Expected Length of Stay with a Multistate Model

Bachelor thesis by

Mimmo Lentz

Supervisor
Dr. Cristian Spitoni

Student Number
5549922

Date
June 14, 2018

Utrecht University

Abstract
A time-inhomogeneous Markov process can be used to model the effect of a hospital infection on the expected length of stay. It also allows for modelling the time-dependency of the hospital infection. The Markov assumption provides us with the Nelson-Aalen estimators for cumulative hazards, a powerful tool to quantify the effect of hospital infection on the expected length of stay. Using a summary measure, we have found that over patients in the ICU, CMV reactivation does not prolong the stay in ICU, but instead shortents the length of stay due to a high death hazard. By assuming the state transitions to follow a Cox proportional hazards model, we have also found that CMV reactivation does have a prolonging effect on length of stay for patients with a low APACHE IV score but not for patients with a high APACHE IV score.

Contents

1 Introduction

Cytomegalovirus (CMV) is a virus that has no symptoms and long-term health consequences for most healthy people. Once acquired, the virus will stay in your body indefinitely and may be reactivated. Reactivation of CMV has been reported in high rates with critically ill CMV seropositve subjects [11]. It has already been associated with increased length of stay in the ICU, and is additionally associated with increased mortality [11]. It is of high importance for medical decision making to get an accurate grasp on the effects of the hospital acquired infection (HI) CMV reactivation on susceptible patients. A prolonging effect of an infection on hospital stay can have an economic impact, while an increase in mortality due to infection is unwishful as well.

The current study attempts to find out what the effect is on expected length of stay in the intensive care unit (ICU). A data set of 271 patients has been provided by the Utrecht Medical Centre containing patients time-to-event data. Time of infection acquistion and time of death or discharge have been recorded, as well as many characteristics of the patients. Typical analysis to deal with time-to-event data is survival analysis, with a generalization to multistate analysis, allowing for more flexibility. We will model the effect of CMV reactivation with a multistate model containing the infection as a state, to include the time-dependency of the infection. Not including the time-dependency of the infection will lead to a bias [4, p. 222] [6]. It would ignore the fact that it takes time for the infection to be acquired.

Additionally, we assume the process of the multistate model to be an inhomogeneous Markov process, i.e. each state transition probability only depends on the current time and not when the current state was entered, it will provide us with forceful tools. The tools that it will provide us are the nonparametric Nelson-Aalen estimator and Aalen-Johansen estimator, estimating the cumulative transition hazards and the transition probability matrix respectively. By plotting these estimators we obtain insight in the risk of making specific transitions. More specifically, we can observe the risk of death or discharge given the infection state of a patient.

On top of observing plots, it is also useful to quantify the observations, such as the conditional probability of an infected patient staying up to a certain time t. The multistate model allows for better quantification than more simplistic models. In this study we will consider a quantity proposed by Schulgen and Schumacher [13] giving the difference in expected subsequent hospital stay given the infection state at a time t. A time-averaged quantity can be computed by also attaching a weight that depends on the distribution of the infection.

Additionally to the previous estimates, we can make predictions for patients with different characteristics. By assuming that the transitions in the multistate model follow a Cox model, we can include covariates into the model and perform a Cox regression. This will allow a physician to make predictions for a specific group of interest what the effects of HI are.

In this article we start by introducing theory on survival analysis, the most basic form of a multistate model, in Section 2. The notion of survival and hazard will be introduced, as well as both nonparametric and semiparametric estimation. In Section 1 we will generalize these all results to a multistate model. Then in Section 4 we use the results from the preceding section to quantify the effect of HI on the expected length of stay. Ultimately we will present the results, and elaborate on them, for the effect of CMV reactivation on the stay in the ICU in Section 5. The last section, 6 is reserved for a discussion of this study.

2 Survival analysis

Survival analysis is the analysis interested in the time it takes for an event to occur. It is the most simple form of a multistate model, consisting of only two states: a state where the event has not yet occured, and another in which it has. The familiar reader can skip this section and move to Section 1 on multistate models. The random variable T represents the time it takes for the event to occur. The following function tells us the probability of surviving, that is, still being in the initial state, upto time $t > 0$:

$$S(t) := P(T \geq t). \tag{2.1}$$

It is conveniently called the survival function and is of great significance in this field. With this function comes the hazard function defined as

$$\alpha(t) := \frac{P(T \in [t, t+dt) \,|\, T \geq t)}{dt}, \tag{2.2}$$

describing the probability of entering the absorbing state at time t, given it survived upto t. Both the survival and the hazard function are a way of describing the distribution of the random variable T. The relationship between the two functions is of high importance because, as it will soon turn out, we will need a notion of probability for computing the expected length of stay, and we can estimate this using the notion of hazard. In the following paragraph we will discuss the relation between the survival function and the hazard function.

The cumulative hazard function is defined as

$$A(t) := \int_0^t \alpha(s)\, ds. \tag{2.3}$$

This function will be the form that we will mostly use the hazard function in. Even more so, this is the form for which we have an estimator, the Nelson-Aalen estimator. In the following section we will rewrite $S(t)$ in terms of increments of $A(t)$ so that we perform estimation in Section 2.2.

2.1 Survival function and product integration

Using that we can alternatively write the hazard function as $\alpha(t) \cdot dt = \frac{f(t)}{S(t)}$ and the fact that $-\frac{d}{dt}S(t) = f(t)$, we establish the relation

$$A(t) = -\int_0^t \frac{dS(s)}{S(s-)},$$

or alternatively

$$dS(t) = -S(t-)dA(t),$$

where the notation $t- = \lim_{\nearrow t} s$ means the limit to t from beneath. If we discretize the time interval into $0 = t_0 < ... < t_K = t$, we get the approximation to the above equation

$$S(t_k) - S(t_{k-1}) \approx -S(t_{k-1})\big(A(t_k) - A(t_{k-1})\big),$$

or

$$S(t_k)/S(t_{k-1}) \approx 1 - \big(A(t_k) - A(t_{k-1})\big),$$

Keeping this in mind, we can use conditional probabilities to rewrite $S(t)$. Note that $P(T > t_k \,|\, T > t_{k-1}) = S(t_k)/S(t_{k-1})$, resulting in

$$\begin{aligned} S(t) &= P(T > t_K \,|\, T > t_{K-1}) \cdots P(T > t_2 \,|\, T > t_1) \cdot P(T > t_1) \\ &= \prod_{k=1}^{K} S(t_k)/S(t_{k-1}) \\ &\approx \prod_{k=1}^{K} 1 - \big(A(t_k) - A(t_{k-1})\big) \end{aligned}$$

As the time grid gets finer, we get a better approximation. If we let $d := \max_k\{t_k - t_{k-1}\}$, then by the theory of product integration, we are assured that the limit in the above approximation exists if we let $d \to 0$. The result is

$$\begin{aligned} S(t) &= \lim_{d\to 0} \prod_{k=1}^{K} 1 - \big(A(t_k) - A(t_{k-1})\big) \\ &:= \prod_{0}^{t} (1 - dA(t)), \end{aligned}$$

an infinite product over one minus the increments of the cumulative hazard. As with any data set, we don't have this type of data available, and therefore use a finite product as an approximation

$$S(t) \approx \prod_{k=0}^{K} (1 - dA(t_k)) \tag{2.4}$$

2.2 Estimation

We next consider estimation of the cumulative hazard function and subsequently the survival function. Due to form of the survival function in (2.1), the estimator for the survival function will follow quite naturally from the one for the cumulative hazard. First we will consider a nonpparametric model for a homogeneous population, with the Nelson-Aalen and Kaplan-Meier estimators for the cumulative hazard and survival function, respectively. Following that, we will discuss the estimators for a semiparametric model, where we will include covariates. The semiparametric model that we will discuss is the Cox proportional hazards model, a model widely used in survival analysis. The model assumes that the hazards for subjects with different covariates differ proportional to a common baseline hazard which can take any shape of function. It includes regression coefficients displaying the covariates effects, that will be estimated by maximizing the *partial likelihood*, a particular type of likelihood maximization.

