The time-dependent "cure-death" model investigating several endpoints simultaneously in trials treating high-risk patients with severe infections

Dissertation zur Erlangung des Doktorgrades

vorgelegt von

Harriet Sommer

an der Fakultät für Mathematik und Physik

der Albert-Ludwigs-Universität Freiburg im Breisgau

Dezember 2017

Dekan:	Prof. Dr. Gregor Herten Physikalisches Institut, Albert-Ludwigs-Universität Freiburg Hermann-Herder-Straße 3 79104 Freiburg, Deutschland
1. Referent:	Prof. Dr. Martin Schumacher Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Freiburg, Medizinische Fakultät, Albert-Ludwigs-Universität Freiburg Stefan-Meier-Straße 26 79104 Freiburg, Deutschland
2. Referent:	Prof. Dr. Leonhard Held Institut für Epidemiologie, Biostatistik und Prävention, Universität Zürich Hirschengraben 84 8001 Zürich, Schweiz
Datum der Promotion:	13. März 2018

Danksagung

Mein besonderer Dank gilt an dieser Stelle Martin Schumacher für das Ermöglichen dieser Doktorarbeit, die tolle Betreuung währenddessen und die Bereitstellung eines Raumes mit der dazugehörigen Ausstattung. Ich danke auch Martin Wolkewitz, der mich in das EU-Projekt COMBACTE aufnahm und mir bei allen Fragen stets zu Seite stand. Jan Beyersmann und Tobias Bluhmki danke ich für den produktiven Aufenthalt in Ulm und Thomas Gerds für die Zeit in Kopenhagen, Jean-François Timsit sowie Dieter Hauschke für die vielen hilfreichen Diskussionen und Tipps. Außerdem bedanke ich mich herzlich bei den Mitarbeitern des Instituts für Medizinische Biometrie und Statistik für zahlreiche Ratschläge und Hilfestellungen bei Problemen, insbesondere bei Maja von Cube und Nadine Binder. Aber vor allem danke ich meinem Mann, Göran Köber, und meinen Büro-Nachbarn Anne-Sophie Stöhlker und Sam Dörken für ihre Unterstützung und den gemeinsamen Weg in den letzten Jahren.

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

Teile dieser Arbeit wurden bereits in Fachzeitschriften veröffentlicht:

- Teile der veröffentlichten Arbeit von Sommer et al. [1] finden Sie in Kapitel 2, 4.1, 4.3, 4.4, 5 und 6.1
- Einige Teile von Kapitel 4.2 und 6.1 überlappen mit Sommer et al. [2]
- Kapitel 6.2 beruht auf dem Leserbrief Sommer et al. [3]

Ich bin der jeweilige Erstautor dieser Artikel und verantwortlich für Grundkonzept, Analysen und Manuskripte.

Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater/-beraterinnen oder anderer Personen) in Anspruch genommen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Ich habe mich bisher an keiner anderen in- oder ausländischen Hochschule um die Promotion beworben.

Die Bestimmungen der Promotionsordnung der Universität Freiburg für die Fakultät für Mathematik und Physik sind mir bekannt; insbesondere weiß ich, dass ich vor der Aushändigung der Doktorurkunde nicht berechtigt zur Führung des Doktrogrades bin.

Freiburg, den 19. März 2018

Harriet Sommer

Summary

In clinical trials for the development of antibacterial drugs, diverse primary endpoints have been used and treatment effects are usually assessed at the end of follow-up which varies between studies. A highly patient-relevant statement would be an assessment over the entire follow-up period with cure and death as co-primary endpoints. We emphasise to examine the time-dependent multistate endpoint "get cured and stay alive over time", since this might be most relevant from the patients' perspective and can capture different "cure patterns" over the treatment period. Such time-dynamic endpoints provide valuable additional information such that potentially hidden treatment effects can be revealed that might be overlooked when only presenting incidence proportions. Based on a "cure-death" multistate model, simple and sophisticated possibilities are introduced and compared to evaluate a treatment difference in probabilities to be cured and alive over time. As an example, non-inferiority is studied by means of one-sided confidence bands provided by a flexible resampling technique, or, an innovative regression method is used for a risk ratio of being cured and alive. These methods are further evaluated via a simulation study and applied to three topical data examples, a randomised controlled trial for the treatment of patients with hospital-acquired pneumonia, a randomised controlled trial for the prevention of recurrent *Clostridium difficile* infection, and a cohort study to investigate the effect of inadequate treatment for patients with ventilator-associated pneumonia due to the pathogen *Pseudomonas aeruginosa*. Multistate methodology, already entrenched in applied infection control

literature for analysing observational data, is highly beneficial and easily applicable for clinical trials as well to examine patient-relevant endpoints.

Contents

1 INTRODUCTION

Antimicrobial resistance and hospital-acquired infections are growing worldwide problems, and, with few innovative drugs making it to the market there is an urgent need for new drugs to treat such (resistant) infections [4, 5]. Severe bacterial diseases occurring in hospitalised patients include, for example, hospital-acquired pneumonia (HAP), particularly ventilator-associated pneumonia (VAP), which is associated with increased morbidity, mortality, length of stay, and costs [6]. VAP occurs in around $9\% - 27\%$ of patients receiving mechanical ventilation, making it the most common nosocomial infection among ventilated patients [7]. It accounts for about half of all antibiotics given in the intensive-care unit (ICU) [8], and a high attributable mortality is estimated in a recent prominent meta-analysis of individual patient data [9]. Thus, these severe bacterial diseases put an immense burden on health care resources and novel antimicrobial agents to treat them are urgently needed. The use of well-defined endpoints in randomised controlled trials (RCTs) of novel antibiotic agents is vital to properly gauge their effectiveness in treating severe hospital-acquired infections. Ideally, these endpoints should be a direct measure of how patients feel, function, and survive [10, 11]. Unfortunately, however, there is a lack of universal, widely accepted endpoints, particularly for severe hospital-acquired infections [12].

1.1 Endpoints in clinical trials for antimicrobial drugs

Especially in severely ill patients, both cure and mortality endpoints have associated challenges. All-cause mortality is the most robust endpoint; it is the most severe one and can be measured objectively. RCTs traditionally reported ICU-, hospital-, or 28-day-mortality, partly as a regulatory requirement, partly in an attempt to balance the

time needed for a drug to show its effects and the time in which other disease processes could obscure this effect [12]. Also, mortality is often related to the underlying illnesses and severity of disease [13].

In many trials, clinical or microbiological cure is used as efficacy endpoint in the development of antimicrobial treatments [14]. While microbiological cure is defined as eradication of the infection pathogen, a definition for clinical cure is difficult to find. Clinical signs and symptoms may vary depending on the infectious process studied, but also because of concurrent adverse events during the stay in the ICU [15]. Mostly, it is defined as complete resolution of signs and symptoms of the infection or no further antimicrobial treatment being needed [16].

In current and former clinical trials for the treatment of, e.g., HAP or VAP, a variety of primary endpoints have been used [16, 17, 18, 19]. Even the existing guidelines are not consistent in their recommendations. The European Medicines Agency (EMA) proposes clinical cure, the clinical outcome measured at a fixed timepoint called the "test-of-cure" (TOC) visit, as an acceptable primary endpoint [20]. In contrast, the Food and Drug Administration (FDA) [21] suggests all-cause mortality evaluated at a fixed timepoint at any time between day 14 and day 28 as the primary efficacy endpoint.

For *Clostridium difficile* infection, the most common infectious diarrhea, clinical cure is recommended by the EMA [20] as primary endpoint. However, one major problem in treating *Clostridium difficile* infection is the high recurrence rate, which is why many clinical trials for new treatments of *Clostridium difficile* infection aim at preventing a recurrent *Clostridium difficile* infection [22]. Also sustained cure (clinical cure and no recurrent infection) is considered as a useful measure of treatment outcome when comparing agents for which the initial clinical responses are similar [23].

Multiple endpoints

However, often a single endpoint does not adequately capture the entire treatment effect impacting patients in various aspects. The assessment of treatment effect on the basis of multiple endpoints is challenging, both in terms of selecting an appropriate test statistic and interpreting the results. Many strategies have been proposed in the literature to handle multiple endpoints, e.g., [24, 25, 26, 27].

Röhmel et al. [27] discussed an application of two co-primary endpoints when it is sufficient to show that one endpoint is superior and the other one non-inferior compared to a control. Logan et al. [25] generalised the three-step procedure proposed by Röhmel et al. within a closed testing formulation. Bloch et al. [26] developed a non-parametric approach to multiple-endpoint testing based on a bootstrap procedure that can be used to demonstrate non-inferiority of a new treatment for all endpoints and superiority for some endpoints. Ramchandani et al. [24] proposed a test based on a simple scoring system to summarise treatment effects across multiple endpoints.

For serious illnesses, it is strongly recommended that a composite endpoint should include both mortality as well as a clinical endpoint [10, 28, 29]. However, construction of such an endpoint is challenging and interpretation can be misleading especially when the intervention appears to affect individual outcomes differently [30]. Thus, up to now, most trials focus on cure or death as endpoints to be analysed separately, and there are few examples in this field combining mortality with a clinical endpoint. Pocock et al. [31] have suggested the win ratio as a new effect measure where pairs of patients from the innovative and control treatment are grouped into winners and losers based on whether the most / least favourable event was experienced first. The win ratio is then calculated as the total number of winner pairs divided by the total number of loser pairs. Evans et al. [32] recently proposed a similar design using superiority considerations combined with the desirability of outcome ranking and a response adjusted for the

duration of antibiotic risk. A composite score is designed, tailored to compare strategies of antibiotic use by weighting its benefits and potential harms at an individual level. The ranked ordinal clinical outcome is categorised into clinical benefit, clinical benefit with some adverse events, survival without clinical benefit, survival without clinical benefit but adverse events, and death.

1.2 The probability of being cured and alive (PCA)

In most clinical trials for novel antibiotics, measures of efficacy are the proportions of patients in the respective treatment groups achieving microbiological or clinical cure. As a measure of safety the proportions of death cases are taken into account [14], such that treatment effects are assessed at the end of follow-up. We recommend combining these two standard measures of clinical benefit by including both cure and death as co-primary endpoint and analysing them as a time-to-event endpoint. Death cases have to be treated as competing events to cure when cure is measured as a time-to-event endpoint [33]. Also, death following shortly after cure should be considered since cure then does not benefit the patient. Recently, Doshi pointed to a situation where patients were considered cured but died on the same day [34]. We strongly suggest to examine the time-dependent endpoint "get cured and stay alive over time", since this might be most relevant from the patients' perspective and can capture different "cure patterns" over the treatment period. The use of this time-to-event endpoint would incorporate both the cure (and death) and the time at which it occurred and can increase power to detect differences between the test and (active) control group. Furthermore, they provide valuable additional information and potentially hidden treatment effects can be revealed that might be overlooked when only presenting incidence proportions. As also stated by Muscedere et al. [11], the use of a time-to-event endpoint combined with mortality may be the best option, especially for HAP / VAP trials. To understand how

the new treatment influences the whole etiological cure process, we utilise a multistate "cure-death" model with states 0 as initial, 1 as cure and alive, and 2 as absorbing death state [1, 2, 35]. The outcome of interest, to get cured and stay alive over time, is given as the probability of being cured and alive (PCA) function and is estimated by the Aalen-Johansen estimator [36] of the transition probability from state 0 to state 1 (in this case it equals the occupation probability for state 1). The latter generalises the Kaplan-Meier [37] estimator to multistate settings. The proposed model accounts for the time-dependency of cure and death, the presence of competing risks, and potential censoring. In particular, it allows for the fact that patients might die, either before or after cure.

Multistate models are appropriate to take into account the time-dependency of such endpoints by modelling events as transitions between states. Well-established statistical methodology is available to adequately analyse multistate data [38, 39, 40, 41, 35, 42], and multistate methodology has already found its way into applied infection control literature for analysing observational data [43, 44, 45, 46, 47, 48, 49, 50]. So far, multistate endpoints besides endpoints represented as estimands arising from a simple survival or competing risks model were only rarely utilised in an RCT setting. In this work, we show that multistate models are highly beneficial and easily applicable for clinical trials as well to examine patient-relevant endpoints [12, 2, 3, 1].

In terms of studying the efficacy of cancer treatment trials, the aforementioned transition probability was first proposed by Temkin [51] as the probability of being in response function (PBRF). The PBRF was then sometimes used as an outcome measure in the context of bone marrow transplant studies [52], when estimating current leukaemia-free survival [53, 54], or for the estimation of being alive without relapse and immunosuppression for graft-versus-host disease in a population of patients with acute

lymphoblastic leukemia [55].

The purpose of such a function is to synthesise the different summary statistics commonly used, the proportion who respond (here: cure) and the average duration of response. This function rises with the occurrence of a response, falls with each recurrence, and can distinguish between a treatment producing a high response rate but generally short-lived responses and another treatment with a low response rate but longer response durations. It provides a complete summary and an attractive visual display of the given data [56]. Temkin demonstrated that the distribution of the sojourn times of each transition could be estimated by a modified version of the well-known Kaplan-Meier estimator that provide, when combined, an estimate of the PBRF at each time of one possible event.

A further approach to estimate the PBRF was proposed by Pepe et al. [57, 58] based on the difference between Kaplan-Meier estimators.

Cure models in oncology—Disambiguation

Since the term *cure model* is widely used in oncolocy, it is necessary to point out the differences to the *cure-death model* we are dealing with in this thesis. Cure models in oncology, where overall survival or progression-free survival are the major outcomes of interest, were first proposed more than 50 years ago [59, 60, 61]. They are mostly divided into the two classes of mixture and nonmixture models [62]. Mixture models, e.g., model survival as a mixture of patients who are cured and those who are not cured where the probability for cure is examined via logistic regression. A latent cure state is incorporated for the proportion of patients that will never experience recurrence due to cure to account for the fact that recurrence is known to influence survival [63, 64]. Recurrence is then considered as an auxiliary variable to enhance the efficiency of the analysis of overall survival [65].

1.3 Statistical methods for PCA

To show that a novel treatment performs better compared to placebo or an active control, a statistical test is employed that is called superiority test. In a non-inferiority test, the aim is to show that a test treatment is not (much) worse than an active control treatment. In a two-sample parallel design, the problem of testing non-inferiority and superiority can be unified by the following hypotheses

$$H_0 : I^A - I^B \leq \delta \quad \text{versus} \quad H_1 : I^A - I^B > \delta,$$

where $I^A - I^B$ is the difference between the true treatment-specific response proportions of a test drug (I^A) and a control drug (I^B). The idea is that statistically significant differences between the proportions may not be of interest unless the difference is greater than a threshold. Consider $I^A - I^B > 0$ an indication of improvement and $I^A - I^B < 0$ an indication of worsening. Then, if the pre-specified margin $\delta < 0$, the rejection of the null-hypothesis indicates non-inferiority, if $\delta \geq 0$, superiority.

Non-inferiority concepts are more complex both in the design and analysis phase and there are several challenges to address [66]. One of the most difficult issues is the specification of the non-inferiority margin that determines the null hypothesis [67, 68]. A variety of statistical methods can be used [66]. Generally, the margin should be based on estimates of the effect of the active comparator out of previous studies or meta-analyses where guidelines recommend the lower bound of the 95% confidence interval of the treatment effect [69]. However, in the case of anti-infective agents, historical data are not always available such that the margin has to be justified using an anticipated benefit of the experimental drug [66]. The margin also depends on a specific proportion of control failures at a landmark time. However, problems arise if these proportions cannot be maintained at this pre-specified landmark time such that the non-inferiority margin becomes questionable. As a solution, Fay and Follmann [70] propose a variable

margin non-inferiority test that does not require any knowledge about the proportion of control failures.

Another aspect worth mentioning is that patients with a treatment crossover may bias an intention-to-treat analysis, which states that patients are analysed according to the group they were randomised to regardless of the treatment they received, towards a conclusion of non-inferiority [71]. But, a per-protocol analysis, where the analysis set comprises only patients who fully comply with their assigned treatment, may be biased as well when baseline characteristics are not balanced anymore. Mauri et al. [66] recommend to analyse both data sets. Without a bias, this should lead to similar results, but, nevertheless, careful consideration may be needed before drawing final conclusions. Moreover, methods were already developed to deal with treatment switching [72].

Lastly, caution is advised with composite endpoints, especially if the respective components are discordant regarding benefits and risks [73].

Why non-inferiority trials are indispensable in antibacterial drug development

Superiority trials are the preferred design for drug development [74]. However, despite aforementioned challenges, almost all antibacterial drugs that we rely on today were based on non-inferiority trial designs [70, 75]. Actually, non-inferiority analyses are indispensable in this medical field [74, 70, 76, 77]. The reasons are many and diverse.

Nowadays, the benefit of newly developed treatments in terms of efficacy is often only marginal over existing treatments. They might be rather advantageous in reducing costs via having fewer adverse effects, a better compliance, or a broader spectrum of activity. Then, it is often not expected to be able to demonstrate superiority when an active therapy exists that is effective and accepted by ethics committees and patients. Furthermore, placebo trials are considered unethical for more serious infections with

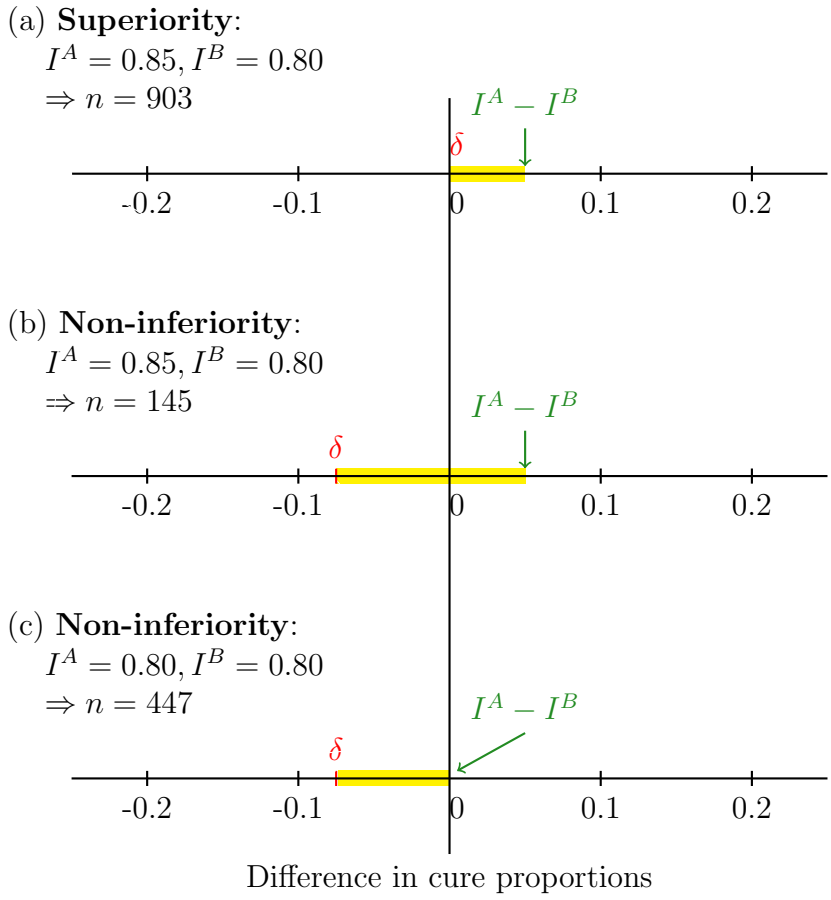

Figure 1.1: Illustration of exemplary superiority and non-inferiority concepts for different margins δ (in a, $\delta = 0$, in b and c, $\delta = -0.075$) and different assumptions for cure proportions I^A and I^B. These values are further used for sample size calculations.

significant morbidity and mortality [66]. Moreover, in practice, at the time a treatment is needed immediately, neither the type of pathogen nor the definite diagnosis of a bacterial infection is known. Unfortunately, technologies to identify pathogens are far from perfect [78]. As a consequence, only having an alternative treatment, a good substitute, can be, under some circumstances, substantially beneficial for a patient

[76]. Resistance to current drugs may develop fast and a long wait before starting the development of a suitable drug could lead to large time gaps.

The main advantage of a non-inferiority trial is that a considerably smaller number of patients is needed to achieve a pre-specified power in comparison to a superiority trial for the same assumed true proportions. Only in this way it is possible to make feasible studies of novel antibiotics, especially for the treatment of rare pathogens [74]. Let us examine this aspect in more detail. For each assumed true difference $I^A - I^B$ one can establish a sample size according to, e.g., [79, 80]. In a placebo-controlled superiority trial, a clinically important difference is set up for the new treatment that results, if large, in a relatively small required sample size. However, as mentioned above, if a standard treatment exists, an active-controlled trial should be preferred for ethical reasons [74]. Then, the concept of a clinically important difference is inappropriate since any amount of improvement is clinically important.

For this, consider the following example: Assuming a cure rate of 85% for the test drug and 80% for the control drug. Using a non-inferiority margin of $\delta = 7.5\%$, as can be seen in Figure 1.1 b, a sample size of 290 patients, 145 in each arm, is sufficient to prove non-inferiority for the endpoint cure using an asymptotic normally distributed test statistic of proportions between two groups with $1 - \beta = 80\%$ power and a level of significance of $\alpha = 2.5\%$. Proving superiority with $\delta = 0$, as can be seen in Figure 1.1 a, would require 1806 patients, 903 in each arm. Assuming equal cure rates in a non-inferiority trial, as can be seen in Figure 1.1 c, would enlarge the sample size, resulting, in this case, in 894 patients, 447 in each arm.

A white paper by the Infectious Diseases Society of America [29] provides examples of ways superiority trials could be implemented if resistant pathogens were sufficiently frequent. They mention, e.g., a hierarchical nested design proposed by Huque et al. [81]. In this design, the primary endpoint is first tested for non-inferiority in the subgroup

of patients who have infections caused by pathogens susceptible to the control drug, following by a superiority test for patients with infections caused by resistant pathogens once non-inferiority is confirmed on the same endpoint. However, so far no RCT has implemented a hierarchical nested design to determine treatment efficacy for infections caused by multi-drug resistant organisms [12].

Assessing the difference of treatment-specific PCAs

Returning to the endpoint we focus on, being cured and alive over time, a much more convincing statement than merely demonstrating non-inferiority or superiority at a single point in time by, e.g., comparing proportions, is to demonstrate non-inferiority or superiority over the entire follow-up period. From the patients' perspective it is highly relevant how the active treatment performs over the complete cure process, and not only at the end of follow-up [33, 11]. The question that arises is how to compare two probability curves of being cured and alive over time and how to simultaneously perform non-inferiority or superiority analyses.

General tests for comparing probability curves

Generally, when considering the comparison of two probability curves, a statistical hypothesis test that the two curves are the same has to be applied. One choice for such a test would be to compare estimates of probability curves at a fixed time using the estimated standard errors to construct a test statistic. This has the obvious disadvantage of an arbitrary selection of a timepoint and, furthermore, is not a particularly powerful test [82]. For the classic survival situation when no competing event is present and no transition from cure to death is possible, generalised linear rank tests, such as the well-known log-rank test [83], have become standard tools for comparing survival or cumulative hazard functions using arguments based on sets of 2×2 contingency tables.

The log-rank test can also be seen as a special case of the general regression method proposed by Cox [84]. A similar class of test statistics for making comparisons between cumulative incidence functions was given by Gray [85] based on the Fine and Gray model [86]. The first test to compare functions as the PBRF was proposed by Pepe [57]. Hsieh et al. [87] proposed an interesting test technique based on log-rank tests for inferences on treatment effects over a whole time frame and not only at one timepoint, suitable for the cure-death model. A restricted version of the test technique is sensitive to a prolonged time to death and duration of cure and a shorter time to cure of the new treatment.

Pseudo-value regression

Pseudo-value regression was proposed by Andersen et al. [88] and Andersen and Klein [89] and provides a simple and generalisable method of modelling complex time-to-event data. It enables a direct regression model for general transition probabilities in a multistate setting avoiding the Markov assumption. The idea is to obtain pseudo-values from a jackknife statistic constructed from a consistent estimator of the probability of interest that are further utilised as outcome variables in a generalised linear model. A direct interpretation of covariate effects is possible, such that, with a treatment indicator as covariate, a test of treatment difference can be performed on the probability of being cured and alive. It results in a relative effect measure comparing two probabilities to be cured and alive whose confidence interval can be used for non-inferiority and superiority statements. Furthermore, in observational studies, when randomisation cannot be employed, it is necessary to adjust for potential confounders, such as, e.g., the propensity score.

Time-simultaneous confidence bands

RCTs often focus on the absolute difference as effect measure, e.g., with proportions of cured patients or mortality cases [90]. To assess, e.g., time-simultaneous non-inferiority, a so-called "confidence band" is then required, in which the absolute difference in probabilities of being cured and alive for active treatment A minus control B over a relevant time interval and not just at a single timepoint lies with a probability of, e.g., 95%. Such a confidence band generalises the concept of a confidence interval to an entire time interval of interest. We will construct these bands adopting a resampling procedure known as "wild bootstrap", which has been established for competing risks settings [91, 92]. This technique can also be applied to other research fields as, e.g., machine learning [93, 94]. Also, Liu et al. [53] adapt this approach to make inference for current leukaemia-free survival curves in a more complex multistate model.

1.4 Scope of thesis

This dissertation is organised as follows: To begin, Section 2 outlines the mathematical background. We will introduce basic statistical techniques in survival analysis, present multistate models, and specify the cure-death model in a multistate framework. The non-parametric estimators for estimands of interest are illustrated in Section 3. Based on the cure-death model, we will focus on several possibilities for a treatment comparison in Section 4. Besides the simplest method, to compare risk differences with proportions of patients cured and alive, time-simultaneous one-sided confidence bands, pseudo-value regression, and a restricted log-rank-based test of treatment effect are presented. These methods are further evaluated via a simulation study in Section 5 to examine how they handle simple and complex treatment imbalances. In Section 6 we will consider three data examples. Our major data example will be the recently

published ceftobiprole trial [95], where the new regimen ceftobiprole is compared to the two-drug regimen ceftazidime / linezolid for the treatment of patients with HAP and VAP. The second example is based on our letter [3] we addressed to the recent article by Wilcox et al. [96]. It is about hospitalised patients with *Clostridium difficile* infection, examining the safety and efficacy of actoxumab and bezlotoxumab. As a third example, we will use a subset of the OUTCOMEREA research data base, a French multicenter study from the OUTCOMEREA research group, that includes data collected in 23 ICUs. As this is an observational study, many other aspects as, e.g., analysis methods using the propensity score, are examined. Here, we will compare adequate and inadequate treatment for patients with VAP due to the pathogen *Pseudomonas aeruginosa*. A discussion and conclusion in Section 7 finally summarises all findings of this thesis and gives some leads for further research.

2 MODEL SPECIFICATION

2.1 Mathematical background

Survival analysis, as one of the oldest fields in statistics, has a much broader meaning today than merely analysing survival in the sense of death rates or mortality [97]. It is about observing a group of individuals from some entry timepoint until a certain event happens—an event of any kind as, e.g., blindness, graduation, employment, etc. It can encompass the study of duration between any two events. These events can be good or bad, such as recovery or relapse, marriage or divorce, which is worth mentioning since the jargon of survival analysis suggests the events to be unpleasant. The basic example is still a transition from an initial state, as alive, to an absorbing state, as death, in Figure 2.1. Absorbing means that once entered this state, an individual cannot move out anymore.

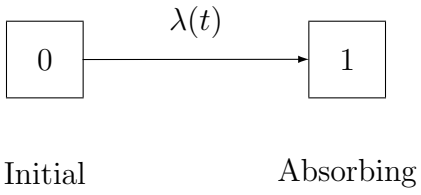

Figure 2.1: Standard survival model with hazard function $\lambda(t)$.

Hazards, survival function, and Cox regression

Let T denote a non-negative random variable which models the failure or survival time of individuals of a population, with distribution function F and density function f. One of the important objects of interest is the survival function

$$S(t) := Pr(T > t) = 1 - Pr(T \leq t) = 1 - F(t) = \int_t^\infty f(u)du, \tag{1}$$

the probability for the occurrence of a special event at time T after a certain time t of an arbitrarily selected individual out of the population and its distribution, which has to be estimated and compared between different groups. S is generally estimated by the Kaplan-Meier estimator [37], a special case of the Aalen-Johansen estimator [98] introduced in Section 3.1. Another important issue is to model the hazard function which is the threshold value of the probability that an individual will experience an event within a small time interval given that it has survived up to the beginning of that interval. It can also be interpreted as the risk of dying at time t or the momentary force of mortality. For a continuous survival time T the hazard function λ is defined as

$$\lambda(t) := \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2}$$

A short and more intuitive notation of (2) is

$$\lambda(t) \cdot dt := Pr(T \in dt \mid T \geq t), \tag{3}$$

where we can write dt for the length of the infinitesimal small time interval $[t, t + \Delta t)$ and the interval itself [41]. While S is monotonically decreasing with $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$, λ can be any non-negative function. The cumulative hazard Λ is defined as

$$\Lambda(t) := \int_0^t \lambda(u)du \tag{4}$$

and is estimated by the Nelson-Aalen estimator [99, 100], introduced in Section 3.1. The survival function in (1) can be expressed in terms of the hazard function as

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right) = \exp\left(-\Lambda(t)\right) \tag{5}$$

since

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{Pr(T \geq t)} \\
&= \lim_{\Delta t \to 0} \frac{Pr(T < t + \Delta t) - Pr(T < t)}{\Delta t} \frac{1}{Pr(T \geq t)} \\
&= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{Pr(T > t)} \\
&= \frac{F'(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt} \log S(t).
\end{aligned}$$

The most interesting survival model examines the relationship between survival and one or more predictors, usually termed covariates or explanatory variables. One famous regression method introduced is the Cox proportional hazards regression [84]. The hazard rate for a subject i with covariate vector $Z_i = (Z_{i1}, \ldots, Z_{ip})$ is described as

$$\lambda(t \mid Z_i) = \lambda_0(t) \exp(\beta' Z_i), \tag{6}$$

with non-negative baseline hazard function $\lambda_0(t)$ and linear predictor $\beta' Z_i = \sum_r \beta_r Z_{ir}$, $r \in \{1, \ldots, p\}$. Hazard rates for all subjects are assumed to be proportional. The hazard ratio (HR) $\exp(\beta_r)$ is associated with an increase of one unit for the rth covariate Z_r.

