Aus dem Institut für Medizinische Biometrie
Universität Heidelberg
Geschäftsführender Direktor: Prof. Dr.sc.hum. Meinhard Kieser

ESTIMATION, TESTING AND SAMPLE SIZE CALCULATION WITHIN THE RESPONDER STRATIFIED EXPONENTIAL SURVIVAL MODEL

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum
an der Medizinischen Fakultät Heidelberg
der Ruprecht-Karls-Universität

vorgelegt von
Samuel Kilian
aus
Wiesbaden
2023

Dekan: Prof. Dr. med. Hans-Georg Kräusslich

Doktorvater: Prof. Dr.sc.hum. Meinhard Kieser

Contents

Notation

Here is a summary of the notation used in this thesis.

Abbrevations	
FDA	Food and Drug Administration
MLE	Maximum Likelihood estimator
pCR	pathological complete response
RSES	responder stratified exponential survival
TOE	Type I error
Pow	power
+resp	response benefit
+surv	survival benefit
+resp +surv	response benefit and survival benefit
L	treatment with Lapatinib
T	treatment with Trastuzumab
L+T	treatment with Lapatinib and Trastuzumab

General math	
$\forall$	for all
$\exists$	exists
arctanh	inverse hyperbolic tangent
tanh	hyperbolic tangent
ψ	Digamma function
$\psi^{(1)}$	Polygamma function of order 1
Γ	Gamma function
B	Beta function
z_γ	γ-quantile of standard normal distribution
$\mathrm{P}(A)$	probability of some event A
$\mathrm{E}[W]$	expected value of some random variable W
$\mathrm{Var}(W)$	variance of some random variable W
$\mathrm{Cov}(W_1, W_2)$	covariance of W_1 and W_2
$\mathrm{Cor}(W_2, W_2)$	Pearson correlation coefficient of X and Y
$W\|_A$	random variable W conditional on event A
$\mathrm{P}(A_1 \mid A_2)$	probability of event A_1 conditional on event A_2
$\mathrm{E}[W \mid A]$	conditional expectation of random variable W on event A

$\mathrm{E}[W_1 \mid W_2]$	conditional expectation of random variable W_1 on random variable W_2
$\mathrm{Var}(W \mid A)$	conditional variance of random variable W on event A
$\mathrm{Var}(W_1 \mid W_2)$	conditional variance of random variable W_1 on random variable W_2
$\mathrm{Cov}(W_1, W_2 \mid A)$	conditional covariance of random variables W_1 and W_2 on event A
$\mathrm{Cov}(W_1, W_2 \mid W_3)$	conditional covariance of random variables W_1 and W_2 on random variable W_3
$\mathrm{Cor}(W_1, W_2 \mid A)$	conditional correlation coefficient of random variables W_1 and W_2 on event A
$\mathrm{Cor}(W_1, W_2 \mid W_3)$	conditional correlation coefficient of random variables W_1 and W_2 on random variable W_3
$W \sim D$	The random variable W is distributed with distribution D.
$W \overset{\text{appr}}{\sim} D$	The random variable W is approximately distributed with distribution D.
$N(\mu, \sigma^2)$	normal distribution with expectation μ and variance σ^2
$\mathrm{Exp}(\lambda)$	exponential distribution with rate/hazard λ
$\Gamma(\alpha, \beta)$	Gamma distribution with shape α and rate β
$\mathrm{Weibull}(b, s)$	Weibull distribution with scale parameter b and shape parameter s
$\beta'(\alpha, \beta, q)$	Beta prime distribution with shape parameters α, β and scale factor q
$I(\vartheta)$	Fisher information matrix of multi-dimensional parameter ϑ
$\overset{D}{\to}$	convergence in distribution
$\overset{P}{\to}$	convergenve in probability

Symbols and letters

f	density of a probability distribution
F	distribution function of a probability distribution
S	survival function of a survival distribution
X	binary response random variable
p	response probability
T_S	survival time random variable
U	censoring time random variable
D	binary event indicator random variable
L	likelihood
n	sample size in one-group model
k	number of responders in one-group model
j	response stratum index $j = 0, 1$
l_j	number of uncensored observations in response stratum j in one-group model
λ_j	hazard of response stratum j in one-group model

θ_j	logarithmic hazard $\log(\lambda_j)$ of response stratum j in one-group model
η_j	inverse hazard $1/\lambda_j$ of response stratum j in one-group model
ϑ	three-dimensional parameter $(p, \lambda_1, \lambda_0)$ or (p, θ_1, θ_0)
$\hat{\lambda}_j$	MLE for λ_j in one-group model
$\hat{\theta}_j$	MLE for θ_j in one-group model
$\hat{\eta}_j$	MLE for η_j in one-group model
u	placeholder for one of the parameters $\lambda_j, \theta_j, \eta_j$
$\hat{u}$	estimator for u
$\hat{\sigma}_{\hat{u}}$	estimator of standard deviation of $\hat{u}$
CI_u	confidence interval for parameter u
E	denotes experimental group
C	denotes control group
i	group index $i = E, C$
S_i	survival function of group i
p_i	response probability in group i
$\lambda_{j,i}$	hazard in response stratum j in group i
$\theta_{j,i}$	logarithm of $\lambda_{j,i}$
H_0	global null hypothesis that RSES model parameters are equal in experimental and control group
$H_{p,0}$	local null hypothesis that response probabilities are equal in experimental and control group
$H_{\theta_j,0}$	local null hypothesis that survival in response stratum j is equal in experimental and control group
T_p	test statistic for testing $H_{p,0}$
T_{θ_j}	test statistic for testing $H_{\theta_j,0}$
α	significance level
$\tilde{\alpha}$	local level for testing the local null hypotheses
n_i	sample size in group i
k_i	number of responders in group i
$l_{j,i}$	number of uncensored observations in response stratum j in treatment group i
$q_{j,i}$	probability of observing an event for one specific patient in response stratum j and treatment group i
I_p	random variable of accepting $H_{p,0}$
Δ_p	difference between response rates: $p_E - p_C$
Δ_{θ_j}	logarithmic hazard ratio of groups in stratum j: $\theta_{j,E} - \theta_{j,C}$
$\hat{\Delta}_u$	estimator of Δ_u
$\tilde{T}_{\theta_j}$	transformation of test statistic T_{θ_j}
p_u	p-value of local test of $H_{u,0}$
R_u	random variable of the rejection of $H_{u,0}$

ω	true coverage probability or true rejection probability
n_{sim}	number of simulations
$p'_i, \theta'_{j,i}$	specified parameter values under the assumed alternative hypothesis for treatment group i and response stratum j
${\sigma'_u}^2$	variance of parameter difference $u_E - u_C$ under specified alternative hypothesis
β_u	acceptance probability of $H_{u,0}$ under specified alternative hypothesis
r	sample size ratio n_E/n_C
T_{LR}	test statistic of logrank test
T_{sLR}	test statistic of stratified logrank test

Chapter 1

Introduction

Endpoints of clinical trials should be appropriate for answering the research question, objectively measurable, and relevant for patients. Thus, for proving efficacy of a new oncological therapy, the primary endpoint is usually overall survival. The new therapy is compared to the present gold standard. However, as therapies get better and diagnoses are made in earlier stages, differences between therapies may only be observable after many years. This may considerably delay approval of new treatments and their application in practice.