Both the nonpparametric and the semiparametric model will use empirical probabilities (the semiparametric model also including weights), for which we will need to introduce two counting processes in order to obtain the frequency sets. At every point t in time we want to know how many subjects are at risk for the event and how many subjects, if any, undergo the event. We write T for the time of the event, and furthermore define C as the censoring time, by which we mean the time the observation of a subject stops. Censoring, or right-censoring, occurs often in medical studies, or other survival studies where the time of study is not large relative to the event times. It can happen that for some subject $C < T$, in which case we don't know the time of event for this subject. Now, to get more formally, we say that a person is at risk for the event at time t if the time of the event and the censoring time is greater than or equal to t, i.e. $\min\{T, C\} \geq t$. Furthermore, a subject is observered to undergo an event at time t if $T = t$ and $C \geq T$. We can now define the two processes that keep track of this information. Define

$$N_n(t) = \mathbf{1}(T \leq t, T \leq C) \tag{2.5}$$

for whether we observe subject n to have undergone the event. We see that $N_n(t)$ equals 1 if the event has been observed and 0 if the event hasn't occured or the subject is no longer under observation. To count the total amount of events upto time t among all subjects, we define

$$N(t) = \sum_{n=1}^{N} N_n(t). \tag{2.6}$$

To describe how many transitions are made exactly at time t, we write $dN_i(t) = N_i(t) - N_i(t-)$, where $t- = \lim_{s \nearrow t} s$, the limit to t from beneath. Usually we only have finitely many time intervals when applying it to a data set. When a subject undergoes an event at time t, this means that in the previous timestep the event has not yet occured, and it has at the current. Hence, we might just as well discretize the timeline and use the notation

$$\Delta N_n(t_k) = N_n(t_k) - N_n(t_{k-1}) \tag{2.7}$$

for subject n to undergo the event at time t_k, with $k \in \{1, ..., K\}$ an event time. Naturally follows the total number of transitions at time t_k:

$$\Delta N(t_k) = N(t_k) - N(t_{k-1}) \tag{2.8}$$

We next define the at-risk process:

$$Y_n(t) = \mathbf{1}(\min\{T, C\} \geq t), \tag{2.9}$$

which equals 1 if subject n is at risk for the event, that is, if the event hasn't occured yet and the subject is still under observation. The total subjects at risk is obtained by summing over n:

$$Y(t) = \sum_{n=1}^{N} Y_n(t). \tag{2.10}$$

2.2.1 Nonparametric estimation with Nelson-Aalen and Kaplan-Meier

An estimate for the increments of A in the finite product (2.4) is given by the fraction of observed events at t and the number at risk just prior to t [4, p. 16]:

$$\Delta \hat{A}(t) = \frac{\Delta N(t)}{Y(t)}. \tag{2.11}$$

An individual is *observed* to fail at t if $\big(T \in [t, t + dt)\big) \wedge \big(T \leq C\big)$, that is if T is in the infinitesimal increment, and T occurs before or during the censoring time C. Because dt is so small, we say it's also *observed* if only $T \in [t, t + dt)$, that is $T \leq C$ is not necessary.

By adding the increments of (2.11) upto a time t, we now obtain an estimate for the cumulative hazard at t, the Nelson-Aalen [4, p. 16]:

$$\hat{A}(t) := \sum_{\min\{T_k, C_k\} \leq t} \frac{\Delta N(\min\{T_k, C_k\})}{Y(\min\{T_k, C_k\})}. \tag{2.12}$$

We can use the increments of the Nelson-Aalen estimator to estimate the survival funtion in the form of (2.4). This is called the Kaplan-Meier [4, p. 16] estimator:

$$\hat{S}(t) := \prod_{\min\{T_k, C_k\} \leq t} \big(1 - \Delta \hat{A}(\min\{T_k, C_k\})\big). \tag{2.13}$$

The Nelson-Aalen estimator is assymptotically normal, the distribution of the Kaplan-Meier follows deterministically (cf. Section 3.2 for variance estimate).

2.2.2 Semiparametric estimation with a Cox proportional hazard model

In this section we introduce a model proposed by Cox (1972) that adds covariates into the model. It assumes that when changing the covariates, the hazard function changes proportionally to a baseline hazard function. Let $\mathbf{Z} = (Z_1, ..., Z_p)$ be the vector of p covariates and $\beta = (\beta_1, ..., \beta_p)$ be the vector of p regression coefficients that display the effects of the covariates on the hazard function. The *Cox proportional hazard* is then given by

$$\alpha(t \,|\, \mathbf{Z}) = \alpha_0(t) \cdot e^{\beta^T \mathbf{Z}}. \tag{2.14}$$

Here the baseline hazard is denoted as $\alpha_0(t)$, from which can be seen that it only depends on time and is the same for all possible values for the covariates. We see that by increasing some covariate Z_k, $k = 1, ..., p$, by a single unit, i.e. add the vector e_k, it will change the hazard function $\alpha(t)$ by a factor e^{β_k}:

$$\begin{aligned} \alpha(t \,|\, \mathbf{Z} + e_k) &= \alpha_0(t) \cdot e^{\beta^T (\mathbf{Z} + e_k)} \\ &= \alpha_0(t) \cdot e^{\beta^T \mathbf{Z}} \cdot e^{\beta_k} \\ &= \alpha(t \,|\, \mathbf{Z}) \cdot e^{\beta_k}. \end{aligned}$$

Hence the name proportional hazard model makes sense.

The Cox model is a regression model where we can estimate the regression coeffecients contained in β by maximizing the partial likelihood function. The partial likelihood function is different from the likelihood function that we're used to in maximum likelihood estimation, but still mostly has the same large sample theory as the standard maximum likelihood estimation [4, p. 95]. The theory behind the large sample theory and the partial likelihood lies in the field of martingale theory and is beyond the scope of this article. For now, the partial likelihood can be expressed in terms of the counting process for the risk set, $Y_n(t)$, and the counting process for the transition set, $N_n(t)$. The risk set will be given a weight necessary for the construction of the partial likelihood, given by

$$S^{(0)}(\beta, t) := \sum_{n=1}^{N} Y_n \cdot e^{\beta \mathbf{Z}_n}. \tag{2.15}$$

Subsequently, the partial likelihood function[1] will be expressed as

$$L(\beta) = \prod_{k=1}^{K} \prod_{n=1}^{N} \Big(\frac{e^{\beta^T \mathbf{Z}_n}}{S^{(0)}(\beta, t_k)} \Big)^{\Delta N_n(t_k)}, \tag{2.16}$$

where $N_n(t)$ is the counting process of subject n, and t_k for $k = 1, ..., K$ are all event times [4, p. 94].

In order to interpret this likelihood, we will take a closer look at the terms. Recall from (2.2) that the hazard function times dt is the probability of failing at time t given that one has survived upto t. From this we can deduce that if subject m has $Y_m(t) = 1$, i.e. subject m is at risk, then the probability of failing is given by $\alpha(t \,|\, \mathbf{Z}_m)\, dt$; and if $Y_m(t) = 0$, then the subject is not at risk, hence has probability zero of failing. Equivalently, using the event times from the data set to replace the infinitesimal time intervals dt with $\Delta t_k = t_k - t_{k-1}$, the probability of failing at time t_k for subject m is *approximately* $\alpha(t \,|\, \mathbf{Z}_m)\, \Delta_{t_k}$ if $Y_m(t_k) = 1$ and zero otherwise. From this it follows that the probability that a specific subject m fails at time t_k, given that a subject fails at time t_k is approximately given by

$$P(\text{subject } m \text{ fails at } t \,|\, \text{a subject fails at time } t) \approx \frac{Y_m(t) \cdot \alpha(t \,|\, \mathbf{Z}_m)}{\sum_{n=1}^{N} Y_n(t) \cdot \alpha(t \,|\, \mathbf{Z}_n)}.$$

By writing the hazard function according to the Cox model in (2.14), we see that this can be also be written as

$$\frac{Y_m(t) \cdot e^{\beta \mathbf{Z}_m}}{\sum_{n=1}^{N} Y_n(t) \cdot e^{\beta \mathbf{Z}_n}},$$

where the baseline hazard $\alpha_0(t)$ has canceled out. To finish the interpretation of the partial likelihood in (2.16), one should now see that the above quantity is equal to the product terms in (2.16). The denominator in the quantity above is the weighted risk set present in the denominator of the product terms. Additionally whenever $\Delta N_m(t) = 1$, we have that subject m fails at time t, and so was at risk at time t, i.e. $Y_m(t) = 1$, meaning that also $Y_m(t) \cdot e^{\beta \mathbf{Z}_m} = (e^{\beta \mathbf{Z}_m})^{\Delta N_m(t)}$; whenever $\Delta N_m(t) = 0$, the term will equal 1 and so doesn't affect the product.

Hence the partial likelihood function is the probability of having all events in that specific order that they have occurerd in the data set. The estimator $\hat{\beta}$ of the regression coefficients is now obtained by maximizing $L(\beta)$.