To compare the survival distributions of two samples, the log-rank test can be used. This standard non-parametric statistical hypothesis test was first proposed by Nathan Mantel [83] and was named the log-rank test by Richard and Julian Peto [101]. The log-rank test is equivalent to the partial likelihood score test for the Cox proportional hazards regression model [102].

Cause-specific hazard, cumulative incidence function, and Fine and Gray regression

If a time-to-event endpoint is analysed that is not all-encompassing, competing risks have to be considered to include multiple endpoints. The competing risks scenario, as can be seen in Figure 2.2, includes a set $\mathcal{S} = \{1, 2, \ldots, J\}$ of states that can be reached out of state 0. The cause-specific hazard rate

$$\lambda_{0j}(t) := \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t;\ \text{cause}\ j \mid T > t)}{\Delta t}, \quad j \in \mathcal{S}$$

models the transition intensity or instantaneous risk per time unit of going from state 0 to state j. It is the probability that an individual in the initial state just prior to time t will pass to state 1 or 2, respectively, in the small time interval $[t, t + \Delta t)$. The probability

$$P_{00}(0, t) = S(t) = \exp\left(- \int_0^t \sum_{j=1}^{J} \lambda_{0j}(u) du \right) \tag{7}$$

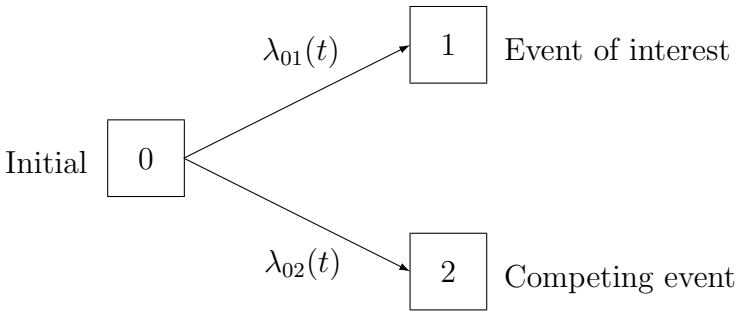

Figure 2.2: Competing risks model with cause-specific hazards $\lambda_{01}(t)$ and $\lambda_{02}(t)$. This restriction to two competing events is only for ease of presentation. For studying the cumulative probability of the event of interest (state 1), all other competing events are combined into state 2.

of staying in state 0 includes all event-specific hazards. It can be seen that (7) is of the same form than (5), the survival probability in the simple survival setting. The cumulative incidence function (CIF)

$$P_{0j}(0,t) := \int_0^t P_{00}(0,u)\lambda_{0j}(u)du, \tag{8}$$

the probability of a transition from state 0 to state j, in the presence of all the competing risks, therefore depends on every cause-specific hazard. It is estimated by the Aalen-Johansen estimator [98], introduced in Section 3.1.

In a simple survival setting, the probability of an event can be estimated as 1 minus the Kaplan-Meier estimator, so there is a simple relationship between the hazard rate and the survival function as can be seen in (5). In a competing risks setting, this relation does not hold. Here, Fine and Gray [86] suggested a Cox-type model for the so-called subdistribution hazard. While in a competing risks setting the cause-specific hazard function is the instantaneous rate of occurence of the given type of event in subjects who are currently event-free, the subdistribution hazard function considers subjects who have not yet experienced an event of *that* type (such that both subjects who are event-free as well as subjects who experienced a competing event are considered) [103]. One can show that 1 minus the Kaplan-Meier estimator where as hazard the subdistribution hazard is inserted equals the Aalen-Johansen estimator of the CIF, so the subdistribution hazard gives a one-to-one correspondence between hazard and CIF [104]. Hence, the score test from fitting the Fine and Gray model can be used to compare the CIFs of two samples directly [105]. This test is due to Gray and therefore often called Gray's test [85]. Since the subdistribution hazard ratio (SHR) describes the relative effect of covariates on the subdistribution hazard function, covariates in the Fine and Gray model can also be interpreted as having an effect on the CIF [103].

2.2 Multistate models

The quantities introduced in previous chapters can also be expressed in terms of multistate models [36, 41]. In general, multistate models make an interpretation of more complex endpoints of interest more accessible than a hazard-based analysis alone. The multistate framework models events as transitions between states, thus, the simplest multistate model is a transition from an initial state to an absorbing state as in Figure 2.1. In general, let us consider a Markov process $(X(t), t \in [0, \infty))$ with a finite state space $\mathcal{S} = \{1, 2, \ldots, J\}$ and a $(J+1) \times (J+1)$ transition matrix

$$\mathbf{P}(s,t) := (P_{lj}(s,t))_{l,j \in \mathcal{S}}$$

with entries

$$P_{lj}(s,t) := Pr(X(t) = j \mid X(s) = l), \quad s \leq t \quad \text{and} \quad l, j \in \mathcal{S}.$$

The survival function is therefore $S(t) = P_{00}(X(t) = 0)$, where $Pr(X(0) = 0) = 1$. The matrix of cumulative hazards

$$\mathbf{\Lambda}(t) := (\Lambda_{lj}(t))_{l,j \in \mathcal{S}}$$

has entries

$$\Lambda_{lj}(s,t) := \int_0^t \lambda_{lj}(u)du, \quad l, j \in \mathcal{S}$$

with transition hazards

$$\lambda_{lj}(t) \cdot dt := Pr(X(t + \Delta t) - = j \mid X(t-) = l), \quad l, j \in \mathcal{S} \quad \text{and} \quad l \neq j.$$

We typically assume $(X(t), t \in [0, \infty))$ to be Markov, a key assumption for estimation techniques in Section 3, which means that

$$Pr(X(t) = j \mid X(s) = l, \text{Past}) = Pr(X(t) = j \mid X(s) = l), \quad s \leq t \quad \text{and} \quad l, j \in \mathcal{S}, \tag{9}$$

that is, the transition probabilities only depend on the past via the current time s and the currently occupied state [41]. "Past" is written for knowledge about the process' history up to time s consisting of the observation of the process in the interval $[0, s]$, also known als *σ-algebra*. As time increases, this makes up an increasing sequence of σ-algebras, a so-called *filtration*. The process is time-inhomogeneous, that means that the transition hazards do depend on time interval $[s, t]$. A homogeneous process assumes that these probabilities are identical regardless the length $t - s$ of the interval. It follows that a homogeneous Markov model is a parametric model with constant transition hazards, whereas in an inhomogeneous Markov model the transition hazard can be any integrable non-negative function [41].

Only for simple Markov processes it is possible to give explicit expressions for the transition probabilities in terms of transition hazards. The product integral

$$\begin{aligned}
\mathbf{P}(s,t) &= \prod_{u\in(s,t]} (\mathbf{I} + d\mathbf{\Lambda}(u)) && (10)\\
&= \lim_{\Delta\mathbf{\Lambda}(t_k)\to 0} \prod_{k=1}^{K} (\mathbf{I} + \Delta\mathbf{\Lambda}(t_k)) && (11)\\
&= \lim_{\Delta\mathbf{\Lambda}(t_k)\to 0} [(\mathbf{I} + \mathbf{\Lambda}(t_K) - \mathbf{\Lambda}(t_{K-1})) \cdots (\mathbf{I} + \mathbf{\Lambda}(t_1) - \mathbf{\Lambda}(t_0))]
\end{aligned}$$

has to be used to express the transition probability matrix $\mathbf{P}(s, t)$ in terms of the matrix of transition intensities $\mathbf{\Lambda}(t)$ for ever finer partitions $s = t_0 < t_1 < \ldots < t_{K-1} < t_K = t$ [98, 106]. The product integral $\prod$ is defined as the limit of approximating finite products

$$\prod_a^b (1 + f(x)dx) = \lim_{\Delta x\to 0} \prod (1 + f(x_i)\Delta x)$$

in a similar manner as the integral $\int$ is defined as limit of approximating sums

$$\int_a^b f(x)dx = \lim_{\Delta x\to 0} \sum f(x_i)\Delta x.$$

2.3 The cure-death model

In order to suitably account for the time-dynamic pattern of cure and death after randomisation, we focus on the "illness-death model without recovery" embedded in the flexible and powerful multistate model framework [104, 42]. Since the context here is hospital acquired infection, we call this model "cure-death model" [1], see Figure 2.3. This is a Markov process $(X(t), t \in [0, \infty))$ with state space $\mathcal{S} = \{0, 1, 2\}$, with 0 as initial, 1 as cure, and 2 as absorbing death state. According to the study protocol, all patients are in the initial state 0 at the start of follow-up, that is randomisation to treatment, immediately after infection, such that $Pr(X(0) = 0) = 1$. The timescale of interest is "time since randomisation" in days. The outcome most relevant for patients is "being cured and alive" (state 1), however, all patients, whether cured or not, are permanently at risk for death (state 2) during the entire follow-up.

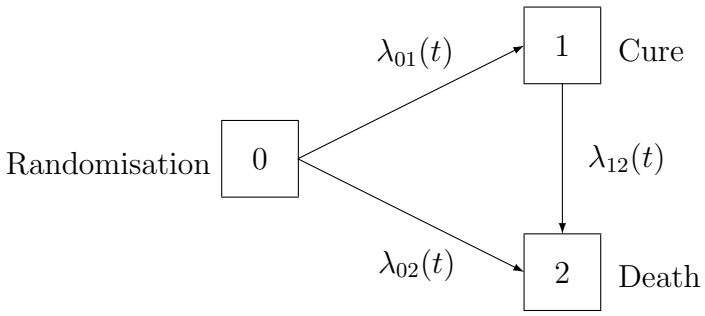

Figure 2.3: The cure-death model for comparing two antimicrobial therapies with an initial infection / randomisation state, a cure state, and a death state. The direction of arrows illustrates the potential transition between the states determined by a transition hazard $\lambda_{01}(t)$, $\lambda_{02}(t)$, or $\lambda_{12}(t)$.

The arrows in Figure 2.3 illustrate the possible transitions within our multistate model determined by transition hazards. They can be interpreted as momentary forces that pull a subject out of one state to another. Here, the matrix of transition probabilities is defined as

$$\mathbf{P}(s,t) := (P_{lj}(s,t))_{l,j}, \quad l,j \in \mathcal{S}$$

with transition probabilities

$$P_{lj}(s,t) := Pr(X(t) = j | X(s) = l), \quad s \leq t,$$

where entry (0,0) is

$$\begin{aligned} P_{00}(0,t) &= Pr(X(t) = 0 \mid X(0) = 0) \\ &= \exp\left(- \int_0^t \lambda_{01}(u) + \lambda_{02}(u) du \right), \end{aligned}$$

entry (0,1) is

$$\begin{aligned} P_{01}(0,t) &= Pr(X(t) = 1 \mid X(0) = 0) \\ &= \int_0^t P_{00}(0,u) \lambda_{01}(u) P_{11}(u,t) du, \end{aligned} \tag{12}$$

entry (0,2) is

$$\begin{aligned} P_{02}(0,t) &= Pr(X(t) = 2 \mid X(0) = 0) \\ &= 1 - (P_{00}(0,t) + P_{01}(0,t)), \end{aligned}$$

entry (1,1) is

$$\begin{aligned} P_{11}(s,t) &= Pr(X(t) = 1 \mid X(s) = 1) \\ &= \exp\left(- \int_s^t \lambda_{12}(v) dv \right), \end{aligned}$$

and entry (1,2) is

$$\begin{aligned} P_{12}(s,t) &= Pr(X(t) = 2 \mid X(s) = 1) \\ &= 1 - P_{11}(s,t). \end{aligned}$$

Entries (1,0), that is $P_{10}(0,t) = Pr(X(t) = 0 \mid X(0) = 1)$, (2,0), that is $P_{20}(0,t) = Pr(X(t) = 0 \mid X(0) = 2)$, and (2,1), that is $P_{21}(0,t) = Pr(X(t) = 1 \mid X(0) = 2)$, are zero and (2,2), that is $P_{22}(0,t) = Pr(X(t) = 2 \mid X(0) = 2)$ is one. Our interest focuses on the probability to be cured and alive in equation (12), the expected proportion of patients cured and alive, that is $P_{01}(0,t)$. The first two terms in the integrand are the same for the cumulative incidence function. The difference is now that state 1 is not absorbing as in the competing risks case, but an intermediate state. Therefore, we model a $1 \rightarrow 2$ transition as well by including $P_{11}(u,t)$ to ensure that individuals stay in state 1 until time t after a 01 transition at time u. Note that $P_{01}(s,t)$ is the same as the state occupation probability $Pr(X(t) = 1)$ since the initial distribution is degenerate in state 0, that is $Pr(X(0) = 0) = 1$.

The Markov assumption

As before, we assume that our multistate model is time-inhomogeneous Markov, see equation (9), which means that the future course of a cured patient through the multistate model in Figure 2.3 depends on the time since randomisation and the fact that the patient is cured, but not on the time the patient has already been cured. Formally, it is

$$P_{12}(s,t) = Pr(X(t) = 2 \mid X(s) = 1, \text{Past}) = Pr(X(t) = 2 \mid X(s) = 1),$$

that means the past and the future of the process are independent given the present time s [41]. Further, $\lambda_{12}(t)$ depends on the transition type and on the current time t

but not on the entry time into the cure state $\tilde{t}$. In a non-Markov model, this transition hazard would be $\lambda_{12}(\tilde{t}, t)$ and may be different for individual times $\tilde{t}$. Such a restriction is not necessary in the competing risks situation in Figure 2.2 since there is, except the initial state, no transient state.

To check if the Markov assumption is fulfilled, we may study the influence of the time to cure, the $0 \to 1$ transition, on the $1 \to 2$ mortality transition including it as a time-dependent variable in a Cox model for the 12 hazard [41]. Alternatively, in this model, another estimator for the PCA function not relying on the Markov assumption can be used [57, 107]. It is based on a decomposition of the probability of interest, $P_{01}(0, t)$, into components that can be estimated by Kaplan-Meier-type estimators, respectively. More information about this alternative can be found in Section 3.2. The idea is to compare the estimate derived by the Aalen-Johansen estimator [36], that relies on the Markov assumption, to the estimate constructed without the necessity of the Markov assumption. A non-existent discrepancy may indicate that the Markov assumption is fulfilled or not indispensable.

A hazard-based analysis applying, for instance, the well-known Cox proportional hazards model, does not allow direct statements regarding the probability of interest, because the latter is a complex functional of all involved hazards. Instead, the Aalen-Johansen estimator [36], introduced in Section 3.1 and 3.2, is employed for non-parametric estimation of this transition probability, which generalises the well-known Kaplan-Meier estimator to general multistate settings. An alternative method for regression analyses, pseudo-value regression, is introduced in Section 4.3.

2.4 Motivation for estimation and simulation techniques

Each transition hazard λ_{lj} in a multistate model can be interpreted as an instantaneous risk per time unit to go from state l to state j. By assuming the hazard rates to be constant over time, they can be estimated as

$$\hat{\lambda}_{lj} = \frac{\#\text{ patients with an } lj \text{ transition}}{\#\text{ patient-days at risk in state } l}.$$

In a simple competing risks setting out of Figure 2.2 with state space $\mathcal{S} = \{0, 1, 2\}$, e.g., the estimated overall risk of going from state 0 to state 1 at the end of follow-up τ is related to the cause-specific rate as follows

$$\hat{P}_{01}(0, \tau) = \frac{\hat{\lambda}_{01}}{\hat{\lambda}_{01} + \hat{\lambda}_{02}}$$

and for the 02 transition

$$\hat{P}_{02}(0, \tau) = \frac{\hat{\lambda}_{02}}{\hat{\lambda}_{01} + \hat{\lambda}_{02}}.$$

This is due to the fact that the CIF out of equation (8) can be expressed as

$$\begin{aligned}
P_{01}(0,t) &= \int_0^t P_{00}(0,u)\lambda_{01}(u)du \\
&= \int_0^t \exp\left(-\int_0^u (\lambda_{01}(v) + \lambda_{02}(v))dv\right)\lambda_{01}(u)du \\
&= \lambda_{01}\int_0^t \exp\left(-(\lambda_{01}+\lambda_{02})u\right)du \\
&= \lambda_{01}\left[-\frac{1}{\lambda_{01}+\lambda_{02}}\exp\left(-(\lambda_{01}+\lambda_{02})u\right)\right]_0^t \\
&= \frac{\lambda_{01}}{\lambda_{01}+\lambda_{02}}(1-\exp(-(\lambda_{01}+\lambda_{02})t))
\end{aligned}$$

when assuming constant hazards, and

$$P_{02}(0,t) = \frac{\lambda_{02}}{\lambda_{01}+\lambda_{02}}(1-\exp(-(\lambda_{01}+\lambda_{02})t)).$$

Cure-death model

In a cure-death model out of Figure 2.3, the probabilities for constant hazards are of the following form:

$$\begin{aligned}
P_{01}(0,t) &= \int_0^t P_{00}(0,u)\lambda_{01}(u)P_{11}(u,t)du \\
&= \int_0^t \exp\left(-\int_0^u (\lambda_{01}(v)+\lambda_{02}(v))dv\right)\lambda_{01}(u)\exp\left(-\int_u^t \lambda_{12}(v)dv\right)du \\
&= \lambda_{01}\int_0^t \exp\left(-(\lambda_{01}+\lambda_{02})u-\lambda_{12}(t-u)\right)du \\
&= \lambda_{01}\left[-\frac{1}{-(\lambda_{01}+\lambda_{02}-\lambda_{12})}\exp\left(-(\lambda_{01}+\lambda_{02}-\lambda_{12})u-\lambda_{12}t\right)\right]_0^t \\
&= \frac{\lambda_{01}}{\lambda_{01}+\lambda_{02}-\lambda_{12}}(-\exp(-(\lambda_{01}+\lambda_{02})t)+\exp(-\lambda_{12}t)),
\end{aligned}$$

$$\begin{aligned}
&P_{02}(0,t) = 1-(P_{00}(0,t)+P_{01}(0,t)) \\
&= 1-\left(\exp(-(\lambda_{01}+\lambda_{02})t)+\frac{\lambda_{01}}{\lambda_{01}+\lambda_{02}-\lambda_{12}}(-\exp(-(\lambda_{01}+\lambda_{02})t)+\exp(-\lambda_{12}t))\right),
\end{aligned}$$

and

$$\begin{aligned}
P_{12}(u,t) &= 1-P_{11}(s,t) \\
&= 1-\exp(-\lambda_{12}(t-u)).
\end{aligned}$$

Extended cure-death models

The cure-death model can easily be extended and aforementioned considerations about the shape of transition probabilities in the constant hazard case are applicable to fit to a specific data situation or a particular clinical question [39], as Figure 6.1 in Section 6.1, Figure 6.6 in Section 6.2, or Figure 6.9 in Section 6.3.

Since multistate models are realised as a nested sequence of competing risks experiments that are regulated by the transition hazards [41], such basic estimation techniques also work in an extended version in more complex multistate situations. How estimation is done in detail is discussed in the following section. Furthermore, with the help of the connection between risk and rate, it is possible to reconstruct transition hazards and simulate respective data even if only summary data are given in a publication. This is done in Section 6.2.2.

3 NON-PARAMETRIC ESTIMATION

In this chapter, we will introduce the well-known Nelson-Aalen and Aalen-Johansen estimator, as often applied non-parametric estimators for estimands introduced the chapter before.

The data on which survival and multistate models are fit are often censored, a problem that does not generally arise with other types of data. It means that the event of interest may not necessarily happen in the time window of observation for all individuals and thus, these times are called right-censored. A typical example is an RCT in which time zero corresponds to treatment randomisation. Patients experiencing no event before the administrative closing of the trial will be right-censored.

So, let C denote the end of follow-up, such that instead of event time T we observe only the censored event time $T = \min(T, C)$. Because T and C are independent, censoring does not disturb the hazard and

$$\lambda(t) \cdot dt = Pr(T \in dt \mid T \geq t) = Pr(T \in dt, T \leq C \mid \min(T, C) \geq t). \tag{13}$$

Consequently, the cumulative cause-specific hazard can be estimated also from censored data that leads us to the Nelson-Aalen estimator. We will show in the following, Section 3.1, that the Kaplan-Meier estimator for the survival function and the Aalen-Johansen estimator for an arbitrary transition probability can be expressed in terms of the Nelson-Aalen estimator. We will introduce these estimators using counting process formulation.

Another important issue is that there are situations where individuals enter the study at a time later than time origin 0. Such delayed entry times are said to be left-truncated. Neither of them will disturb the concept of hazards.

3.1 Nelson-Aalen and Aalen-Johansen estimator

A non-parametric estimator of the cumulative hazard in (4) is given by the Nelson-Aalen estimator proposed by Nelson [99, 100], by Altshuler [108], and Aalen [109].

We assume that there are n individuals under study, $i \in \{1, \ldots, n\}$, with data arising from n independent replicates of a multistate process with state space $\mathcal{S}$ as in Section 2.2, respectively subject to a right-censoring time C_i. Let $Y_{l,i}(t) = 1$ if patient i is in state l before time t, so the at-risk process

$$Y_l(t) := \sum_{i=1}^{n} Y_{l,i}(t), \quad l \in \mathcal{S}$$

counts the number of patients at risk just prior to time t. Let $N_{lj,i}(t) = 1$ if individual i moves directly from state l to state j until time t, so the counting process

$$N_{lj}(t) := \sum_{i=1}^{n} N_{lj,i}(t), \quad l, j \in \mathcal{S} \quad \text{and} \quad l \neq j$$

counts the number of lj transitions in time interval $[0, t]$. We will write

$$dN_{lj}(t) := N_{lj}(t) - N_{lj}(t-)$$

for the increments of $N_{lj}(t)$, the number of lj transitions observed exactly at time t. We have seen in (2.2) that $\lambda_{lj}(t)dt$ is an infinitesimal conditional transition probability. If no transition is observed at time t, $dN_{lj}(t)$ equals zero and, consequently, the estimate of $\lambda_{lj}(t)dt$ as well. If we do observe lj transitions at time t, $dN_{lj}(t)$ is greater than zero and we can estimate $\lambda_{lj}(t)dt$ as the number of transition divided by the number at risk just prior to t, that is $Y_l(t)$. The Nelson-Aalen estimate $\hat{\mathbf{\Lambda}}(t)$ of the cumulative hazard matrix $\mathbf{\Lambda}(t) = (\Lambda_{lj})_{l,j \in S}$ has entries that sum up these increments

$$\hat{\Lambda}_{lj}(t) := \sum_{s \leq t} \frac{dN_{lj}(s)}{Y_l(s)}.$$

An estimator of the variance of $\hat{\Lambda}_{lj}(t)$ is given by

$$\hat{V}_{lj}(t) := \sum_{s \leq t} \frac{dN_{lj}(s)}{Y_l^2(s)}$$

that can be used to construct pointwise confidence intervals.

The estimation of transition probabilities is more complicated since, in general, they are a complex function of transition hazards and the state occupied at time t may result from a nested sequence of competing risks experiments [41]. In (2.2) we have seen that $P(s,t)$ can be approximated based on partitions of the interval $[s,t]$. Partitioning further approaches a limit as in (11) that can be expressed by product integration. Consequently, the Aalen-Johansen estimator, independently established by Aalen and Johansen [98] and Fleming [110, 111], is a finite matrix product

$$\hat{\mathbf{P}}(s,t) := \prod_{u \in (s,t]} (\mathbf{I} + d\hat{\mathbf{\Lambda}}(u)), \tag{14}$$

as in (10) where the Nelson-Aalen estimate is plugged in. Thus, the Kaplan-Meier estimator of the survival function is a special version of the Aalen-Johansen estimator

$$\hat{S}(t) = \hat{P}_{01}(0,t) = \prod_{u \in (0,t]} (1 + d\hat{\Lambda}_{01}(u)), \tag{15}$$

when there are only two states as in Figure 2.1.

For estimation of the covariance for the Aalen-Johansen estimator, a Greenwood-type variance estimator is available, but, unfortunately, with a complicated structure (Equation 4.4.17 in [112]). For facilitation, Andersen et al. develop a recursion formula for such an estimator as well (Equation 4.4.19 in [112]). Generally, the Greenwood-type variance estimator is preferred in the setting having competing risks data with left-truncation [113]. This can again be used to construct pointwise confidence intervals.

3.2 Non-parametric estimation of the PCA function

An estimator of the PCA function

$$P_{01}(0,t) = \int_0^t P_{00}(0,u)\lambda_{01}(u)P_{11}(u,t)du,$$

in model out of Figure 2.3 as in equation (12), is given by the Aalen-Johansen estimator that was introduced in Section 3.1. $P_{00}(0,u)$ and $P_{11}(u,t)$ are estimated by Kaplan-Meier-type estimators $\hat{P}_{00}(0,u)$ and $\hat{P}_{11}(u,t)$, $\lambda_{01}(u)du$ by the increment of the Nelson-Aalen estimator $d\hat{\Lambda}_{01}(u)$ for the cumulative cure hazard $\Lambda_{01}(u) = \int_0^t \lambda_{01}(u)du$, such that

$$\hat{P}_{01}(0,t) = \sum_{0<u\leq t} \hat{P}_{00}(0,u-)\frac{dN_{01}(u)}{Y_0(u)}\hat{P}_{11}(u,t),$$

with

$$\hat{P}_{00}(0,u) = \prod_{s\leq u}\left(1 - \frac{d(N_{01}(s)+N_{02}(s))}{Y_0(s)}\right)$$

and

$$\hat{P}_{11}(u,t) = \prod_{s<u\leq t}\left(1 - \frac{dN_{12}(u)}{Y_1(u)}\right).$$

Note that in the absence of right-censoring, the Aalen-Johansen estimator is identical to the relative proportion of patients being in state 1 of Figure (2.3) at time t.

Alternative estimator not relying on the Markov assumption

As discussed in Pepe [57], an alternative estimator for the PCA function not relying on the Markov assumption can be used, originally proposed by Tsai et al. [107, 114]. As the equation

$$P_{00}(0,t) + P_{01}(0,t) + P_{02}(0,t) = 1$$

holds, $P_{01}(0,t)$ can be expressed as

$$P_{01}(0,t) = 1 - P_{00}(0,t) - P_{02}(0,t),$$

where $P_{00}(0,t)$ can be estimated as the Kaplan-Meier estimator for the first event and $P_{02}(0,t)$ as 1 minus the Kaplan-Meier estimator for death. With the condition that censoring is independent of the state, the alternative non-parametric estimator for the PCA function is then given by

$$\begin{aligned}\hat{P}_{01}(0,t) &= 1 - \prod_{u \leq t}\left(1 - \frac{d(N_{01}(u) + N_{02}(u))}{Y_0(u)}\right) - \left(1 - \prod_{u \leq t}\left(1 - \frac{d(N_{02}(u) + N_{12}(u))}{Y_0(u) + Y_1(u)}\right)\right) \\ &= \prod_{u \leq t}\left(1 - \frac{d(N_{02}(u) + N_{12}(u))}{Y_0(u) + Y_1(u)}\right) - \prod_{u \leq t}\left(1 - \frac{d(N_{01}(u) + N_{02}(u))}{Y_0(u)}\right).\end{aligned}$$

In Figure 6.1 out of Section 6.1, $P_{01}(0,t)$ can be expressed as

$$P_{01}(0,t) = 1 - P_{00}(0,t) - P_{02}(0,t) - P_{03}(0,t).$$

Again, $P_{00}(0,t)$ can be estimated as the Kaplan-Meier estimator for the first event and $P_{02}(0,t)$ as 1 minus the Kaplan-Meier estimator for death. However, $P_{03}(0,t)$ can not be estimated as as 1 minus the Kaplan-Meier estimator due to the presence of the competing event "death". An option is to combine state 2 and 3 for the alternative estimator as this does not affect the examination of $P_{01}(0,t)$, such that

$$\hat{P}_{01}(0,t) = \prod_{u \leq t}\left(1 - \frac{d(N_{02}(u) + N_{03}(u) + N_{12}(u))}{Y_0(u) + Y_1(u)}\right) - \prod_{u \leq t}\left(1 - \frac{d(N_{01}(u) + N_{02}(u))}{Y_0(u)}\right).$$

In Figure 6.6 out of Section 6.2, $P_{01}(0,t)$ can be expressed as

$$P_{01}(0,t) = 1 - P_{00}(0,t) - P_{02}(0,t) - P_{03}(0,t) - P_{04}(0,t).$$

As above, $P_{00}(0,t)$ can be estimated as the Kaplan-Meier estimator for the first event. However, $P_{02}(0,t)$, a competing event transition *before* potential cure, can not be estimated as 1 minus the Kaplan-Meier estimator due to the presence of the competing event "cure", the same applies to $P_{03}(0,t)$ and $P_{04}(0,t)$. Again, an option is to combine

state 2, 3, and 4 for the alternative estimator as this does not affect the examination of $P_{01}(0,t)$, such that

$$\hat{P}_{01}(0,t) = \prod_{u \leq t} \left(1 - \frac{d(N_{02}(u) + N_{13}(u) + N_{14}(u))}{Y_0(u) + Y_1(u)}\right) - \prod_{u \leq t} \left(1 - \frac{d(N_{01}(u) + N_{02}(u))}{Y_0(u)}\right).$$

If the original data of the MODIFY I and II trial [96] were available, this method could be used to check if the Markov assumption is appropriate. In Figure 6.9 out of Section 6.3, the estimand of interest is the combination of $P_{01}(0,t) + P_{03}(0,t)$, the probability of being extubated alive (but still in hospital) plus the probability of being discharged from hospital. $P_{01}(0,t)$ can be expressed as

$$P_{01}(0,t) = 1 - P_{00}(0,t) - P_{02}(0,t) - P_{03}(0,t).$$

For estimation, we may combine state 2 and 3, such that

$$\hat{P}_{01}(0,t) = \prod_{u \leq t} \left(1 - \frac{d(N_{02}(u) + N_{12}(u) + N_{13}(u))}{Y_0(u) + Y_1(u)}\right) - \prod_{u \leq t} \left(1 - \frac{d(N_{01}(u) + N_{02}(u))}{Y_0(u)}\right).$$

$P_{03}(0,t)$ can be expressed as

$$P_{03}(0,t) = 1 - P_{00}(0,t) - P_{02}(0,t) - P_{01}(0,t).$$

Because information about the $3 \to 2$ transition is incomplete (incomplete mortality follow-up for death after discharge), it is not possible to apply the "trick" as before. Here, a 13 transition constitutes a competing event for the $1 \to 2$ transition. Since only a component of our quantity of interest is estimable using the alternative method, we will check if the Markov assumption is fulfilled by studying the influence of the time to cure, the $0 \to 1$ transition, on the $1 \to 2$ reintubation or mortality transition and on the 13 discharge transition by including it as a time-dependent variable in a Cox model for the 12 and 13 hazard, respectively. If a complete mortality follow-up was available, such an alternative analysis may be repeated.