To avoid withholding of a promising therapy, the Food and Drug Administration (FDA) provides four different programs for expedited development of new therapies (Wallach et al. 2018). One of them is the *Accelerated Approval* pathway, where the approval is based on a surrogate endpoint. The approval is preliminary and has to be confirmed later when the main endpoint can be assessed. Between 1992 and 2021, 278 preliminary accelerated approvals were granted by the FDA after a median processing time of 6 months (Food and Drug Administration 2022). Of these, 50% were confirmed later, 10% were withdrawn, and 40% are still ongoing.

In 2014, the FDA published a more detailed guidance (which was updated in 2020) regarding the use of pathological complete response (pCR) as a surrogate endpoint when approving a novel neoadjuvant treatment of high-risk early-stage breast cancer (Food and Drug Administration 2020). Although the appropriateness of pCR as surrogate endpoint is disputed (Conforti et al. 2021), the guidance illustrates the use of a binary surrogate endpoint for survival. When planning and analysing trials in this context, the correlation between surrogate endpoint and survival has to be taken into account. The relationship between surrogate endpoint and survival can be modeled by means of a conditional survival model proposed by Xia et al. (2014). They investigated the correlation coefficient between the surrogate endpoint and the survival endpoint. Furthermore, they assessed the power of the logrank test and stratified logrank test in various scenarios of the conditional survival model. However, they did not present methods for parameter estimation, statistical testing, and sample size calculation within the conditional survival model. This gap constitutes a major hurdle for the application of this approach.

The aim of this thesis is to investigate the conditional survival model of Xia et al. (2014) and to develop and evaluate methods for parameter estimation, hypothesis testing, and sample size calculation within the model. For parameter estimation, point estimators are derived and their distribution is assessed. Furthermore, confidence intervals for the parameters are derived and evaluated. For hypothesis testing, this thesis focusses on the survival endpoint and does not aim to develop a complete analysis strategy for studies where an interim decision is made after assessing the surrogate endpoint. However, future research may use the findings of this thesis for the development of such methods. A hypothesis test of the difference between two treatment groups regarding the parameters of the conditional survival model is developed. Furthermore, confidence intervals for parameter differences are derived. One consequence of the conditional survival model is that the hazards of the treatment groups are non-proportional. Thus, standard methods like the logrank test may not be the most powerful (Dormuth et al. 2022). Hence, when assessing the characteristics of the developed test, they are compared to the logrank test and the stratified logrank test. Furthermore, an approximate and exact sample size calculation method for the developed test are derived and evaluated. The presented methods are applied to a clinical example (Huober et al. 2019).

This thesis is structured as follows. In Chapter 2, the methods used for obtaining the results of this research are described. This comprises the description of the conditional survival model (Section 2.1) as well as the general description of how formulas, estimators, hypothesis tests, and sample size calculation methods are derived and evaluated (Sections 2.2, 2.3, 2.4, 2.5, and 2.6). Furthermore, four simulation studies used for evaluating the methods are described.

The corresponding results, i.e. the specific derivation of formulas, estimators, hypothesis tests, and sample size calculation methods, as well as the results of their evaluation are given in Chapter 3. Most subsections in Chapter 2 correspond to a subsection in Chapter 3. In Section 3.1, possible relations between the survival distributions of two treatment groups within the conditional survival model are derived. Estimators for the model parameters are derived and investigated in Section 3.2. The subsections of Section 3.2 comprise the assessment of the asymptotic distribution of the estimators, alternative parameterizations of model parameters, approximate normality of the estimators for different parameterizations, correlation of the estimators, approximate confidence intervals for the model parameters, and the exact distribution of the estimators. In Section 3.3, testing the difference between two treatment groups is investigated. The subsections of Section 3.3 comprise the development of an approximate and an exact test for the difference between model parameters, the derivation of formulas for the exact calculation of rejection probabilities of these tests, and the construction of approximate confidence intervals for the differences of model parameters. In Section 3.4, the derived approximate test is investigated. The subsections of Section 3.4 comprise the assessment and comparison to the logrank test and the stratified logrank test regarding Type I error rate and power, and the evaluation of coverage probabilities of the derived confidence intervals for parameter differences. Additionally, since the approximate testing procedure consists in simultaneously testing three local hypotheses, the correlation of the local test decisions is evaluated. In Section 3.5, an approximate and an exact sample size calculation method are derived for the derived approximate test. In

Section 3.6, the derived approximate sample size calculation method and the derived approximate test are applied to a clinical example. Section 4 discusses the presented research and Section 5 contains a summary of this thesis in German and English. Appendix A presents technical details and further results that are not shown in the main part due to lower importance. Appendix B contains the R code of all functions used for the calculations in this thesis.

Chapter 2

Methods

In this chapter, the conditional survival model is described and notation is introduced. Furthermore, the methods deriving estimators and assessing their distribution are presented. Then, the development of an approximate and an exact hypothesis testing procedure as well as the assessment of test characteristics is described. Additionally, the methods for deriving an approximate and exact sample size calculation procedure and for assessing performance of these procedures are presented. Lastly, the application of these methods to a clinical example is described.

All calculations and simulations are done in R, version 4.2.0 (R Core Team 2022). The used R-functions are given in Appendix B.

2.1 Statistical model

The conditional survival model was proposed by Xia et al. (2014). They did not specifically name their model back then, in this thesis, however, it is denoted as the *responder stratified exponential survival* (RSES) model. Its application is motivated in modelling the survival of patients that received a cancer therapy. The therapy may have an effect on an early detectable binary surrogate endpoint like tumor response. Then, the response status may affect the survival of patients. The RSES model describes the survival time of a patient as an exponentially distributed random variable with the parameter depending on the response status of the patient. Formally, the random variable X distinguishes the responders ($X = 1$) from the non-responders ($X = 0$) and is Bernoulli distributed with probability p. The survival time $T_S|X = 1$ of a responder follows a $\text{Exp}(\lambda_1)$-distribution and the survival time $T_S|X = 0$ of a non-responder follows a $\text{Exp}(\lambda_0)$-distribution. The common density function is dependent on the three parameters p, λ_1, λ_0 and is given by

$$f_{p,\lambda_1,\lambda_0}(x,t) = x \cdot p \cdot \lambda_1 \cdot \exp(-\lambda_1 t) + (1-x) \cdot (1-p) \cdot \lambda_0 \exp(-\lambda_0 t).$$

By summing over x, the marginalized distribution function of the survival time is obtained:

$$\tilde{F}_{p,\lambda_1,\lambda_0}(t) = p \cdot (1 - \exp(-\lambda_1 t)) + (1 - p) \cdot (1 - \exp(-\lambda_0 t))$$

Thus, the survival function is

$$\begin{aligned} S_{p,\lambda_1,\lambda_0}(t) &= 1 - \tilde{F}_{p,\lambda_1,\lambda_0}(t) \\ &= p \cdot \exp(-\lambda_1 t) + (1 - p) \cdot \exp(-\lambda_0 t). \end{aligned} \tag{2.1}$$

In a clinical trial, two cancer therapies may be compared regarding their survival and RSES model parameters. Figure 1 visualizes the model in two treatment groups, an experimental group E and a control group C. Differences within the three-parameter model may not be easily interpretable. Specifically, a difference in parameter sets does not imply a global survival benefit of one group neither does it imply a survival difference between groups. In Section 3.1, possible relations between the survival distributions as a function of the RSES model parameters are derived by comparing the derivatives and the asymptotic behavior of the survival functions.