After having estimated β by maximizing the partial likelihood (2.16), we can use this estimate for the proportional hazard model (2.14). To do so want to establish an estimate for the baseline hazard in (2.14) in order to have an estimate for the hazard. This can be done by estimating the cumulative hazard using Breslow's estimator:

$$\hat{A}_0(t) := \sum_{T_k \wedge C_k \leq t} \frac{\Delta N(T_k \wedge C_k)}{S^{(0)}(\hat{\beta}, T_k \wedge C_k)}. \tag{2.17}$$

[1] This is Breslow's approximation to the partial likelihood function for handeling tied event times. If we have K event times, the exact partial likelihood function is $L(\beta) = \prod_{k=1}^{K} \frac{e^{\beta^T \mathbf{Z}_k}}{S^{(0)}(\beta, t_k)}$, where $\mathbf{Z}_k$ corresponds to the covariates of the individual that makes a transition at t_k. The exact method requires that all event times are different on a continuous time scale, something that is mostly costly when acquiring data (e.g., acquiring data on a daily basis costs much less than on an exact time basis).

It is basically the Nelson-Aalen estimator for the covariate case, where the risk set in the Nelson-Aalen estimator has been replaced with the weighted risk set $S^{(0)}(\hat{\beta}, t)$, taking into account the proportional effects of the covariates. The estimator for the cumulative cause-specific hazard follows naturally by integrating (2.14) and plugging in Breslow's estimator:

$$\hat{A}(t \mid \mathbf{Z}_n) = \hat{A}_0(t) e^{\beta^T \mathbf{Z}_n}. \tag{2.18}$$

3 Multistate model

We now turn from survival analysis, the most basic form of a multistate model, to more complex multistate models. They allow us to analyse processes that go through different intermediate states and finish in different absorbing states. In medical research this is very useful for modeling diseases that go through different phases. In this section we will use the tools from survival analysis to make inferences from a multistate model. Most procedures and results in this section are similar to the ones in survival analysis, only performed in matrix form. Additionally the assumption of Markovianity, which will be discussed shortly hereafter, is needed in the type of multistate models discussed in this article. Because a multistate model exists of multiple transitions, extra effort is needed in checking which transitions follow a Cox model.

We consider a multistate model $(X_t)_{t\geq 0}$ with finitely many states $\{0, ..., J\}$, $J \in \mathbb{N}$. All state transitions are right-continuous, meaning in our case the largest interval on which X_t is constant is left-closed and right-open. The possible transitions in the multistate model are reflected in the transition probability matrix

$$\mathbf{P}(s,t) = \big(P_{ij}(s,t)\big)_{ij}, \qquad i,j \in \{1, ..., J\}. \tag{3.1}$$

Here $P_{ij}(s,t)$ is the probability that the process is in state j at time t, given that it's in state i at $s < t$, that is

$$P_{ij}(s,t) = P(X_t = j \,|\, X_s = i), \qquad s < t. \tag{3.2}$$

A key assumption that we now make is the Markov assumption, necessary for the results we will discuss. It states that the above probability that the process is in j at t only depends on the state at s and not what happened before s[2], that is conditioning on the current state is the same as conditioning on the current state and the past. In notation this means

$$P(X_t = j \,|\, X_s = i) = P(X_t = j \,|\, X_s = i, X_u = l), \qquad \forall u < s, \forall l \in \{1, ..., J\}. \tag{3.3}$$

We will talk more about the Markov assumption after we introduced the hazards for the multistate model.

In standard survival analysis the hazard was the probability that the event would occur given that the event hasn't occurred upto time t. In a multistate model we don't speak of events having or not having occured, but instead speak of *state transitions*. As a consequence we talk about transition hazards, hazards that correspond to a specific transition. The hazard corresponding to the transition from $i \to j$ is given by

$$\alpha_{ij}(t) = P(X_{(t+dt)-} = j \,|\, X_{t-} = i), \qquad i \neq j, \tag{3.4}$$

where the notation $t- = \lim_{\nearrow t} s$ means the limit to t from beneath. The condition of the above probability states that the process is in state i just prior to t; the event of interest in the above probability states that the process is in j within the infinitesimal interval $[t, t + dt)$. This notation ensures us that the process won't visit an intermediate state between i and j, and that consequently the transition hazard represents the direct force of the process from i to j. Additionally we define

$$\alpha_{ll}(t) = -\sum_{j \neq l} \alpha_{lj}(t), \tag{3.5}$$

which will make more sense in the next section where we will write $\mathbf{P}(s,t)$ in terms of hazards. Finally, we define the cumulative transitions hazard $A_{ij}(t) := \int_0^t \alpha_{ij}(u)du$, together with its matrix form

$$\mathbf{A}(t) = \big(A_{ij}(t)\big)_{ij}, \qquad i,j = 0, ..., J. \tag{3.6}$$

3.1 Transition probability matrix and product integration

The transition probability matrix and the increments of the cumulative hazard matrix can be related in a form of product integration, just as in the previous section. This relation relies entirely on the

[2] In the case of time-dependent covariates, also conditioning on the covariates before time s should not affect the probability on the left hand side.

Markov assumption by using the Chapman-Kolmogorov equations. In order not to get side-tracked too much, this relation is discussed in appendix A in Section 7. The relation is given by

$$\mathbf{P}(s,t) = \prod_{s}^{t} (\mathbf{I} + d\mathbf{A}(u))\, du \tag{3.7}$$

$$\approx \prod_{k=0}^{K} (\mathbf{I} + d\mathbf{A}(t_k)). \tag{3.8}$$

3.2 Estimation

In this section we will discuss both non- and semi-parametric estimation for the multistate model. The estimation for the multistate model is done in similar fashion to the estimation in survival analysis in 2.2. The nonparametric estimation in the multistate case will involve the Nelson-Aalen estimator for cumulative transition hazards and the Aalen-Johansen estimator for the transition probability matrix. The semiparametric will again include covariates and will additionally assume that transitions of interest follow a Cox model. For both the nonparametric and semiparametric, we will have to introduce transition specific processes.

Consider a transition $i \to j$ of the multistate model with $i, j \in \{0, ..., J\}$. Let T be the time that a subject makes the transition without visiting an intermediate state. Furthermore let C be the censoring time for this transition. A study always has an end-of-study time, and the censoring is always less than or equal to the end-of-study time. We say that a transition is also censored when a state other than j is the first state visited after i. Then subject n is said to have made the transition $i \to j$ by time t if

$$N_{ij;n}(t) = \mathbf{1}(L_n \leq T_n \leq C_n,\, T_n \leq t,\, X_{T_n}^{(n)} = j,\, X_{(T_n)-}^{(n)} = i). \tag{3.9}$$

This is probably better understood when explained in words. We have $N_{ij;n}(t) = 1$ when the time of the event T_n is within the observation interval, i.e. between L_n and C_n; the event has occurred before t; the state visited at time T_n is j; just prior to T_n the subject is in state i. The latter is necessary to ensure that $N_{ij;n}(t)$ counts a direct transition of $i \to j$ at t. In short, it equals 1 if subject n is observed to have made a transition $i \to j$ upto time t. The total number of observed transitions $i \to j$ is then given by

$$N_{ij}(t) = \sum_{n=1}^{N} N_{ij;n}(t). \tag{3.10}$$

The quantity $\Delta N_{ij}(t) = N_{ij}(t) - N_{ij}(t-)$ will tell us the exact number of transitions $i \to j$ occur at time t.

One would presume that we would now define a risk set for a transition $i \to j$, but instead we use

$$Y_{i;n}(t) = \mathbf{1}(L_n \leq T_n \leq C_n,\, X_{t-}^{(n)} = i), \tag{3.11}$$

only telling us whether subject n is in state i just prior to t and is under observation. Note that this quantity doesn't contain a state the transition moves into. The total number of subjects under observation in state i just prior to t is given by

$$Y_i(t) = \sum_{n=1}^{N} Y_{i;n}(t). \tag{3.12}$$

3.2.1 Nonparametric estimation with Nelson-Aalen and Aalen-Johansen

The Nelson-Aalen estimator for transition $i \to j$ in the multistate model is defined in a similar fashion as the estimator for cumulative hazard (2.12) in survival analysis. An estimate for the increments of the cumulative hazard matrix, $\Delta\mathbf{A}(t)$ is given by $\hat{\mathbf{A}} = \left(\hat{A}_{ij}(t)\right)_{ij}$, $i, j = 0, ..., J$, where the elements are

$$\Delta\hat{A}_{ij}(t) = \frac{\Delta N_{ij}(t)}{Y_{ij}(t)}. \tag{3.13}$$

The Nelson-Aalen estimator is given by the sum over these increments, $\hat{\mathbf{A}}(t) = \big(\hat{A}_{ij}(t)\big)_{ij}$, with elements

$$\hat{A}_{ij}(t) = \sum_{\min\{T_k, C_k\} \leq t} \frac{\Delta N_{ij}(\min\{T_k, C_k\})}{Y_i(\min\{T_k, C_k\})}, \tag{3.14}$$

where $i, j = 0, ..., J$. As mentioned in Section 2.2, the Nelson-Aalen estimator is asymptotically normally distributed in the survival analysis case. This is also the case for the multistate model. An estimate for the variance of the Nelson-Aalen estimate is given by

$$\hat{\sigma}^2_{ij} := \sum_{\min\{T_k, C_k\} \leq t} \frac{\Delta N_{ij}(\min\{T_k, C_k\})}{Y_i^2(\min\{T_k, C_k\})}, \tag{3.15}$$

where $i, j = 0, ..., J$.