4 TREATMENT COMPARISON

When focusing on clinical cure as primary endpoint for comparison of an active treatment A to a control B, the traditional procedure is to estimate risk (incidence) differences with corresponding confidence intervals using treatment-specific incidence proportions of cured patients given by

$$\hat{I}_A - \hat{I}_B = \frac{\#\text{ cured with treatment } A}{n_A} - \frac{\#\text{ cured with treatment } B}{n_B} \tag{16}$$

at a pre-specified timepoint, say τ, mostly end of follow-up, to demonstrate non-inferiority or superiority. We will write # for the number of patients with a specific characteristic. n_A and n_B correspond to the treatment specific number of patients. In a complete data situation without censoring, these quotients are the correct estimators for the treatment-specific cumulative incidence functions of being cured at τ and the difference in (16) should coincide with the difference

$$\widehat{\text{CIF}}_{01}^{A}(\tau) - \widehat{\text{CIF}}_{01}^{B}(\tau).$$

In HAP and VAP trials, often non-inferiority analyses are applied [20], as can be seen in, e.g., Awad et al. [95]. For this, it is examined if the lower bound of confidence interval

$$\hat{I}_A - \hat{I}_B \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{I}_A\left(1-\hat{I}_A\right)}{n_A} + \frac{\hat{I}_B\left(1-\hat{I}_B\right)}{n_B}} \tag{17}$$

exceeds a protocol-defined non-inferiority margin. Here, $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution, for $\alpha = 5\%$ it is 1.96. Alternative methods to construct the confidence interval are discussed in Altman et al. [115].

The cure-death model in Figure 2.3 provides an analysis strategy that includes two

co-primary endpoints simultaneously. The timing of the events cure and death as well as their chronological sequence are modelled with an initial state 0 (randomisation), a cure state 1, and a death state 2. Based on this model, we will focus on the following possibilities for a treatment comparison:

1. Risk differences with proportions of patients cured *and* alive (focus on a pre-specified timepoint only),

2. Time-simultaneous one-sided confidence bands (extension of point 1 to a time interval of interest),

3. Pseudo-value regression techniques (time-simultaneous examination allowing for additional covariate adjustment), and

4. Restricted log-rank-based test of treatment effect (alternative approach based on the well-known log-rank test).

We will discuss advantages and disadvantages of the proposed procedures in showing non-inferiority or superiority.

4.1 Risk differences cured and alive

A simple procedure to examine both cure and death is to use risk differences with incidence proportions of patients cured *and* alive given by

$$\frac{\#\text{ cured and alive with treatment } A}{n_A} - \frac{\#\text{ cured and alive with treatment } B}{n_B} \tag{18}$$

that is tantamount with the treatment-specific proportions of patients in state 1 of model 2.3 at time τ. In a complete data situation without censoring, these quotients are the correct estimators for the treatment-specific probabilities to be cured and alive at τ and the difference is given by

$$\hat{P}_{01}^{A}(\tau) - \hat{P}_{01}^{B}(\tau).$$

The confidence interval for (18) is calculated analogous to (17) that enables statements concerning non-inferiority and superiority. Also, a chi-squared test for equality of proportions can be applied.

However, we can make inferences about one timepoint only and, consequently, the result strongly depends on this selected timepoint. By not considering the time-dynamic process, lots of information that may be highly relevant from the patients' perspective can get lost. A time-simultaneous solution is provided in Section 4.2.

4.2 Time-simultaneous confidence bands

A much more convincing statement than merely demonstrating non-inferiority or superiority at a single point in time is to consider the entire follow-up period. From the patients' perspective it is a highly relevant information how the active treatment performs over the complete cure process [33, 11]. To assess, e.g., time-simultaneous non-inferiority, a so-called "confidence band" is required, in which the difference in probabilities of being cured and alive for active treatment A minus control B,

$$P_{01}^{A}(0,t) - P_{01}^{B}(0,t),$$

over a relevant time interval, e.g. $[0,\tau]$, and not just at a single timepoint lies with, e.g., a probability of 95%. Such a confidence band generalises the concept of a confidence interval to an entire time interval of interest. Adapting the principles of confidence interval inclusion as discussed in [116] to time-simultaneous confidence bands, treatment A can be deemed non-inferior to B if the confidence band for the difference of the treatment-specific probability curves,

$$\hat{P}_{01}^{A}(0,t) - \hat{P}_{01}^{B}(0,t) - q_{\alpha},$$

lies above the protocol-defined non-inferiority margin over the whole time period considered [2]. The statement for superiority analyses works analogously while using a reference value of zero.

We will construct q_{α} adopting a resampling procedure known as "wild bootstrap", originally proposed by Wu [117], that has first been applied in regression analyses when there are heteroskedastic errors. Nonparametric analysis of transition probabilities in multistate models based on asymptotic theory is challenging due to the complicated structure of the limiting covariance process. In a simple competing risks setting, Lin [91] proposes to work with a martingale representation for the limit process of the

Aalen-Johansen estimator of the cumulative incidence function minus the true quantity. This technique keeps the data fixed while martingales are substituted by multipliers that involve simulated normal variates. Lin also proposes a confidence band and a Kolmogorov-Smirnov type test for comparing cumulative incidence functions. Beyersmann et al. [92] provide a general conditional central limit theorem for the wild bootstrap in Lin's approach and also show that other multipliers, as e.g., centred Poissons, may lead to better finite sample performance. Bluhmki et al. [118] extend Lin's resampling technique to general multistate situations where the aim is to estimate the matrix of transition probabilities. The latter work is applied here.

To construct q_α, let us first define the difference

$$D(0,t) := k(t)\left[\left(\hat{P}_{01}^A(0,t) - P_{01}^A(0,t)\right) - \left(\hat{P}_{01}^B(0,t) - P_{01}^B(0,t)\right)\right],$$

with positive weight function k. Different weight functions result into different types of confidence bands. Following Lin [91], we can choose $k(t) = \mathbf{I} \quad \forall t \in [0,\tau]$, resulting into linear confidence bands where the width does not depend on time t.

Many other ways to construct such a confidence band are conceivable, e.g., depending on the precision of the respective estimate or, as examined in Hieke et al. [119] for the ordinary competing risks setting, using weights according to the standard log-rank test, that is $k(t) = \frac{Y_A(t)Y_B(t)}{Y_A(t)+Y_B(t)} \quad \forall t \in [0,\tau]$, where Y_A and Y_B are the risk processes for treatment group A and B. Here, for each day the weight is the product of the risk sets in each group divided by the sum of these risk sets. The uncertainty of the estimated transition probability does depend on time and so does the uncertainty of the difference. The confidence band with weight function as in the log-rank test or depending on the estimate precision might appear as a reasonable choice since it gets wider with decreasing risk sets or decreasing precision, respectively. However, we choose a confidence band with equal width over time due to several reasons: A confidence band that gets wider

over time would never allow to reject the null hypothesis of non-inferiority even if an innovative treatment would be substantially better than an alternative. A possibility is to restrict to a more narrow time frame of interest or to keep the same width over time. For our situation, the "patient flow" through the model carries important information independent of the estimate precision or the size of the risk set. A decreasing risk set in the initial state, e.g., could imply an increasing cure proportion while a decreasing risk set in state 1 points to frequent mortality cases after cure. Such information is highly important for a treatment comparison and motivates the use of a quantile q_α independent of time.

The challenge is now to approximate the distribution of $D(0,t)$. The idea is to approximate the limit distribution of the Nelson-Aalen estimate in a first step and to use this information for approximating the limit distribution of the Aalen-Johansen estimate a the second step:

Step 1

Let the counting processes $Y_{l,i}(t)$, $Y_l(t)$, $N_{lj,i}(t)$, and $N_{lj}(t)$ as well as the Nelson-Aalen and Aalen-Johansen estimate be defined as in Section 3.1. According to Andersen et al. [112], it is shown that

$$M_{lj,i}(t) := N_{lj,i}(t) - \int_0^t Y_{l,i}(u)\lambda_{lj}(u)du$$

are martingales and that $\sqrt{n}\left(\hat{\Lambda}_{lj}(u) - \Lambda_{lj}(u)\right)$ has a martigale representation

$$\sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{\mathbf{1}(Y_l(u)>0)}{Y_l(u)}dM_{lj,i}(u). \tag{19}$$

Using these arguments and the martigale central limit theorem (Theorem II.5.1 in [112]),

convergence in distribution to a zero mean Gaussian limit process

$$\sqrt{n}\left(\hat{\mathbf{\Lambda}}(t) - \mathbf{\Lambda}(t)\right) \xrightarrow{d} \mathbf{U} = (U_{lj})_{l,j\in\mathcal{S}}, \tag{20}$$

the central limit theorem for the Nelson-Aalen estimator (Theorem IV.1.2 in [112]), is proven, while $\mathbf{U}$ contains independent Gaussian martingales as non-diagonal entries. To approximate $\mathbf{U}$ and to follow the idea of Lin [91], the unknown martignales in (19) are substituted by known quantities $G_{lj,i}dN_{lj,i}(u)$, the so-called multipliers, while conditioning on the data. In Lin, $G_{lj,i}$ is chosen to be standard normal, but also other variables are possible, as shown in Beyersmann et al. [92]. This results in a wild bootstrap version $\boldsymbol{\xi}(t) = (\xi_{lj}(t))_{l,j\in S}$, a 3×3 matrix process with non-diagonal entries

$$\xi_{lj}(t) = \sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{\mathbf{1}(Y_l(u) > 0)}{Y_l(u)} G_{lj,i}dN_{lj,i}(u)$$

of the left-hand side of (20) whose distribution may be approximated by simulating a large number of $G_{lj,i}$ replicates. To approximate the limit distribution, the wild bootstrap resampling is mathematically not required in the first step since the limit process has independent increments. Nevertheless, this is needed afterwards when making inference on transition probabilities.

Step 2

The matrix of transition probabilities can be expressed in terms of cumulative hazards using product integration

$$\mathbf{P}(s,t) = \prod_{u\in(s,t]} (\mathbf{I} + d\mathbf{\Lambda}(u)),$$

as in (10). Here, $\mathbf{I}$ is the 3×3 identity matrix. This gives the Aalen-Johansen estimator by including the Nelson-Aalen estimator for the cumulative hazard, as in (14). Again, we know out of martingale theory and the functional delta method (Theorem II.8.1 in

[112]) that the stochastic process on the left-hand side converges in distribution to a zero mean Gaussian limit process

$$\sqrt{n}\left(\hat{\mathbf{P}}(s,t)-\mathbf{P}(s,t)\right)\xrightarrow{d}\int_s^t \mathbf{P}(s,u)d\mathbf{U}(u)\mathbf{P}(u,t), \tag{21}$$

the central limit theorem for the Aalen-Johansen estimator (Theorem IV.4.2 in [112]). The problem is that in (21), a martignale representation is much more complicated compared to (20) since the Gaussian limit process lacs independent increments resulting into an enormously complicated covariance structure. The process cannot be approximated by a Brownian bridge. The trick is to plug in $\boldsymbol{\xi}(u)$ for $\mathbf{U}(u)$ into (21) and to receive convergence in distribution

$$\int_s^t \mathbf{P}(s,u)d\boldsymbol{\xi}(u)\mathbf{P}(u,t)\xrightarrow{d}\int_s^t \mathbf{P}(s,u)d\mathbf{U}(u)\mathbf{P}(u,t),$$

while the left-hand side is approximated by

$$\boldsymbol{\zeta}(s,t):=\int_s^t \hat{\mathbf{P}}(s,u)d\boldsymbol{\xi}(u)\hat{\mathbf{P}}(u,t). \tag{22}$$

For the cure-death model, (22) is a 3×3 matrix where entries (1,0) and (2,0) are zero and (2,2) is 1. Entries (0,0) and (1,1) are Kaplan-Meier-type estimators $\hat{P}_{00}(s,t)\xi_{00}(t)$ and $\hat{P}_{11}(s,t)\xi_{11}(t)$. Entry $(0,1)$ is

$$\int_s^t \hat{P}_{00}(s,u)d\xi_{00}(u)\hat{P}_{01}(u,t)+\int_s^t \hat{P}_{00}(s,u)d\xi_{01}(u)\hat{P}_{11}(u,t)+\int_s^t \hat{P}_{01}(s,u)d\xi_{11}(u)\hat{P}_{11}(u,t),$$

entry $(0,2)$ is

$$\int_s^t \hat{P}_{00}(s,u)d\xi_{00}(u)\hat{P}_{02}(u,t)+\int_s^t \hat{P}_{00}(s,u)d\xi_{01}(u)\hat{P}_{12}(u,t)+\int_s^t \hat{P}_{01}(s,u)d\xi_{11}(u)\hat{P}_{12}(u,t)$$
$$+\int_s^t \hat{P}_{00}d\xi_{02}(u)+\int_s^t \hat{P}_{01}(s,u)d\xi_{12}(u),$$

and entry $(1,2)$ is

$$\int_s^t \hat{P}_{11}(s,u)d\xi_{11}(u)\hat{P}_{12}(u,t)+\int_s^t \hat{P}_{11}(s,u)d\xi_{12}(u).$$

The distribution of $D(0,t)$ can now be approximated by

$$\hat{D}(0,t) := \frac{\zeta_{01}^A(0,t)}{\sqrt{n_A}} - \frac{\zeta_{01}^B(0,t)}{\sqrt{n_B}},$$

where $\zeta_{01}^A(0,t)$ and $\zeta_{01}^B(0,t)$ are the treatment-specific functions. Finally, the boundary value q_α is evaluated through simulation such that

$$P\left(\max_{t\in[0,\tau]} \hat{D}(0,t) > q_\alpha\right) = \alpha.$$

In other words, and for $\alpha = 5\%$, the procedure determines q_α via resampling such that

$$\begin{aligned} 0.95 &= P\left(\max_{t\in[0,\tau]} \hat{D}(0,t) \leq q_\alpha\right) \\ &\approx P\left(\left(\hat{P}_{01}^A(t) - P_{01}^A(t)\right) - \left(\hat{P}_{01}^B(t) - P_{01}^B(t)\right) \leq q_\alpha \quad \forall t \in [0,\tau]\right) \\ &= P\left(\hat{P}_{01}^A(t) - \hat{P}_{01}^B(t) - q_\alpha \leq P_{01}^A(t) - P_{01}^B(t) \quad \forall t \in [0,\tau]\right) \end{aligned}$$

guaranteeing that the true difference in the treatment-specific probabilities of being cured and alive lies above the lower margin $\hat{P}_{01}^A(t) - \hat{P}_{01}^B(t) - q_\alpha$ of the derived one-sided time-simultaneous confidence band for all $t \in [0,\tau]$ with a probability of approximately 95%. This induces an α-level non-inferiority test. In practice, it is recommended to decide for an interval $[0,\tau]$ that covers the first and last failure time since asymptotic approximations tend to be poor beyond these times [91].

Non-inferiority and superiority hypotheses

While the test for a non-inferiority analysis is based on the hypotheses

$$H_0: \quad P_{01}^A(0,t) - P_{01}^B(0,t) \leq \delta_{abs} \quad \text{versus} \quad H_1: \quad P_{01}^A(0,t) - P_{01}^B(0,t) > \delta_{abs},$$

with δ_{abs} as margin suitable for this absolute effect measure, the test for a superiority analysis is based on the hypotheses

$$H_0: \quad P_{01}^A(0,t) - P_{01}^B(0,t) \leq 0 \quad \text{versus} \quad H_1: \quad P_{01}^A(0,t) - P_{01}^B(0,t) > 0.$$

4.3 Pseudo-value regression

The transition probability of being cured and alive introduced in Section 3.2 is of main interest, where the Aalen-Johansen estimate is given by a product integral of transition intensities. However, plugging in the estimated transition intensities does not directly give estimates of covariate effects on the transition probability of interest since these are complex non-linear functions [120]. For this, Scheike et al. [121] proposed a binominal modelling approach to construct direct regression models for transition probabilities based on the inverse probabilities of censoring weighting technique. This approach was further extended by Azarang et al. [122]. Pseudo-value regression, proposed by Andersen et al. [88] and Andersen and Klein [89], provides an alternative technique for direct regression modelling of transition probabilities and thus, to test for treatment difference [123, 53]. It was first applied to an illness-death model in the setting of graft-versus-host disease after bone marrow transplantation [88] but found its application also in other settings, as, e.g., in Liu et al. [53] to study current leukaemia-free survival or in Grand and Putter [124] to explore the impact of socio-economic factors on the expected length of stay in health and disability. The idea is to obtain pseudo-values from a jackknife statistic constructed from a consistent estimator of the probability of interest that are further utilised as outcome variables in a generalised linear model. Model parameters are further estimated using generalised estimating equations (GEEs) [125].

Obtain pseudo-values

We begin with selecting a set of K timepoints s_k, $k \in \{1, \ldots, K\}$, on which we want to perform regression. The pseudo-values for every patient i, $i \in \{1, \ldots, n\}$, and every

timepoint k are computed as

$$\hat{\theta}_i(s_k) = n \cdot \hat{P}_{01}(0, s_k) - (n-1) \cdot \hat{P}_{01}^{(-i)}(0, s_k), \quad i \in \{1, \ldots, n\},$$

where $\hat{P}_{01}$ is the estimated transition probability using the complete sample and $\hat{P}_{01}^{(-i)}$ the one based on the sample without the ith observation, the so-called "leave-one-out estimator". To continue, a consistent estimator of the transition probability is needed, provided by the Aalen-Johansen estimator. The pseudo-values $\hat{\theta}_i = \left(\hat{\theta}_i(s_1), \ldots, \hat{\theta}_i(s_k)\right)$ can be seen as the contribution of subject i to the estimate of interest [120]. The data of some patients might be right-censored when incompletely observed and so is the estimator $\hat{P}_{01}$. The idea is to replace the incomplete observed data with the pseudo-value and treat it as it was raw data. The advantage is that pseudo-values are still defined for all individuals at all timepoints event if right-censoring is present. In the absence of censoring, this approach is equivalent to using the raw data and the pseudo-value reduces to an indicator of whether the patient is in state 1 at time s_k [126].

Estimate regression coefficients

Once pseudo-values for the selected timepoints and each patient are obtained, we continue with a generalised linear model

$$g(P_{01}(s_k \mid Z)) = \beta' Z,$$

with log link function g, parameter vector β, and covariate vector Z. The choice of the link function is important for the interpretation of the regression parameters [127]. Finally, we estimate the regression parameters using a generalised estimating equation [125]

$$U(\beta) = \sum_i \left(\frac{dg^{-1}(\beta' Z_i)}{d\beta} \right) W_i^{-1} \left(\hat{\theta}_i - g^{-1}(\beta' Z_i) \right),$$

with identity matrix for the working covariance matrix W_i. While an identity matrix is often used for simplicity, Andersen and Klein [89] showed via simulations that an empirical working covariance matrix has a slightly smaller mean squared error. The covariance matrix of $\hat{\beta}$ is estimated via an ordinary sandwich variance estimator. Graw et al. [128] provided proofs regarding the asymptotic properties of this approach in the competing risks setting, e.g., showing asymptotic equivalence of the uncensored observations and the pseudo-values with respect to their conditional expectations given covariates. Overgaard et al. [129] provided more refined and general proofs for the competing risks case, extending the result of consistency and asymptotic normality given by Jacobsen and Martinussen [130] for the survival case. Their approach may be further extended for general multistate models.

Choice of timepoints

As mentioned before, a selection of timepoints has to be made. When focusing on one timepoint only, the pseudo-value regression model can be compared to a censored data logistic regression model. Using all event times leads to large matrices in the GEE and consequently to a loss of efficacy such that most researchers have proposed to restrict to 5–10 timepoints equally spaced on the event time scale to capture a greatest possible information about the event time distribution without losing efficacy [126, 120, 131].

Non-inferiority and superiority hypotheses

To examine a treatment effect, $\exp(\hat{\beta})$ for treatment as covariate can be interpreted as a cure risk ratio (CRR) and tests for non-inferiority analyses can be based on the hypotheses

$$H_0: \quad \text{CRR} \leq \delta_{rel} \quad \text{versus} \quad H_1: \quad \text{CRR} > \delta_{rel},$$

with δ_{rel} as margin suitable for this relative effect measure. The test for a superiority analysis is based on the hypotheses

$$H_0: \quad \text{CRR} \leq 1 \quad \text{versus} \quad H_1: \quad \text{CRR} > 1.$$

4.4 Restricted log-rank-based test

Hsieh et al. [87] propose several tests comparing treatments on the basis of a semi-Markov model. The restricted log-rank-based test, e.g., is a non-parametric method to manage ordered categories of responses and to integrate information on duration of response that fits to the situation present in the cure-death model.

To begin, let us have a look at a proportional transition-specific hazard model which assumes the hazard of each transition $j \in \{01, 02, 12\}$ in the cure-death model to follow a Cox model [84]

$$\lambda_j(t \mid Z_i) = \lambda_{j;0}(t) \exp(\beta_j' Z_i), \tag{23}$$

with non-negative baseline hazard function $\lambda_{j;0}(t)$ and linear predictor $\beta_j' Z_i$. The partial likelihood, originally introduced by Cox [132], is used for estimation of the regression coefficients and can be written as

$$\begin{aligned} &L(\beta_{01}, \beta_{02}, \beta_{12}) = L_{01}(\beta_{01}) \times L_{02}(\beta_{02}) \times L_{12}(\beta_{12}) \\ =&\prod_{k=1}^{K_{01}} \frac{\exp(\beta_{01}' Z_{(k)})}{\sum_{r \in R_{t_{01(k)}}} \exp(\beta_{01}' Z_r)} \times \prod_{k=1}^{K_{02}} \frac{\exp(\beta_{02}' Z_{(k)})}{\sum_{r \in R_{t_{02(k)}}} \exp(\beta_{02}' Z_r)} \times \prod_{k=1}^{K_{12}} \frac{\exp(\beta_{12}' Z_{(k)})}{\sum_{r \in R_{t_{12(k)}}} \exp(\beta_{12}' Z_r)}, \end{aligned}$$

where $R_{t_{01(k)}}$ and $R_{t_{02(k)}}$ are the sets of individuals that are still in state 0 at transition time $t_{01(k)}$ or $t_{02(k)}$ and at risk for transition 01 or 02, respectively. $R_{t_{12(k)}}$ is the set of individuals that are still alive at the transition time to death $t_{12(k)}$. K_{01} is the total number of individuals reaching state cure, K_{02} the total number of individuals reaching state death without being cured, and K_{12} the total number of individuals reaching death after being cured. The likelihood can be factorised for each j so that we can formally analyse each transition separately by treating the others as censored. A test of $\beta_j = 0$ is given as a simple score statistic with the score function

$$\left[\frac{\partial \log L_{01}(\beta_{01})}{\partial \beta_{01}}, \frac{\partial \log L_{02}(\beta_{02})}{\partial \beta_{02}}, \frac{\partial \log L_{12}(\beta_{12})}{\partial \beta_{12}} \right]$$

and the negative expected value of the second derivative of the partial likelihood function as information matix.

As mentioned in Section 2.1, the score test statistic for the Cox partial likelihood is the same as the log-rank test statistic when the data consists of failure time data in two groups. This is due to the fact that the numerator of the score test for a test of $\beta_j = 0$ turns out to be identical to the numerator, # observed minus # expected, of the log-rank test [102]. Moreover, the estimated variance obtained from the Cox model is nearly identical to the denominator in the log-rank test. Let us have a deeper look into the construction of the log-rank test statistic. It compares estimates of the hazard functions of two (treatment) groups A and B at each time l where there is an event the following way: For each time $l \in \{1, \ldots, L\}$, let R_{Al} and R_{Bl} be the number of subjects at risk and $R_l = R_{Al} + R_{Bl}$. Let O_{Al} and O_{Bl} be the observed number of events in the groups respectively at time l, and define $O_l = O_{Al} + O_{Bl}$. Under the null hypothesis of treatment equality and given that O_l events happened at time l, O_{Al} is hypergeometrically distributed with parameters R_l, R_{Al}, and O_l. This distribution has expected value $E_{Al} = \frac{O_l}{R_l} R_{Al}$ and variance $V_l = \frac{O_l(R_{Al}/R_l)(1-R_{Al}/R_l)(R_l-O_l)}{R_l-1}$. Finally, the log-rank test statistic compares each observed value O_{Al} to its expectation value E_{Al} and is defined as

$$\frac{(O-E)^2}{V} := \frac{\left(\sum_{l=1}^{L} O_{Al} - \sum_{l=1}^{L} E_{Al}\right)^2}{\sum_{l=1}^{L} V_l}.$$

General log-rank-based test

Such a test is sensitive to deviations from the null hypothesis of $\beta_j = 0$ of any type. It is constructed by computing the observed and expected number of events (O and E) for each transition 01, 02, or 12, in one of the groups at each observed event time and then adding these to obtain an overall summary across all timepoints where there is an event, divided by the variance V. The log-rank-based test (LT) results in the sum of

three under H_0 asymptotically independent log-rank test statistics

$$\begin{aligned}\chi^2_{LT} &:= \chi^2_{01} + \chi^2_{02} + \chi^2_{12} \\ &= \frac{(O_{01} - E_{01})^2}{V_{01}} + \frac{(O_{02} - E_{02})^2}{V_{02}} + \frac{(O_{12} - E_{12})^2}{V_{12}} \quad \overset{H_0}{\sim} \quad \chi^2(3).\end{aligned}$$

Restricted log-rank-based test

Yet, the model is not adapted to the request that a transition to cure is preferred over a transition to death. The overall aim is that a patient passes into state 2 (death) as late as possible and remains in state 1 (cure) as long as possible. With a restriction to the regression coefficients in the partial likelihood ($\beta_{01} = -\beta_{12} = -\beta_{02}$), the restricted log-rank-based test (RLT) respects that ordered response and results in

$$\chi^2_{RLT} := \frac{(O_{RL} - E_{RL})^2}{V_{RL}} \quad \overset{H_0}{\sim} \quad \chi^2(1),$$

a test with an embedded structure where $O_{RLT} := O_{02} - O_{01} + O_{12}$, $E_{RLT} := E_{02} - E_{01} + E_{12}$, and $V_{RLT} := V_{02} + V_{01} + V_{12}$. It incorporates all required aspects into one single statistic being χ^2-distributed with one degree of freedom. This restricted version is sensitive to deviations from the null hypothesis if a transition to cure dominates a direct death transition and if cured, a patient remains in the cure state. Hsieh et al. [87] showed that this test statistic achieves the highest power when one treatment is better than the other for all three transitions in the desired way (more patients are cured, less patients die directly, and less patients die after cure).

The restricted log-rank-based test achieves only high power if one treatment is better than the other for all three transitions [87]. Another point is that the direction of difference cannot be discerned by a test statistic based on a quadratic form such that non-inferiority analyses are not possible. Strictly speaking, it only fits to tests for equality.

Remark

These tests can easily be extended if an additional state needs to be included. In Section 6.1, Figure 6.1, they will be of the form

$$\begin{aligned}\chi^2_{LT} &= \chi^2_{01} + \chi^2_{02} + \chi^2_{12} + \chi^2_{03} + \chi^2_{32} \\ &= \frac{(O_{01} - E_{01})^2}{V_{01}} + \frac{(O_{02} - E_{02})^2}{V_{02}} + \frac{(O_{12} - E_{12})^2}{V_{12}} + \frac{(O_{03} - E_{03})^2}{V_{03}} + \frac{(O_{32} - E_{32})^2}{V_{32}} \\ &\overset{H_0}{\sim} \quad \chi^2(5)\end{aligned}$$

and

$$\chi^2_{RLT} = \frac{(O_{RLT} - E_{RLT})^2}{V_{RLT}} \quad \overset{H_0}{\sim} \quad \chi^2(1),$$

where $O_{RLT} := O_{02} - O_{01} + O_{12} + O_{03} + O_{32}$, $E_{RLT} := E_{02} - E_{01} + E_{12} + E_{03} + E_{32}$ and $V_{RLT} := V_{02} + V_{01} + V_{12} + V_{03} + V_{32}$.

For multistate model 6.6 in Section 6.2, e.g., they will be of the form

$$\begin{aligned}\chi^2_{LT} &= \chi^2_{01} + \chi^2_{02} + \chi^2_{13} + \chi^2_{14} \\ &= \frac{(O_{01} - E_{01})^2}{V_{01}} + \frac{(O_{02} - E_{02})^2}{V_{02}} + \frac{(O_{13} - E_{13})^2}{V_{13}} + \frac{(O_{14} - E_{14})^2}{V_{14}} \\ &\overset{H_0}{\sim} \quad \chi^2(4)\end{aligned}$$

and

$$\chi^2_{RLT} = \frac{(O_{RLT} - E_{RLT})^2}{V_{RLT}} \quad \overset{H_0}{\sim} \quad \chi^2(1),$$

where $O_{RLT} := O_{02} - O_{01} + O_{13} + O_{14}$, $E_{RLT} := E_{02} - E_{01} + E_{13} + E_{14}$ and $V_{RLT} := V_{02} + V_{01} + V_{13} + V_{14}$.