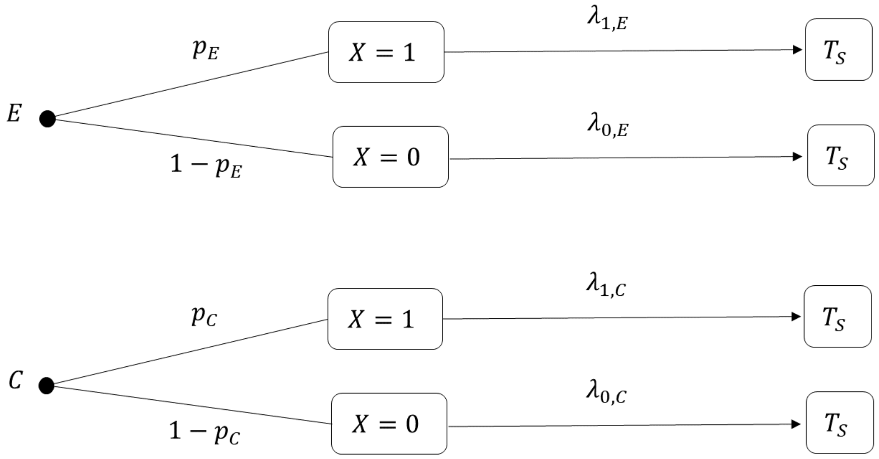


Figure 1: Two-group RSES model. Experimental (E) and control (C) group are each defined by a set of three parameters. X denotes the response status and T_S the survival time.

2.2 Estimation of model parameters

Estimators for the model parameters p, λ_1, λ_0 are derived in Section 3.2 by the Maximum Likelihood method. The maximum of the likelihood is found by finding the root of the derivative of the logarithmic likelihood. The derived estimators are investigated in the subsections of Section 3.2.

2.2.1 Asymptotic distribution of MLEs

In Section 3.2.1, a convergence theorem from the literature is applied to derive the asymptotic joint distribution of the Maximum Likelihood estimators (MLEs). The asymptotic variance and covariance structure of the MLEs is derived by calculating the inverse of the Fisher information matrix. Estimators of the variance of the MLEs are constructed by plugging the MLEs into the derived variance formulas.

2.2.2 Alternative parameterization of survival parameters

In Section 3.2.2, the same procedure as described in Section 2.2.1 is repeated for two further parameterizations of the exponential survival within the response strata $j = 0, 1$: using $\theta_j := \log(\lambda_j)$ or $\eta_j := 1/\lambda_j$. This is done to choose the parameterization with the best normal approximation for the construction of approximate confidence intervals, hypothesis tests, and power formulas.

2.2.3 First simulation study

A simulation study is used to investigate the approximate normality of the standardized MLEs, the correlation of the MLEs, and the coverage probability of approximate confidence intervals for the model parameters. It is called *first* simulation study throughout this thesis to differentiate it from further simulation studies. For the first simulation study, seven sample sizes ($n = 50, 70, 100, 200, 300, 400, 500$) and two different response probabilities ($p = 0.2, 0.5$) are considered. λ_0 is set to 0.037 and two different values of λ_1 $(0.037, 0.02)$ are considered. The choice of these values is based on the example in Section 3.6. There, the RSES model parameters are extracted from a clinical study (Huober et al. 2019). In each of the resulting 28 scenarios, 10^5 studies are simulated. In each study, four different censoring distributions are considered: no censoring, Weibull$(1/0.018, 2)$, Exp(0.02), Exp(0.04), in descending order of extent of censoring. These censoring distributions are chosen to cover different extents and types of censoring. Table 1 shows the probability to observe an event for one specific patient for these censoring distributions and values of λ.

The binary response variable X is generated by the function `rbinom()` of the R package `stats` with the sample size and response probability specified in the scenario. The event times T_S within the response strata are generated by the function `rexp()` of the R package `stats` with the hazard rates specified in the

Table 1: Probability of observing an event for different censoring distributions and Exp(λ)-distributed survival time.

Censoring distribution	$\lambda = 0.037$	$\lambda = 0.02$
no cens.	1.00	1.00
Exp(0.04)	0.77	0.58
Weibull(1/0.018, 2)	0.65	0.50
Exp(0.02)	0.48	0.33

scenario. The exponential censoring times U are also generated by the function `rexp()`. The Weibull censoring times are generated by the function `rweibull()` of the R package `stats`. All random variables are generated with a seed to assure reproducibility.

2.2.4 Approximate normality of MLEs

In Section 3.2.3, the approximate normality of the standardized MLEs for λ_j, θ_j and η_j is investigated by the first simulation study. For investigating approximate normality, the MLEs are standardized with the Wald-method (Lehmann and Romano 2010, p. 508) by $(\hat{u} - u)/\hat{\sigma}_{\hat{u}}$ with u denoting the parameter λ_1, η_1, or θ_1, $\hat{u}$ being the MLE of u and $\hat{\sigma}_{\hat{u}}$ being the square root of the variance estimators derived in Sections 3.2.1 and 3.2.2. This standardization is chosen as it is also used for constructing approximate confidence intervals, hypothesis tests of parameter difference, and power formulas. In every trial of the first simulation study, the MLEs $\hat{\lambda}_j, \hat{\theta}_j$ and $\hat{\eta}_j$ and their standardized versions are calculated. The distribution of the standardized MLEs is visualized by quantile-quantile plots, using the standard normal distribution as reference distribution. For the following calculations and considerations, the parameterization (p, θ_1, θ_0) of the RSES model is used as the distribution of $\hat{\theta}_j$ shows the best normal approximation. Using the results of the first simulation study, the estimated bias and root mean squared error (RMSE) of $\hat{\theta}_j$ is calculated and visualized with respect to sample size by line plots.

2.2.5 Correlation of MLEs

In Section 3.2.4, the pairwise Pearson correlation coefficients of $\hat{p}, \hat{\theta}_1$ and $\hat{\theta}_0$ are estimated by the sample correlation coefficients within the first simulation study. For this, the R function `cor()` of the R package `stats` is used. Furthermore, confidence intervals for the sample correlation coefficient are estimated by the R function `cor.test()` of the R package `stats`. In this function, the sample correlation coefficient $\widehat{\mathrm{Cor}}$ is transformed by the Fisher transformation with the inverse hyperbolic tangent arctanh. The transformed value is approximately normally distributed with standard error $\sqrt{n-3}$, where n is the sample size. Thus, the approximate 95% confidence interval on the transformed scale is given by $\operatorname{arctanh}(\widehat{\mathrm{Cor}}) \pm z_{0.975} \cdot \sqrt{n-3}$, where $\widehat{\mathrm{Cor}}$ is the sample correlation coefficient. This interval is back-transformed by tanh to obtain a confidence interval for the true correlation coefficient. Estimated correlation coefficients and confidence intervals with respect to sample size are shown with line plots for each of the three pairwise comparisons of $\hat{p}, \hat{\theta}_1$ and $\hat{\theta}_0$.

2.2.6 Approximate confidence intervals for model parameters

In Section 3.2.5, approximate two-sided $1 - \alpha$ confidence intervals for the parameters p, θ_1, and θ_0 are derived by using the approximate normality of the standardized MLEs $(\hat{u} - u)/\hat{\sigma}_{\hat{u}}$. For the confidence interval for p, a formula for the exact coverage probability is given and the exact coverage probability is visualized with respect to sample size for $\alpha = 0.05, p = 0.1, 0.2, 0.3, 0.4, 0.5$ and $n = 50, 70, 100, 200, 300, 400, 500$ by line plots. For θ_1 and θ_0, the coverage probability of the derived approximate 95% confidence intervals is estimated by the first simulation study described in Section 2.2.3. Here, the coverage probability conditional on the existence of $\hat{\theta}_j$ is estimated. This probability can be interpreted in the following way: Under the condition that $\hat{\theta}_j$ exists, how likely will the confidence interval of θ_j cover the true value? If ω is the true coverage probability, the estimated coverage probability has a standard error of $\sqrt{\omega(1-\omega)/n_{\text{sim}}}$, where $n_{\text{sim}} = 10^5$ is the number of simulations per scenario. This results in a maximal standard error of 0.0016 for $\omega = 0.5$. Assuming a true coverage probability of $\omega \approx 0.95$ yields a standard error of 0.0007.