The increments of the Nelson-Aalen estimator can be used to estimate the transition probability matrix in its form of (3.8) by approximating it with a finite product over all K transition times. The result is the empirical transition matrix, or Aalen-Johansen estimator,

$$\hat{\mathbf{P}}(s,t) = \prod_{s=t_1,...,t_K=t} \big(\mathbf{I} + \Delta\hat{\mathbf{A}}(t_k)\big). \tag{3.16}$$

This is a key estimator for later, when making predictions for specific covariates on the length of stay. Estimation of the variance of the Aalen-Johansen estimator becomes more difficult, and is not discussed here. The next section will cover how to compute the hazards for a semiparametric multistate model, and subsequently (3.16) will then also be used for computing probabilities.

3.2.2 Semiparametric estimation with a Cox proportional hazards model

In Section 2.2.2 the Cox model has already been introduced for an 'alive'-'death' model. We can just as easily use the Cox model for transitions of a multistate model, as long as we assume that such a transition follows a Cox model. The transition hazard for transition $i \to j$, with $i, j \in \{0, ..., J\}$ and $i \neq j$, dictated by this model are given by

$$\alpha_{ij}(t \,|\, \mathbf{Z}_n) = \alpha_{ij;0}(t) e^{\beta_{ij}^T \mathbf{Z}_n}, \tag{3.17}$$

where β_{ij} is the vector with transition-specific regression coefficients and $\alpha_{ij;0}(t)$ the transition-specific baseline hazard. We assume specifically that each transition follows a Cox model on its own, and so we cannot assume these transition-specific hazards share a common baseline hazard or share the same regression coefficients. We can however cleverly model a common effect of the covariates on different transitions, while at the same time leaving transition-specific effects of the covariates intact.

To take into account a that different transition hazards can be affected differently by covariates, we can construct a new vector of regression coefficients β that contains all vectors β_{ij}. The corresponding vector of the same length as β containing the covariates, is given by $\mathbf{Z}_{ij;n}$. This vector only contains the covariates for the same element positions of β_{ij} in β; all other elements are zero.

In the proportional hazards model in the multistate case one estimates the regression coefficients using the partial likelihood in a similar fashion

$$L(\beta) = \prod_t \prod_{n=1}^{N} \prod_{j=0}^{J} \prod_{i=0}^{J} \Big(\frac{e^{\beta^T \mathbf{Z}_{ij;n}}}{S_{ij}^{(0)}(\beta, t)} \Big)^{\Delta N_{ij;n}(t)}, \tag{3.18}$$

where $S_{ij}^{(0)}(\beta, t) := \sum_{n=1}^{N} e^{\beta \mathbf{Z}_{ij;n}} \cdot Y_{0;n}$ is the transition-specific weighted risk set. The power term $\Delta N_{ij;n}(t)$ equals one whenever individual n makes a transition $i \to j$ at time t. The estimate $\hat{\beta}$ of the regression coefficients is obtained by maximizing $L(\beta)$.

After having estimated β by maximizing the partial likelihood (3.18), we can use this estimate for the proportional transition-specific hazard model (3.17). We next want to establish an estimate for the transition-specific *baseline* hazard in (3.17) in order to have an estimate for the transition-specific

hazard. We can do so by estimating the cumulative transition-specific hazard model using Breslow's estimator:

$$\hat{A}_{ij;0}(t) := \sum_{T_k \wedge C_k \leq t} \frac{\Delta N_{ij}(T_k \wedge C_k)}{S_{ij}^{(0)}(\hat{\beta}, T_k \wedge C_k)}. \tag{3.19}$$

The estimator for the cumulative transition-specific hazard for an individual with covariates $\mathbf{Z} = z$ follows naturally by integrating (3.17) and plugging in Breslow's estimator:

$$\hat{A}_{ij}(t \,|\, z) = \hat{A}_{ij;0}(t) e^{\beta_{ij}^T z}. \tag{3.20}$$

In matrix form we then obtain

$$\hat{\mathbf{A}}(t \,|\, z) = \big(\hat{A}_{ij}(t \,|\, z)\big)_{ij}, \qquad i, j = 1, ..., J+1. \tag{3.21}$$

The increments of $\hat{\mathbf{A}}(t \,|\, z)$ can subsequently be used to compute an emperical transition probability matrix similar to the Aalen-Johansen for specific covariates. We obtain

$$\hat{\mathbf{P}}(s, t \,|\, z) = \prod_{s=t_1,...,t_K=t} \big(\mathbf{I} + \Delta \hat{\mathbf{A}}(t_k \,|\, z)\big). \tag{3.22}$$

The statistical properties of these predictions will not be discussed here (cf. [3]).

4 Expected length of stay

Thus far we have introduced what a multistate model is and how we can perform estimation on it. We will now discuss how to quantify the effect of a HI on the expected length of stay in the ICU using such a multistate model. The multistate model that we will use is visualized in Figure 1, and is mathematically defined by the process $(X_t)_{t\geq 0}$, with $X_t \in \{0, 1, 2, 3\}$. The states correspond exactly to those in Figure 1, making state 2 and 3 two absorbing states that terminate the hospital stay. Furthermore, we will make the assumption that the multistate model is a Markov process[3]. Let the random variable

$$T := \inf\{t > 0 \mid X_t \in \{2, 3\}\} \tag{4.1}$$

be the time that the patient enters one of the absorbing, i.e. the time of end of stay. We are now equipped to use the results from the previous section to quantify the effect of a HI on the expected length of stay. First we will quantify the effect in both a non- and semi-parametric model as a function of time; and after we will provide a scalar summary measure by additionally adding the subdistribution of the infection as a weight.

4.1 Subsequent expected length of stay given infectious state as function of time

The method is first given in the form of a nonparametric model. The method for the semiparametric case is exactly the same, except that in that case it is a prediction for a patient with specific covariates instead of estimation. The quantity we use is proposed by Schulgen and Schumacher [13] and is given by

$$\phi(s) = E(T \mid X_s = 1) - E(T \mid X_s = 0). \tag{4.2}$$

It describes the difference in expected length of stay between patients who are infected at time s and patients who are not infected at time s. The two terms on the right-hand side are given by

$$E(T \mid X_s = 0) = s + \int_s^\infty P_{00}(s, t) + P_{01}(s, t)\, dt, \tag{4.3}$$

and

$$E(T \mid X_s = 1) = s + \int_s^\infty P_{11}(s, t)\, dt. \tag{4.4}$$

To estimate the quantities (4.3) and (4.4), we have to take two things into consideration. First, we don't know if the integral converges in both cases. Second, the Nelson-Aalen (and thereby also the Aalen-Johansen) estimator is based on the obsersvations corresponding to event times. Because we only have observations during study time, we can only estimate up to the end-of-study time. Our best approximation for $\phi(s)$ is then given by

$$\phi(s) = \int_s^\tau P_{00}(s, t) + P_{01}(s, t)\, dt - \int_s^\tau P_{11}(s, t)\, dt, \tag{4.5}$$

where τ is the end-of-study time. In the case of a semiparametric model we can use Breslow's estimator (3.21) to predict the transition probabilities. These can subsequently be used to predict $\phi(s)$ deterministicly. In this case we would obtain

$$\phi(s; z) = \int_s^\tau P_{00}(s, t; z) + P_{01}(s, t; z)\, dt - \int_s^\tau P_{11}(s, t; z)\, dt. \tag{4.6}$$

Obtaining variance estimators for the estimators of $\phi(s)$ and $\phi(s, z)$ has not been done in this article. Bootstrapping could be performed to obtain variability measures, but has not been done in this article due to the lack of time.

[3]The practical context of the Markov assumption will be discussed in the next section.

4.2 Summary measure for expected length of stay

The quantity $\phi(s)$ is always dependent on the time s, and so only tells us something about the effect of the infection on a specific time. Additionally we can provide a physician with a time-averaged summary that can be interpreted more directly. To construct such a summary measure, we would have to weight $phi(s)$ with a distribution depending on s. One such a weight, proposed by Schulgen and Schumacher [13], is the *subdistribution* function of the infection state. Defining $T_0 := \inf\{t > 0 \,|\, X_t \neq 0\}$, the subdistribution function is given by

$$P(T_0 \leq t \,|\, X_{T_0} = 1). \tag{4.7}$$

The interpretation of this weight is that of weighting by attributable risk based on stratifiction, if each day s would be considered as a stratum [2, p. 190].

