}_2 = \{t^{(2)}_1, \ldots, t^{(2)}_{J2}\}$ the set of ordered observed times of transition out of state 2, and $\mathcal{T}_3 = \{t^{(3)}_1, \ldots, t^{(3)}_{J3}\}$ the set of ordered observed times of transition out of state 3. The above probabilities can be estimated as follows:

$$
\widehat{P}_{11}(s,t) = \prod_{s<t_j\le t} \left(1 - \hat{\lambda}_{12}(t_j) - \hat{\lambda}_{14^{(1)}}(t_j) - \hat{\lambda}_{15^{(1)}}(t_j)\right) \tag{S6}
$$

$$
\widehat{P}_{14^{(1)}}(s,t) = \sum_{s<t_j\le t} \hat{\lambda}_{14^{(1)}}(t_j) \prod_{s<u<t_j} \left(1 - \hat{\lambda}_{12}(u) - \hat{\lambda}_{14^{(1)}}(u) - \hat{\lambda}_{15^{(1)}}(u)\right) \tag{S7}
$$

$$
\begin{aligned}
\widehat{P}_{14^{(2)}}(s,t) = \sum_{s<t_j\le t} \sum_{0<t^{(2)}_j<t-t_j} & \left\{ \hat{\lambda}_{12}(t_j) \prod_{s<u<t_j} \left(1 - \hat{\lambda}_{12}(u) - \hat{\lambda}_{14^{(1)}}(u) - \hat{\lambda}_{15^{(1)}}(u)\right) \right\} \\
& \times \left\{ \lambda^{(2)}_{24^{(2)}}(t^{(2)}_j) \prod_{0<v<t^{(2)}_j} \left(1 - \hat{\lambda}_{23}(v)\hat{\lambda}_{24^{(2)}}(v) - \hat{\lambda}_{25^{(2)}}(v)\right) \right\}
\end{aligned} \tag{S8}
$$

$$
\begin{aligned}
\widehat{P}_{14^{(3)}}(s,t) = \sum_{s<t_j\le t} \sum_{0<t^{(2)}_j<t-t_j} \sum_{0<t^{(3)}_j<t-t_j-t^{(2)}_j} & \left\{ \hat{\lambda}_{12}(t_j) \prod_{s<u<t_j} \left(1 - \hat{\lambda}_{12}(u) - \hat{\lambda}_{14^{(1)}}(u) - \hat{\lambda}_{15^{(1)}}(u)\right) \right\} \\
& \times \left\{ \lambda^{(2)}_{24^{(2)}}(t^{(2)}_j) \prod_{0<v<t^{(2)}_j} \left(1 - \hat{\lambda}_{23}(v) - \hat{\lambda}_{24^{(2)}}(v) - \hat{\lambda}_{25^{(2)}}(v)\right) \right\} \\
& \times \left\{ \lambda^{(2)}_{34^{(3)}}(t^{(3)}_j) \prod_{0<w<t^{(3)}_j} \left(1 - \hat{\lambda}_{34^{(3)}}(w) - \hat{\lambda}_{35^{(3)}}(w)\right) \right\}
\end{aligned} \tag{S9}
$$

S1.2 Conditional distributions of time spent in state 2 and state 3

Conditional probabilities that describe the distribution time spent in state 2 conditional on the pathway are $P_{2|124^{(2)}}(t)$, $P_{2|125^{(2)}}(t)$, $P_{2|1234^{(3)}}(t)$, $P_{2|1235^{(3)}}(t)$. These can be written in terms of of $P_{22}(s,t)$, $P_{24^{(2)}}(s,t)$, $P_{24^{(3)}}(s,t)$, $P_{25^{(2)}}(s,t)$, $P_{25^{(3)}}(s,t)$. For example:

$$P_{2|124^{(2)}}(t) = \Pr(X^{(2)}(t) = 2|X^{(2)}(\infty) = 4^{(2)}) = \frac{P_{24^{(2)}}(t,\infty)P_{22}(0,t)}{P_{24^{(2)}}(0,\infty)} \tag{S10}$$