Or for multistate model 6.9 in Section 6.3, e.g., they will be of the form

$$\begin{aligned}\chi^2_{LT} &= \chi^2_{01} + \chi^2_{02} + \chi^2_{12} + \chi^2_{13} \\ &= \frac{(O_{01} - E_{01})^2}{V_{01}} + \frac{(O_{02} - E_{02})^2}{V_{02}} + \frac{(O_{12} - E_{12})^2}{V_{12}} + \frac{(O_{13} - E_{13})^2}{V_{13}} \\ &\overset{H_0}{\sim} \quad \chi^2(4)\end{aligned}$$

and

$$\chi^2_{RLT} = \frac{(O_{RLT} - E_{RLT})^2}{V_{RLT}} \quad \overset{H_0}{\sim} \quad \chi^2(1),$$

where $O_{RLT} := O_{02} - O_{01} + O_{12} - O_{13}$, $E_{RLT} := E_{02} - E_{01} + E_{12} - E_{13}$ and $V_{RLT} := V_{02} + V_{01} + V_{12} + V_{13}$.

5 SIMULATION

The purpose of the following simulation is to demonstrate how the cure-death model and the proposed methods for a treatment comparison handle simple and complex treatment imbalances. For ease of illustration, we assume the transition hazards to be constant. Data is generated according to Beyersmann et al. [41], where for the standard treatment, "treatment B", we choose time constant hazard rates $\lambda_{01}(t)^B = 0.07$, $\lambda_{02}^B(t) = 0.04$, and $\lambda_{12}(t)^B = 0.02$ and for the innovative treatment, "treatment A", several scenarios are examined. Also, no additional censoring is generated.

We compare the methods introduced in Section 4 in a non-inferiority and superiority setting. However, the restricted log-rank-based test out of Section 4.4 cannot be used for non-inferiority analyses. Furthermore, we will apply the confidence band procedure out of Section 4.2 only in the non-inferiority setting, since, for superiority analyses, a particular part of the time frame has to be chosen dependent on the clinical setting. Also, to make a difference- versus ratio-based effect measure comparable for non-inferiority analyses, let us make the following consideration:

$$\begin{aligned} & H_0 : I^A - I^B \leq \delta_{abs} = f \cdot I^B \\ \Leftrightarrow\ & H_0 : I^A \leq I^B \cdot (f+1) \\ \Leftrightarrow\ & H_0 : \frac{I^A}{I^B} \leq f + 1. \end{aligned}$$

For δ, let us assume a hypothetical margin of -12.5%, as proposed in the EMA guideline for the comparison of clinical cure rates [20]. For I^B, let us assume the end of follow-up to be at day 30 and generate 1000 data sets with 300 individuals each using the transition rates mentioned above. This gives a mean proportion of patients being cured

and alive of approximately 42%, our "baseline risk". Thus,

$$\begin{aligned} -12.5\% &= f \cdot 42\% \\ \Rightarrow f &= -0.3 \\ \Rightarrow H_0 : \frac{I^A}{I^B} &\leq -0.3 + 1 = 0.7 = \delta_{rel}, \end{aligned}$$

such that a comparable cure risk ratio for a non-inferiority margin of -12.5% is 0.7 with a baseline risk of being cured and alive of 42%.

5.1 Simulation scenarios

Several simulation scenarios for the new treatment "treatment A" were examined for a treatment comparison with 50 and 300 individuals in each treatment group and 1000 simulated studies. The cause-specific hazard ratio is given for transition 01 and 02 and the hazard ratio for transition 12; treatment differences are marked in bold:

- Scenario 1: Treatment A is better in the cure rate (more cure cases),
  $\boldsymbol{\lambda_{01}^A = 0.14}$ (CSHR=2),
  $\lambda_{02}^A = 0.04$ (CSHR=1),
  and $\lambda_{12}^A = 0.02$ (HR=1)

- Scenario 2: Treatment A is better for death after cure (less death cases),
  $\lambda_{01}^A = 0.07$ (CSHR=1),
  $\lambda_{02}^A = 0.04$ (CSHR=1),
  and $\boldsymbol{\lambda_{12}^A = 0.005}$ (HR=0.25)

- Scenario 3: Treatment A is better for death without being cured,
  $\lambda_{01}^A = 0.07$ (CSHR=1),
  $\boldsymbol{\lambda_{02}^A = 0.01}$ (CSHR=0.25),
  and $\lambda_{12}^A = 0.02$ (HR=1)

- Scenario 4: Treatment A better in both mortality rates,
  $\lambda_{01}^{A} = 0.07$ (CSHR=1),
  $\boldsymbol{\lambda_{02}^{A} = 0.01}$ (CSHR=0.25),
  and $\boldsymbol{\lambda_{12}^{A} = 0.005}$ (HR=0.25)

- Scenario 5: Treatment A is better in the cure rate but worse in mortality rates,
  $\boldsymbol{\lambda_{01}^{A} = 0.14}$ (CSHR=2),
  $\boldsymbol{\lambda_{02}^{A} = 0.06}$ (CSHR=1.5),
  and $\boldsymbol{\lambda_{12}^{A} = 0.03}$ (HR=1.5).

An overview of all simulation scenarios can be found in Table 5.1.

Scenario	λ_{01}^{A}	CSHR	λ_{02}^{A}	CSHR	λ_{12}^{A}	HR
1	0.14	2	0.04	1	0.02	1
2	0.07	1	0.04	1	0.005	0.25
3	0.07	1	0.01	0.25	0.02	1
4	0.07	1	0.01	0.25	0.005	0.25
5	0.14	2	0.06	1.5	0.03	1.5
λ_{01}^{B}	0.07		0.04		0.02	

Table 5.1: Overview of all five simulation scenarios with λ_{01}^{A} as transition hazard for the treatment group and λ_{01}^{B} as transition hazard for the control group. CSHR = Cause-specific hazard ratio, HR = Hazard ratio, effect measures for treatment divided by control.

5.2 Results

Risk differences cured and alive

In Figure 5.1, risk differences with 95% confidence intervals can be seen using the overall proportions of patients cured and using the proportions of patients cured and alive at day 30 with a hypothetical non-inferiority margin of -12.5%. For both, mean values over 1000 studies are presented. In Scenario 1, treatment A is superior concerning cure. The overall risk difference for cured patients is 14.8%, significantly favouring treatment A. Using only patients cured and alive at day 30, the confidence interval is wider and covers zero (value of no effect). In Scenario 2, more patients stay alive after being cured for treatment A. While there is no difference in the analysis using only patients cured, the proposed analysis using patients cured and alive shows a significant effect favouring treatment A. In Scenario 3, there is no huge difference among the analysis strategies but both measures show a positive effect for treatment A because the competing event death without being cured has less impact. In Scenario 4, where treatment A is better in both mortality hazards, the analysis strategies differ substantially since patients rather stay cured and alive, comparable to Scenario 2. It is interesting that in Scenario 5, when analysing proportions of patients cured and alive at day 30, the risk difference is negative and non-inferiority is not given anymore, while a positive effect is present using only patients cured. This scenario is motivated by trials with extremely high mortality rates or when a microbiological cure is examined (the pathogen is eradicated but the patient dies due to a toxic treatment).

To be cured and alive over time

In the left part of Figure 5.2, the probability to be cured and alive for treatment A and B is displayed to get an overview how the different scenarios perform over the interval

[0, 40]. The difference in probabilities is given in the right part of Figure 5.2. The curves represent mean values over the Aalen-Johansen estimates from 1000 simulated studies. In Scenario 1, the superiority of treatment A concerning cure can be seen especially during the first phase. After a while, when transitions from cure to death occur with the same rate, the difference between treatmens becomes less. In Scenario 2, more patients stay alive after being cured for treatment A. Since transitions from cure to death occur mostly later in time, a treatment difference is present only after a while. A similar picture can be seen in Scenario 3 and 4. In Scenario 5, a positive effect can be seen during the first days but curves cross since the $1 \to 2$ transition is more exposed for treatment A.

Additionally, plots representing the overall probability to die are given in Figure 5.3. Here, the curves show mean values over the Kaplan-Meier estimates from 1000 simulated studies.

Power estimates

Power estimates for the non-inferiority and superiority null hypotheses are summarised in Table 5.2. Results are given for the chi-squared test for equality of proportions cured and alive at day 30, the test based on confidence bands, the pseudo-value regression results using ten times equally distributed over the whole time frame and at day 30, and the restricted log-rank-based test for the difference of two transition probabilities.

The chi-squared test for equality of proportions cured and alive at day 30 is very similar to the pseudo-value regression results comparing transition probabilities at day 30, especially with a larger group of individuals. Pseudo-value regression using ten times equally distributed over the whole time frame represents the difference between the curves displayed in Figure 5.2 best, when interest is focused on a relative effect measure. The

procedure using confidence bands is stricter in comparison to pseudo-value regression, especially when there are only few individuals available, resulting in a wide confidence band. Using a larger amount of individuals, the performance improves promptly due to much more narrow confidence bands. The restricted log-rank-based test detects given treatment differences satisfactory and is not limited to examining one timepoint only. In Scenario 5, interpretation of a combined risk over the whole time frame is possible only to a limited extent since the treatment advantage changes over time.

5.3 Discussion

For a complete picture, transition probabilities should be taken into account. Pseudo-value regression using ten times equally distributed over the whole time frame provides a possibility to analyse effects at several timepoints simultaneously. If intended, a specific time interval of interest can be examined in more detail. Thus, it could be helpful in settings like in Scenario 5, where the effect changes over time and interpretation may depend on which risk is more important at what time. This technique results in a relative effect measure for treatment difference, comparable to a relative risk over time. However, absolute effect measures are often of interest, where the method using confidence bands is better suited. Looking at the absolute difference over time in combination with confidence bands could provide a valuable statistical tool for such analyses, especially when treatment effects vary over time. Although the restricted log-rank-based test performs quite well, it does not directly concern the transition probability of interest. Moreover, it is only applicable for superiority analyses and thus, not recommended in this setting.

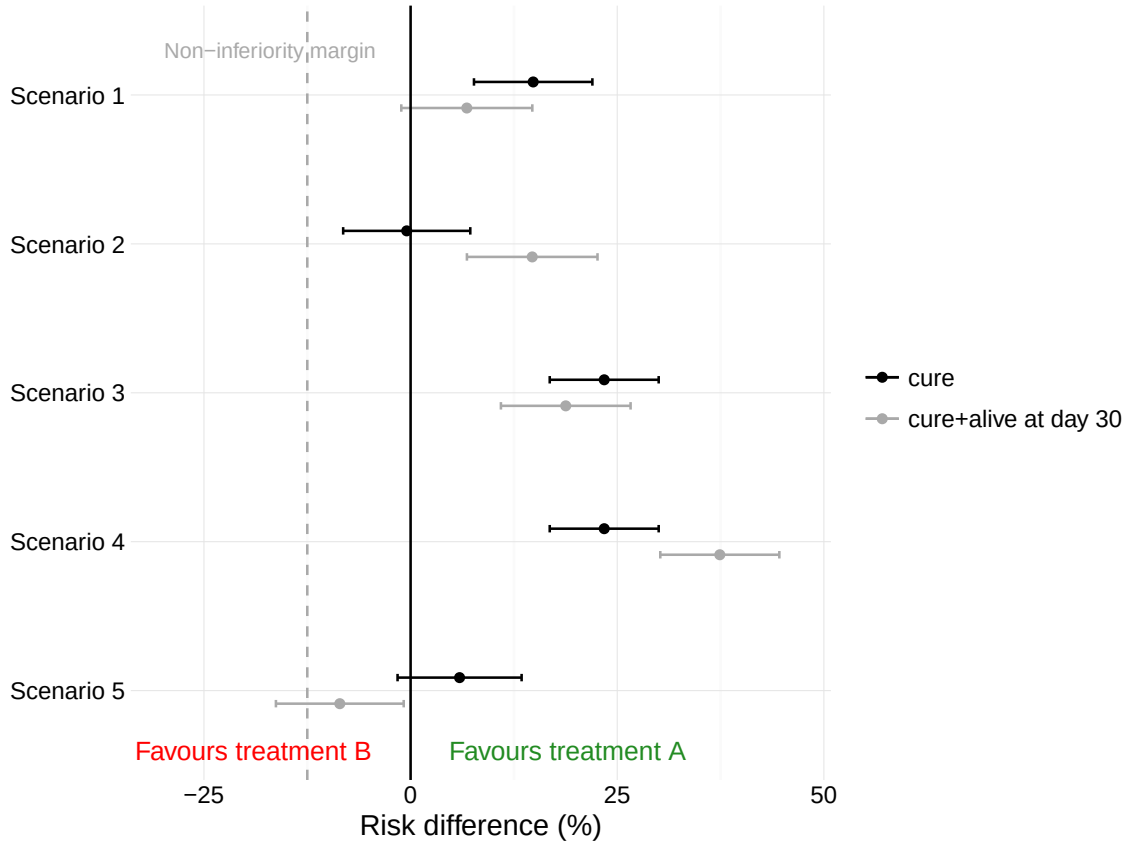

Figure 5.1: Risk differences with 95% confidence intervals for the comparison of two antimicrobial therapies. Hypothetical non-inferiority margin was set to -12.5%. Risk differences using the overall proportions of patients cured are displayed in black, risk differences using the proportions of patients cured and alive at day 30 are displayed in grey.

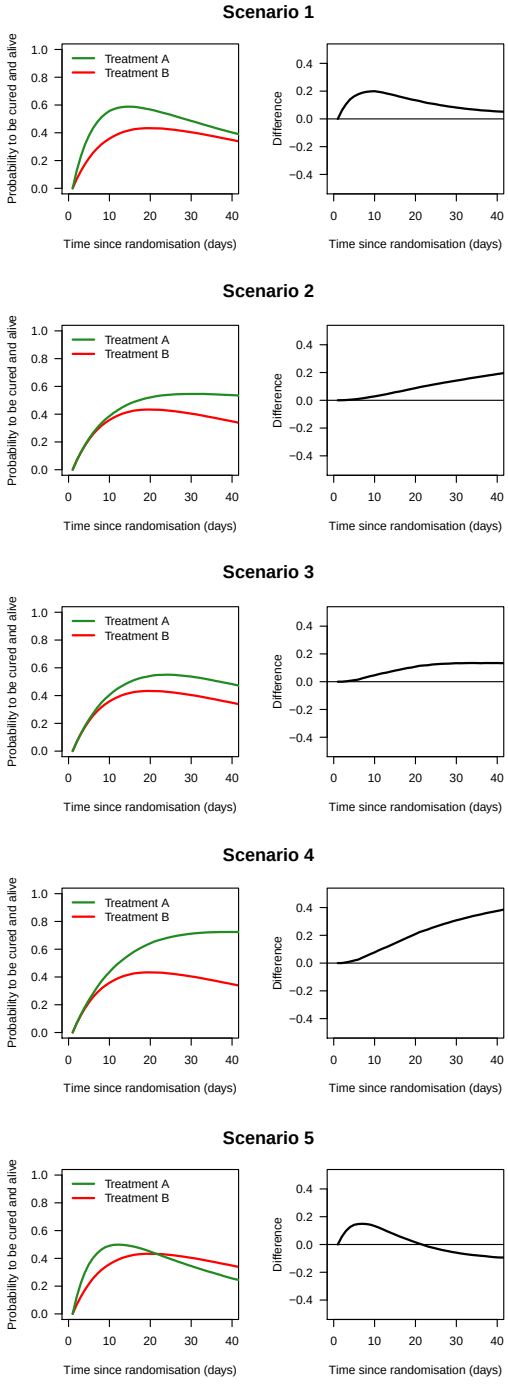

Figure 5.2: Mean transition probabilities and their difference for the simulation scenarios with 300 individuals in each treatment group and 1000 independent data sets.

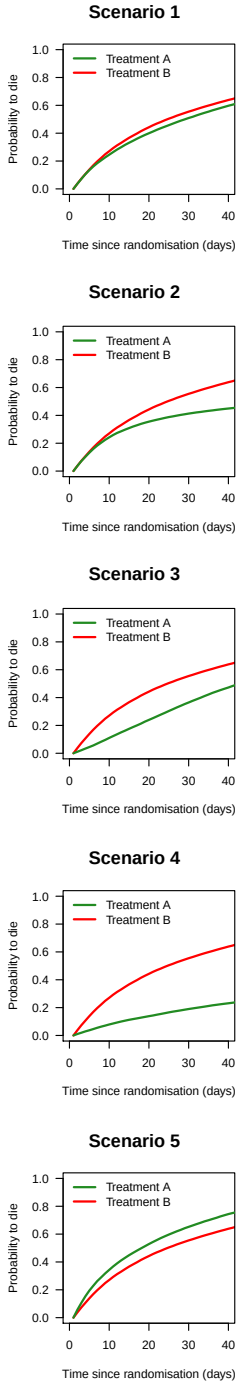

Figure 5.3: Mean probabilities to die for the simulation scenarios with 300 individuals in each treatment group and 1000 independent data sets.

		Non-inferiority					**Superiority**				
n	Scenario	χ^2 cure+alive	CB	Pseudo all	Pseudo 30	RLT	χ^2 cure+alive	CB	Pseudo all	Pseudo 30	RLT
	1	47.6	14.4	93.7	63.8		10.4		31.9	10.3	59.1
	2	79.6	5.6	85.7	88.4		33.4		19.9	28.0	20.3
50	3	87.7	3.8	91.1	85.3	–	49.1	–	21.6	24.7	25.4
	4	99.9	9.6	99.9	100		97.2		79.9	90.0	84.0
	5	63.0	1.1	47.2	8.0		0		2.3	0	9.0
	1	98.7	97.2	100	100		31.7		98.0	49.7	100
	2	100	72.7	100	100		95.9		86.6	98.0	56.5
300	3	100	70.5	100	100	–	99.6	–	90.4	94.4	99.8
	4	100	82.5	100	100		100		100	100	100
	5	17.4	3.2	99.8	38.8		0		6.7	0	23.7

Table 5.2: Power estimates for the chi-squared test for equality of proportions cured and alive at day 30 (χ^2 cure+alive), the confidence band procedure (CB), the pseudo-value regression using ten times equally distributed over the whole time frame (Pseudo all) and at day 30 (Pseudo 30), and the restricted log-rank-based test for the difference of two transition probabilities (RLT). For non-inferiority analyses, a hypothetical margin of -12.5% is assumed, corresponding to a CRR of 0.7. Results are given for $n = 50$ individuals and $n = 300$ individuals in each group, out of 1000 simulated studies.

6 APPLICATION

6.1 Ceftobiprole trial

The cure-death model provides a suitable framework for analysing the data of the recently published ceftobiprole trial [95]. As in this data example, one of the problems that usually arise is that for patients who achieved the cure state after the TOC, this is not recorded anymore. A consequential special feature of the such data is that patients after the TOC or patients who experienced a clinical failure are no longer under risk for transition from randomisation to cure in Figure 2.3. Thus, the log-rank-based test statistics have to be extended to be suitable to the model in Figure 6.1 as explained in Section 4.4. Here, an additional state "failure" is used for patients where systemic nonstudy antibiotics between baseline and the TOC visit for the treatment of pneumonia were received or an adverse event occurred.

6.1.1 The trial

The present non-inferiority trial [95] compared the new regimen ceftobiprole, established to combat a wide range of gram-positive bacteria, such as, e.g., *Staphylococcus aureus*, to the two-drug regimen ceftazidime / linezolid. It was a double-blind, randomised, multicentre phase III comparison in 781 patients with HAP, among them 210 with VAP, conducted during April 2005 and May 2006 in 157 sites in Europe, North and South America, and the Asia-Pacific region.

Clinical cure diagnosed at the TOC visit, mostly held within a time frame of 7 up to 14 days after the end of treatment, served as primary endpoint, all-cause mortality as secondary. With the protocol-defined non-inferiority margin of -15%, risk differences of proportions of cured patients at the TOC showed that ceftobiprole is non-inferior

to ceftazidime / linezolid for the entire study population of patients with HAP (−2.9 [−10.0, 4.1]) and HAP excluding VAP (0.8 [−7.3, 8.8]). But, non-inferiority was not demonstrated in VAP patients (−13.7 [−26.0, −1.5]) since the previous 95% confidence interval is not completely included in [−15, ∞). These results were given in [95] and refer to ceftobiprole minus ceftazidime / linezolid.

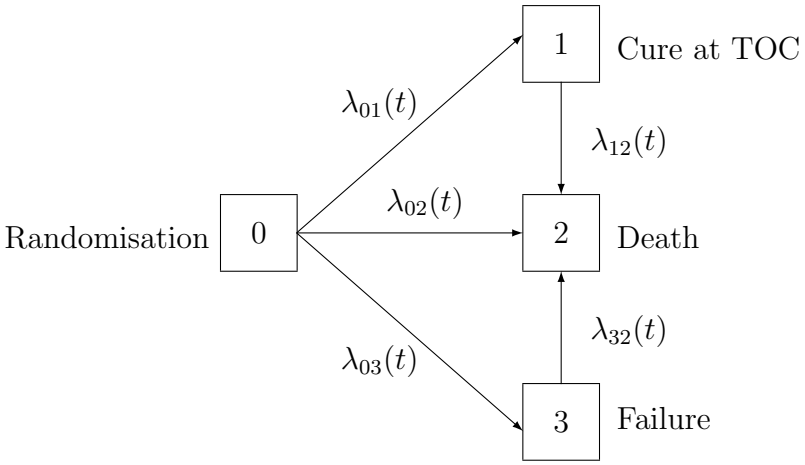

Figure 6.1: The extended cure-death model for the comparison of two antimicrobial therapies in the ceftobiprole trial [95]. Let $\lambda_{01}(t)$ be the cure rate for cure at the test of cure (TOC) visit, $\lambda_{02}(t)$ the mortality rate without being cured at TOC, $\lambda_{03}(t)$ the failure rate for failure at TOC, $\lambda_{12}(t)$ the rate to die after deemed cured at TOC, and $\lambda_{32}(t)$ the rate to die after deemed as a failure at TOC.

6.1.2 Results

The data visualisation in Figure 6.3 provides an illustration of the time course of events for the ceftobiprole and the ceftazidime / linezolid group. On the x-axis, time from randomisation, which equals time from treatment, is shown. Individuals are ordered according to their time on treatment. Clinical cure is displayed in the form of grey filled dots after the grey lines describing the time on treatment. The follow-up time was more than 30 days for the majority of patients and it can be seen that many patients die shortly after randomisation (bottom left black filled dots), probably due to their underlying disease. Censoring is rare before day 28 (few unfilled dots on the left-hand side) and, obviously, death from any cause acts as a competing event (bottom left black filled dots). Patients dying after cure are marked with a cross.

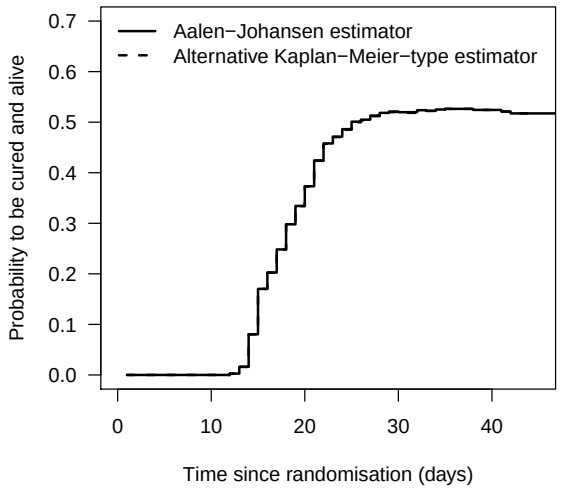

Figure 6.2: Aalen-Johansen estimator and alternative Kaplan-Meier-type estimator for the PCA function including all patients of the ceftobiprole trial [95].

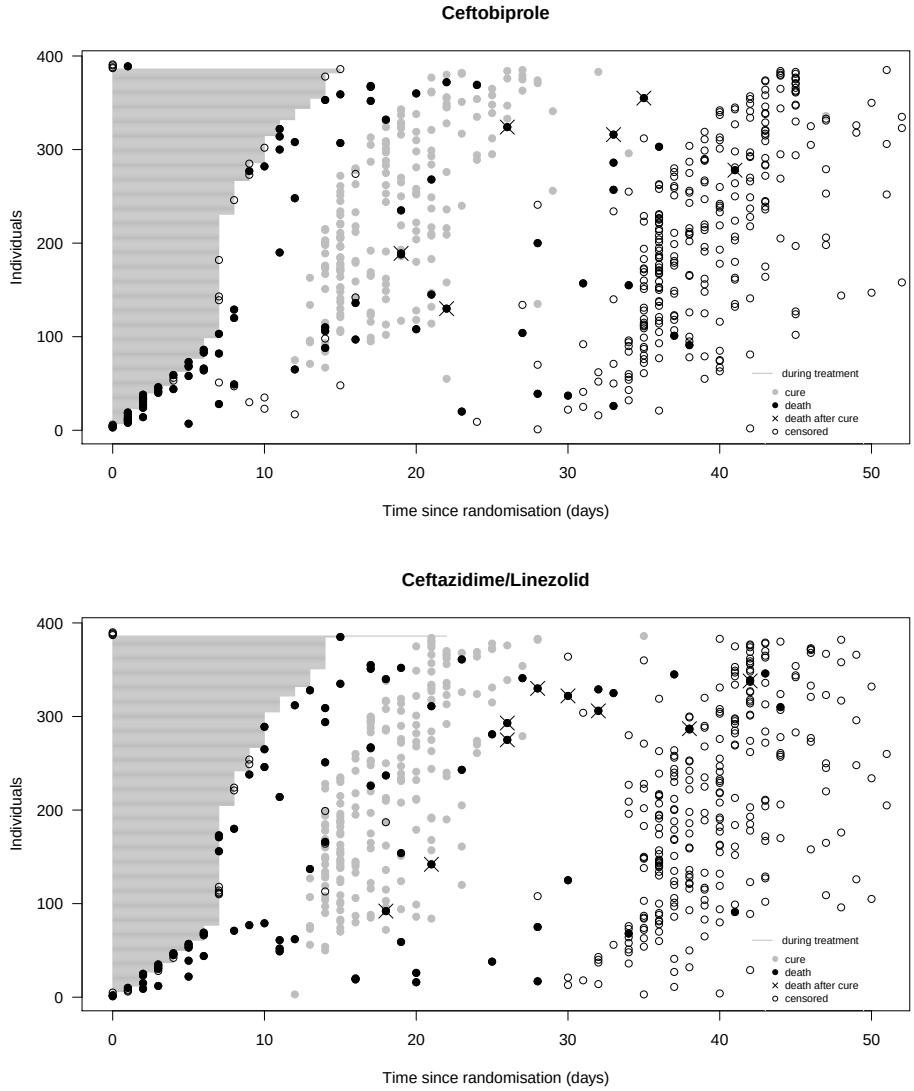

Figure 6.3: Data visualisation for the treatment groups of the ceftobiprole trial [95]. On the x-axis, time from randomisation, which equals time from treatment, is displayed. Cure at TOC is displayed in the form of grey filled dots after the grey lines describing the time on treatment. The black filled dots represent death cases, patients dying after cure are marked with a cross. Censoring can be seen via unfilled dots.

The Markov assumption

To check if the Markov assumption is fulfilled, we studied the influence of the intermediate event time, the time to cure, on the mortality transition of cured patients using a Cox proportional hazards model [84]. In other words, we kept the waiting time in state 1 in a regression model for the $1 \to 2$ hazard [41]. This model reported a non-significant coefficient ($p = 0.73$) for the time to cure. We further examined the alternative Kaplan-Meier-type estimator for the PCA function not relying on the Markov assumption that was introduced in Section 3.2. Figure 6.2 shows that both the Aalen-Johansen estimator and the alternative Kaplan-Meier-type estimator are equal such that we considered the Markov assumption to be appropriate to analyse this data.

Risk differences cured and alive

Repeating the analysis with risk differences and confidence intervals for the proportion of patients cured and alive at day 30 in comparison to risk differences and confidence intervals for the proportion of cured patients showed consistent results due to very few transitions from cure to death: The risk difference concerning only patients cured and alive at day 30 for the entire sample results in -2.43 $[-9.44, 4.58]$. In HAP excluding VAP patients, we calculated 1.51 $[-6.62, 9.63]$ and in the subset of patients with VAP -13.77 $[-25.64, -1.90]$.

Transition probability and simultaneous confidence bands

The Aalen-Johansen estimator of the probability to be cured and alive over the whole time frame of interest is displayed in the left part of Figure 6.4. For the entire sample and the HAP excluding VAP group, the probability curves show a similar course across treatment groups. In the VAP only group, there is a clear distinction between treatments, favouring ceftazidime / linezolid.

The right part of Figure 6.4 illustrates the difference in probabilities of being cured and alive together with the 95% one-sided simultaneous confidence band (dashed black line) on the time interval of interest [0, 47]. The interval is chosen such that all observed transition times are covered. The boundary values q for the construction of the respective confidence band are also displayed. It can be seen that for both the entire sample and the HAP excluding VAP group, the confidence band lies above the non-inferiority margin of -15% (grey solid line) for the entire interval [0, 47], but not for the group of VAP patients. Hence, for the entire sample and the HAP excluding VAP group, non-inferiority concerning cure and alive over the time period of interest is shown. All results do support the original analysis [95], which showed non-inferiority of overall cure proportions for both the entire sample and the HAP excluding VAP group.

Pseudo-value regression

For the pseudo-value regression, a comparable non-inferiority margin to -15% in the sense of a CRR would be 0.7, calculated as in Section 5. We investigated the effect over the whole time frame including ten times equally distributed ($s_k = \{12, 15, 18, \ldots, 39\}$) and at day 30 ($s_k = 30$). For the subgroup of patients with HAP, non-inferiority can be shown. It results in CRR = 0.92 [0.80, 1.05] (whole time frame) and CRR = 0.92 [0.80, 1.06] (day 30). Also for HAP excluding VAP since we obtain CRR = 1.00 [0.87, 1.15] (whole time frame) and CRR = 1.00 [0.88, 1.15] (day 30). For the subgroup of patients with VAP, non-inferiority is not demonstrated because CRR = 0.55 [0.35, 0.87] (whole time frame) and CRR = 0.58 [0.38, 0.89] (day 30).