Another weight, proposed by Allignol et al. [2, p. 190], is to weight according to $P(T_0 \leq t)$. With the Markov assumption, the interpretation is that group membership, i.e. either acquiring HI during stay or not, becomes definite at T_0.

The latter weight only provides a description of whether a patient gets infected during his stay, while the subdistribution (4.7) describes the probability of acquiring a HI. In this article we want to know what the effect is of the HI on the length of stay, and so it makes more sense to weight with according to attributable risk. Hence we will use the subdistribution (4.7) as a weight, yielding in the summary measure

$$\theta := \int_0^\tau \phi(s) \cdot P(T_0 \leq t \,|\, X_{T_0} = 1). \tag{4.8}$$

The estimation of the subdistribution of the infectious state is as follows. We first define the new random variable $\vartheta := T_0 \cdot \mathbf{1}(X_{T_0} = 1)$, which is equal to T_0 if the infection state is visited and infinite if the patient goes directly to an absorbing state. In a medical study, we will in the latter case censor this with the end-of-study time, τ. We can now estimate $P(T_0 \leq t \,|\, X_{T_0} = 1) = P(\vartheta \leq t)$ using results from Section 3.2.1.

For a semiparametric model, estimation of θ follows naturally. Estimation of $\phi(s)$ is performed by predicting the empirical transition probability matrix and computing the expected values for the end-of-stay time. Preferably, the subdistribution for infection is predicted as well for specific covariates, however due to lack of time this has not succeeded. Instead we used the homogeneous subdistribution for infection. This results in the summary measure

$$\theta(z) := \int_0^\tau \phi(s; z) \cdot P(T_0 \leq s \,|\, X_{T_0} = 1)\, ds. \tag{4.9}$$

Also the quantities (4.8) and (4.9) lack variance estimates for there estimators. Bootstrapping would be an option, but has not been done in this article due to the lack of time.

5 Effect of CMV reactivation on the expected length of stay in ICU

The purpose of this article is to find out what the effect of CMV reactivation is on the expected length of stay in the ICU. One way to do so is to include CMV reactivation as a baseline covariate (cf. Table 2). However, this would ignore the fact that acquiring the infection is time-dependent. To include the time-dependency of the infection we decided to model the infection and the end of stay (death or discharge) as a multistate model. The model is defined in similar fashion as in the introduction of Section 4 and Figure 1. For clarity we provide the model again, now formally focused on the effect of CMV reactivation on death and discharge. We define the multistate process as

$$(X_t)_{t\geq 0} \qquad X_t \in \{0, 1, 2, 3\}, \tag{5.1}$$

where the states $\{0, 1, 2, 3\}$ correspond to the labels in schematic representation in Figure 1. It is

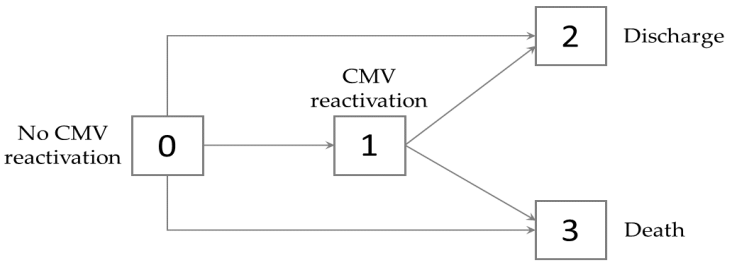


Figure 1: A schematic visualisation of the multistate model with CMV reactivation state.

assumed that the model satisfies the Markov assumption, and so the probabilities of making a transition only depend on the current time and not the time one has spend in the current state (cf. Section 3.1). Furthermore let

$$T := \inf\{t > 0 \,|\, X_t \in \{2, 3\}\} \tag{5.2}$$

be the time that the patient either dies or is discharged, i.e. the time of end of stay.

In this section we will analyse the effect of CMV reactivation on the expected length of stay using a data set from the ICU of the Utrecht Medical Centre. First we will provide a description of the problem and the data set. We will then perform nonparametric estimation and present the cumulative hazards of all the transitions and the transition probabilities, estimated by the Nelson-Aalen and Aalen-Johansen estimators repectively. These will describe the process of hospital stay and already allow us to draw decent conclusions. Next, we attempt to quantify these results so that a physician is able to interpret these. Ultimately, we will additionally assume the transitions follow a Cox proportional hazards model so that we can check the effect of CMV reactivation for patients with different covariates.

5.1 Problem description

In the data set provided by the Utrecht Medical Centre, we have a record of 271 patients on the Intensive Care Unite (ICU). Patients are followed for a maximum of 35 days, which is the end-of-study time. During stay there is a group of 67 patients (24.7%) who acquire the CMV reactivation, the HI of interest, before having left the ICU with discharge or death. In Table 1 we see the characteristics of the patients that are included in the current research. Both age and BMI seem to be similar across both groups. The APACHE IV score[4] differs significantly ($p = 0.003$) between the two groups, with a median

[4]Acute Physiology and Chronic Health Evaluation IV is a scoring system to quantify the severity of the illness of patients.

score of 91 for CMV reactivation group and 76 for the non-reactivation group. Mechanical ventilation is given as a baseline characteristic for patients who have mechanical ventilation on admission day. Of the CMV reactivation group, every patient has mechanical ventilation on day 1.

Table 1: Characteristics for patients by CMV reactivation status

	Reactivation ($n = 67$)[a]	Non-reactivation ($n = 204$)	p-value
Age	63	64.5	0.841
BMI[b]	25.99	25.42	0.462
APACHE IV Score	91	76	0.003
Mechanical ventilation	67	194	0.141

The second and third column contains medians (mechanical ventilation frequencies); the last column provides p-values from the nonparametric Mann-Whitney test (mechanical ventilation Chi-squared test).

[a] Six patients who had reactivation after death or discharge were put in the non-reactivation group.

[b] Two patients' BMI have missing values.

By lookig at Table 2 we see that a crude analysis finds that CMV reactivation has a significant effect on the length of stay, death and discharge. We find that the CMV reactivation group stays longer in the ICU (16 as opposed to 8 days), have more relative deaths[5] (23/67 as opposed to 29/204), and have less relative discharges (36/67 as opposed to 167/204) than the non-reactivation group. The test for the differences in deaths and discharges is done with a Chi-squared test. Additionally it becomes clear that it takes time for the CMV reactivation to take place (median reactivation time = 9).

Table 2: Crude clinical outcomes of patients by CMV reactivation status

	Reactivation ($n = 67$)[a]	Non-reactivation ($n = 204$)	p-value
Death [b]	23	29	< 0.001
Discharge	36	167	< 0.001
Time of stay	16	8	< 0.001
Reactivation time	9	-	-

The second and third column contain frequencies (median for time of stay); the last column provides p-values from the Chi-squared test (and Mann-Whitney test for time of stay).

[a] Six patients who had reactivation after death or discharge were put in the non-reactivation group.

[b] Three patients had death after discharge or end-of-study, and were chosen to be censored for death.

5.2 Nelson-Aalen and Aalen-Johansen estimates

To describe the effect of CMV reactivation in the multistate model we computed the Nelson-Aalen estimates (3.14) and Aalen-Johansen estimates (3.16).

The Nelson-Aalen estimates of the cumulative hazards for all the transition but the infection transition are plotted in Figure 2. The variance estimates that are used for the confidence bands are Aalen-types, as in (3.15). The confidence intervals could have been improved by using a log-minus-log transformation on the pointwise confidence intervals, but this has not been done in this article. If we compare the discharge hazards, we see that the Nelson-Aalen estimates have a pretty similar graph, but the variability for the CMV reactivation group is larger.

If we compare the death hazards between the two groups, we see that the Nelson-Aalen estimates for the CMV reactivation-to-death transition is above the one for the No CMV reactivation-to-death transition. This implies that over the whole study, the CMV reactivation causes a higher death hazard. Furthermore we see that the variability for the CMV reactivation group is larger, probably due to the lower risk set in this group.

The Nelson-Aalen estimate for the cumulative infection hazard is presented in Figure 3a. Also this confidence band is constructed using (3.15) as a variance estimator; a log-minus-log transformation could have been performed on the pointwise confidence intervals, but has not been done. We see that up to day 7 the Nelson-Aalen estimates increase only twice, and that starting from day 7 infection takes

[5] Three patients had died after having been discharged or after the end of study. It was decided to censor their deaths.