We have the following expressions for $P_{22}(s,t)$, $P_{24^{(2)}}(s,t)$, $P_{24^{(3)}}(s,t)$:

$$P_{22}(s,t) = e^{-\int_s^t(\lambda_{23}^{(2)}(x)+\lambda_{24^{(2)}}^{(2)}(x)+\lambda_{25^{(2)}}^{(2)}(x))dx} \tag{S11}$$

$$P_{24^{(2)}}(s,t) = \int_s^t \lambda_{24^{(2)}}^{(2)}(u)e^{-\int_s^{u^-}(\lambda_{23}^{(2)}(x)+\lambda_{24^{(2)}}^{(2)}(x)+\lambda_{25^{(2)}}^{(2)}(x))dx}du \tag{S12}$$

$$P_{24^{(3)}}(s,t) = \int_s^t \int_0^{t-u} \left\{\lambda_{23}^{(2)}(u)e^{-\int_s^{u^-}(\lambda_{23}^{(2)}(x)+\lambda_{24^{(2)}}^{(2)}(x)+\lambda_{25^{(2)}}^{(2)}(x))dx}\right\}\left\{\lambda_{34^{(3)}}^{(3)}(v)e^{-\int_0^{v^-}(\lambda_{34^{(3)}}^{(3)}(x)+\lambda_{35^{(3)}}^{(3)}(x))dx}\right\}dvdu \tag{S13}$$

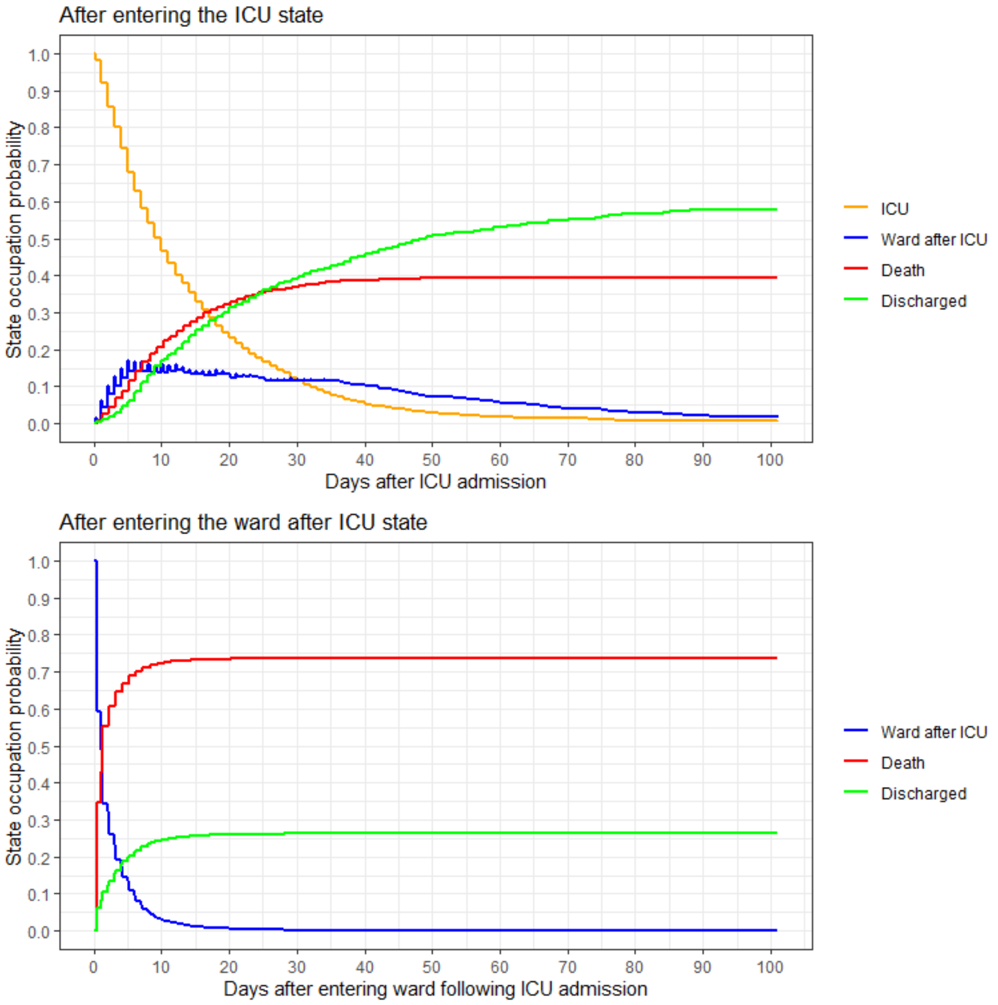
Conditional probabilities that describe the distribution time spent in state 3 conditional on the pathway are $P_{3|1234^{(3)}}(t)$, $P_{2|1235^{(3)}}(t)$. These can be written in terms of of $P_{23}(s,t)$, $P_{34^{(3)}}(s,t)$, $P_{35^{(3)}}(s,t)$. We have the following expressions for $P_{33}(s,t)$, $P_{34^{(3)}}(s,t)$:

$$P_{33}(s,t) = e^{-\int_s^t(\lambda_{34^{(3)}}^{(3)}(x)+\lambda_{35^{(3)}}^{(3)}(x))dx} \tag{S14}$$

$$P_{34^{(3)}}(s,t) = \int_s^t \lambda_{34^{(3)}}^{(2)}(u)e^{-\int_s^{u^-}(\lambda_{34^{(3)}}^{(3)}(x)+\lambda_{35^{(3)}}^{(3)}(x))dx}du \tag{S15}$$

Similar expressions can be written for $P_{25^{(2)}}(s,t)$, $P_{25^{(3)}}(s,t)$, and $P_{35^{(3)}}(s,t)$, and the probabilities can be estimated non-parametrically in a similar way as described above for state 1.

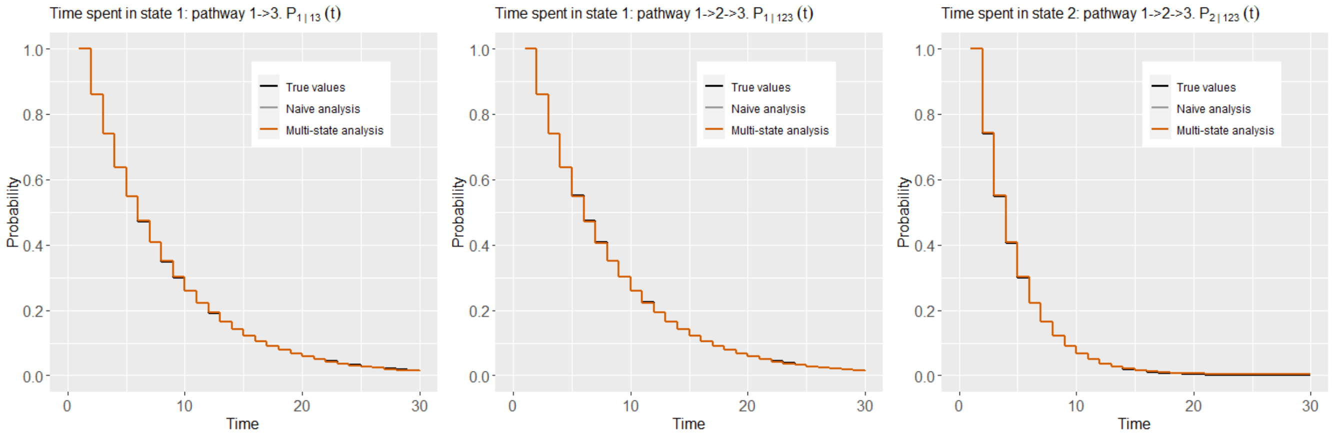
Figure S1: Estimated state occupation probabilities up to 100 days after entering each state, applied to Covid-19 hospitalised patients using the CO-CIN data. For individuals admitted directly to ICU.