General and restricted log-rank-based test

The general log-rank-based test gives a non-significant difference between the treatments for all patients ($p = 0.75$) and the subset of patients with HAP excluding VAP ($p =$

0.90). For VAP patients the analysis does indicate some, albeit weak evidence for a treatment difference ($p = 0.06$). The restricted log-rank-based test gives a non-significant difference between the treatments for all patients ($p = 0.25$) and the subset of patients with HAP excluding VAP ($p = 0.91$) and a significant difference for VAP patients ($p = 0.01$).

In order to only examine the overall probability to die we used 1 minus Kaplan-Meier estimator plots, as can be seen in Figure 6.5. Only in VAP patients, a treatment difference can be seen. We performed a simple log-rank test, yielding a p-value of 0.62 for the whole group, $p = 0.54$ for the subset of patients with HAP excluding VAP, and $p = 0.07$ for the VAP patients.

6.1.3 Discussion

Our analysis illustrates the advantageous feature of the cure-death model. Here, it does not only confirm non-inferiority of ceftobiprole as found in Awad et al. [95], but provides a stronger and more patient-benefiting non-inferiority result: First, the endpoint "get cured and stay alive over time" is a highly relevant outcome in the context of antimicrobial trial data because patients only benefit from cure when staying alive for a certain time. Second, we demonstrate non-inferiority of being cured and alive over the complete treatment process and not only at the end of follow-up as intended in the original analysis. Third, in contrast to the traditional proportion comparisons, the comprehensive cure-death multistate model captures the complex timing of cure and death, competing events, and also allows for different follow-up times due to death or right-censoring. Moreover, the Markov assumption, that is required for the Aalen-Johansen estimator, was appropriate to analyse the data.

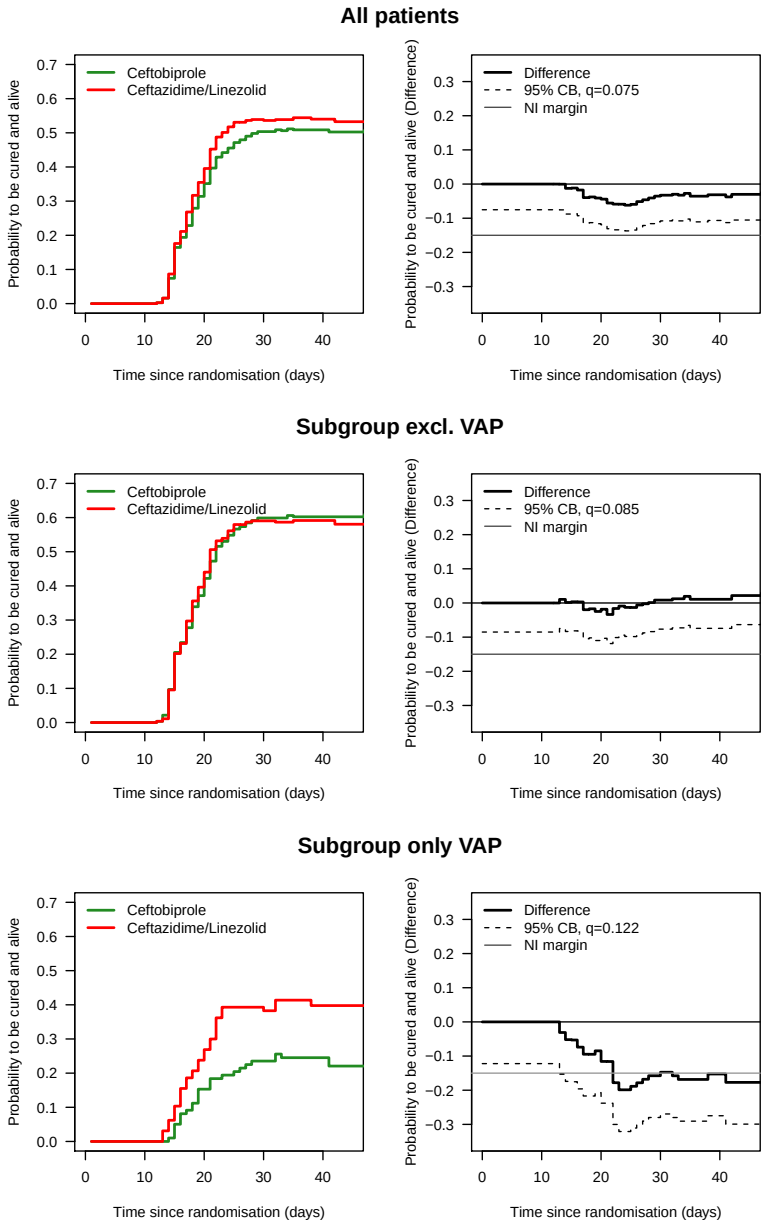

Figure 6.4: Transition probabilities derived from the Aalen-Johansen estimator for subgroups in the ceftobiprole trial [95]. Left: probability to be cured and alive. Right: estimated difference of probabilities with 95% one-sided simultaneous confidence bands (CB), corresponding boundary value q, and protocol-defined non-inferiority (NI) margin of -15% on the time interval [0, 47].

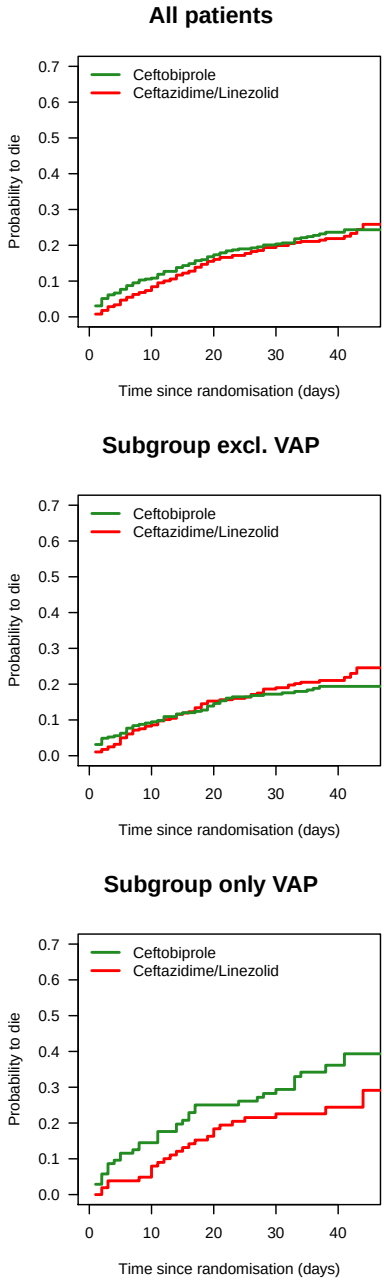

Figure 6.5: Probability to die over time using 1 minus Kaplan-Meier estimates for subgroups in the ceftobiprole trial [95].

6.2 MODIFY I+II trial

The cure-death framework can be adapted to more complex disease histories as, e.g., patients with *Clostridium difficile*, the most common cause of infectious diarrhea in hospitalised patients. Avoiding a recurrent *Clostridium difficile* infection (rCDI) is often of major interest, such as in [96]. However, the analysis of prevention effects on rCDI has two challenges. First, infected patients need to be initially cured before acquiring an rCDI. Second, patients might die during follow-up; thus, death is a classical competing event for cure and rCDI. Moreover, it is a highly relevant interest from the patients' perspective how an active treatment performs over the complete cure process. Therefore, we strongly recommend the time-dependent endpoint "clinical cure, free of rCDI, and alive over time" as key secondary endpoint of interest by applying the cure-death model. This represents a much stronger endpoint than comparing sustained cure (clinical cure and no rCDI) proportions at the end of follow-up [3].

A suitable multistate model is able to account for the time-dynamic pattern of CDI cure, death, and rCDI. We will use an extended cure-death multistate model with an initial infusion state, a cure state, an rCDI state, and competing risks states as in Figure 6.6. According to the study protocol, all patients start in state 0, that is infusion, immediately after randomisation to a treatment. The timescale of interest is "time since infusion" in weeks.

6.2.1 The trial

In a recent article, Wilcox et al. [96] present results of the **MO**NOCOLONAL ANTIBO**DI**ES **F**OR C DIFFICILE THERAP**Y** (MODIFY) I and II trial, examining the safety and efficacy of actoxumab and bezlotoxumab. These were two double-blind, randomised, placebo-controlled, multicentre phase III trials conducted at 322 sites in 30

countries from 2011 to 2015. In MODIFY I and II, a total of 2655 patients randomised and 2580 being treated for *Clostridium difficile* infection received a standard treatment in combination with either actoxumab / bezlotoxumab (779 patients), bezlotoxumab (788 patients), actoxumab (235 patients), or placebo (778 patients). A total of 2560 patients were included in the efficacy analyses, that is the modified intention-to-treat (mITT) population, namely actoxumab / bezlotoxumab (773 patients), bezlotoxumab (781 patients), actoxumab (232 patients), and placebo (773 patients). Actoxumab was not evaluated anymore in MODIFY II since earlier results indicated a lack of efficacy [133] that is why we focus on three treatment groups, actoxumab / bezlotoxumab, bezlotoxumab, and placebo.

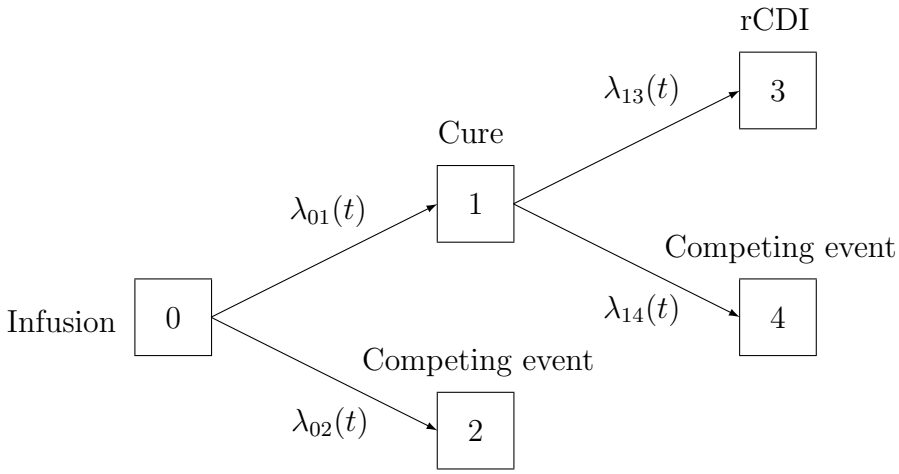

Figure 6.6: A modified cure-death multistate model with an initial infusion state, a cure state, a state of recurrent *Clostridium difficile* infection (rCDI), and competing event states (including, e.g., treatment failure and death). The direction of arrows illustrates the potential transition between the states determined by a transition hazard.

The primary endpoint was rCDI, defined as a new episode of infection after initial clinical cure within 12 weeks after infusion. Also sustained cure was analysed at the end of week 12 as secondary endpoint. Consistent results were seen in both MODIFY I and MODIFY II, demonstrating superior efficacy of bezlotoxumab and actoxumab / bezlotoxumab in the prevention of rCDI compared to placebo. The rates of initial clinical cure were 80% with bezlotoxumab, 73% with actoxumab / bezlotoxumab, and 80% with placebo and the rates of sustained cure were 64%, 58%, and 54%, respectively.

6.2.2 Reconstruction of transition rates

In accordance with the information presented in Wilcox et al. [96], we simulate an "artificial" data set under the Markov assumption with the aim to be as close as possible to the original data of the MODIFY I and II trial. No additional censoring is generated. The following information are used to reconstruct transition rates that are further utilised for simulation:

- Proportions initial clinical cure

  625/781 = 80% for bezlotoxumab

  568/773 = 73% for actoxumab / bezlotoxumab

  621/773 = 80% for placebo

- Proportions rCDI of full analysis set

  129/781 = 17% for bezlotoxumab

  119/773 = 15% for actoxumab / bezlotoxumab

  206/773 = 27% for placebo

- Proportions sustained cure (cure and no rCDI)

  $(568 - 119)/773 = 58\%$ for bezlotoxumab

  $(625 - 129)/781 = 64\%$ for actoxumab / bezlotoxumab

  $(621 - 206)/773 = 54\%$ for placebo

- Proportions rCDI using patients with cure

  $129/625 = 21\%$ for bezlotoxumab

  $119/568 = 21\%$ for actoxumab / bezlotoxumab

  $206/621 = 33\%$ for placebo

- Under risk after 12 weeks

  343 for bezlotoxumab

  301 for actoxumab / bezlotoxumab

  272 for placebo

- Treatment failure (inicial clinical cure not reached)

  $(781 - 625)/781 = 20\%$ for bezlotoxumab

  $(773 - 568)/773 = 27\%$ for actoxumab / bezlotoxumab

  $(773 - 621)/773 = 20\%$ for placebo

Transition rates are calculated as number of patients divided by number of patient-days at risk. For transition $0 \to 1$ and $0 \to 2$ we take $7 + 2$ days as mean duration at risk since cure was recorded within $14 + 2$ days. For transition 13 we take $12 \times 7 - (7 + 2)$ days as mean duration at risk since most recurrences occurred early after infusion. For transition 14 we take $10 \times 7 - (7 + 2)$ days as mean duration at risk. It results in the following rates:

- Bezlotoxumab

$$\lambda_{01} = 625/(781 \times (7+2)) = 0.089$$

$$\lambda_{02} = (781-625)/(781 \times (7+2)) = 0.022$$

$$\lambda_{13} = (625-129-343)/(625 \times (12 \times 7-(7+2))) = 0.003$$

$$\lambda_{14} = 129/(625 \times (10 \times 7-(7+2))) = 0.003$$

- Actoxumab / bezlotoxumab

$$\lambda_{01} = 568/(773 \times (7+2)) = 0.082$$

$$\lambda_{02} = (773-568)/(773 \times (7+2)) = 0.029$$

$$\lambda_{13} = (568-119-301)/(568 \times (12 \times 7-(7+2))) = 0.003$$

$$\lambda_{14} = 119/(568 \times (10 \times 7-(7+2))) = 0.003$$

- Placebo

$$\lambda_{01} = 621/(773 \times (7+2)) = 0.089$$

$$\lambda_{02} = (773-621)/(773 \times (7+2)) = 0.021$$

$$\lambda_{13} = (621-206-272)/(621 \times (12 \times 7-(7+2))) = 0.003$$

$$\lambda_{14} = 206/(621 \times (10 \times 7-(7+2))) = 0.005$$

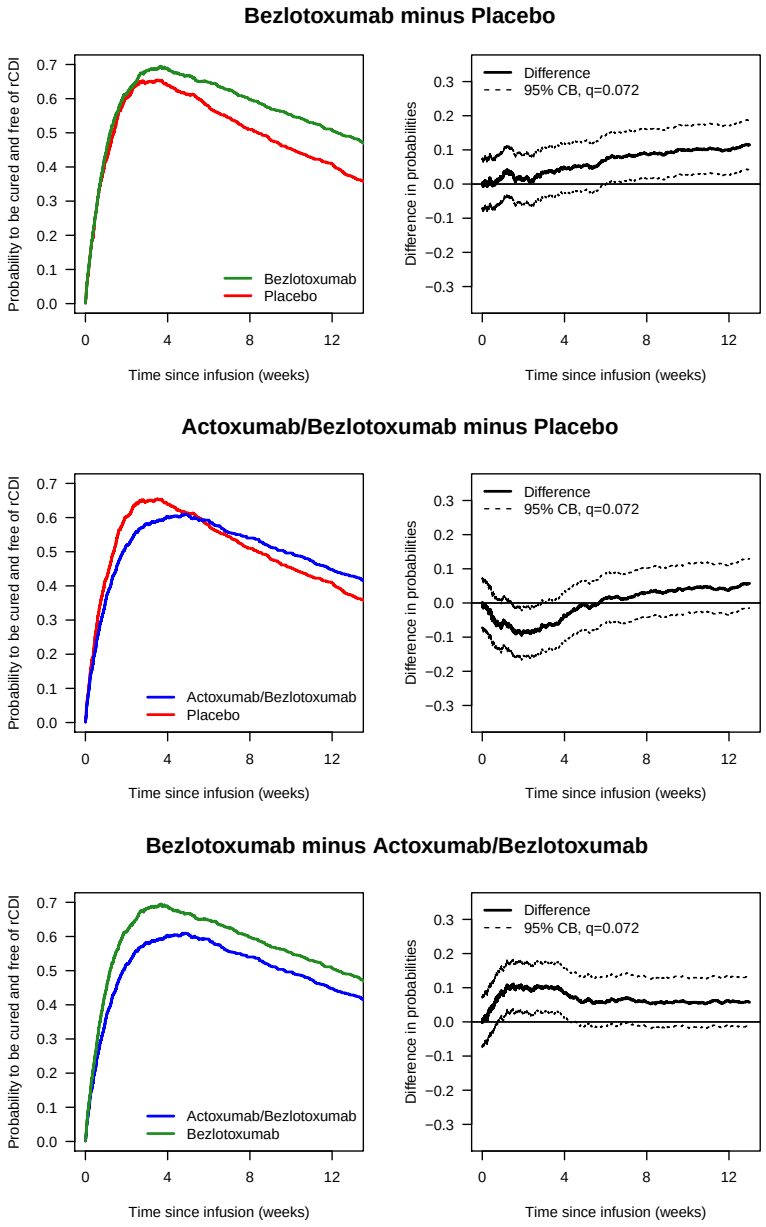

Figure 6.7: Transition probabilities derived from the Aalen-Johansen estimator for the comparison of treatment groups in the MODIFY I and II trial. Left: probability to be cured and free of recurrent *Clostridium difficile* infection (rCDI). Right: estimated difference of probabilities with 95% two-sided simultaneous confidence bands (CB).

6.2.3 Results

Transition probability and simultaneous confidence bands

The left part of Figure 6.7 illustrates the probabilities of being cured, alive, and free of rCDI over the time interval of interest until week 12, that is the estimated transition probability $\hat{P}_{01}(0,t)$ in Figure 6.6. The right part illustrates the difference in probabilities together with the 95% one-sided simultaneous confidence band (dashed black line). Comparing bezlotoxumab and placebo, the probability of being cured, alive, and free of rCDI is significantly higher for bezlotoxumab from week 6 on. Comparing actoxumab / bezlotoxumab and placebo, this probability remains higher in the placebo group for approximately the first five weeks, significantly around week 2 and 3. Bezlotoxumab performs considerably better than actoxumab / bezlotoxumab for the entire time period, significantly around week 1 until 4.

Pseudo-value regression

For the pseudo-value regression, we investigated the effect over the whole time frame including ten times equally distributed (day $s_k = \{4, 14, 24, \ldots, 84\}$) and at week 12 (day $s_k = 84$). For the comparison of bezlotoxumab with placebo, this technique yields a significant difference CRR = 1.12 [1.05, 1.21] (whole time frame) and CRR = 1.24 [1.12, 1.38] (week 12), also for the comparison of bezlotoxumab with actoxumab / bezlotoxumab, CRR = 1.13 [1.05, 1.22] (whole time frame) and CRR = 1.14 [1.03, 1.25] (week 12). For the comparison of actoxumab / bezlotoxumab with placebo, it results in CRR = 0.99 [0.92, 1.07] (whole time frame) and CRR = 1.09 [0.98, 1.22] (week 12).

General and restricted log-rank-based test

For the comparison of bezlotoxumab with placebo, the general log-rank-based test gives a significant difference ($p < 0.01$) and the restricted log-rank-based test as well ($p < 0.01$), also for the comparison of bezlotoxumab with actoxumab / bezlotoxumab. For the comparison of actoxumab / bezlotoxumab with placebo, the general log-rank-based test gives a significant difference ($p < 0.01$) but the restricted log-rank-based not ($p = 0.66$). This is due to the fact that the probability curves of interest cross, see Figure 6.7 (middle left).

6.2.4 Considerations about a suitable estimand for rCDI prevention

We noticed that the proportions of patients experiencing the primary endpoint rCDI, given in, e.g., Figure 1 in [96], do not coincide with the cumulative risk of rCDI estimated by the Kaplan-Meier method in Figure 2 in [96]. Wilcox et al. censored patients at the date of medication infusion who fail to achieve clinical cure. This results in a classical competing risks bias, unfortunately still common in leading medical journals [90, 134, 135]. Experiencing, e.g., death as competing event precludes the occurrence of clinical cure and therefore also the occurrence of rCDI. Conditioning on the future violates the first principle in time-to-event analysis and ignoring competing events leads to an overestimated cumulative rCDI risk, a biased result. Here, an appropriate analysis can only be ensured based on a multistate model as in Figure 6.6.

The impact of bezlotoxumab is thought to be solely on preventing rCDI which is why this was the focus for the primary endpoint in [96]. Since it is not an anitibiotic and cannot exert its pharmacologic effect in a setting where no toxin is present, it is clinically not expected to impact the proportion of patients who attain cure of the initial CDI. However, bezlotoxumab could have theoretically an indirect impact on mortality as well besides the prevention of rCDI. This has to be taken into account since a possibly

differential effect by chance on competing risks (clinical failure and / or mortality) can potentially mask the treatment difference for reducing rCDI. For this, several marginal and conditional probability functions are possible to examine for assessing the effect of a treatment on the probability of requiring an rCDI.

Marginal probability functions

The probability for an rCDI among all patients is the transition probability to go from state 0 (infusion) to state 3 (rCDI), that is

$$P_{03}(t),$$

estimated by $\frac{\#\text{ rCDI}}{\#\text{ mITT}}$ at τ. A possibly differential effect by chance on the competing risk transition 02 or 04 can potentially mask the treatment difference for reducing rCDI when only the 03 transition is considered. As a consequence, one must consider all marginal probability curves simultaneously to interpret them appropriately and to get a complete picture. Here, $P_{02}(t)$ and $P_{04}(t)$, the probabilities for competing events before and after cure have to be taken into account. When analysing the probability for an rCDI among patients with cure, that is the transition probability to go from state 1 (cure) to state 3 (rCDI),

$$P_{13}(t),$$

estimated by $\frac{\#\text{ rCDI}}{\#\text{ cure}}$ at τ, $P_{14}(t)$, the probabilities for a competing event after cure has to be taken into account.

Using these *marginal* probability functions, several probability curves have to be examined for a comprehensive picture of a treatment effect, even if the focus lies on one event type only.

Conditional probability functions

An alternative possibility that enables to focus on the probability of one event type without considering the incidence curve of the competing events are *conditional* probability functions [136]. They are restricted to a specific study population of interest and "share the same flavour as the Kaplan-Meier function" [136]. The difference to marginal probability functions is best illustrated with an example.

Example. Let us assume two treatment groups, T and C, 100 cured patients in each group. At the end of follow-up, 20 have an rCDI in treatment group T and 40 experience a competing event. In group C also 20 patients have an rCDI but the treatment was more toxic such that 60 out of 100 experience a competing event as, e.g., death. The probability of a recurrent infection among cured patients at the end of follow-up, $\hat{P}_{13}(\tau)$, is $\frac{20}{100} = 20\%$ for both treatment groups, although the occurrence of more competing events may have prevented the event rCDI from occuring in group C. Thus, the probability of a competing event $\hat{P}_{14}(t)$ has to be considered, where $\hat{P}_{14}(\tau) = \frac{40}{100} = 40\%$ for group T and $\hat{P}_{14}(\tau) = \frac{60}{100} = 60\%$ for group C. Alternatively, the probability of a recurrent infection among cured patients conditional to the fact that they have not experienced a competing event after cure is $\frac{20}{100-40} = \frac{20}{60} = 33\%$ for treatment T and $\frac{20}{100-60} = \frac{20}{40} = 50\%$ for treatment C, such that the actual treatment difference appears. This function is also nicely described by Pepe and Mori [136].

In our model, when considering only cured patients, the probability of a recurrent infection among cured patients conditional to the fact that they have not experienced a competing event after cure, we have

$$\frac{P_{13}(t)}{1 - P_{14}(t)} = \frac{P_{13}(t)}{P_{11}(t) + P_{13}(t)},$$

estimated by $\frac{\#\text{ rCDI}}{\#\text{ cure and no comp. event}}$ at τ. When considering all patients and the probability of a recurrent infection conditional to the fact that they have not experienced a competing event before is of interest, we have

$$\frac{P_{03}(t)}{1 - P_{02}(t) - P_{04}(t)} = \frac{P_{03}(t)}{P_{00}(t) + P_{01}(t) + P_{03}(t)},$$

estimated by $\frac{\#\text{ rCDI}}{\#\text{ cure and no comp. event}}$ at τ. The only difference is that the risk set of patients in state 1 develops over time, whereas the aforementioned probability treats these patients as left-truncated. The probability of a recurrent infection conditional to the fact that patients are cured and have not experienced a competing event before gives

$$\frac{P_{03}(t)}{1 - P_{00}(t) - P_{02}(t) - P_{04}(t)} = \frac{P_{03}(t)}{P_{01}(t) + P_{03}(t)},$$

estimated by $\frac{\#\text{ rCDI}}{\#\text{ cure and no comp. event}}$ at τ. Again, the difference is how to handle the risk set. In the latter two probabilities, $P_{00}(t)$ becomes zero over time and in the first conditional probability, $P_{00}(t)$ is assumed to be zero from time zero. It is also possible to condition only on being cured, as in Schumacher et al. [137], such that the probability for an rCDI conditional to the fact that for experiencing a recurrent infection a patient has to be cured first, gives

$$\frac{P_{03}(t)}{1 - P_{00}(t) - P_{02}(t)} = \frac{P_{03}(t)}{P_{01}(t) + P_{03}(t) + P_{04}(t)},$$

estimated by $\frac{\#\text{ rCDI}}{\#\text{ cure}}$ at τ. An overview of all probability functions considered is given in Table 6.1 and for illustration purposes, the aforementioned transition probabilities are estimated and plotted in Figure 6.8 using the simulated data of the bezlotoxumab study group.

We recommend to focus on the probability of a recurrent infection conditional to the fact that patients are cured and have not experienced a competing event before, $\frac{P_{03}(t)}{P_{01}(t)+P_{03}(t)}$,

since this may answer the initial research question best. Moreover, when no effect is present on the competing events, there may be only negligible upto no differences between the estimated estimands when comparing treatments.

6.2.5 Discussion

Most clinical trials for new treatments of CDI have used inter alia clinical cure as the primary endpoint [22] that is also recommended by the European Medicines Agency [20]. Moreover, it is a highly relevant interest from the patients' perspective how an active treatment performs over the complete cure process. When the focus lies on rCDI prevention, we strongly recommend the time-dependent endpoint "clinical cure, free of rCDI, and alive over time" as key secondary endpoint of interest that complements the use of the endpoint "sustained cure" [3]. In this example, our analysis indicates the possibility that although both active treatment groups decreased the rCDI risk, the probability of being cured, alive, and free of rCDI remains higher in the placebo group compared to actoxumab / bezlotoxumab for approximately the first five weeks.

Further, in the presence of competing events, Kaplan-Meier risk estimates as used in Wilcox et al. [96] are biased. Here, an appropriate analysis can only be ensured based on a multistate model where several possible estimands are conceivable to assess rCDI prevention.

Probability function	Interpretation	Reference population	Proportions at τ
	Marginal probability functions		
$P_{03}(t)$	Pr(rCDI)	all patients	$\frac{\#\text{ rCDI}}{\#\text{ mITT}}$
$P_{13}(t)$	Pr(rCDI)	patients with cure	$\frac{\#\text{ rCDI}}{\#\text{ cure}}$
	Conditional probability functions		
$\frac{P_{13}(t)}{1-P_{14}(t)}$ $= \frac{P_{13}(t)}{P_{11}(t)+P_{13}(t)}$	Pr(rCDI | no comp. event)	patients with cure	$\frac{\#\text{ rCDI}}{\#\text{ cure and no comp. event}}$
$\frac{P_{03}(t)}{1-P_{02}(t)-P_{04}(t)}$ $= \frac{P_{03}(t)}{P_{00}(t)+P_{01}(t)+P_{03}(t)}$	Pr(rCDI | no comp. event)	all patients	$\frac{\#\text{ rCDI}}{\#\text{ cure and no comp. event}}$
$\frac{P_{03}(t)}{1-P_{00}(t)-P_{02}(t)-P_{04}(t)}$ $= \frac{P_{03}(t)}{P_{01}(t)+P_{03}(t)}$	Pr(rCDI | cure and no comp. event)	all patients	$\frac{\#\text{ rCDI}}{\#\text{ cure and no comp. event}}$
$\frac{P_{03}(t)}{1-P_{00}(t)-P_{02}(t)}$ $= \frac{P_{03}(t)}{P_{01}(t)+P_{03}(t)+P_{04}(t)}$	Pr(rCDI | cure)	all patients	$\frac{\#\text{ rCDI}}{\#\text{ cure}}$

Table 6.1: Marginal and conditional probability functions suitable as an estimand to assess rCDI prevention, their interpretation, the study population that is used for calculation (reference population), and the values at end of follow-up (τ).