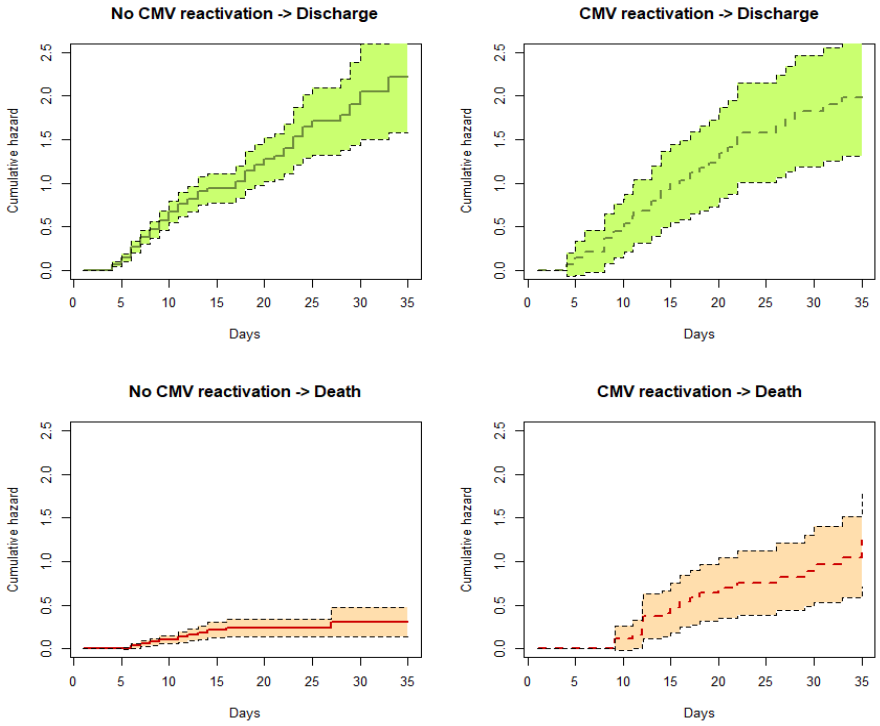
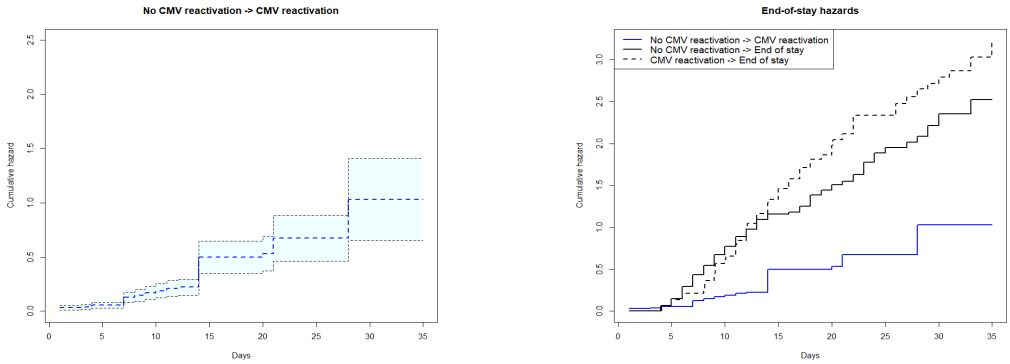


Figure 2: The Nelson-Aalen estimates of the cumulative hazards with pointwise confidence bands.

place more frequently. Starting from day 14 infection happens less frequently, but when it occurs, it occurs to more patients causing the high jumps in the plot. As time progresses, patients move to the infection state or an absorbing state, causing the risk set for the infection transition to decline. As a consequence we see the confidence intervals get wider for larger time values.



(a) The Nelson-Aalen estimates of the cumulative infection hazard with pointwise confidence bands.

(b) Nelson-Aalen estimates for the end-of-stay hazards and the infection hazard.

Aggregated over the two groups, we see the end-of-stay hazards and the infection hazard in Figure 3b. We see readily that the estimate for the CMV reactivation group is above the non-infected group, and so we expect the infected group to leave the ICU faster than the non-infected group.

This can also be observed by plotting the Aalen-Johansen estimates. These can be found in Figure 4. We see that the red area in the right plot is slightly more convex after day 14 than the left plot. As a result we will see that the expected length of stay given infection state at day 10 will be higher for the infected group. Note that the expected length of stay given infection state at time s is the area of the complement of the red area (Figure 4: left for not infected at day 10; right for infected at day 10).

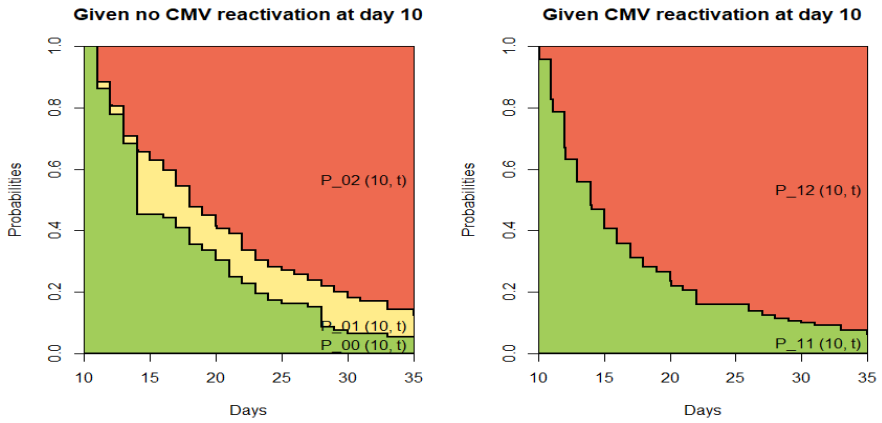


Figure 4: Transition probabilities given infection state at day 10 from non-infected state (left) and infected state (right).

5.3 Quantification

In the previous section we could informally deduce from the Nelson-Aalen estimate plots that over the whole study time, i.e. 35 days, the end-of-stay hazard is larger for the infection state. We now use the estimate (4.5) to quantify this suspicion in Figure 5 We see that for most days s, we expect a negative

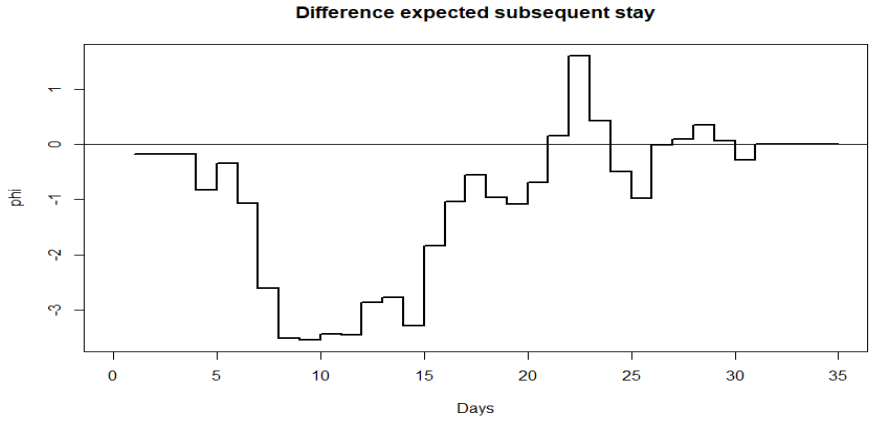


Figure 5: Estimate of the difference in expected subsequent stay given infection state.

difference $phi(s)$. This means that if we know a patient has CMV reactivation on such a day, we expect the patient to leave the hospital faster than a patient without CMV reactivation.

Additionally we provide a summary measure that attaches a time-averaged weight on $\phi(s)$. The weight that is attached is the subdistribution of the infection, given by $P(T_0 < t \,|\, X_{T_0} = 1)$, where $T_0 := \inf\{t > 0 \,|\, X_t \neq 0\}$ (Figure 6). The weight represents the attributable risk of the infection per

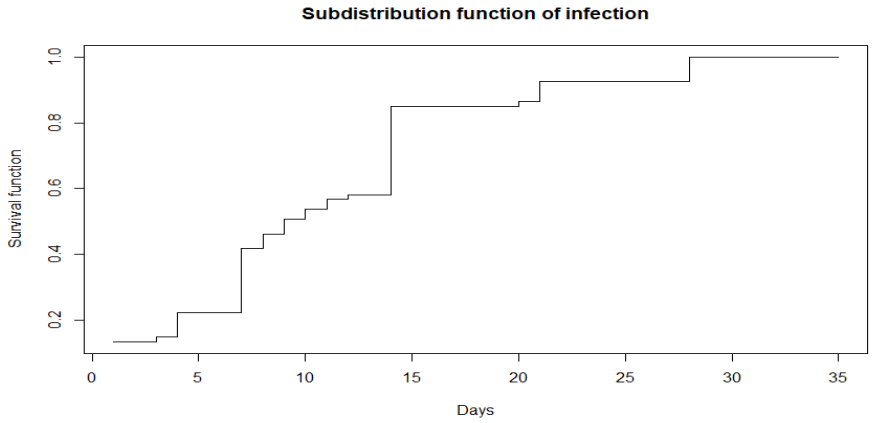


Figure 6: The subdistribution function of CMV reactivation, i.e. $P(T_0 \leq t \,|\, X_{T_0} = 1)$, with $T_0 := \inf\{t > 0 \,|\, X_t \neq 0\}$.

day. We obtain

$$\begin{aligned}\hat{\theta} &= \int_0^\tau \hat{\phi}(s) \cdot \hat{P}(T_0 < t \,|\, X_{T_0} = 1) \\ &= -19.7.\end{aligned}$$

The summary measure should not be interpreted as the expected prolongation due to the infection.