2.2.7 Exact distribution of MLEs

In Section 3.2.6, the exact distribution of θ_1 and θ_0 in the case of no censoring or exponential censoring distribution is derived by using known distributions of the transformation and combination of exponentially distributed random variables. The asymptotic behavior of the expected value and variance is derived by using theorems from the literature like the Continuous Mapping Theorem, Hoeffding's inequality, or the laws of total expectation, variance, and covariance. In the same way, the asymptotic behavior of the pairwise covariance and correlation coefficient of $\hat{p}, \hat{\theta}_1$ and $\hat{\theta}_0$ is derived. These findings are used to derive formulas for calculating the exact coverage probability of the approximate confidence intervals for $\hat{\theta}_1$ and $\hat{\theta}_0$ in the case of no censoring or exponential censoring. These formulas are validated by the following procedure: In each scenario, the exact coverage probability ω is calculated with the derived formula. With the large number of simulations $n_{\text{sim}} = 10^5$, the estimated coverage probability is almost exactly normally distributed with mean ω and standard error $\sqrt{\omega(1-\omega)/n_{\text{sim}}}$. Note that the calculation of coverage probability is conditional on the existence of $\hat{\theta}_j$ and thus the number of effective simulations is smaller than n_{sim}. Hence, when investigating the deviation of estimated coverage probabilities from exactly calculated coverage probabilities, the use of the larger value n_{sim} constitutes a conservative approach. However, the cases where $\hat{\theta}_j$ doesn't exist are very rare. Based on the normal distribution, 2.5% and 97.5% quantiles of the distribution of the estimated coverage probability are calculated. These quantiles and the exact coverage probability are visualized together with the estimated coverage probabilities. If the calculations are correct, the estimated coverage probabilities should lie between the two quantiles in most cases.

2.3 Hypothesis testing

In Section 3.3, an approximate test for the difference between two treatment groups within the RSES model is developed. Furthermore, formulas for the exact calculation of rejection probabilities in the case of no censoring or exponential censoring are derived. Additionally, approximate confidence intervals for the differences of model parameters are constructed. Lastly, an exact testing procedure is derived.

There are different ways to formulate a null hypothesis regarding the RSES model. Xia et al. (2014) used the logrank test and stratified logrank test to assess differences between survival distributions. In this thesis, the difference of parameter sets is tested. Consider the comparison of an experimental group

E and a control group C with respective parameter triples $(p_E, \theta_{1,E}, \theta_{0,E})$ and $(p_C, \theta_{1,C}, \theta_{0,C})$. In Section 3.3, a test is constructed for the null hypothesis

$$H_0 : p_C = p_E,\ \theta_{1,C} = \theta_{1,E} \text{ and } \theta_{0,C} = \theta_{0,E}.$$

Note that this is not a necessary condition for equality of the marginal survival distributions as is shown in Section 3.1. This procedure aims at detecting any group difference within the parametric model. In particular, it is not meant to make inference about differences of single parameters because they can only be interpreted and translated to survival difference when considered as a triple.

The global null hypothesis H_0 is an intersection of three local null hypotheses:

$$\begin{aligned} H_{p,0}&\text{: } p_C = p_E \\ H_{\theta_1,0}&\text{: } \theta_{1,C} = \theta_{1,E} \\ H_{\theta_0,0}&\text{: } \theta_{0,C} = \theta_{0,E} \end{aligned}$$

2.3.1 Approximate RSES test

In Section 3.3.1, for the local hypotheses $H_{p,0}, H_{\theta_1,0}, H_{\theta_0,0}$, test statistics $T_p, T_{\theta_1}, T_{\theta_0}$ similar to Wald test statistics (Lehmann and Romano 2010, p. 508) are constructed by standardizing the difference between the MLEs of both groups. The difference is divided by an estimator of the standard deviation of the difference. Here, the asymptotic variance formulas derived in Section 3.2 are used. For variance estimation, estimates of the true parameters have to be plugged into the formulas. For this, the true parameters are estimated under the local null hypothesis, i.e. under equality of the respective parameter in both treatment groups. This is done to provide more accurate variance estimation under the local null hypothesis so that the distribution of the test statistics is closer to the standard normal distribution under the local null hypothesis. This approximate normality is then used to derive the tests of the local hypotheses. Then, a testing strategy is derived to test the global null hypothesis. This testing procedure is called *approximate RSES test*.

2.3.2 Exact calculation of rejection probability

In Section 3.3.2, formulas for the calculation of exact rejection probabilities of the approximate RSES test for the cases of no censoring and exponential censoring are derived. This is done by using the exact distributions of the MLEs derived in Section 3.2.6, and applying some analytical calculations. The derived formulas are validated by the results of the second simulation study described in Section 2.4. The validation of the formulas is done similar to the approach described in Section 2.2 for validating the formulas for the exact coverage probability of approximate confidence intervals. This means that in different scenarios, exact rejection probabilities are calculated to derive the distribution of the estimated rejection probabilities. Then, the estimated rejection probabilities are compared visually to the exactly calculated rejection probabilities.

2.3.3 Approximate confidence intervals for parameter differences

The canonical effect measures for the local tests of $H_{p,0}, H_{\theta_1,0}, H_{\theta_0,0}$ are the parameter differences corresponding to the numerators of the test statistics. For $H_{p,0}$, this is the response rate difference

$$\Delta_p = p_E - p_C.$$

For $H_{\theta_1,0}$, this is the logarithmic responder hazard ratio

$$\Delta_{\theta_1} = \theta_{1,E} - \theta_{1,C}.$$

For $H_{\theta_0,0}$, this is the logarithmic non-responder hazard ratio

$$\Delta_{\theta_0} = \theta_{0,E} - \theta_{0,C}.$$

In Section 3.3.3, approximate two-sided $1 - \alpha$ confidence intervals for the parameter differences are derived by using the approximate normality of the MLEs and the asymptotic variance formulas derived in Section 3.2. For variance estimation, estimates of the true parameters have to be plugged into the formulas. Here, in contrast to the approach used for the local test statistics, the true parameters are not estimated under the assumption of equality of the parameters. Instead, the true parameters are estimated separately in both groups. This is done to provide good coverage probability of the confidence intervals over the whole range of parameter constellations. However, this means that the confidence intervals are only approximately equivalent to the test decisions.

2.3.4 Exact RSES test

In Section 3.3.4, an exact test of the global null hypothesis H_0 in the case of no censoring is constructed. Since the case of no censoring is very rare in practice, this test won't be useful for application but is given in this thesis for completeness. The exact distribution of monotone transformations of the test statistics T_{θ_1} and T_{θ_0} conditional on the numbers of responders k_E and k_C in each treatment group is derived. On this basis, an exact testing strategy based on exact p-values is derived. Furthermore, the exactness of the derived test is proven. A formula to calculate the exact rejection probability of the exact RSES test is derived. With this formula, Type I error rate and power is calculated in the scenarios of the second simulation study described in Section 2.4. However, due to the computationally extensiveness of the calculations, only smaller sample sizes of $n_E = n_C = 50, 60, 70, 80, 90, 100$ are considered.