After entering the ICU state

After entering the ward after ICU state

Figure S2: Simulation results for scenario 1 (exponential data generating model), with no censoring. Summary of the distribution of time spent in different states conditional on the pathway taken, estimated using the naive analysis (ignoring censored observations) and the multi-state analysis.

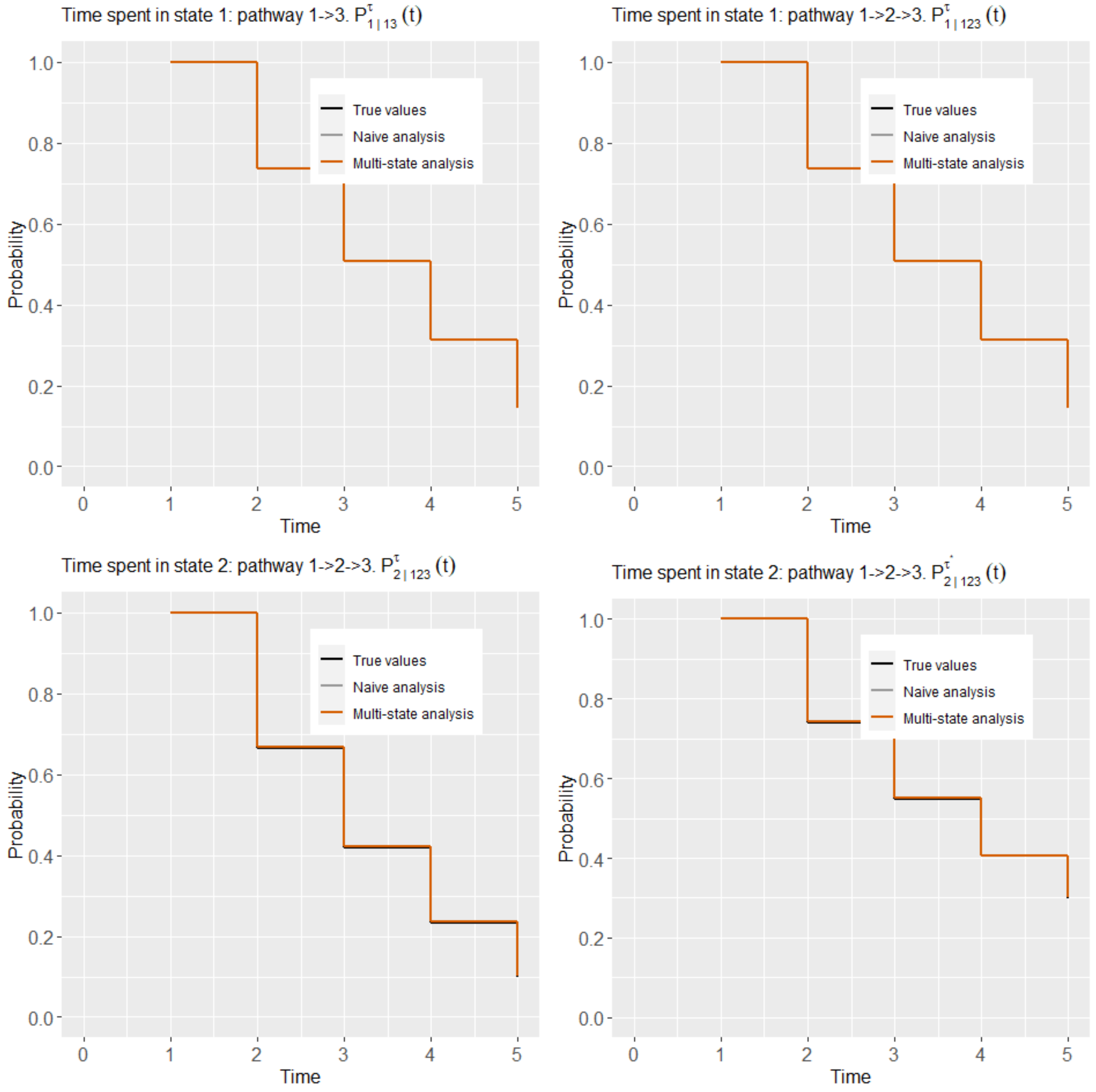


Time spent in state 1: pathway 1->3. $P_{1|13}(t)$

Time spent in state 1: pathway 1->2->3. $P_{1|123}(t)$

Time spent in state 2: pathway 1->2->3. $P_{2|123}(t)$

Figure S3: Simulation results for scenario 1 (exponential data generating model), with no censoring and time horizon $\tau = 5$. Summary of the distribution of time spent in different states conditional on the pathway taken, estimated using the naive analysis (ignoring censored observations) and the multi-state analysis.


Time spent in state 1: pathway 1->3. $P^{\tau}_{1|13}(t)$

Time spent in state 1: pathway 1->2->3. $P^{\tau}_{1|123}(t)$

Time spent in state 2: pathway 1->2->3. $P^{\tau}_{2|123}(t)$

Time spent in state 2: pathway 1->2->3. $P^{\tau^*}_{2|123}(t)$

Figure S4: Simulation results for scenario 1 (exponential data generating model), with censoring. Summary of the distribution of time spent in different states conditional on the pathway taken, estimated using the naive analysis (ignoring censored observations) and the multi-state analysis.