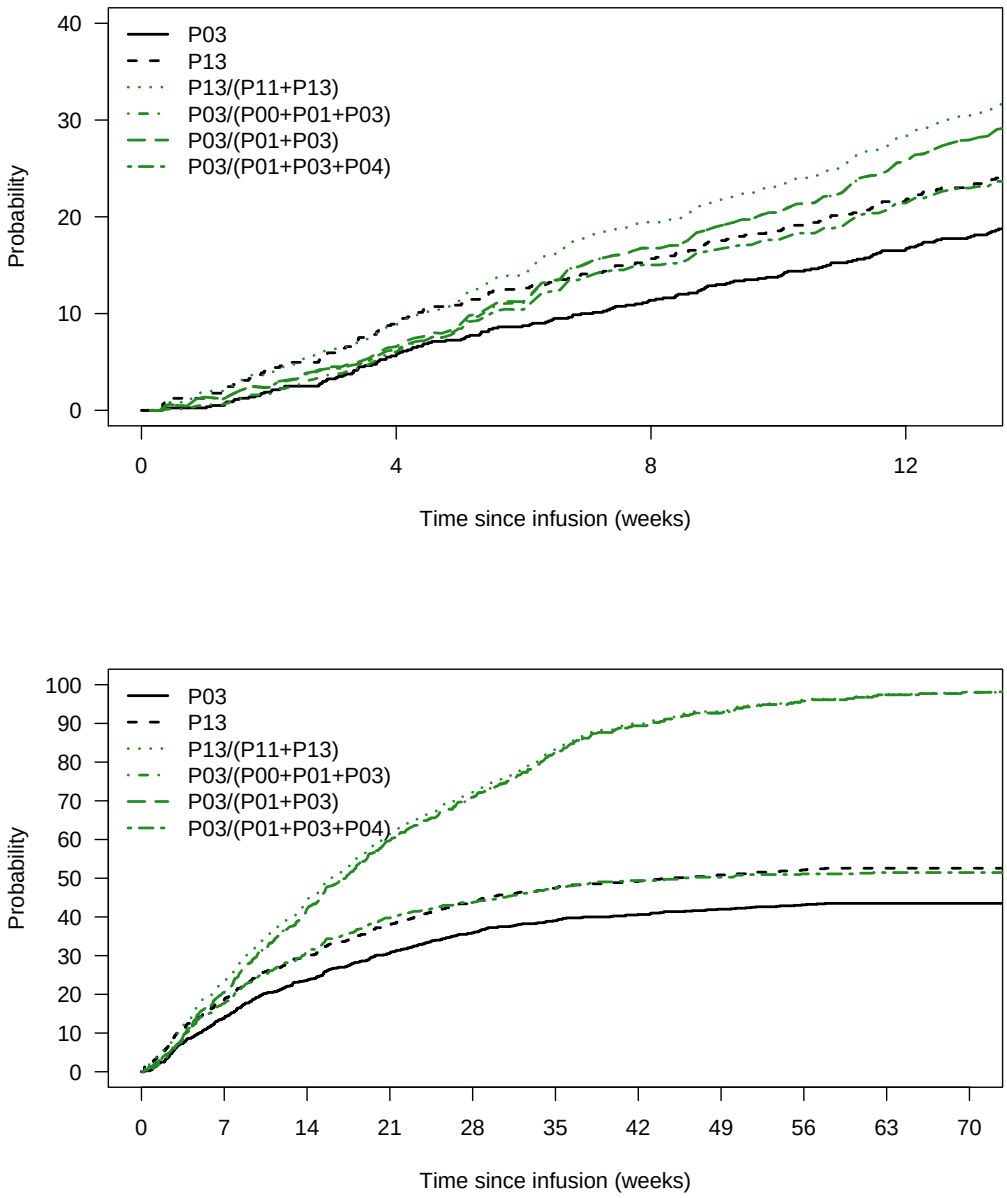

Figure 6.8: Estimated marginal and conditional probability functions for the bezlotoxumab study group to assess rCDI prevention. The lower figure shows an extended time frame to illustrate the values at end of follow-up.

6.3 OUTCOMEREA study

As mentioned in the introduction, VAP is the most common nosocomial infection amongst ventilated patients, the majority caused by the organism *Pseudomonas aeruginosa* [7]. Patients with VAP require immediate treatment but pathogens are frequently resistant to many antimicrobial agents [138]. A treatment is classified as adequate if one or more antibiotics initiated for VAP were active against the causative *Pseudomonas aeruginosa* on the basis of the antibiotic susceptibility profile of the strain [139]. However, due to limited diagnostic test opportunities, the adequacy of antimicrobial therapy can only be determined after 24 hours subsequent to the begin of treatment resulting in a considerable amount of patients that receive inadequate initial therapy. In addition, the complete absence of antimicrobial treatment is generally considered as inadequate antimicrobial treatment [140].

Impact of inadequate treatment

The extent of negative impact that inadequate immediate treatment could have is disputed. Existing research has not provided consistent evidence on whether inadequate treatment is associated, e.g., with increased mortality. A broad systematic review of general bacteraemic patients has generated conflicting findings [141]. The general assumption that infections caused by antibiotic-resistant bacteria are associated with an increased mortality rate is based on the possibility that due to limited diagnostic test opportunities adequate treatment for such infections might be initiated later than for infections caused by antibiotic-susceptible bacteria [140].

On the one hand, concerning patients suffering from a VAP, some studies have found out that delayed or inadequate treatment is associated with an increased mortality [142, 143, 144, 145, 146], others found an effect only in the subgroup of patients with a high disease severity [147]. Bloos et al. [148, 149] showed that an early recognition

followed by immediate initiation of adequate therapy is important to improve survival in septic patients, which is among the most common causes of death in hospital. Beneath patients with gram-negative bacteremia, appropriate antimicrobial therapy has been shown to reduce mortality [150] and, when initiated early, to have a favorable effect on outcome in critically ill patients with bacteremia or other serious infections [151, 152, 153, 154, 155]. Kang et al. [156] showed that inadequate immediate treatment is associated with an adverse outcome in antibiotic-resistant gram-negative bacteremia, particularly in patients with a high-risk source of bacteremia.

On the other hand, several reports have noted that inadequate immediate treatment did not result in a considerable difference in the outcomes of patients with severe infections [157, 158, 159]. No effect of inadequate treatment on mortality and discharge was found in context withb ICU-acquired *Enterobacteriaceae bacteraemia* [160]. Some studies found that initial inadequate treatment is not associated with treatment failure [139, 147, 161].

We will consider VAP patients and study the association of adequate treatment in comparison to inadequate treatment on being cured and alive making use of a multistate model. Here, we will define cure as successful extubation or discharge alive from hospital. Such a cure-death model was already proposed to gain a better understanding of how a new treatment influences the time-dynamic cure and death process in a randomised clinical trial setting [1, 12] and can be easily applied to the observational OUTCOMEREA study, using the multistate model in Figure 6.9. The outcome of interest can be expressed as

$$P_{01}(0,t) + P_{03}(0,t) = Pr(X(t) = 1 \mid X(0) = 0) + Pr(X(t) = 3 \mid X(0) = 0),$$

the probability of being extubated alive (but still in hospital) plus the probability of being discharged from hospital. It is a summary measure where all of the four relevant

transition hazards out of Figure 6.9 have influence, directly or indirectly.

Randomised controlled trials may not be feasible due to ethical considerations and thus, methods to correct for possible confounders have to be applied. The propensity score (PS) as multivariable scoring system is a suitable method to collapse several predictors for adequate treatment into a single value. Here, the PS is defined as the conditional probability of receiving adequate treatment given patients' covariates and is used to balance the distribution of possible confounders between patients with adequate and inadequate treatment. Several PS methods will be applied.

6.3.1 The study

We examined patients of the French prospective observational OUTCOMEREA research data base, where data was collected between 1997 and 2014 from 23 ICUs. Patients with VAP due to *Pseudomonas aeruginosa* were included who had received at least 48 hours of mechanical ventilation. A total of 465 patients were included where 308 received immediate adequate treatment and 157 inadequate treatment. Patients in both treatment groups were similar in baseline characteristics, ICU length of stay, day-30, and in-hospital mortality.

6.3.2 Propensity score

RCTs provide the highest level of evidence when estimating the effects of interventions on outcomes since a random treatment allocation ensures that treatment status is not confounded with patients' baseline characteristics. In observational (non-randomised) trials, treatment allocation (= exposure) is often influenced by patients' characteristics and often, as a consequence, baseline characteristics in treatment groups differ systematically.

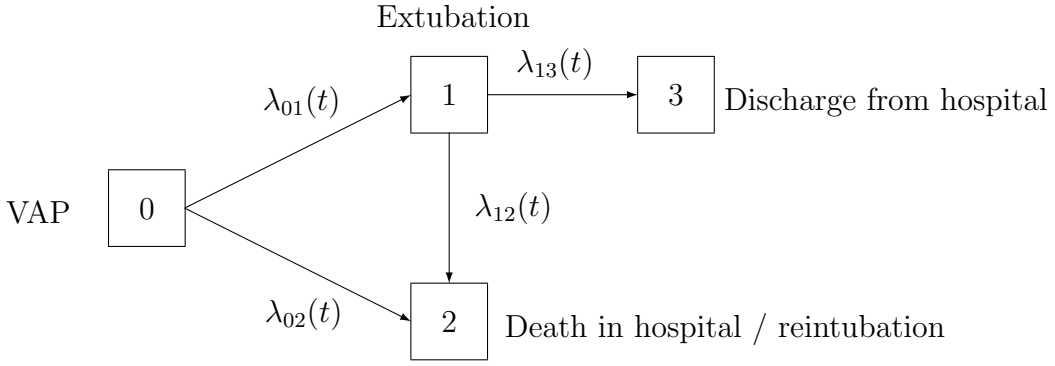

Figure 6.9: The cure-death model for comparing adequate and inadequate treatment with an initial VAP state, an extubation state, a death in hospital / reintubation state, and a discharge from hospital state. The direction of arrows illustrates the potential transition between the states determined by a transition hazard $\lambda_{01}(t)$, $\lambda_{02}(t)$, $\lambda_{12}(t)$, or $\lambda_{13}(t)$.

The concept

The propensity score (PS) aims to balance covariate distributions such that conditional on the PS, the distribution of observed baseline covariates will be similar in treatment groups [162, 163]. It is the conditional probability of receiving a certain treatment given a set of patients' pre-treatment covariates.

In a first step, the PS is estimated from the existing data, for example in a logistic regression model. The question which covariates to include to estimate the PS is often subject of debate. Some researchers suggest to include covariates associated with treatment, others advocate to focus on covariates associated with outcome, and some favour covariates associated with both treatment and outcome. Brookart et al. [164], e.g., suggest that covariates unrelated to the treatment but related to the outcome should always be included in a PS model. Austin et al. [165] showed that including covari-

ates related to the treatment but unrelated to the outcome does not improve variable balance and therefore should not be included in the PS model. Furthermore, they recommend to include all measured confounders (covariates associated with treatment *and* outcome) since, otherwise, this can lead to biased estimation of the treatment effect. But, in the same vein, they alert of using such a model as panacea for unmeasured confounders. Austin et al. [166] advice that the choice of covariates should be based on subject-matter expertise rather than formal statistical hypothesis testing in the study sample. Including the ones about which one would be concerned if baseline imbalance existed is a good choice [167, 166].

In a second step, the estimation of the treatment effect of interest follows with the aid of the PS. In doing so, four methods are available: PS matching, inverse probability of treatment weighting (IPTW), regression adjustment for the PS, or stratification according to the PS:

- PS matching: Every treated patient is assigned an untreated patient (1:1 matching) or several (1:n matching) with the same or a similar PS. The maximum permitted difference between matched subjects, also called the "caliper", is chosen as a width equal to 0.2 of the standard deviation of the logit of the propensity score, as this caliper has been shown to be optimal in a range of settings [168]. In the matched collective, the therapy effect is estimated taken into account of the matching. In practice, a univariable Cox model, e.g., is fitted, allowing the baseline hazard function to vary across matched sets [169].

- IPTW: Every patient receives as weight the reciprocal of treatment probability that belongs to its actual treatment status. A patient in the treatment group receives the weight $\frac{1}{PS}$, a patient in the control group weight $\frac{1}{1-PS}$. So, a treated patient with a low PS (for treatment) gets a high weight because his baseline

characteristics are similar of those in the untreated group and the other way round.

- Regression adjustment for the PS: A conventional regression model is used with the outcome of interest as dependent variable and the treatment allocation as well as the PS as independent variables. The influence of treatment on outcome is adjusted for the PS and therefore as well for all covariates used to estimate the PS.

- Stratification according to the PS: This is a coarsened version of PS matching. Here, the entire sample is partitioned in equal parts (for example quintiles) with respect to the estimated PS. In each of these parts the treatment effect is estimated and the overall effect is then summarised using meta-analytical methods.

Each of these methods has its advantages and disadvantages, PS matching and IPTW are considered to be the preferred procedures [169]. However, it has to be beared in mind that PS methods can only adjust for known and actually measured patients' characteristics and are not able to replace RCTs.

Variables to include

We fit a logistic regression model to predict adequate treatment assignment as a function of baseline covariates and had a look at the CSHRs for different transitions $0 \to 1$, $0 \to 2$, $1 \to 2$, and $1 \to 3$ adjusted for these covariates. We select variables based on their association with treatment and outcome or only outcome using an inclusion criterion of $p \leq 0.157$ for at least one of the CSHRs or treatment from the univariable analysis [170], stratified by center. This reference value corresponds to the well-established Akaike information criterion for model selection [171]. It results in including the following variables into the PS model:

- Sex
- Age
- Resistant VAP
- SOFA score
- Multiple antibiotics
- Smoking
- Alcohol abuse
- Substance abuse
- Dialysis
- Sepsis
- Days from admission to VAP
- Days from mechanical ventilation to VAP

Subject-matter experts advise to include variables like “resistant VAP”, “multiple antibiotics”, and “SOFA score” that are among the variables above. The PS distribution can be seen in Figure 6.10.

An important issue is to check whether baseline characteristics differ substantially between adequately and inadequately treated patients in the propensity score matched sample in comparison to in the original sample. In Table 6.2, we examined these variables using standardised differences. Acceptable balance has been reached as systematic differences between adequately and inadequately treated subjects in the original sample concerning prognostically important variables (as, e.g., sex, SOFA score, multiple antibiotics) have been substantially reduced or eliminated in the matched sample.

	Unmatched sample $n = 465$			**Matched sample** $n = 335$		
Baseline covariate	Inadequate $n = 157$	Adequate $n = 308$	SDiff	Inadequate $n = 112$	Adequate $n = 223$	SDiff
Male (%)	107 (68.2)	226 (73.4)	0.12	77 (68.8)	158 (71.2)	0.05
Age (SD)	63.40 (16.57)	63.19 (14.98)	0.01	63.42 (16.30)	63.98 (15.14)	0.04
Resistant VAP (%)	58 (36.9)	95 (30.8)	0.13	50 (44.6)	85 (38.3)	0.13
Medical (%)	26 (16.6)	72 (23.4)	0.17	19 (17.0)	52 (23.4)	0.16
SOFA (SD)	5.80 (3.70)	6.39 (3.82)	0.16	6.14 (3.69)	6.33 (3.98)	0.05
Sepsis	152 (96.8)	298 (96.8)	0.00	108 (96.4)	214 (96.4)	0.00
Diabetes	18 (11.5)	47 (15.3)	0.11	8 (7.1)	36 (16.2)	0.28
Multiple antibiotics (%)	95 (60.5)	275 (89.3)	0.70	93 (83.0)	189 (85.1)	0.06
Smoking (%)	46 (29.3)	91 (29.5)	0.01	32 (28.6)	60 (27.0)	0.03
Alcohol (%)	29 (18.5)	63 (20.5)	0.05	21 (18.8)	41 (18.5)	0.01
Substance abuse (%)	5 (3.2)	2 (0.6)	0.19	2 (1.8)	2 (0.9)	0.08
BMI (SD)	26.23 (6.97)	25.87 (6.52)	0.05	26.08 (7.40)	25.75 (5.88)	0.05
Dialysis (%)	10 (6.4)	25 (8.1)	0.07	6 (5.4)	16 (7.2)	0.08
Days from admission to VAP (SD)	21.51 (22.21)	20.54 (19.93)	0.05	21.21 (23.00)	21.34 (21.50)	0.01
Days from mechanical ventilation to VAP (SD)	13.84 (14.03)	13.57 (11.53)	0.02	13.02 (14.16)	13.36 (11.06)	0.03

Table 6.2: Comparison of baseline characteristics between adequately and inadequately treated patients in the original sample and in the PS matched sample. SD = Standard deviation, SDiff = standardised difference.

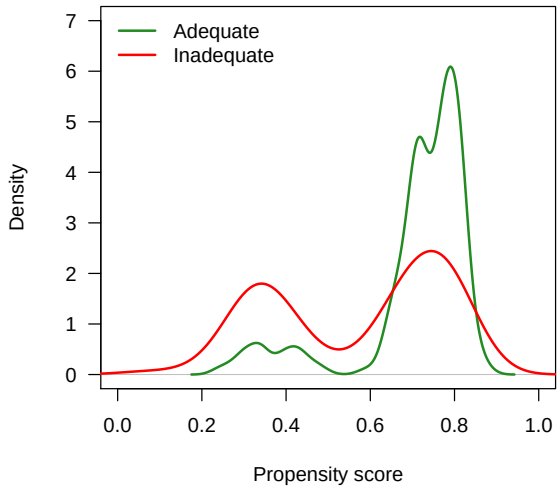

Figure 6.10: The propensity score distribution for adequate and inadequate treatment.

6.3.3 Results

First of all, transition frequencies for adequately and inadequately treated patients can be found in Table 6.3. The data visualisation in Figure 6.12 provides an illustration of the time course of events for adequate and inadequate treatment.

The Markov assumption

To check if the Markov assumption is fulfilled, we studied the influence of the time to cure, the $0 \to 1$ transition, on the $1 \to 2$ reintubation / mortality transition and on the $1 \to 3$ discharge transition by including it as a time-dependent variable in a Cox model for the $1 \to 2$ and $1 \to 3$ hazard, respectively. The model for the discharge transition reported a non-significant coefficient ($p = 0.65$) for the time to cure. Although it gives a significant coefficient ($p = 0.02$) regarding the reintubation / mortality transition, a

hazard ratio of 0.98 is very close to 1 and thus, we consider the Markov assumption to analyse these data [172]. Moreover, if censoring is independent, the Nelson-Aalen estimator is still an appropriate estimator for the cause-specific cumulative hazards and the Aalen-Johansen estimator for the state occupation probabilities is consistent even in the absence of the Markov property [173, 174]. The Aalen-Johansen estimator in general is less sensitive to violations of the Markov assumption as originally thought [175, 176]. We further examined the alternative Kaplan-Meier-type estimator for the probability of being extubated alive (but still in hospital) not relying on the Markov assumption that was introduced in Section 3.2. Figure 6.11 shows that the Aalen-Johansen estimator is similar to the alternative Kaplan-Meier-type estimator. Because information about the $3 \rightarrow 2$ transition is incomplete (incomplete mortality follow-up for death after discharge), it is not possible to apply this alternative estimator for the probability of being discharged from hospital.

Cause-specific hazards and Cox regression

Univariable Cox regression (all regression results are given in Table 6.4 and Table 6.5) showed an increased risk of 1.65 [0.95, 2.86] in the transition from cure to death ($1 \rightarrow 2$) for patients with adequate treatment, with albeit weak evidence against the null hypothesis ($p = 0.076$). The hazard ratio decreased with the adjustment for covariates in a multivariable regression model, also for the $0 \rightarrow 2$ transition. A slight increase can be seen in the 01 transition by adjusting for covariates in the multivariable regression. With the application of PS methods, the PS matching and PS IPTW yields a significant result for this $1 \rightarrow 2$ transition. A difference in the $1 \rightarrow 2$ transition can also be seen having a look at the cumulative cause-specific hazards in Figure 6.13, here also for the transition from cure to discharge ($1 \rightarrow 3$). For all other transitions, no evidence was found for a difference between adequate and inadequate treatment.

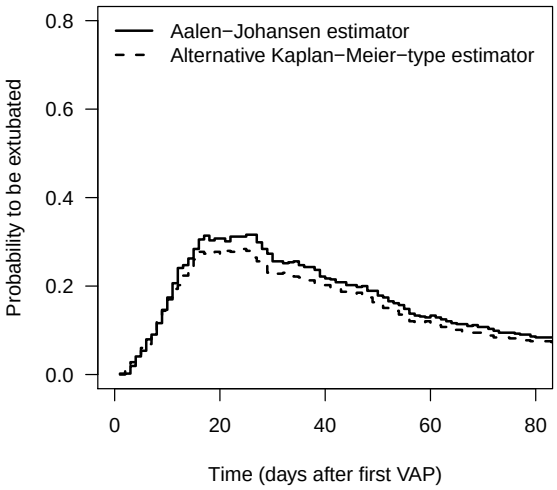

Figure 6.11: Aalen-Johansen estimator and alternative Kaplan-Meier-type estimator for the probability of being extubated alive (but still in hospital) including all patients of the OUTCOMEREA study.

Transition probability and simultaneous confidence bands

No evidence was found for a difference between adequate and inadequate treatment, neither in the 01 transition and 03 transition in Figure 6.14, nor in the sum of these transitions, using simultaneous confidence bands for the difference in Figure 6.15.

Pseudo-value regression

For the transition probability of interest, we investigated the effect over the whole time frame using ten times equally distributed ($s_k = \{1, 11, 21, \ldots, 61\}$), and at day 20, 40, and 60 ($s_k = 20$, 40, 60), in a univariable, multivariable, and several PS adjusted analyses (see Table 6.4 and Table 6.5). Again, no evidence was found for a difference between adequate and inadequate treatment.

General and restricted log-rank-based test

The general log-rank-based test gives a non-significant difference between adequate and inadequate treatment ($p = 0.65$), the restricted log-rank-based test as well ($p = 0.63$).

6.3.4 Post hoc analyses and discussion

No evidence was found for an unfavourable effect of inadequate treatment on being extubated or discharged alive. There may be a marginal effect on the transition from extubation to reintubation / death, to the disadvantage of adequate treatment, that has to be given further consideration. These counter-intuitive results of a comparable performance of immediately adequate and inadequate treatment encouraged to perform subsequent post hoc analyses to be able to explain the surprising equivalent effect of inadequate treatment.

For this, in order to have a deeper look into which covariates have an influence on treatment, we used a classification and regression tree (CART) procedure. CART is one popular way to form trees from data, where the aim is to divide patients into subgroups defined by patients' characteristics in terms of their covariate values [177, 178]. The concept of a binary tree-structure is sometimes called "recursive partitioning" and was first proposed by Morgan and Sonquist [179]. The split yielding the maximal test statistic, which represents the greatest possible separation of the patients with respect to the outcome variable, is performed. The subgroups should be disparate and internally homogeneous.

We split the whole sample according to the variable with the highest impact on treatment, that is multiple antibiotics. One further split is done according to diabetes in the subset of patients with multiple antibiotics. As can be seen in Figure 6.16, in the smaller group with no multiple antibiotics, the majority, 74% of patients, are classified into the inadequate group, although in total, only 34% belong to the inadequate group.

Also, these patients, with a mean SOFA score of 4.9 (median 4), are healthier than the larger group with multiple antibiotics, with a mean SOFA score of 6.5 (median 6).

In an ICU setting, critically ill patients have to be treated immediately. For patients that are not critically ill, doctors often wait a couple of days before a treatment is given. In fact, during this phase, these patients are allocated to the inadequate treatment group. Thus, 62 out of 157 patients (39%) with no multiple antibiotics and a lower SOFA score are present in the inadequate treatment group. In other words, almost half of the patients in the inadequate treatment group are patients with a considerably better health condition. This imbalance could be a possible explanation for the fact that no evidence was found for an effect of adequate and inadequate treatment on being extubated or discharged alive. As a consequence, reflections should take place on whether the definition of adequate treatment or rather the assignment to the adequate treatment group is appropriate or whether a third category of "delayed / no treatment" would be reasonable.

	Total	Adequate	Inadequate
$0 \rightarrow 1$	303	201	102
$0 \rightarrow 2$	162	107	55
$1 \rightarrow 2$	66	49	17
$1 \rightarrow 3$	237	152	85

Table 6.3: Comparison of transition frequencies between adequately and inadequately treated patients.

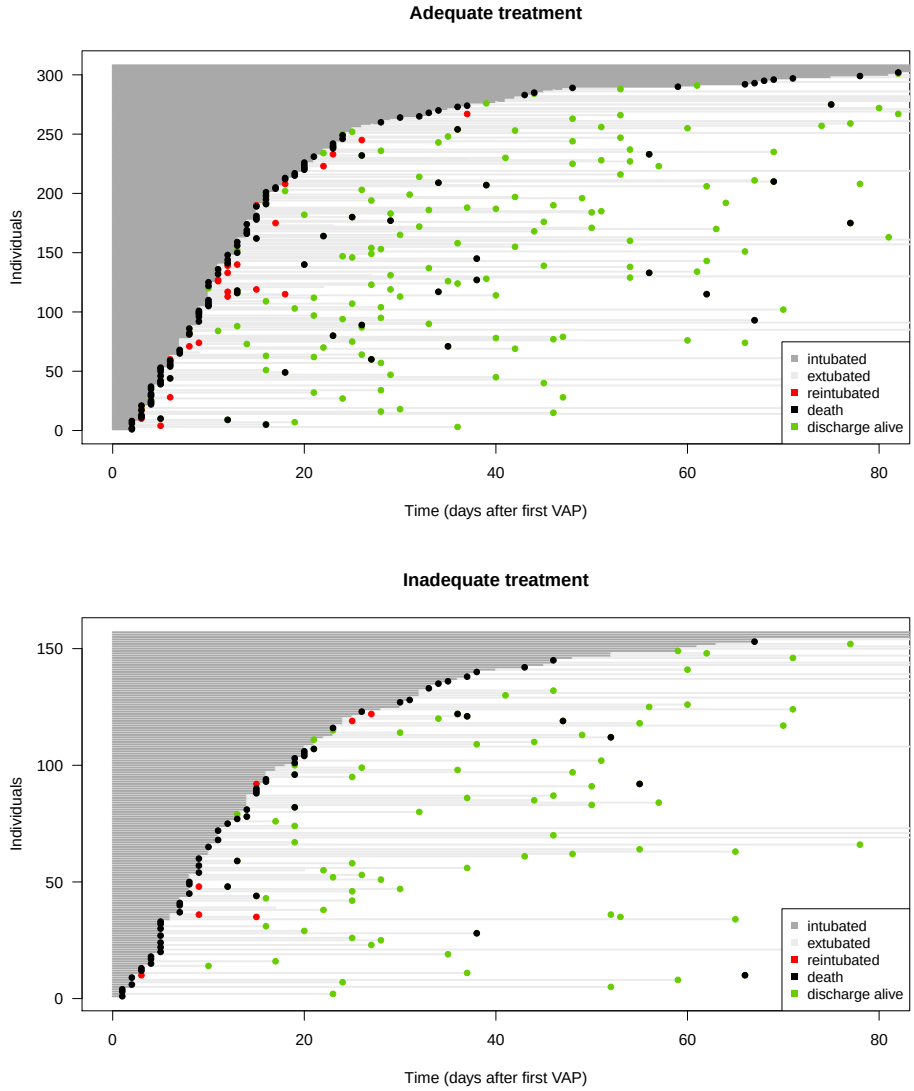

Figure 6.12: Data visualisation for adequate and inadequate treatment. On the x-axis, time from first ventilator-associated pneumonia (VAP) is given. Being extubated is displayed in the form of light grey lines after dark grey lines describing intubation. The black filled dots represent death cases, patients reintubated after cure are marked with a red dot. Patients discharged alive are marked with green dots.

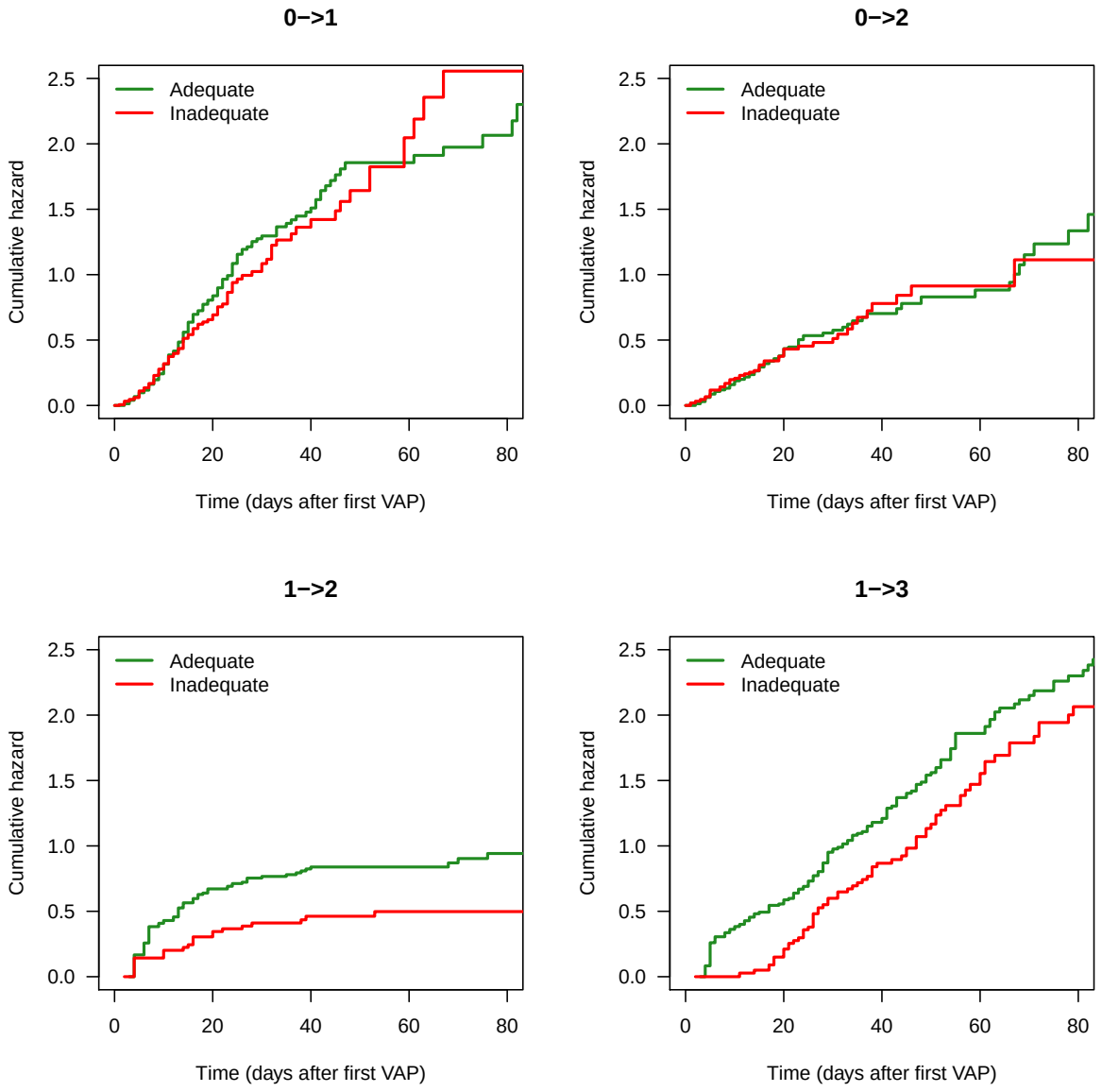

Figure 6.13: Cumulative hazards given by the Nelson-Aalen estimator for the OUTCOMEREA data. Transition to extubation ($0 \to 1$), to death ($0 \to 2$), from extubation to reintubation / death ($1 \to 2$), and from extubation to discharge ($1 \to 3$) are displayed.