2.4 Assessment of test characteristics

In Section 3.4, Type I error rate and power of the approximate test are analysed and compared to the logrank test and the stratified logrank test. See Appendix A.2 for details of the used logrank test statistics. Note that stratifying the logrank test for response status deliberately ignores a survival benefit originating from a response benefit. Thus, the stratified logrank test is not appropriate in this setting. However, it is included for comparison purposes because Xia et al. (2014) considered it as well. Furthermore, coverage probabilities of the approximate confidence intervals for parameter differences are evaluated. Lastly, the correlation of the local test decisions is evaluated.

2.4.1 Second simulation study

Type I error rate, power, coverage probabilities, and the correlation of the local RSES tests are estimated by a simulation study which is called *second* simulation study throughout this thesis. Seven sample sizes ($n_E = n_C = 50, 70, 100, 200, 300, 400, 500$), six Type I error scenarios TOE1.1, ..., TOE2.3, and five power scenarios Pow1.1, ..., Pow3 are considered. The Type I error scenarios and power scenarios are described in more detail in Sections 2.4.2 and 2.4.3. Like in the first simulation study described in Section 2.2.3, four different censoring distributions are considered: no censoring, Weibull$(1/0.018, 2)$, Exp(0.02), Exp(0.04), in descending order of extent of censoring.

The data generation in the second simulation study is similar to the first simulation study. In each of the 11 scenarios, 10^5 studies are simulated. In each study, the binary response variable is generated in both treatment groups. Then, the event times within the treatment groups and response strata are generated. For each of the four censoring distributions, censoring times are generated. All random variables are generated with a seed to assure reproducibility. For each study and censoring distribution, the approximate RSES test, logrank test, and stratified logrank test are applied at a significance level of $\alpha = 0.05$. For the application of the approximate RSES test, the local tests of the local hypotheses $H_{p,0}, H_{\theta_1,0}, H_{\theta_0,0}$ are applied at a local level $\tilde{\alpha}$ as described in more detail in Section 3.3.1. The rejections of these local tests at level $\tilde{\alpha}$ are assessed to investigate the independence of the local tests, as described in Section 2.4.5. Furthermore, the approximate confidence intervals for parameter difference derived in Section 3.3.3 are calculated. As described in Section 2.2, the standard error of the estimated probabilities can be calculated by $\sqrt{\omega(1-\omega)/n_{\text{sim}}}$, where ω is the true probability and $n_{\text{sim}} = 10^5$ the number of simulations per scenario. This results in a maximal standard error of 0.0016 if $\omega = 0.5$. For Type I error rate and confidence interval coverage probability, a true probability of $\omega \approx 0.05$ or $\omega \approx 0.95$ can be assumed. This yields a standard error of 0.0007.

2.4.2 Assessment of Type I error rate

Type I error rate is investigated in Section 3.4.1. In all six Type I error scenarios, parameters in experimental and control group are equal. In the first three scenarios, denoted by TOE1.1, TOE1.2, and TOE1.3, the survival of responders and non-responders is equal with $\lambda_1 = \lambda_0 = 0.071$. The response probability varies over the three values $p = 0.13, 0.26, 0.52$. In the last three scenarios, denoted by TOE2.1, TOE2.2, and TOE2.3, the responder survival is given by $\lambda_1 = 0.0284$ and the non-responder survival by $\lambda_0 = 0.071$. This corresponds to a hazard ratio of 0.4 between responders and non-responders. Again, the response probability varies over the three values $p = 0.13, 0.26, 0.52$. These assumptions for p, λ_1, λ_0 were chosen for comparability because Xia et al. (2014) used these values for their simulation study. They based these choices on the findings of a clinical trial that investigated the effect of different combinations of chemotherapy, followed by surgical tumor removal, on breast cancer response rates, disease-free survival, and overall survival (Bear et al. 2006). Figure 2 shows the marginal survival distributions in both scenarios. Scenarios TOE1.1, TOE1.2, and TOE1.3 have equal marginal survival distributions since responders and non-responders have equal survival.

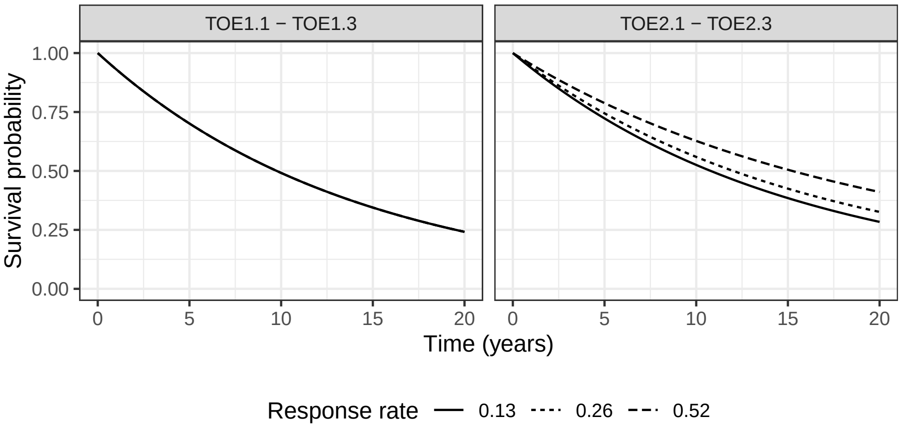


Figure 2: Marginal survival distribution of six scenarios for assessing Type I error rate. Survival in scenarios TOE1.1 - TOE1.3 is equal. Survival in scenarios TOE2.1 - TOE2.3 is better for higher response rates.

In Section 3.4.1, the estimated Type I error rates for the six Type I error scenarios are visualized with respect to sample size and censoring distribution, and compared between the tests for each scenario.

2.4.3 Assessment of Power

Power is assessed in Section 3.4.2. In all five power scenarios, responder survival is better than non-responder survival with a hazard ratio of $\lambda_{1,i}/\lambda_{0,i} = 0.4, i = E, C$. In the first two scenarios (+resp, Pow1.1: $p_E = 0.26$, Pow1.2: $p_E = 0.52$), survival benefit of the experimental group is solely due to a higher response probability ($p_E = 0.26, 0.52$ vs. $p_C = 0.13$). In scenarios 3 and 4 (+resp +surv, Pow2.1: $p_E = 0.26$, Pow2.2: $p_E = 0.52$), there is additionally a survival benefit of the experimental group within the strata. That means that responders in the experimental group have a better survival than responders in the control group and non-responders in the experimental group have a better survival than non-responders in the control group. The hazard ratio between the treatment groups within the response strata was chosen to be $\lambda_{j,E}/\lambda_{j,C} = 0.8, j = 0, 1$, as this value was also used in the simulation study of Xia et al. (2014). In Scenario 5 (+surv, Pow3: $p_E = 0.13$), the response probabilities are equal in both groups. The survival benefit of the experimental group is solely due to a better survival of both responders and non-responders. The hazard ratio was again chosen to be $\lambda_{j,E}/\lambda_{j,C} = 0.8, j = 0, 1$. Figure 3 shows the marginal survival distributions in experimental and control group in all five scenarios.