5.4 Predictions

We will now make the additional assumption to our model that it follows a Cox proportional hazards model. This means that each transition hazard in our model behaves in a proportional manner to changes in the covariates. The covariates that are included in this study are the characteristics age, BMI and APACHE IV score, presented in Table 1. Only mechanical ventilation has been left out, because every infected patient had mechanical ventilation on admission, leading to unreliable results. To take into account possible different effects of a covariate on different transitions, we define the vector with regression coefficients as such

$$\beta = (\beta_{01}, \beta_{02}, \beta_{03}, \beta_{12}, \beta_{13})^T, \tag{5.3}$$

where

$$\beta_{ij} = (\beta_{ij;age}, \beta_{ij;BMI}, \beta_{ij;APACHE}, \beta_{ij;MV}), \qquad i = 0, 1,\ i < j = 1, 2, 3. \tag{5.4}$$

Corresponding to this vector, we construct the vector $\mathbf{Z}_{ij}$ with covariates for a specific transition $i \to j$, $i = 0, 1$ and $i < j = 1, 2, 3$. The vector contains as many zero's as the length of β, only on the location of β_{ij} we place the values of age, BMI, APACHE_IV and MV respectively.

By maximizing Breslow's approximation of the partial likelihood function (3.18), we obtain the coefficients for β, presented in Table 3.

We see that age has a significant effect on the hazard for transition $0 \to 3$, i.e. on the death hazard for non-infected patients. Furthermore, the APACHE IV score has a significant effect on the hazards of the transitions of non-infected patients. For non-infected patients, a higher APACHE IV score results in a higher infection hazard, a lower discharge hazard and a higher death hazard.

In Figure 7 we plotted the Nelson-Aalen estimates of predictions for a patient with a high APACHE IV score (120) and a patient with a low APACHE IV score (20) for comparing the two. The age and BMI were in both cases given the mean value across the whole data set. Care must be taken when comparing the two patient types. For the non-infected state, the APACHE IV score was a good predictor for all three transitions from that state. However, for the infected state non of the covariates age, BMI and APACHE IV were considered a good predictor. Hence, the differences in the discharge and death hazards for the low and high APACHE IV patients should not be trusted.

Table 3: Cox regression coefficients

	coef	exp(coef)	se(coef)	z	p	
age_{01}	0.01	1.01	0.01	1.11	0.27	
age_{02}	0.01	1.01	0.01	0.91	0.36	
age_{03}	0.04	1.04	0.02	2.58	0.01	***
age_{12}	-0.01	0.99	0.01	-1.03	0.30	
age_{13}	0.02	1.02	0.02	0.75	0.45	
BMI_{01}	0.00	1.00	0.02	0.09	0.93	
BMI_{02}	0.01	1.01	0.01	0.49	0.62	
BMI_{03}	0.01	1.01	0.03	0.41	0.68	
BMI_{12}	0.04	1.04	0.04	1.04	0.30	
BMI_{13}	-0.02	0.98	0.05	-0.38	0.70	
APACHE IV_{01}	0.01	1.01	0.00	1.87	0.06	*
APACHE IV_{02}	-0.01	0.99	0.00	-3.73	0.00	****
APACHE IV_{03}	0.01	1.01	0.01	1.68	0.09	*
APACHE IV_{12}	-0.00	1.00	0.01	-0.35	0.72	
APACHE IV_{13}	0.01	1.01	0.01	0.65	0.52	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$

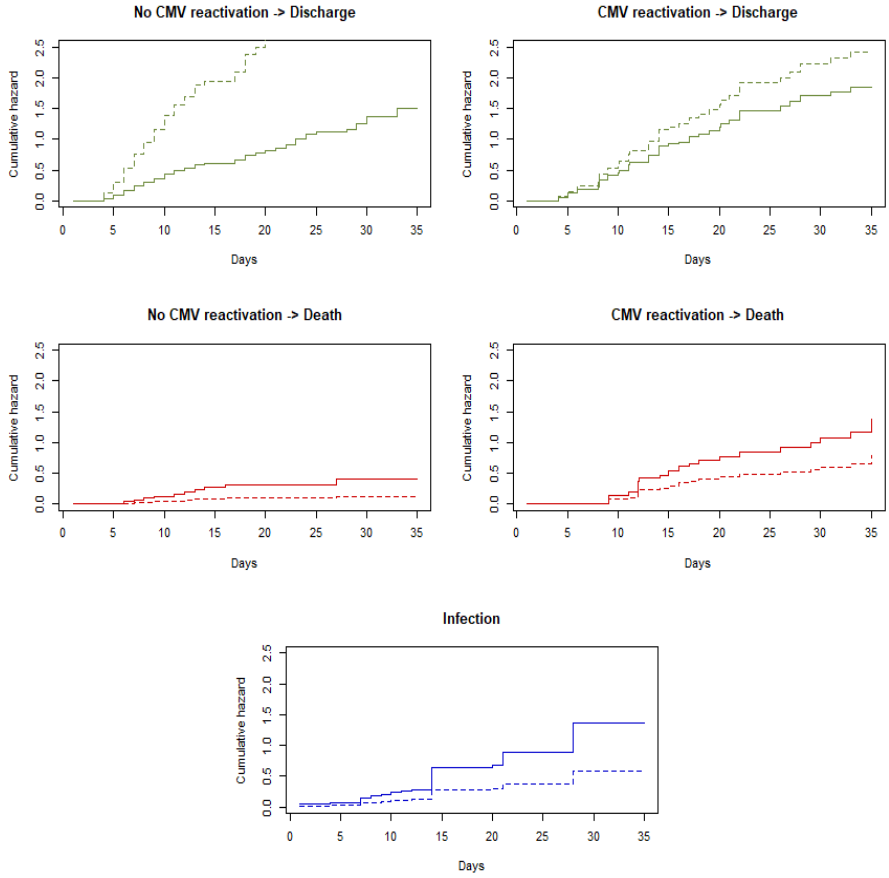


Figure 7: Nelson-Aalen estimates of the cumulative hazards for patient with high APACHE (solid line) and low APACHE (dotted line).

The estimates for the difference in expected subsequent stay given infection state, $\phi(s)$, is plotted in Figure 8 for the patient with a high APACHE IV score (120) and the patient with a low APACHE IV score (20). We can see that for the patient with the low APACHE (dotted line), the effect of the

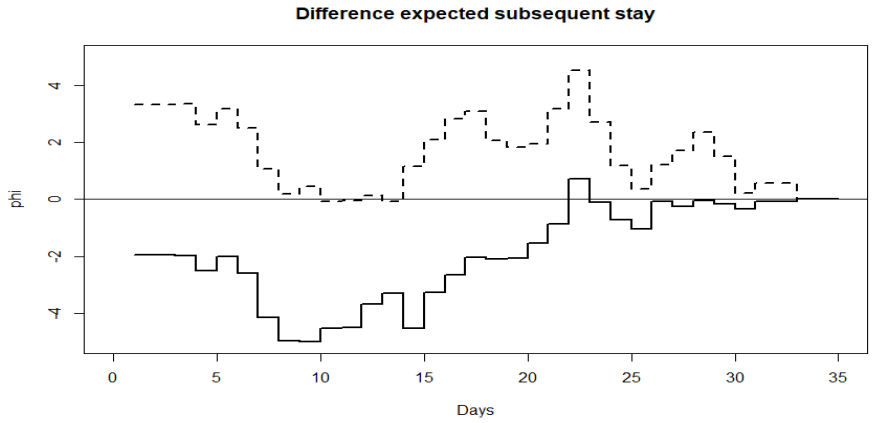


Figure 8: Estimate of the difference in expected subsequent stay given infection state for patient with high APACHE (solid line) and low APACHE (dotted line).

infection on the length of stay is prolonging. This can be explained by the high discharge hazard for non-infected low APACHE patients (Figure 2, top-left, dotted line). For a patient with low APACHE the discharge hazard drops enormously after being infected. In comparison, the difference in death hazards for the patient with low APACHE is less. This results in the prolongation of ICU stay due to infection.