Regression type	UNI **exp(coef)**	MULTI **exp(coef)**
CSHR		
$0 \to 1$	1.04 [0.82, 1.32]	1.12 [0.85, 1.47]
$0 \to 2$	1.02 [0.74, 1.42]	0.94 [0.65, 1.35]
$1 \to 2$	1.65 [0.95, 2.86]	1.49 [0.83, 2.67]
$1 \to 3$	0.98 [0.75, 1.28]	0.93 [0.69, 1.26]
CRR		
whole time frame	1.00 [0.82 1.23]	0.96 [0.80 1.17]
around $t = 20$	1.08 [0.85 1.38]	1.00 [0.79 1.26]
around $t = 40$	0.98 [0.80 1.19]	0.94 [0.77 1.14]
around $t = 60$	0.95 [0.79 1.15]	0.91 [0.76 1.09]

Table 6.4: Regression results without propensity score methods. The respective effect measure is given with 95% confidence interval. CSHR = Cause-specific hazard ratio, CRR = Cure risk ratio.

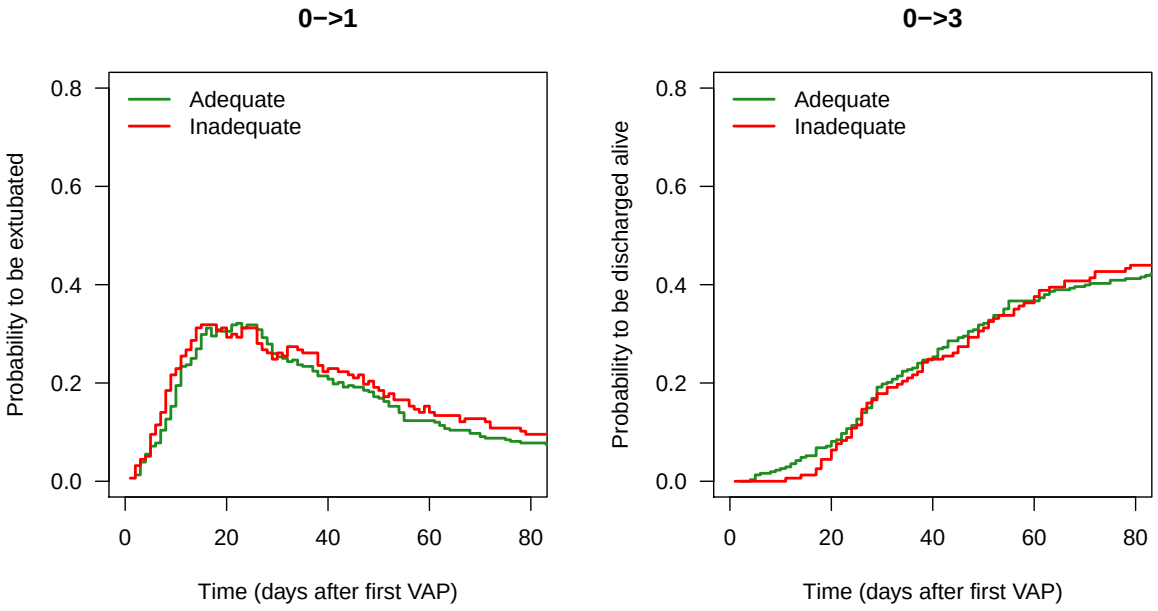

Figure 6.14: Transition probabilities derived from the Aalen-Johansen estimator for the OUTCOMEREA data. Left: probability to be extubated (stay extubated but still in hospital, that is the state occupation probability for state 1). Right: probability to be discharged alive (that is the state occupation probability for state 3).

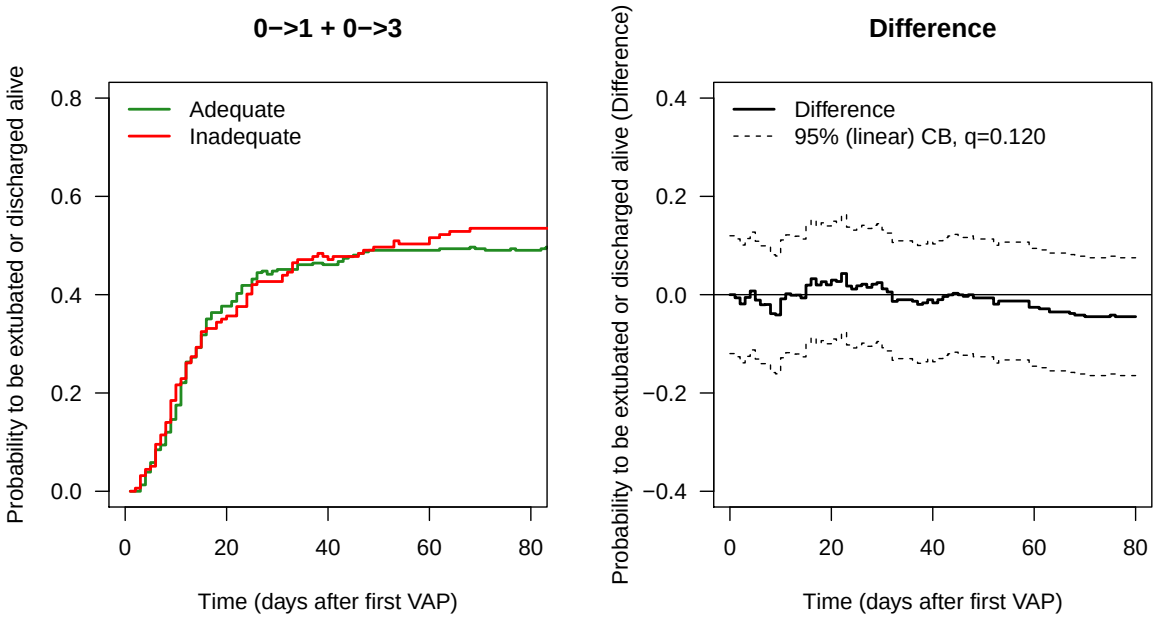

Figure 6.15: Transition probabilities derived from the Aalen-Johansen estimator for the OUTCOMEREA data. Left: probability to be extubated and / or discharged alive. Right: estimated difference of probabilities with 95% two-sided simultaneous confidence band (CB) and corresponding boundary value q.

Regression type	PS matching **exp(coef)**	PS IPTW **exp(coef)**	PS reg adj **exp(coef)**	PS strat **exp(coef)**
CSHR				
$0 \to 1$	1.04 [0.78, 1.39]	1.01 [0.86, 1.18]	1.02 [0.79, 1.33]	1.04 [0.81, 1.35]
$0 \to 2$	0.92 [0.63, 1.35]	1.01 [0.80, 1.26]	0.97 [0.69, 1.39]	0.93 [0.65, 1.34]
$1 \to 2$	**2.26 [1.02, 5.01]**	**1.55 [1.09, 2.22]**	1.56 [0.87, 2.80]	1.24 [0.68, 2.29]
$1 \to 3$	1.30 [0.89, 1.89]	0.89 [0.75, 1.07]	0.91 [0.68, 1.23]	0.88 [0.66, 1.18]
CRR				
whole time frame	1.00 [0.82, 1.23]	0.99 [0.80, 1.23]	1.00 [0.81, 1.25]	1.47 [0.48, 4.54]
around $t = 20$	1.08 [0.84 1.37]	1.04 [0.80, 1.35]	1.06 [0.82, 1.38]	0.98 [0.34, 2.84]
around $t = 40$	0.98 [0.80„ 1.19]	0.95 [0.77, 1.19]	0.97 [0.78, 1.20]	0.76 [0.32, 1.85]
around $t = 60$	0.95 [0.79, 1.15]	0.92 [0.76, 1.13]	0.94 [0.77, 1.15]	0.76 [0.36, 1.64]

Table 6.5: Regression results using different propensity score adjustments. The respective effect measure is given with 95% confidence interval. CSHR = Cause-specific hazard ratio, CRR = Cure risk ratio.

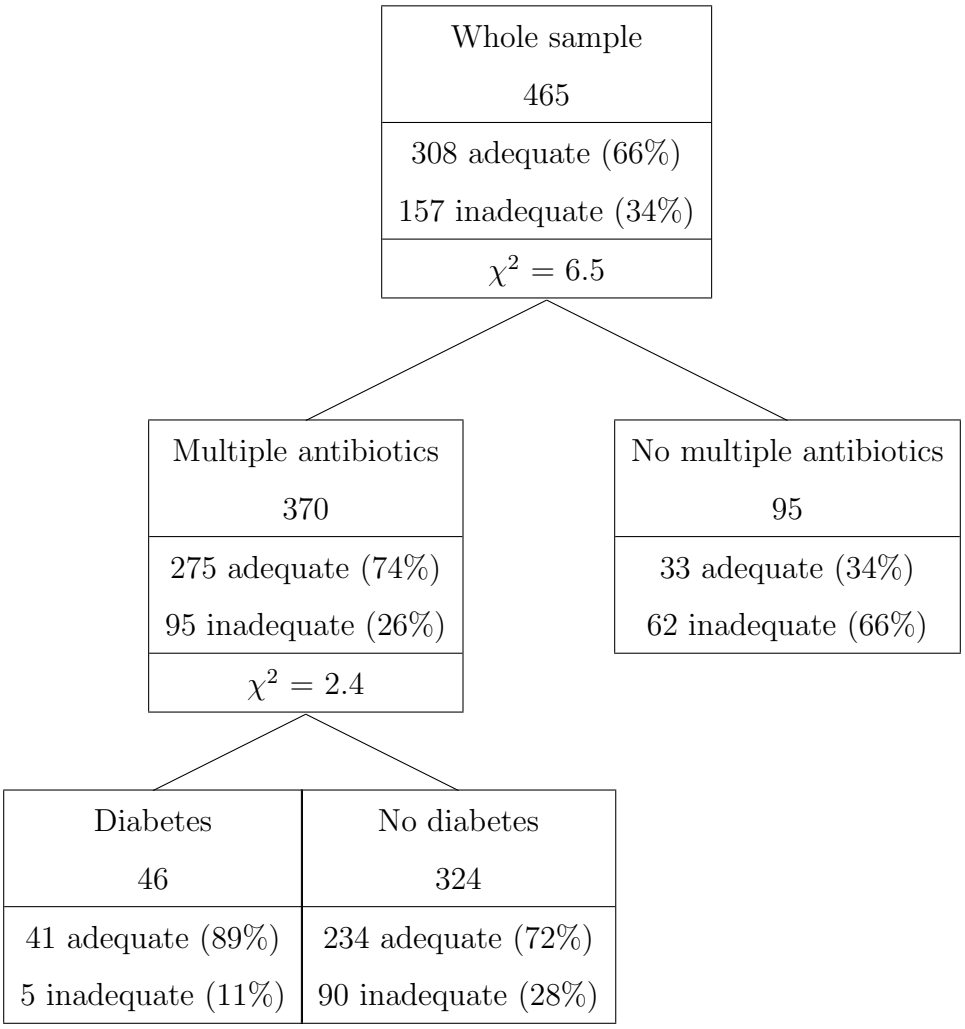

Figure 6.16: Interaction tree for the OUTCOMEREA data presenting covariates with an impact on treatment.

7 DISCUSSION AND CONCLUSION

In trials treating high-risk patients with severe infectious diseases, a proper analysis of the patient-relevant endpoint cure requires accounting for the time-dependent nature of the cure status. Patients are usually interested in how long it will take to get better and when they might get sick again [33]. For this, we presented a comprehensive multistate Markov model to examine the impact of a treatment on cure and death over time. This "cure-death model" is usually known as "illness-death model". It takes account of the time-dependency of each event by considering transitions between certain states. The basic version of this model contains three relevant transitions that can be analysed separately and with the help of a patient-relevant summary measure, the probability of being cured and alive function. One advantage of this summary measure is the more direct interpretation of probabilities in comparison to intensities. While transition intensities give a local description of the model dynamics, transition probabilities give a global description of what has accumulated over time. This function can be highly valuable when analysing data of randomised clinical trials, as in Section 6.1 and 6.2 but is also interesting at the planning stage. Furthermore, it can be applied in infection control and hospital epidemiology as in Section 6.3.

The Markov assumption

The applied cure-death multistate model requires to be time-inhomogeneous Markov. In the application chapter in Section 6.1 and 6.3 we could show that the Markov assumption was appropriate to analyse the data. Inter alia, we applied an alternative estimator for the PCA function not relying on the Markov assumption [107, 114]. It is based on a decomposition of the probability of interest into components that can be

estimated by Kaplan-Meier-type estimators, respectively.

The Markov assumption is also used, e.g., in Temkin [51] where the risk for a $1 \to 2$ transition could be measured solely in terms of time for the $0 \to 1$ transition, that is $\lambda_{12}(\tilde{t}, t) = \lambda_{12}(t)$. However, there are several possible models characterised by different assumptions for transition hazards. In Lagakos [180, 181] it is assumed that hazards are constant over time. A semi-Markov model is used in Lagakos et al. [182], Hsieh et al. [87], and Voelkel [183] where the $1 \to 2$ transition depends on the duration in state 1, that is $\lambda_{12}(\tilde{t}, t) = \lambda_{12}(t - \tilde{t})$. When time is measured from treatment, as in the time-inhomogeneous Markov model, patients are not at risk for death after cure until they are cured. If time is measured from cure, as in a semi-Markov model, a different time scale is used. All cured patients are collected in state 1 and simultaneously exposed to the risk of death.

Generally, one can test which assumption is more reasonable for a given data set. In some settings, the Markov assumption might be inappropriate / violated since it ignores the disease history. Then, estimation of general transition probabilities becomes problematic. For this, alternative estimators not relying on the Markov assumption were already proposed [184, 176, 185, 121, 186]. Pepe et al. [57, 58] proposed estimating the probability of an intermediate condition as prevalence functions in a non-Markov model, partly based on the estimator proposed by Tsai et al. [107, 114]. In a recently published work, Azarang et al. [122] addressed the problem of estimating the transition probabilities in a possibly non-Markov illness-death model in the presence of covariates using a binominal approach, analogous to that of Scheike et al. [121] for competing risks.

Moreover, under the independent censoring assumption, the Aalen-Johansen estimator for the state occupation probabilities as the probability to be cured and alive is consistent even if the Markov property is violated [176, 175, 174, 173].

Strengths and limilations

Having the outcome of interest, to get cured and stay alive over time, estimated by the Aalen-Johansen estimator of the probability to be in state 1, we examined several possibilities to assess a treatment effect in Section 4 and compared them in a simulation study in Section 5. Besides simple methods comparing risk differences with proportions of patients cured and alive and a log-rank-based test, we introduced an advanced technique to construct time-simultaneous confidence bands (when the absolute treatment effect is of interest) and an innovative regression technique to directly make inference on probability functions in a multistate setting (when the relative treatment effect is of interest).

Whereas a test simply investigates the question if the treatments are different at all, the confidence band for the difference in probability curves tells us where they are different or non-inferior and by how much. Pseudo-value regression can also be used to examine a specific time interval of interest. Moreover, only the pseudo-value approach accommodates the inclusion of additional explanatory variables in the analysis, that may be necessary not only within observational studies as in Section 6.3. The model could also adjust for the duration of antibiotic treatment, e.g., as a prolonged duration of antibiotic treatment is associated with the development of antimicrobial resistance [187].

However, both the pseudo-value regression as well as the wild bootstrap resampling are computationally cumbersome for large samples. Also, as mentioned in Section 4.3, for pseudo-value regression, a restriction to some timepoints has to be made. An approach for including all event times was presented by Liu et al. [53] in the setting with current leukemia-free survival. They proposed a score test with a closed form expression for the two-sample situation.

Future work / open questions

A first aspect worth considering for future research is the timing and definition of cure. In most studies, time of death is observed exactly unless it is right censored. However, this is not the case for clinical or microbiological cure. Often, the observation of such an intermediate event may only be registered at scheduled examination times, e.g., when the patient is seen by the general practitioner. In many trials, clinical cure is measured when the clinical study investigator performs the TOC. In the ceftobiprole trial, this was mostly performed within a time frame of 7 up to 14 days after the end of treatment. Hence, strictly speaking, we do not know the exact onset time of the intermediate condition cure, the time is therefore interval censored. More advanced statistical methods are required to address this problem; see Sun [188] for a general review and Commenges [189] for a more specialised review of methods for multistate models and interval censored data. More frequent data capture would allow for an improved analysis of well-defined cure endpoints, which is not possible adequately when endpoints are captured at a limited number of fixed timepoints [33]. Peace [14] stated that prospectively, it is not possible to classify a patient as cured or not during the treatment period but retrospectively, the exact cure time, theoretically, can be located where it occurred. Nevertheless, as timepoint of cure the timepoint of the TOC is utilised.

Furthermore, there is no unique definition of cure available. An efficacy endpoint is mostly based on resolution and improvement of signs and symptoms of infection at a timepoint after completion of therapy [10], but, a systematic review of Weiss et al. [16] shows that no agreement has been reached neither in the definition of cure nor in the time the TOC is performed.

A second point is how to find the appropriate non-inferiority margin. In testing non-

inferiority of novel anti-infective drugs of, e.g., the binary endpoint cure, a pre-specified margin is used and the difference in proportions of cure cases between the test and control group are compared at a landmark time. This margin is selected to be reasonable for a certain amount of cure cases in the control group at a fixed landmark time. Since our method avoids pre-specifying a specific time, it therefore requires further considerations on how to chose a suitable margin taken the proportion of cure cases in the control group at every timepoint into account. Upto now, for demonstration purposes, we used the margin applied for the "original" analyses of the RCT in Section 6.1.

A third interesting and important issue is how to obtain sample size for such an endpoint as being cured and alive over time. Generally, study planning for more specific multistate endpoints will typically be simulation-based [41] and methods to address this topic are currently under development. For these types of simulation studies, little information is required. As explained in Allignol et al. [190], information out of previous studies, e.g., could already inform such simulations.

Conclusion

In conclusion, the cure-death model provides a framework that enables a simultaneous analysis of both endpoints, cure and death. For a complete picture of the treatment effect, we recommend to take transition probabilities into account. Hereby, a better understanding on how a new treatment influences the time-dynamic cure process is possible. Crowley and Breslow [82] state that the major use for such probability curves is in graphical presentations. We show that beyond graphical advantages also methods for hypothesis testing are possible. The choice of the method to compare these curves is a matter of clinical preference. When the relative treatment effect is interesting, pseudo-value regression provides a suitable approach. However, absolute effect

measures are often of interest. Then, looking at the absolute difference over time in combination with confidence bands could provide a valuable statistical tool for such analyses, especially when treatment effects vary over time. Then, such time-dynamic patterns are important to be detected from the patients' perspective and may have direct impact on patient care. This may be included into future guidelines containing appropriate recommendations to tackle severe infectious diseases.

8 Notation and Acronyms

$\mathbb{1}$	Indicator function
α	Type-I-error
A, B	Treatments
β	Vector of regression coefficients
$\hat{\beta}$	Vector of estimated regression coefficients
CART	Classification and regression tree
CI	Confidence interval
CIF	Cumulative incidence function
COMBACTE	Combatting Bacterial Restistance in Europe
CRR	Cure risk ratio
δ_{abs}	Non-inferiority margin for absolute effect measure
δ_{rel}	Non-inferiority margin for relative effect measure
EMA	Europeans Medicines Agency
exp	Exponential function
E_{RL}, E_{01}, E_{02} E_{12}	Expected number of events
$\hat{\theta}_i$	Pseudo observation
$f(t)$	Density function
$F(t)$	Distribution function
FDA	Food and Drug Administration
g	Link function
GEE	Generalised estimating equation
H_0, H_1	Null-hypothesis and alternative hypothesis
HAP	Hospital-acquired pneumonia

HR	Hazard ratio
$i \in \{1, \ldots, n\}$	Individual
$\mathbf{I}$	Identity matrix
ICU	Intensive care unit
log	Natural logarithm
$\lambda_{lj}(t)$	Transition hazard
$\Lambda_{lj}(t)$	Cumulative transition hazard
$\hat{\Lambda}_{lj}(t)$	Nelson-Aalen estimator of the cumulative transition hazard
$\mathbf{\Lambda}(t)$	Matrix of cumulative transition hazard
$\hat{\mathbf{\Lambda}}(t)$	Matrix of Nelson-Aalen estimates
$\lambda_0(t)$	Baseline hazard
$\lambda(t \mid Z_i)$	Conditional hazard, given covariate vector Z_i
$M(t)$	Martingale
n, n_A, n_B	Number of individuals, treatment-specific
$N(t)$	Counting process
O_{RL}, O_{01}, O_{02} O_{12}	Observed number of events
I^A, I^B	Proportions in treatment group A and B
PCA	Probability of being cured and alive
PBRF	Probability of being in response function
$Pr(\cdot)$	Probability
$P_{lj}(t)$	Transition probability
$\hat{P}_{lj}(t)$	Aalen-Johansen estimator of the transition probability
$\mathbf{P}(t)$	Matrix of transition probabilities
$\hat{\mathbf{P}}(t)$	Matrix of Aalen-Johansen estimates
RCT	Randomised controlled trial
$\mathcal{S}$	State space

$S(t)$	Survival function
SE	Standard error
SHR	Subdistribution hazard ratio
s, t	Timepoints
T, T_i	Event time, event time for patient i
TOC	"Test-of-cure" visit
τ	End of follow-up
U_{lj}	Gaussian process
$\mathbf{U}$	Matrix of Gaussian processes
$U(\beta)$	Score function
V, V_{RL}, V_{01}, V_{02} V_{12}	Variances
VAP	Ventilator-associated pneumonia
W_i	Working covariance matrix
$Y(t)$	Risk set process
$(X(t), t \in [0, \infty))$	Stochastic process with finite state space
$z_{1-\frac{\alpha}{2}}$	$1 - \frac{\alpha}{2}$ quantile of the standard normal distribution
Z_i	Covariate vector for patient i
$\prod\!\!\!\!\!\int$	Product integral
$\xrightarrow{d}$	Convergence in distribution
#	Number / amount

9 Software

All analyses done in this dissertation were performed using the open-source software `R` [191], with the help of several `R` packages available on CRAN (http://cran.r-project.org/). For the Cox proportional hazard model and all related analyses, the `R` package `survival` [192] was applied. We used the `R` package `mvna` [193] and `etm` [194] for estimation of cumulative hazards and transition probabilities, respectively. For estimation of the regression coefficients in a generalised estimation equation, we used `R` package `geepack` [195]. We worked with the `R` package `nonrandom` [196] for estimation of the propensity score. For data preparation, the `R` packages `dplyr` [197] and `tidyr` [198] were employed; for plots, the `R` package `ggplot2` [199] was sometimes used.

10 Bibliography

[1] H Sommer, M Wolkewitz, and M Schumacher. The time-dependent "cure-death" model investigating two equally important endpoints simultaneously in trials treating high-risk patients with resistant pathogens. *Pharmaceutical Statistics*, 16:267–279, 2017.

[2] H Sommer, T Bluhmki, J Beyersmann, and M Schumacher. Assessing noninferiority in treatment trials regarding severe infectious diseases: An extension to the entire follow-up period using a cure-death multistate model. *Antimicrobial Agents and Chemotherapy*, 62:e01691–17, 2018.

[3] H Sommer, JF Timsit, and M Wolkewitz. Bezlotoxumab and recurrent Clostridium difficile infection. *New England Journal of Medicine (letter to the editor)*, 376:1594–1595, 2017.

[4] T Kostyanev, MJM O'Brien S Bonten, H Steel, S Ross, B Tacconelli E François, M Winterhalter, RA Stavenger, A Karlen, S Harbarth, J Hackett, HS Jafri, C Vuong, A MacGowan, A Witschi, G Angyalosi, JS Elborn, R deWinter, and H Goossens. The Innovative Medicines Initiative's New Drugs for Bad Bugs programme: European public-private partnerships for the development of new strategies to tackle antibiotic resistance. *Journal of Antimicrobial Chemotherapy*, 71:290–295, 2016.

[5] E Bettiol, WC Rottier, MD del Toro, S Harbarth, MJ Bonten, and J Rodríguez-Baño. Improved treatment of multidrug-resistant bacterial infections: Utility of clinical studies. *Future Microbiology*, 9:757–771, 2014.

[6] SE Cosgrove. The relationship between antimicrobial resistance and patient outcomes: Mortality, length of hospital stay, and health care costs. *Clinical Infectious Diseases*, 42:S82–S89, 2006.

[7] JD Hunter. Ventilator associated pneumonia. *British Medical Journal*, 344:e3225, 2012.

[8] JL Vincent, DJ Bihari, PM Suter, HA Bruining, J White, MH Nicolas-Chanoin, M Wolff, RC Spencer, and M Hemmer. The prevalence of nosocomial infection in intensive care units in Europe: Results of the European Prevalence of Infection in Intensive Care (EPIC) Study. *Journal of the American Medical Association*, 274:639–644, 1995.

[9] WG Melsen, MM Rovers, RHH Groenwold, DCJJ Bergmans, C Camus, TT Bauer, EW Hanisch, B Klarin, M Koeman, and WA Krueger. Attributable mortality of ventilator-associated pneumonia: A meta-analysis of individual patient data from randomised prevention studies. *The Lancet Infectious Diseases*, 13:665–671, 2013.

[10] GH Talbot, JH Powers, SC Hoffmann, J Toerner, J Alder, M Ariyasu, S Barriere, H Boucher, C Broom, and M Brunda. Developing outcomes assessments as endpoints for registrational clinical trials of antibacterial drugs: 2015 update from the Biomarkers Consortium of the Foundation for the National Institutes of Health. *Clinical Infectious Diseases*, 62:603–607, 2016.

[11] JG Muscedere, A Day, and DK Heyland. Mortality, attributable mortality, and clinical events as end points for clinical trials of ventilator-associated pneumonia and hospital-acquired pneumonia. *Clinical Infectious Diseases*, 51:S120–S125, 2010.

[12] JF Timsit, MEA de Kraker, H Sommer, E Weiss, E Bettiol, M Wolkewitz, D Wilson, and S Harbarth. Appropriate endpoints for evaluation of new antibiotic therapies for severe infections: A white paper from the COMBACTE network. *Intensive Care Medicine*, 43:1002–1012, 2017.

[13] M Bekaert, JF Timsit, S Vansteelandt, P Depuydt, A Vésin, M Garrouste-Orgeas, J Decruyenaere, C Clec'h, E Azoulay, and D Benoit. Attributable mortality of ventilator-associated pneumonia: A reappraisal using causal analysis. *American Journal of Respiratory and Critical Care Medicine*, 184:1133–1139, 2011.

[14] KE Peace. A survival analysis instead of an endpoint analysis for antibiotic data. *The Philippine Statistician*, 56:9–18, 2007.

[15] M Garrouste-Orgeas, JF Timsit, L Soufir, M Tafflet, C Adrie, F Philippart, JR Zahar, C Clec'h, D Goldran-Toledano, S Jamali, et al. Impact of adverse events on outcomes in intensive care unit patients. *Critical Care Medicine*, 36(7):2041–2047, 2008.

[16] E Weiss, W Essaied, C Adrie, JR Zahar, and JF Timsit. Treatment of severe hospital-acquired and ventilator-associated pneumonia: A systematic review of inclusion and judgment criteria used in randomized controlled trials. *Critical Care*, 21:162, 2017.

[17] LE Arthur, RS Kizor, AG Selim, ML van Driel, and L Seoane. Antibiotics for ventilator-associated pneumonia. *The Cochrane Library*, 2016.

[18] MO Harhay, J Wagner, SJ Ratcliffe, RS Bronheim, A Gopal, S Green, E Cooney, ME Mikkelsen, MP Kerlin, DS Small, and SD Halpern. Outcomes and statistical power in adult critical care randomized trials. *American Journal of Respiratory and Critical Care Medicine*, 189:1469–1478, 2014.

[19] B Blackwood, M Clarke, DF McAuley, PJ McGuigan, JC Marshall, and L Rose. How outcomes are defined in clinical trials of mechanically ventilated adults and children. *American Journal of Respiratory and Critical Care Medicine*, 189:886–893, 2014.

[20] European Medicines Agency. Addendum to the guideline on the evaluation of medicinal products indicated for treatment of bacterial infections. EMA/CHMP/351889/2013. 2013. Available at `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/11/WC500153953.pdf`.

[21] Food and Drug Administration. Guidance for industry: Hospital-acquired bacterial pneumonia and ventilator-associated bacterial pneumonia: Developing drugs for treatment. 2014. Available at `http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm234907.pdf`.

[22] MP Bauer, EJ Kuijper, and JT van Dissel. European Society of Clinical Microbiology and Infectious Diseases. European Society of Clinical Microbiology and Infectious Diseases (ESCMID): Treatment guidance document for Clostridium difficile infection. *Clinical Microbiology and Infection*, 15:1067–1079, 2009.

[23] S Johnson, DN Gerding, TJ Louie, NM Ruiz, and SL Gorbach. Sustained clinical response as an endpoint in treatment trials of Clostridium difficile-associated diarrhea. *Antimicrobial Agents and Chemotherapy*, 56(8):4043–4045, 2012.