In Section 3.4.2, the estimated power values for the five power scenarios are

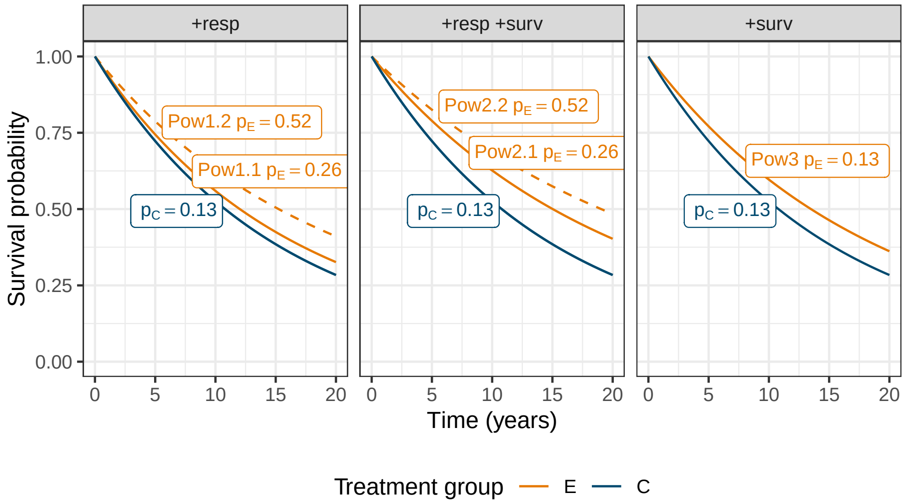


Figure 3: Marginal survival distributions of experimental and control group in five scenarios for assessing power.

visualized with respect to sample size and censoring distribution, and compared between the tests for each scenario.

2.4.4 Assessment of coverage probability of approximate confidence intervals for parameter difference

In Section 3.4.3, coverage probabilities of the 95% confidence intervals for the parameter differences derived in Section 3.3.3 are calculated and visualized in all scenarios of the second simulation study.

2.4.5 Assessment of independence of local test decisions

The derivation of the approximate RSES testing procedure is based on the approximate independence of the three local test statistics. In Section 3.4.4, the results of the second simulation study are used to investigate this assumption. Let R_p, R_{θ_1} and R_{θ_0} be the binary random variables indicating the rejection of the respective local test. For example, $R_p = 1_{|T_p| > z_{1-\alpha/2}}$, with T_p being the test statistic for the local hypothesis $H_{p,0}$. The pair of rejections R_p and R_{θ_1} is independent if and only if

$$\mathrm{P}(R_p = 1 \text{ and } R_{\theta_1} = 1) = \mathrm{P}(R_p = 1) \cdot \mathrm{P}(R_{\theta_1} = 1).$$

This is equivalent to $\mathrm{Cor}(R_p, R_{\theta_1}) = 0$ since

$$\begin{aligned} \mathrm{Cov}(R_p, R_{\theta_1}) &= \mathrm{E}[R_p \cdot R_{\theta_1}] - \mathrm{E}[R_p] \cdot \mathrm{E}[R_{\theta_1}] \\ &= \mathrm{P}(R_p = 1 \text{ and } R_{\theta_1} = 1) - \mathrm{P}(R_p = 1) \cdot \mathrm{P}(R_{\theta_1} = 1). \end{aligned} \tag{2.2}$$

Analogous formulas apply to the pairs (p, θ_0) and (θ_1, θ_0). Thus, for investigating the pairwise independence of the local test decisions, the pairwise correlation coefficients of the rejection random variables are analysed. Within the second simulation study, rejection rates $\hat{p}_{R_p}, \hat{p}_{R_{\theta_1}}$ and $\hat{p}_{R_{\theta_0}}$ are calculated to estimate the probability of $R_p = 1, R_{\theta_1} = 1$ and $R_{\theta_0} = 1$. Furthermore, the pairwise common rejection rates are calculated to estimate the probability of the pairwise intersections of these events. The pairwise covariances are estimated by plugging these estimates into equation (2.2). The variances of $\hat{p}_{R_p}, \hat{p}_{R_{\theta_1}}$ and $\hat{p}_{R_{\theta_0}}$ are estimated by the variance formula for Bernoulli variables (e.g. $\hat{p}_{R_p} \cdot (1 - \hat{p}_{R_p})$). Then, correlation coefficients are estimated by $\mathrm{Cor}(A, B) = \frac{\mathrm{Cov}(A,B)}{\sqrt{\mathrm{Var}(A) \cdot \mathrm{Var}(B)}}$. Note that this approach is equivalent to calculating the *mean square contingency* (Cramér 1974, p. 282), also known as *phi coefficient*. The standard error of the correlation estimation is estimated by bootstrap resampling with $n_{\mathrm{BT}} = 10^5$ draws. This is explained exemplarily for the correlation of R_p and R_{θ_1}: Based on the observed number of events of $\{R_p = 1\}, \{R_{\theta_1} = 1\}$ and $\{R_p = 1 \text{ and } R_{\theta_1} = 1\}$, new numbers of these events are generated by a multinomial distribution using the R function `rmultinom()` of the R package `stats` with the number of trials for one draw being n_{sim}. From these event numbers, the sample correlation coefficient is calculated. This is repeated n_{BT} times. The sample standard deviation of the resulting n_{BT} sample correlation coefficients is used as estimate for the standard error of the correlation estimation. Estimated correlation coefficients and standard errors are visualized with respect to sample size for different scenarios and censoring distributions.

2.5 Sample size calculation

2.5.1 Derivation and assessment of sample size calculation methods

In Section 3.5, an approximate and an exact sample size calculation method are derived for the approximate RSES test. The approximate method is based on the approximate normality of the test statistics of the three local tests. For the three local tests, approximate formulas for the probability to accept the respective null hypothesis are derived. Here, an assumption of the censoring probabilities in the treatment groups and response strata has to be made to take censoring into account. Due to the approximate independence of the local tests, the acceptance probability of the approximate RSES test can then be calculated as the product of the acceptance probabilities of the local tests. After specifying significance level, power, sample size ratio, model parameters, and censoring probabilities, the power formula can be solved numerically for the required sample size. This approach also leaves the possibility of splitting Type I error rate or power differently to weight certain hypotheses. The calculation method can easily be adapted to such changes.

The approximate sample size calculation method is evaluated within a simulation study that is called *third* simulation study throughout this thesis. The third simulation study comprises 29 scenarios and is described in Section 2.5.2.

The exact sample size calculation method is only applicable in the case of no censoring or exponential censoring. It is an iterative procedure where the power is calculated exactly by the formulas derived in Section 3.3.2.

In all scenarios where the approximate sample size calculation method gives sample sizes $n_E, n_C < 100$, the exact sample size is calculated by the exact sample size calculation method. Then, approximate and exact sample sizes are compared visually. Since the iterative exact method is computationally extensive for large sample sizes, exact sample sizes are not calculated in the other scenarios.