Computing the summary measures (4.9) for the patient with low APACHE, $Z = z_L$, and for the patient with high APACHE, $Z = z_H$, results in the following

$$\begin{aligned}\hat{\theta}(z_L) &= \int_0^\tau \hat{\phi}(s; z_L) \cdot \hat{P}(T_0 \leq s \,|\, X_{T_0} = 1)\, ds \\ &= 38.8,\end{aligned}$$

and

$$\begin{aligned}\hat{\theta}(z_H) &= \int_0^\tau \hat{\phi}(s; z_H) \cdot \hat{P}(T_0 \leq s \,|\, X_{T_0} = 1)\, ds \\ &= -36.4.\end{aligned}$$

Note that the homogeneous subdistribution for the infection is used, and that no prediction is performed for the patients' covariates. This could change the subdistribution in Figure 6, and thereby the weights.

6 Discussion

In the current study we wanted to find out what the effect of CMV reactivation is on length of stay for patients in the ICU. It was already found to be associated with prolongation of length of stay in the ICU [11], but we have found an opposite effect. Instead, we found that given infection status of a patient at time s, the expected subsequent stay was lower for the infected patient. By looking at the Nelson-Aalen estimates of the specific transitions, we could conclude that the lower expected subsequent stay for an infected patient is due to a higher death hazard. This result is in line with the association of CMV reactivation with increased mortality, found in [11].

The expected length of stay for a patient who is infected at time s is longer than for a patient who is not infected at time s, for most times s. This could be seen from the quantity proposed by [13]. A subsequent summary measure, which added a weight of attributable risk of the infection per day, was estimated by $\hat{\theta} = -19.7$. Two things should be noted on the summary measure θ. First, there is no direct interpretation from the magnitude of θ, and it might need to be held in proportion to, e.g., the mean stay for patients. Second, this summary measure does not contribute if it's negative and the HI of interest is malignant. Only if the difference between the cumulative discharge hazards for infected patients and non-infected patients is higher than the same difference in cumulative death hazards, only then is the summary measure θ informative on prolongation of stay due to the infection. What not has been done in this study is to split up $\phi(s)$ into contribution of death and contribution of discharge.

Additionally we have found that APACHE IV is a significant predictor for the effect on all hazards for non-infected patients. We have however not found APACHE IV to be a significant predictor for the hazards of infected patients. Having this in mind, we could predict a high difference for θ between patients with a high APACHE IV and low APACHE IV, which is already very promising. The assumption that the model follows a Cox proportional hazards model is a heavy assumption, and needs more care in modelling to be satisfied.

7 Appendix A: Product integration and Chapman-Kolmogorov equations

In this section we will establish the relation (3.7). Using the Markov property, we have $P(X_t = j \,|\, X_s = i) = \sum_{l=0}^{J} P(X_t = j \,|\, X_u = l) \cdot P(X_u = l \,|\, X_s = i)$, resulting in the Chapman-Kolmogorov equation:

$$\mathbf{P}(s,t) = \mathbf{P}(s,u)\mathbf{P}(u,t).$$

This can be used to obtain the Kolmogorov forward differential equation

$$\frac{\partial}{\partial t}\mathbf{P}(s,t) = \mathbf{P}(s,t)\mathbf{a}(t) \tag{7.1}$$

, by noting that

$$\begin{aligned} P(X_t = j \,|\, X_s = i) &= \sum_{l=0}^{J} P(X_t = j \,|\, X_{(s+ds)-} = l) \cdot P(X_{(s+ds)-} = l \,|\, X_{s-} = i) \\ &= \sum_{\substack{l=0 \\ l \neq i}}^{J} P(X_t = j \,|\, X_{(s+ds)-} = l) \cdot \alpha_{il}(s)\, ds \\ &\qquad + P(X_t = j) \,|\, X_{(s+ds)-} = l) \cdot (1 - \alpha_{ii}(t)). \end{aligned}$$

Then (7.1) can be rewritten as

$$\mathbf{P}(s,t) = \mathbf{I} + \int_s^t \mathbf{P}(s,u-) d\mathbf{A}(u) \tag{7.2}$$

Using the Chapman-Kolmogorov equations we can write

$$\mathbf{P}(s,t) = \mathbf{P}(t_0,t_1)\mathbf{P}(t_1,t_2)\cdots\mathbf{P}(t_{K-1},t_K). \tag{7.3}$$

Now, take $\mathbf{P}(s,t_{i+1}) - \mathbf{P}(s,t_i)$ as the approximation for the left-hand side of (7.1). Then we can write

$$\mathbf{P}(s,t_i) - \mathbf{P}(s,t_{i-1}) \approx \mathbf{P}(s,t_{i-1})\big(\mathbf{A}(t_i) - \mathbf{A}(t_{i-1})\big).$$

Rewriting gives

$$\mathbf{P}(s,t_i \,|\, t_{i-1}) \approx \mathbf{I} + \big(\mathbf{A}(t_i) - \mathbf{A}(t_{i-1})\big).$$

Because of the Markov assumption, we have that $\mathbf{P}(s,t_i \,|\, t_{i-1}) = \mathbf{P}(t_{i-1},t_i)$, because $s < t_{i-1}$. This results in

$$\mathbf{P}(t_{i-1},t_i) \approx \mathbf{I} + \big(\mathbf{A}(t_i) - \mathbf{A}(t_{i-1})\big). \tag{7.4}$$

If we plug this into (7.3), we get

$$\mathbf{P}(s,t) \approx \prod_{k=1}^{K} \mathbf{I} + \big(\mathbf{A}(t_k) - \mathbf{A}(t_{k-1})\big). \tag{7.5}$$

For $d := \max_{1\leq k\leq K}\{t_k - t_{k-1}\}$, if we let $d \to 0$, we are assured that limit of products exists (cf. [8]). We then get

$$\begin{aligned} \mathbf{P}(s,t) &= \lim_{d\to 0} \prod_{k=1}^{K} \mathbf{I} + \big(\mathbf{A}(t_k) - \mathbf{A}(t_{k-1})\big) \\ &= \prod_{s}^{t} \big(\mathbf{I} + d\mathbf{A}(u)\big). \end{aligned}$$

References

[1] OO Aalen and S. Johansen. Empirical transition matrix for nonhomogeneous markov-chains based on censored observations. *Scandinavian Journal of Statistics*, 5:141–150, 1978.

[2] Arthur Allignol, Martin Schumacher, and Jan Beyersmann. Estimating summary functionals in multistate models with an application to hospital infection data. *Computational statistics*, 26(2):181–197, 2011.

[3] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[4] Jan Beyersmann, Arthur Allignol, and Martin Schumacher. *Competing risks and multistate models with R*. Springer Science & Business Media, 2011.

[5] Jan Beyersmann, Petra Gastmeier, Hajo Grundmann, Sina Bärwolff, Christine Geffers, Martin Behnke, Henning Rüden, and Martin Schumacher. Transmission-associated nosocomial infections: prolongation of intensive care unit stay and risk factor analysis using multistate models. *American journal of infection control*, 36(2):98–103, 2008.

[6] Jan Beyersmann, Martin Wolkewitz, and Martin Schumacher. The impact of time-dependent bias in proportional hazards modelling. *Statistics in medicine*, 27(30):6439–6454, 2008.

[7] Liesbeth C de Wreede, Marta Fiocco, Hein Putter, et al. mstate: an r package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*, 38(7):1–30, 2011.

[8] Richard D Gill. Lectures on survival analysis. In *Lectures on Probability Theory*, pages 115–241. Springer, 1994.

[9] Mia Klinten Grand and Hein Putter. Regression models for expected length of stay. *Statistics in medicine*, 35(7):1178–1192, 2016.

[10] James R Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

[11] David SY Ong, Cristian Spitoni, Peter MC Klein Klouwenberg, Frans M Verduyn Lunel, Jos F Frencken, Marcus J Schultz, Tom Van der Poll, Jozef Kesecioglu, Marc JM Bonten, and Olaf L Cremer. Cytomegalovirus reactivation and mortality in patients with acute respiratory distress syndrome. *Intensive care medicine*, 42(3):333–341, 2016.

[12] Hein Putter, Marta Fiocco, and Ronald B Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430, 2007.

[13] Gabi Schulgen and Martin Schumacher. Estimation of prolongation of hospital stay attributable to nosocomial infections: new approaches based on multistate models. *Lifetime Data Analysis*, 2(3):219–240, 1996.