[24] R Ramchandani, DA Schoenfeld, and DM Finkelstein. Global rank tests for multiple, possibly censored, outcomes. *Biometrics*, 72:926–935, 2016.

[25] BR Logan and AC Tamhane. Superiority inferences on individual endpoints following noninferiority testing in clinical trials. *Biometrical Journal*, 50:693–703, 2008.

[26] DA Bloch, TL Lai, Z Su, and P Tubert-Bitter. A combined superiority and noninferiority approach to multiple endpoints in clinical trials. *Statistics in Medicine*, 26:1193–1207, 2007.

[27] J Röhmel, C Gerlinger, N Benda, and J Läuter. On testing simultaneously noninferiority in two multiple primary endpoints and superiority in at least one of them. *Biometrical Journal*, 48:916–933, 2006.

[28] DL Price, DB Rubin, and T Valappil. Antimicrobial products: Statistical challenges and opportunities. *Statistics in Biopharmaceutical Research*, 7:325–330, 2015.

[29] Infectious Diseases Society of America. White paper: Recommendations on the conduct of superiority and organism-specific clinical trials of antibacterial agents for the treatment of infections caused by drug-resistant bacterial pathogens. *Clinical Infectious Diseases*, 55:1031–1146, 2012.

[30] J Pogue, PJ Devereaux, L Thabane, and S Yusuf. Designing and analyzing clinical trials with composite outcomes: Consideration of possible treatment differences between the individual outcomes. *PloS one*, 7:e34785, 2012.

[31] SJ Pocock, CA Ariti, TJ Collier, and D Wang. The win ratio: A new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*, 33:176–182, 2012.

[32] SR Evans, D Rubin, D Follmann, G Pennello, WC Huskins, JH Powers, D Schoenfeld, C Chuang-Stein, SE Cosgrove, VG Fowler, et al. Desirability of outcome ranking (DOOR) and response adjusted for duration of antibiotic risk (RADAR). *Clinical Infectious Diseases*, 61:800–806, 2015.

[33] JH Powers, K Howard, T Saretsky, S Clifford, S Hoffmann, L Llorens, and G Talbot. Patient-reported outcome assessments as endpoints in studies in infectious diseases. *Clinical Infectious Diseases*, 63:S52–S56, 2016.

[34] P Doshi. Speeding new antibiotics to market: A fake fix? *British Medical Journal*, 350:h1453, 2015.

[35] H Putter, M Fiocco, and RB Geskus. Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, 26:2389–2430, 2007.

[36] OO Aalen, Ø Borgan, and H Gjessing. *Survival and event history analysis: A process point of view*. Springer, 2008.

[37] EL Kaplan and P Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[38] M Wolkewitz, M von Cube, and M Schumacher. Multistate modeling to analyze nosocomial infection data: an introduction and demonstration. *Infection Control & Hospital Epidemiology*, 2017.

[39] M von Cube, M Schumacher, and M Wolkewitz. Basic parametric analysis for a multi-state model in hospital epidemiology. *BMC Medical Research Methodology*, 17:111, 2017.

[40] M Wolkewitz, BS Cooper, MJM Bonten, AG Barnett, and M Schumacher. Interpreting and comparing risks in the presence of competing events. *British Medical Journal*, 349:g5060, 2014.

[41] J Beyersmann, A Allignol, and M Schumacher. *Competing risks and multistate models with R*. Springer, 2011.

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

[43] LS Munoz-Price, JF Frencken, S Tarima, and M Bonten. Handling time dependent variables: Antibiotics and antibiotic resistance. *Clinical Infectious Diseases*, 62:1558–1563, 2016.

[44] M Schumacher, A Allignol, J Beyersmann, N Binder, and M Wolkewitz. Hospital-acquired infections – Appropriate statistical treatment is urgently needed! *International Journal of Epidemiology*, 42:1502–1508, 2013.

[45] VD Rosenthal, FE Udwadia, HJ Munoz, N Erben, F Higuera, K Abidi, EA Medeiros, EF Maldonado, SS Kanj, S Gikas, AG Barnett, and N Graves. Time-dependent analysis of extra length of stay and mortality due to ventilator-associated pneumonia in intensive-care units of ten limited-resources countries: Findings of the International Nosocomial Infection Control Consortium (INICC). *Epidemiology and Infection*, 139:1757–1763, 2011.

[46] JF Timsit, JR Zahar, and S Chevret. Attributable mortality of ventilator-associated pneumonia. *Current opinion in Critical Care*, 17:464–471, 2011.

[47] J Beyersmann, M Wolkewitz, A Allignol, N Grambauer, and M Schumacher. Application of multistate models in hospital epidemiology: Advances and challenges. *Biometrical Journal*, 53(2):332–350, 2011.

[48] M Wolkewitz, RP Vonberg, H Grundmann, J Beyersmann, P Gastmeier, S Bärwolff, C Geffers, M Behnke, H Rüden, and M Schumacher. Risk factors for the development of nosocomial pneumonia and mortality on intensive care units: Application of competing risks models. *Critical Care*, 12:R44, 2008.

[49] J Beyersmann, P Gastmeier, H Grundmann, S Bärwolff, C Geffers, M Behnke, H Rüden, and M Schumacher. Use of multistate models to assess prolongation of intensive care unit stay due to nosocomial infection. *Infection Control & Hospital Epidemiology*, 27(05):493–499, 2006.

[50] M Samore and S Harbarth. *A methodologically focused review of the literature in hospital epidemiology and infection control. In: Hospital Epidemiology and Infection Control (Ed. G. Mayhall)*, chapter 93, pages 1645–1657. Lippincott Williams & Wilkins, third edition, 2004.

[51] NR Temkin. An analysis for transient states with application to tumor shrinkage. *Biometrics*, 34:571–580, 1978.

[52] JP Klein and Y Shu. Multi-state models for bone marrow transplantation studies. *Statistical Methods in Medical Research*, 11:117–139, 2002.

[53] L Liu, B Logan, and JP Klein. Inference for current leukemia free survival. *Lifetime Data Analysis*, 14:432–446, 2008.

[54] JP Klein, N Keiding, Y Shu, RM Szydlo, and JM Goldman. Summary curves for patients transplanted for chronic myeloid leukaemia salvaged by a donor lym-

phocyte infusion: The current leukaemia-free survival curve. *British Journal of Haematology*, 109:148–152, 2000.

[55] M Eefting, LC de Wreede, CJM Halkes, PA von dem Borne, S Kersting, EWA Marijt, H Veelken, H Putter, J Schetelig, and JHF Falkenburg. Multi-state analysis illustrates treatment success after stem cell transplantation for acute myeloid leukemia followed by donor lymphocyte infusion. *Haematologica*, 101:506–514, 2016.

[56] CB Begg and M Larson. A study of the use of the probability-of-being-in-response function as a summary of tumor response data. *Biometrics*, 38:59–66, 1982.

[57] MS Pepe. Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association*, 86:770–778, 1991.

[58] MS Pepe, G Longton, and M Thornquist. A qualifier Q for the survival function to describe the prevalence of a transient condition. *Statistics in Medicine*, 10:413–421, 1991.

[59] JW Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society*, 11(1):15–53, 1949.

[60] J Berkson and RP Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952.

[61] JL Haybittle. The estimation of the proportion of patients cured after treatment for cancer of the breast. *The British Journal of Radiology*, 32(383):725–733, 1959.

[62] M Othus, B Barlogie, ML LeBlanc, and JJ Crowley. Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, 18(14):3731–3736, 2012.

[63] ASC Conlon, JMG Taylor, and DJ Sargent. Improving efficiency in clinical trials using auxiliary information: Application of a multi-state cure model. *Biometrics*, 71:460–468, 2015.

[64] ASC Conlon, JMG Taylor, and DJ Sargent. Multi-state models for colon cancer recurrence and death with a cured fraction. *Statistics in Medicine*, 33(10):1750–1766, 2014.

[65] ASC Conlon, JMG Taylor, DJ Sargent, and G Yothers. Using cure models and multiple imputation to utilize recurrence as an auxiliary variable for overall survival. *Clinical Trials*, 8(5):581–590, 2011.

[66] L Mauri and RB D'Agostino. Challenges in the design and interpretation of noninferiority trials. *New England Journal of Medicine*, 377(14):1357–1367, 2017.

[67] TR Fleming, K Odem-Davis, MD Rothmann, and Y Li Shen. Some essential considerations in the design and conduct of non-inferiority trials. *Clinical Trials*, 8(4):432–439, 2011.

[68] HM Hung, SJ Wang, and R O'Neill. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal*, 47(1):28–36, 2005.

[69] Food and Drug Administration. Guidance for industry: Non-inferiority clinical trials to establish effectiveness. 2016. Available at `https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf`.

[70] MP Fay and DA Follmann. Non-inferiority tests for anti-infective drugs using control group quantiles. *Clinical Trials*, 13:632–640, 2016.

[71] Y Matsuyama. A comparison of the results of intent-to-treat, per-protocol, and g-estimation in the presence of non-random treatment changes in a time-to-event non-inferiority trial. *Statistics in Medicine*, 29(20):2107–2116, 2010.

[72] JP Morden, PC Lambert, N Latimer, KR Abrams, and AJ Wailoo. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. *BMC Medical Research Methodology*, 11(1):4, 2011.

[73] BL Wiens. Multiple comparisons in non-inferiority trials: Reaction to recent regulatory guidance on multiple endpoints in clinical trials. *Journal of Biopharmaceutical Statistics*, pages 1–11, 2017.

[74] JH Rex, GH Talbot, MJ Goldberger, BI Eisenstein, RM Echols, JF Tomayko, MN Dudley, and A Dane. Progress in the fight against multidrug-resistant bacteria 2005–2016: Modern noninferiority trial designs enable antibiotic development in advance of epidemic bacterial resistance. *Clinical Infectious Diseases*, 2017.

[75] S Nambiar, K Laessig, J Toerner, J Farley, and E Cox. Antibacterial drug development: Challenges, recent developments, and future considerations. *Clinical Pharmacology & Therapeutics*, 96:147–149, 2014.

[76] MJ DiNubile. Noninferior antibiotics: When is "not bad" "good enough"? In *Open Forum Infectious Diseases*, volume 3, page ofw110. Oxford University Press, 2016.

[77] SM Snapinn. Noninferiority trials. *Trials*, 1:19–21, 2000.

[78] AM Caliendo, DN Gilbert, CC Ginocchio, KE Hanson, L May, TC Quinn, FC Tenover, D Alland, AJ Blaschke, RA Bonomo, KC Carroll, MJ Ferraro, LR Hirschhorn, WP Joseph, T Karchmer, AT MacIntyre, LB Reller, and AF Jack-

son. Better tests, better care: Improved diagnostics for infectious diseases. *Clinical Infectious Diseases*, 57(suppl_3):S139–S170, 2013.

[79] SC Chow, H Wang, and J Shao. *Sample size calculations in clinical research*. CRC Press, 2007.

[80] WC Blackwelder. Proving the null hypothesis in clinical trials. *Controlled Clinical Trials*, 3:345–353, 1982.

[81] MF Huque, T Valappil, and GG Soon. Hierarchical nested trial design (HNTD) for demonstrating treatment efficacy of new antibacterial drugs in patient populations with emerging bacterial resistance. *Statistics in Medicine*, 33:4321–4336, 2014.

[82] J Crowley and N Breslow. Statistical analysis of survival data. *Annual Review of Public Health*, 5(1):385–411, 1984.

[83] N Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50:163–170, 1966.

[84] DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34:187–220, 1972.

[85] RJ Gray. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, pages 1141–1154, 1988.

[86] JP Fine and RJ Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94:496–509, 1999.

[87] FY Hsieh, John Crowley, and Douglass C Tormey. Some test statistics for use in multistate survival analysis. *Biometrika*, 70:111–119, 1983.

[88] PK Andersen, JP Klein, and S Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90:15–27, 2003.

[89] PK Andersen and JP Klein. Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies. *Scandinavian Journal of Statistics*, 34:3–16, 2007.

[90] PC Austin and JP Fine. Accounting for competing risks in randomized controlled trials: A review and recommendations for improvement. *Statistics in Medicine*, 36:1203–1209, 2017.

[91] DY Lin. Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine*, 16:901–910, 1997.

[92] J Beyersmann, S Di Termini, and M Pauly. Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics*, 40:387–402, 2013.

[93] NC Oza. Online bagging and boosting. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2340–2345, 2005.

[94] HKH Lee and MA Clyde. Lossless online Bayesian bagging. *The Journal of Machine Learning Research*, 5:143–151, 2004.

[95] SS Awad, AH Rodriguez, YC Chuang, Z Marjanek, AJ Pareigis, G Reis, TWL Scheeren, AS Sánchez, X Zhou, M Saulay, and M Engelhardt. A phase 3 randomized double-blind comparison of ceftobiprole medocaril versus ceftazidime plus linezolid for the treatment of hospital-acquired pneumonia. *Clinical Infectious Diseases*, 59:51–61, 2014.

[96] MH Wilcox, DN Gerding, IR Poxton, NT Shen, A Maw, LL Tmanova, JR Leal, SJ Heitman, JM Conly, EA Henderson, et al. Bezlotoxumab and recurrent Clostridium difficile infection. *New England Journal of Medicine*, 376:305–317, 2017.

[97] OO Aalen, PK Andersen, Ø Borgan, RD Gill, and N Keiding. History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics*, 5(1):1–28, 2009.

[98] OO Aalen and S Johansen. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5:141–150, 1978.

[99] W Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1, 1969.

[100] W Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945–966, 1972.

[101] R Peto and J Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135:185–207, 1972.

[102] E Marubini and MG Valsecchi. *Analysing survival data from clinical trials and observational studies*, volume 15. John Wiley & Sons, 2004.

[103] PC Austin and JP Fine. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Statistics in Medicine*, 2017.

[104] C Schmoor, M Schumacher, J Finke, and J Beyersmann. Competing risks and multistate models. *Clinical Cancer Research*, 19:12–21, 2013.

[105] J Beyersmann, P Gastmeier, and M Schumacher. Incidence in ICU populations: How to measure and report it? *Intensive Care Medicine*, 40:871–876, 2014.

[106] RD Gill and S Johansen. A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, 18:1501–1555, 1990.

[107] WY Tsai, S Leurgans, and J Crowley. Nonparametric estimation of a bivariate survival function in the presence of censoring. *The Annals of Statistics*, pages 1351–1365, 1986.

[108] B Altshuler. Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, 6:1–11, 1970.

[109] OO Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6:701–726, 1978.

[110] TR Fleming. Nonparametric estimation for nonhomogeneous Markov processes in the problem of competing risks. *Annals of Statistics*, 6:1057–1070, 1978.

[111] TR Fleming. Asymptotic distribution results in competing risks estimation. *Annals of Statistics*, 6:1071–1079, 1978.

[112] PK Andersen, Ø Borgan, RD Gill, and N Keiding. *Statistical models based on counting processes*. Springer, 1993.

[113] A Allignol, M Schumacher, and J Beyersmann. A note on variance estimation of the Aalen-Johansen estimator of the cumulative incidence function in competing risks, with a view towards left-truncated data. *Biometrical Journal*, 52(1):126–137, 2010.

[114] WY Tsai. Bivariate survival time and censoring. *Unpublished dissertation, University of Wisconsin, Department of Biostatistics*, 1982.

[115] D Altman, D Machin, T Bryant, and M Gardner. *Statistics with confidence: Confidence intervals and statistical guidelines*. John Wiley & Sons, 2013.

[116] S Wellek. *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press, 2010.

[117] CFJ Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14:1261–1295, 1986.

[118] T Bluhmki, C Schmoor, D Dobler, M Pauly, J Finke, M Schumacher, and J Beyersmann. A two-stage wild bootstrap approach for the Aalen-Johansen estimate in mulistate models. *Biometrics*, 2018.

[119] S Hieke, H Bertz, M Dettenkofer, M Schumacher, and J Beyersmann. Initially fewer bloodstream infections for allogeneic vs. autologous stem-cell transplants in neutropenic patients. *Epidemiology and Infection*, 141:158–164, 2013.

[120] PK Andersen and MP Perme. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19:71–99, 2009.

[121] TH Scheike, MJ Zhang, and TA Gerds. Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95(1):205–220, 2008.

[122] L Azarang, T Scheike, and J de Uña-Álvarez. Direct modeling of regression effects for transition probabilities in the progressive illness–death model. *Statistics in Medicine*, 36(12):1964–1976, 2017.

[123] JP Klein, B Logan, M Harhoff, and PK Andersen. Analyzing survival curves at a fixed point in time. *Statistics in Medicine*, 26:4505–4519, 2007.

[124] MK Grand and H Putter. Regression models for expected length of stay. *Statistics in Medicine*, 35(7):1178–1192, 2016.

[125] K Liang and SL Zeger. Longitudinal data analysis using generalized linear models. *Biometrica*, 73:13–22, 1986.

[126] JP Klein, HC Van Houwelingen, JG Ibrahim, and TH Scheike. *Handbook of survival analysis*. Chapman and Hall/CRC, 2013.

[127] TA Gerds, TH Scheike, and PK Andersen. Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine*, 31:3921–3930, 2012.

[128] F Graw, TA Gerds, and M Schumacher. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2):241–255, 2009.

[129] M Overgaard, ET Parner, and J Pedersen. Asymptotic theory of generalized estimating equations based on Jack-knife pseudo-observations. *The Annals of Statistics*, 45(5):1988–2015, 2017.

[130] M Jacobsen and T Martinussen. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics*, 43(3):845–862, 2016.

[131] JP Klein and PK Andersen. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61(1):223–229, 2005.

[132] DR Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.

[133] BA Leav, B Blair, M Leney, M Knauber, C Reilly, I Lowy, DN Gerding, CP Kelly, K Katchar, and R Baxter. Serum anti-toxin B antibody correlates with protection from recurrent Clostridium difficile infection. *Vaccine*, 28(4):965–969, 2010.

[134] M Schumacher, K Ohneberg, and J Beyersmann. Competing risk bias was common in a prominent medical journal. *Journal of Clinical Epidemiology*, 80:135–136, 2016.

[135] C van Walraven and FA McAlister. Competing risk bias was common in Kaplan–Meier risk estimates published in prominent medical journals. *Journal of Clinical Epidemiology*, 69:170–173, 2016.

[136] MS Pepe and M Mori. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine*, 12(8):737–751, 1993.

[137] M Schumacher, M Wangler, M Wolkewitz, J Beyersmann, et al. Attributable mortality due to nosocomial infections-a simple and useful application of multistate models. *Methods of Information in Medicine*, 46:595–600, 2007.

[138] R Masterton, G Drusano, DL Paterson, and G Park. Appropriate antimicrobial treatment in nosocomial infections – The clinical challenges. *Journal of Hospital Infection*, 55:1–12, 2003.

[139] B Planquette, JF Timsit, BY Misset, C Schwebel, E Azoulay, C Adrie, A Vesin, S Jamali, JR Zahar, et al. Pseudomonas aeruginosa ventilator-associated pneumonia. Predictive factors of treatment failure. *American Journal of Respiratory and Critical Care Medicine*, 188:69–76, 2013.

[140] MH Kollef. Inadequate antimicrobial treatment: An important determinant of outcome for hospitalized patients. *Clinical Infectious Diseases*, 31(Supplement_4):S131–S138, 2000.

[141] JC McGregor, SE Rich, AD Harris, EN Perencevich, R Osih, TP Lodise, RR Miller, and JP Furuno. A systematic review of the methods used to assess the association between appropriate antibiotic therapy and mortality in bacteremic patients. *Clinical Infectious Diseases*, 45:329–337, 2007.

[142] T Herkel, R Uvizl, L Doubravska, M Adamus, T Gabrhelik, MH Sedlakova, M Kolar, V Hanulik, V Pudova, and K Langova. Epidemiology of hospital-acquired pneumonia: Results of a Central European multicenter, prospective, observational study compared with data from the European region. *Biomedical Papers*, 160:448–455, 2016.

[143] J Garnacho-Montero, M Sa-Borges, J Sole-Violan, F Barcenilla, A Escoresca-Ortega, M Ochoa, A Cayuela, and J Rello. Optimal management therapy for Pseudomonas aeruginosa ventilator-associated pneumonia: An observational, multicenter study comparing monotherapy with combination antibiotic therapy. *Critical Care Medicine*, 35:1888–1895, 2007.

[144] A Fraser, M Paul, N Almanasreh, E Tacconelli, U Frank, R Cauda, S Borok, M Cohen, S Andreassen, and A Nielsen. Benefit of appropriate empirical antibiotic treatment: Thirty-day mortality and duration of hospital stay. *The American Journal of Medicine*, 119(11):970–976, 2006.

[145] M Iregui, S Ward, G Sherman, VJ Fraser, and MH Kollef. Clinical importance of delays in the initiation of appropriate antibiotic treatment for ventilator-associated pneumonia. *Chest*, 122:262–268, 2002.

[146] H Dupont, H Mentec, JP Sollet, and G Bleichner. Impact of appropriateness of initial antibiotic therapy on the outcome of ventilator-associated pneumonia. *Intensive Care Medicine*, 27:355–362, 2001.

[147] C Clec'h, JF Timsit, A De Lassence, E Azoulay, C Alberti, M Garrouste-Orgeas, B Mourvilier, G Troche, M Tafflet, and O Tuil. Efficacy of adequate early antibiotic therapy in ventilator-associated pneumonia: influence of disease severity. *Intensive Care Medicine*, 30(7):1327–1333, 2004.

[148] F Bloos. Clinical diagnosis of sepsis and the combined use of biomarkers and culture-and non-culture-based assays. *Sepsis: Diagnostic Methods and Protocols*, pages 247–260, 2015.

[149] F Bloos and K Reinhart. Rapid diagnosis of sepsis. *Virulence*, 5(1):154–160, 2014.

[150] CI Kang, SH Kim, HB Kim, SW Park, YJ Choe, MD Oh, EC Kim, and KW Choe. Bacteremia: Risk factors for mortality and influence of delayed receipt of effective antimicrobial therapy on clinical outcome. *Clinical Infectious Diseases*, 37(6):745–751, 2003.

[151] FX Hanon, T Lund Sørensen, K Mølbak, H Schønheyder, DL Monnet, and G Pedersen. Survival of patients with bacteraemia in relation to initial empirical antimicrobial treatment. *Scandinavian Journal of Infectious Diseases*, 34(7):520–528, 2002.

[152] S Harbarth, J Garbino, J Pugin, JA Romand, D Lew, and D Pittet. Inappropriate initial antimicrobial therapy and its effect on survival in a clinical trial of immunomodulating therapy for severe sepsis. *The American Journal of Medicine*, 115(7):529–535, 2003.

[153] EH Ibrahim, G Sherman, S Ward, VJ Fraser, and MH Kollef. The influence of inadequate antimicrobial treatment of bloodstream infections on patient outcomes in the ICU setting. *Chest Journal*, 118(1):146–155, 2000.

[154] L Leibovici, I Shraga, M Drucker, H Konigsberger, Z Samra, and SD Pitlik. The benefit of appropriate empirical antibiotic treatment in patients with bloodstream infection. *Journal of Internal Medicine*, 244(5):379–386, 1998.

[155] RD MacArthur, M Miller, T Albertson, E Panacek, D Johnson, L Teoh, and W Barchuk. Adequacy of early empiric antibiotic treatment and survival in severe sepsis: experience from the MONARCS trial. *Clinical Infectious Diseases*, 38(2):284–288, 2004.

[156] CI Kang, SH Kim, WB Park, KD Lee, HB Kim, EC Kim, MD Oh, and KW Choe. Bloodstream infections caused by antibiotic-resistant gram-negative bacilli: Risk factors for mortality and impact of inappropriate initial antimicrobial therapy on outcome. *Antimicrobial Agents and Chemotherapy*, 49(2):760–766, 2005.

[157] DW Bates, KE Pruess, and TH Lee. How bad are bacteremia and sepsis? Outcomes in a cohort with suspected bacteremia. *Archives of Internal Medicine*, 155(6):593–598, 1995.

[158] J Vallés, J Rello, A Ochagavía, J Garnacho, and MA Alcalá. Community-acquired bloodstream infection in critically ill adult patients: Impact of shock and inappropriate antibiotic therapy on survival. *Chest*, 123(5):1615–1624, 2003.

[159] R Zaragoza, A Artero, JJ Camarena, S Sancho, R Gonzalez, and JM Nogueira. The influence of inadequate empirical antimicrobial treatment on patients with bloodstream infections in an intensive care unit. *Clinical Microbiology and Infection*, 9(5):412–418, 2003.

[160] KB Pouwels, E Van Kleef, S Vansteelandt, R Batra, JD Edgeworth, T Smieszek, and JV Robotham. Does appropriate empiric antibiotic therapy modify inten-

sive care unit-acquired Enterobacteriaceae bacteraemia mortality and discharge? *Journal of Hospital Infection*, 96:23–28, 2017.

[161] C Pena, S Gomez-Zorrilla, I Oriol, F Tubau, MA Dominguez, M Pujol, and J Ariza. Impact of multidrug resistance on Pseudomonas aeruginosa ventilator-associated pneumonia outcome: Predictors of early and crude mortality. *European Journal of Clinical Microbiology & Infectious Diseases*, 32:413–420, 2013.

[162] PR Rosenbaum and DB Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, pages 41–55, 1983.

[163] PC Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46:399–424, 2011.

[164] MA Brookhart, S Schneeweiss, KJ Rothman, RJ Glynn, J Avorn, and T Stürmer. Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156, 2006.

[165] PC Austin, P Grootendorst, and GM Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, 26:734–753, 2007.

[166] PC Austin. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, 33:1242–1258, 2014.

[167] PR Rosenbaum. Observational studies. In *Observational Studies*, pages 1–17. Springer, 2002.

[168] PC Austin. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10:150–161, 2011.

[169] PC Austin. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine*, 32:2837–2849, 2013.

[170] S Weitzen, KL Lapane, AY Toledano, AL Hume, and V Mor. Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety*, 13:841–853, 2004.

[171] T Teräsvirta and I Mellin. Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics*, pages 159–171, 1986.

[172] A Allignol, M Schumacher, and J Beyersmann. Estimating summary functionals in multistate models with an application to hospital infection data. *Computational Statistics*, 26(2):181–197, 2011.

[173] S Datta and GA Satten. Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. *Statistics & Probability Letters*, 55:403–411, 2001.

[174] DV Glidden. Robust inference for event probabilities with non-Markov event data. *Biometrics*, 58:361–368, 2002.

[175] N Gunnes, Ø Borgan, and OO Aalen. Estimating stage occupation probabilities in non-Markov models. *Lifetime Data Analysis*, 13:211–240, 2007.

[176] A Allignol, J Beyersmann, T Gerds, and A Latouche. A competing risks approach for nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*, 20:495–513, 2014.

[177] Y Fujiwara, T Yamada, Y Naomoto, T Yamatsuji, Y Shirakawa, S Tanabe, K Noma, T Kimura, H Aoki, and H Matsukawa. Multicentred surgical site infection surveillance using partitioning analysis. *Journal of Hospital Infection*, 85(4):282–288, 2013.

[178] C Schmoor, K Ulm, and M Schumacher. Comparison of the Cox model and the regression tree procedure in analysing a randomized clinical trial. *Statistics in Medicine*, 12:2351–2366, 1993.

[179] JN Morgan and JA Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415–434, 1963.

[180] SW Lagakos. A stochastic model for censored-survival data in the presence of an auxiliary variable. *Biometrics*, pages 551–559, 1976.

[181] SW Lagakos. Using auxiliary variables for improved estimates of survival time. *Biometrics*, pages 399–404, 1977.

[182] SW Lagakos, CJ Sommer, and M Zelen. Semi-Markov models for partially censored data. *Biometrika*, 65:311–317, 1978.

[183] JG Voelkel. Multivariate counting processes and the probability of being in response function. *Dissertation Abstracts International Part B: Science and Engineering*, 42:1981, 1981.

[184] J de Uña-Álvarez and L Meira-Machado. Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. *Biometrics*, 71:364–375, 2015.

[185] AP Amorim, J de Uña-Álvarez, and L Meira-Machado. Presmoothing the transition probabilities in the illness–death model. *Statistics & Probability Letters*, 81:797–806, 2011.

[186] L Meira-Machado, J de Uña-Álvarez, and S Datta. Conditional transition probabilities in a non-markov illness-death model. *Discussion Papers in Statistics and Operation Research 12/05*, 11, 2012.

[187] TP van Boeckel, S Gandra, A Ashok, Q Caudron, BT Grenfell, SA Levin, and R Laxminarayan. Global antibiotic consumption 2000 to 2010: An analysis of national pharmaceutical sales data. *The Lancet Infectious Diseases*, 14:742–750, 2014.

[188] J Sun. *The statistical analysis of interval-censored failure time data*, volume 2. Springer, 2006.

[189] D Commenges. Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research*, 11(2):167–182, 2002.

[190] A Allignol, M Schumacher, C Wanner, C Drechsler, and J Beyersmann. Understanding competing risks: A simulation point of view. *BMC Medical Research Methodology*, 11, 2011.

[191] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, A, 2016.

[192] TM Therneau. *survival: Survival analysis*, 2016. R package version 2.40-1.

[193] A Allignol. *mvna: Nelson-Aalen estimator of the cumulative hazard in multistate models*, 2013. R package version 1.2-3.

[194] A Allignol. *etm: Empirical transition matrix*, 2014. R package version 0.6-2.

[195] S Højsgaard, U Halekoh, and J Yan. *geepack: Generalized estimating equation package*, 2016. R package version 1.2-1.

[196] S Stampf. *nonrandom: Stratification and matching by the propensity score*, 2014. R package version 1.42.

[197] H Wickham and R Francois. *dplyr: A grammar of data manipulation*, 2016. R package version 0.5.0.

[198] H Wickham. *tidyr: Easily tidy data with spread() and gather() functions*, 2016. R package version 0.6.0.

[199] H Wickham and W Chang. *ggplot2: An implementation of the grammar of graphics*, 2016. R package version 2.1.0.