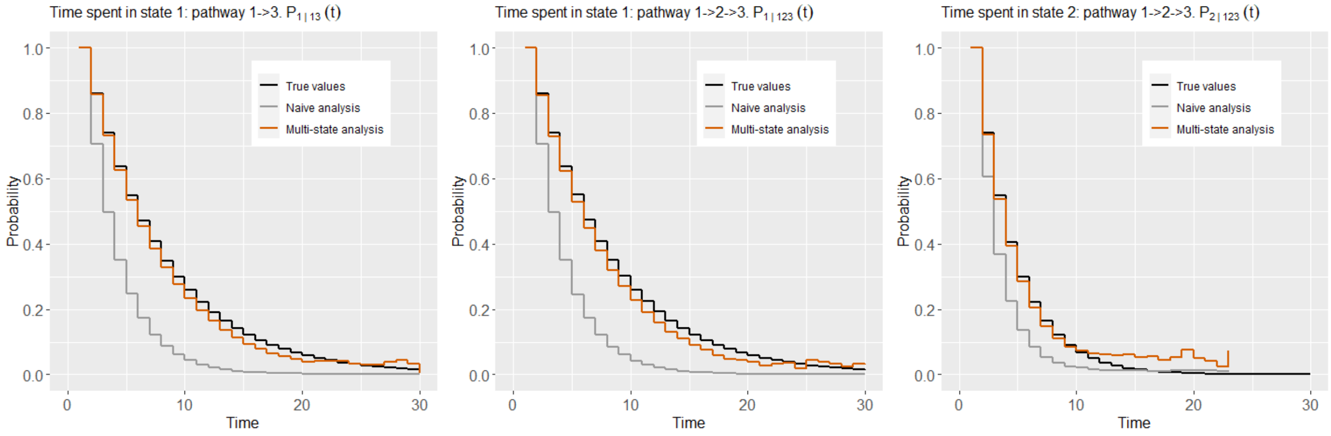
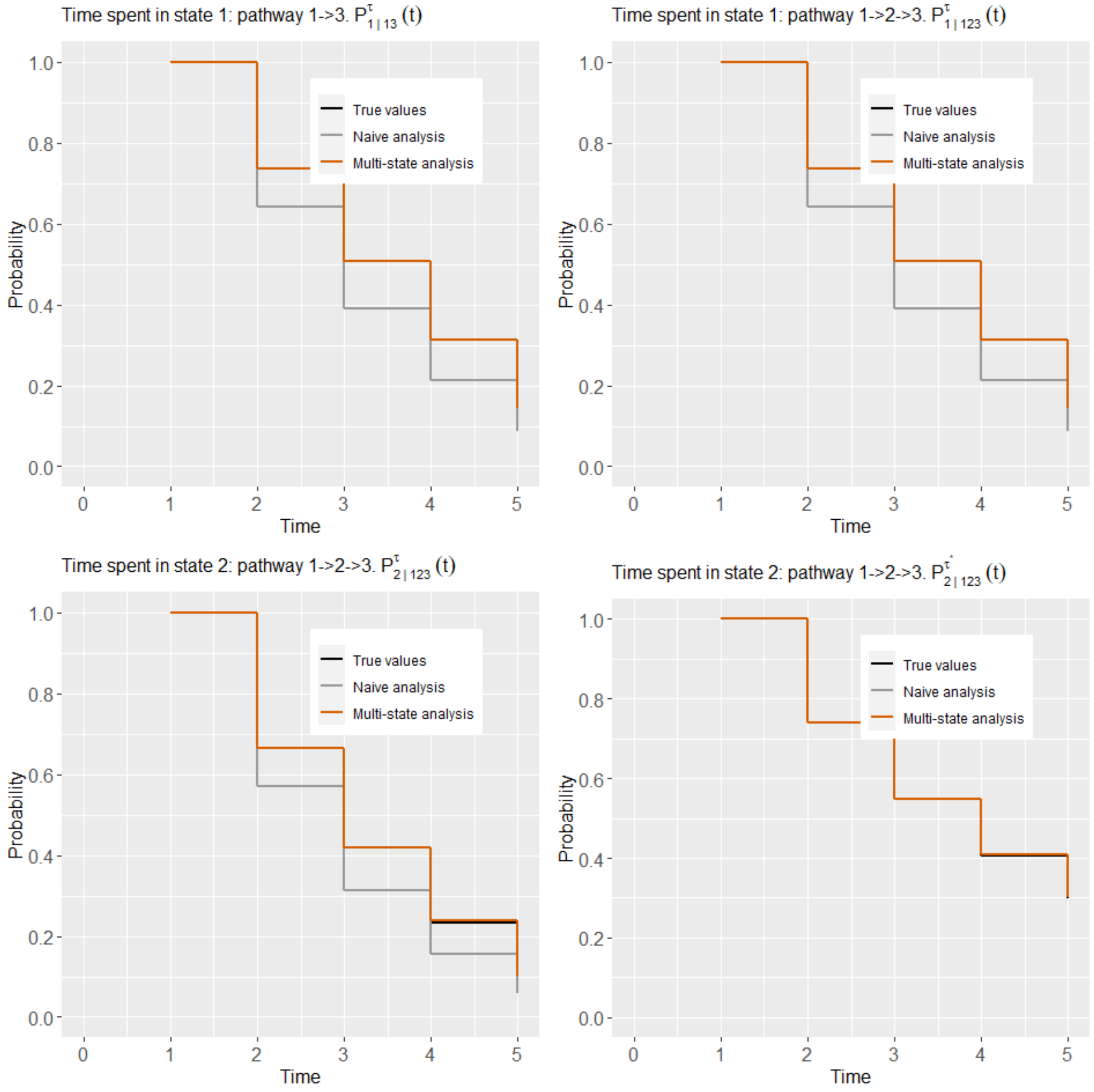


Figure S5: Simulation results for scenario 1 (exponential data generating model), with censoring and time horizon $\tau = 5$. Summary of the distribution of time spent in different states conditional on the pathway taken, estimated using the naive analysis (ignoring censored observations) and the multi-state analysis.


Time spent in state 1: pathway 1->3. $P^{\tau}_{1|13}(t)$
Time spent in state 1: pathway 1->2->3. $P^{\tau}_{1|123}(t)$
Time spent in state 2: pathway 1->2->3. $P^{\tau}_{2|123}(t)$
Time spent in state 2: pathway 1->2->3. $P^{\tau^*}_{2|123}(t)$
True values
Naive analysis
Multi-state analysis
Probability
1.0
0.8
0.6
0.4
0.2
0.0
Time
0
1
2
3
4
5