2.5.2 Third simulation study

The third simulation study comprises 29 different scenarios with four censoring distributions. In all scenarios, the non-responder hazard in the control group is set to $\lambda_{0,C} = \gamma := 0.142$. The variable γ is used for better readibility of the description of the scenarios. The sample size ratio is set to $r := n_E/n_C = 1$. The same censoring distributions as in the first and second simulation study are considered: no censoring, Weibull(1/0.018, 2), Exp(0.02), and Exp(0.04). The following six constellations of survival in treatment groups and response strata are considered:

- Constellation 1: Equal survival in all strata: $\lambda_{0,E} = \lambda_{0,C} = \lambda_{1,E} = \lambda_{1,C} = \gamma$
- Constellation 2: Better survival of responders in experimental group (1): $\lambda_{0,C} = \lambda_{0,E} = \lambda_{1,C} = \gamma$ and $\lambda_{1,E} = \gamma/2$
- Constellation 3: Better survival of responders in experimental group (2): $\lambda_{0,C} = \lambda_{0,E} = \lambda_{1,C} = \gamma$ and $\lambda_{1,E} = \gamma/3$
- Constellation 4: Better survival of responders and non-responders in experimental group: $\lambda_{0,C} = \lambda_{1,C} = \gamma$ and $\lambda_{0,E} = \lambda_{1,E} = \gamma/2$
- Constellation 5: Better survival of non-responders in experimental group, even better survival of responders in experimental group: $\lambda_{0,C} = \lambda_{1,C} = \gamma, \lambda_{0,E} = \gamma/2$ and $\lambda_{1,E} = \gamma/3$
- Constellation 6: Better survival of responders in control group, better survival of non-responders in experimental group, even better survival of responders in experimental group: $\lambda_{0,C} = \gamma, \lambda_{1,C} = \gamma/2, \lambda_{0,E} = \gamma/2$ and $\lambda_{1,E} = \gamma/3$

In each of these constellations, response probability in the control group is $p_C = 0.13$. Five different response probabilities in the experimental group are considered: $p_E = 0.13, 0.26, 0.39, 0.52, 0.8$. The one scenario with equal survival in all strata (Constellation 1) and equal response probabilities $p_E = p_C = 0.13$ is not suitable for sample size calculation since the distribution in both groups is equal. This leaves 29 scenarios in total. In each scenario and for each censoring distribution, censoring probabilities in the treatment groups and response strata are calculated. Based on this and the other scenario parameters, approximate sample sizes are calculated by the derived method to obtain a power of 0.8 at a significance level of 0.05.

Data generation is similar to the second simulation study. In each scenario, $n_{\text{sim}} = 10^5$ studies are simulated using the calculated sample sizes and distribution parameters. The approximate RSES test is applied to each study and the actual power is estimated for each scenario and censoring distribution. As described before, the standard error of the power estimate can be calculated by $\sqrt{\omega(1-\omega)/n_{\text{sim}}}$, where ω is the true power. This results in a maximum standard error of 0.0016. If the true power is $\omega \approx 0.8$, the standard error is approximately 0.0025. The calculated sample sizes are shown and the estimated power is visually compared with the desired value of 0.8 for each scenario and censoring distribution.

2.6 Example

In Section 3.6, the derived methods are applied to a clinical example. Huober et al. (2019) investigated the effect of three treatments on pathological complete response and survival in patients with HER2-positive early breast cancer. They investigated differences in event-free survival and overall survival between the groups by Cox regression. In Section 3.6, the RSES model parameters for the three groups are calculated from the reported results in Huober et al. (2019) by solving the equations for the survival function. The resulting RSES model distributions in the three treatment groups are visually described. Furthermore, the censoring distribution in the trial is derived under the assumption of exponential censoring by analytical calculations. For the following considerations, each of the three pairwise comparisons of the three treatment groups is investigated separately. For every pairwise comparison, the distribution assumptions are used to calculate approximate sample sizes of the approximate RSES test to obtain a power of 0.8 at a significance level of 0.05. For these sample sizes and distribution parameters, power of the approximate RSES test, logrank test, and stratified logrank test is estimated by a simulation study that is called *fourth* simulation study in this thesis. The data generation and power estimation is similar to the third simulation study described in Section 2.5.2. The estimated power is described and compared between the three tests. The approximate RSES test rejects the global null hypothesis H_0 if one of the three local hypotheses $H_{p,0}, H_{\theta_1,0}, H_{\theta_0,0}$ is rejected. To describe the influence of the corresponding local tests on the rejection of the global null hypothesis, the rejection rates of the local hypotheses within the simulation study are described.

Chapter 3

Results

In this chapter, possible relations between survival functions of two treatment groups within the RSES model are investigated. Furthermore, estimators for model parameters are derived and their distribution is assessed. Then, an approximate and an exact hypothesis testing procedure are developed and the test characteristics are assessed. Additionally, an approximate and an exact sample size calculation procedure are developed and the performance of these procedures is assessed. Lastly, some of these methods are applied to a clinical example.

3.1 Survival differences between two treatment groups

When two therapies are compared regarding their survival and RSES model parameters, different kinds of relations can occur. These relations are investigated in this section. Again, E denotes the experimental group, C the control group, and $i = E, C$ the group indicator. Let $p_i, \lambda_{1,i}, \lambda_{0,i}$ be the respective parameter sets of the groups as described in Section 2.1 and shown in Figure 1. For better readability, the survival functions $S_{p_i,\lambda_{1,i},\lambda_{0,i}}$ are abbreviated as S_i. Three cases of the relation of S_E and S_C can be differentiated:

I. Completely equal: $S_E(t) = S_C(t) \quad \forall t \geq 0$
II. Uniformly different: $S_E(t) \neq S_C(t) \quad \forall t > 0$
III. Crossing: not completely equal but $\exists t > 0$ such that $S_E(t) = S_C(t)$

It is easily seen from formula (2.1) that S_E and S_C are completely equal if and only if one of the following conditions hold:

1. $p_E = p_C, \lambda_{1,E} = \lambda_{1,C}$ and $\lambda_{0,E} = \lambda_{0,C}$
2. $p_E = p_C = 0$ and $\lambda_{0,E} = \lambda_{0,C}$
3. $p_E = p_C = 1$ and $\lambda_{1,E} = \lambda_{1,C}$
4. $\lambda_{1,E} = \lambda_{1,C} = \lambda_{0,E} = \lambda_{0,C}$

This can be reformulated as:

1. The parameter sets in both groups are equal.

2. There are no responders in both groups and the non-responder parameters are equal in both groups.
3. There are no non-responders in both groups and the responder parameters are equal in both groups.
4. Survival is equal for responders and non-responders and in both groups.

To investigate whether two survival curves cross for some $t > 0$, the relation at $t = 0$ has to be compared with the relation at $t \to \infty$. If the distributions are not completely equal, the relation at $t = 0$ is determined by the derivations at $t = 0$:

$$S_i'(0) = -\lambda_{1,i}p_i - \lambda_{0,i}(1 - p_i) \tag{3.1}$$

This term can be interpreted as a weighted common hazard of responders and non-responders. If $S_E'(0) < S_C'(0)$, it is $S_E(t) < S_C(t)$ for small $t > 0$ (and vice versa). If $S_E'(0) = S_C'(0)$, the first non-equal pair of the higher derivatives is decisive, e.g. S_i''.

The survival functions S_i are decreasing and convex. So S_E and S_C can only cross two times. One time is at $t = 0$. The other time occurs for $t > 0$ if and only if $S_E(t) < S_C(t)$ for small $t > 0$ and $S_E(t) > S_C(t)$ for sufficiently large t (or vice versa). The first condition can be determined by the derivations at $t = 0$. The second condition can be determined by analysing the asymptotic behavior for $t \to \infty$. Suppose a function

$$g(t) = a\exp(-ct) + b\exp(-dt)$$

with $a, b, c, d > 0$ and $d > c$. In the RSES model, c and d correspond to the parameters λ_1 and λ_0, with c being the smaller of both parameters. It is

$$g(t) = a\exp(-ct) + b\exp(-ct)^{d/c}.$$

Due to $d/c > 1$ and $\exp(-ct) \xrightarrow{t\to\infty} 0$, the asymptotic behavior of g is determined by the first term. For $t \gg 0$ it is

$$g(t) \leq M\exp(-ct)$$

for some constant M. Also,

$$g(t) \geq a\exp(-ct).$$

Hence, the asymptotic relation between two survival curves S_E and S_C is determined by $a\exp(-ct)$, i.e. the hazard and proportion of the fitter stratum. Let $\lambda_{\min,i}$ be the hazard of the fitter stratum in group i, i.e.

$$\lambda_{\min,i} := \min(\lambda_{1,i}, \lambda_{0,i}).$$

In most practical cases, this will be the responder stratum. In the special case $p_i = 0$, set $\lambda_{\min,i} := \lambda_{0,i}$. Analogously, if $p_i = 1$, set $\lambda_{\min,i} := \lambda_{1,i}$. Define $\lambda_{\max,i}$ analogously. Let $p_{i,\lambda_{\min}} = p_i$ if $\lambda_{\min,i} = \lambda_{1,i}$ and $p_{i,\lambda_{\min}} = 1 - p_i$ if $\lambda_{\min,i} = \lambda_{0,i}$ be the proportion of the fitter stratum. If $\lambda_{\min,E} > \lambda_{\min,C}$, then $S_E(t) < S_C(t)$ for $t \gg 0$ (and vice versa). If $\lambda_{\min,E} = \lambda_{\min,C}$, the proportion of the fitter stratum is decisive: If $p_{E,\lambda_{\min}} > p_{C,\lambda_{\min}}$, then $S_E(t) > S_C(t)$ for $t \gg 0$. If the proportions of the fitter stratum are also equal, the maximum hazards are decisive for the asymptotic behavior for $t \to \infty$. In this case, due to $\lambda_{\min,E} = \lambda_{\min,C}$, $p_{E,\lambda_{\min}} = p_{C,\lambda_{\min}}$ and formula (3.1), the derivations of the survival distributions at $t = 0$ are

$$S'_E(0) = -\lambda_{\max,E} \cdot (1 - p_{E,\lambda_{\min}}) - \lambda_{\min,E} \cdot p_{E,\lambda_{\min}}$$

and

$$S'_C(0) = -\lambda_{\max,C} \cdot (1 - p_{E,\lambda_{\min}}) - \lambda_{\min,E} \cdot p_{E,\lambda_{\min}}.$$

Thus, $S'_E(0) < S'_C(0)$ if and only if $\lambda_{\max,E} > \lambda_{\max,C}$. Since that means $S_E(t) < S_C(t)$ for $t \gg 0$, the curves cannot cross in this case and are uniformly different. If the maximum hazards are also equal, i.e. $\lambda_{\max,E} = \lambda_{\max,C}$, the two curves are completely equal.

To summarise, the curves don't cross if and only if one of the following conditions is true (assuming the curves are not completely equal):

1. The first non-equal derivatives at $t = 0$ fulfill $S_E^{(m)}(0) > S_C^{(m)}(0)$ and one of the following statements is true:
   - $\lambda_{\min,E} < \lambda_{\min,C}$
   - $\lambda_{\min,E} = \lambda_{\min,C}$ and $p_{E,\lambda_{\min}} > p_{C,\lambda_{\min}}$
2. The previous condition with E and C exchanged.

This can be reformulated as:

1. Event rate at the beginning is higher in group E and:
   - The fitter stratum in group E has better survival than the fitter stratum in group C or
   - the fitter strata in both groups have equal survival but the fitter stratum in group E is larger than in group C.
2. The previous condition with E and C exchanged.

Two curves cross if and only if they are not completely equal and not uniformly different.

Figure 4 shows two examples. In the first, the experimental group has a higher response probability and better survival of each responders and non-responders. Thus, survival in the experimental group is uniformly better. In the second example, responder survival is better in the control group. This advantage comes into effect after a certain time which leads to crossing survival curves.

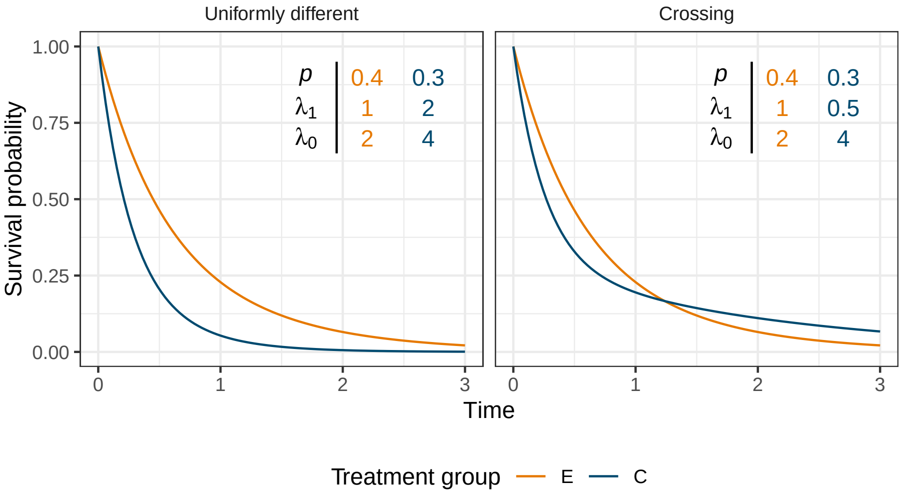


Figure 4: Survival functions of two different two-group models. The corresponding model parameters p, λ_1, λ_0 are given next to the curves.

3.2 Estimation of model parameters

In this section, estimators of the model parameters are derived by the Maximum Likelihood method. The subsections comprise the assessment of the asymptotic distribution of the MLEs, alternative parameterizations of model parameters, approximate normality of the MLEs for different parameterizations, correlation of the MLEs, approximate confidence intervals for the model parameters, and the exact distribution of the MLEs. Let X be the response status, T_S the survival time, U the censoring time, $D = 1_{T_S \leq U}$ the event indicator, and $T = \min(T_S, U)$ the observed event-free time. Assume identical censoring over responders and non-responders and independence of U and T_S. The common density of (X, T, D) is

$$f_{(X,T,D)}(x,t,d) = \begin{cases} p \cdot f_{T_S|_{X=1}}(t) \cdot (1 - F_U(t)) & x = 1, d = 1, \\ p \cdot (1 - F_{T_S|_{X=1}}(t)) \cdot f_U(t) & x = 1, d = 0, \\ (1-p) \cdot f_{T_S|_{X=0}}(t) \cdot (1 - F_U(t)) & x = 0, d = 1, \\ (1-p) \cdot (1 - F_{T_S|_{X=0}}(t)) \cdot f_U(t) & x = 0, d = 0, \end{cases}$$

where f and F are the density and distribution functions of T_S or U. The common density can be written in closed form as

$$\begin{aligned} f_{(X,T,D)}(x,t,d) = p^x (1-p)^{1-x} &\cdot f_{T_S|_{X=1}}(t)^{xd} \cdot (1 - F_{T_S|_{X=1}}(t))^{x(1-d)} \\ &\cdot f_{T_S|_{X=0}}(t)^{(1-x)d} \cdot (1 - F_{T_S|_{X=0}}(t))^{(1-x)(1-d)} \\ &\cdot f_U(t)^{1-d} \cdot (1 - F_U(t))^d. \end{aligned}$$

Since $1 - F_{T_S}(t) = \exp(-\lambda t)$ for an Exp(λ)-distributed variable T_S, the log likelihood of one realization (x, t, d) is

$$\begin{aligned}\log\left(L_{(x,t,d)}(p, \lambda_1, \lambda_0)\right) = {} & x\log(p) + (1-x)\log(1-p) \\ & + xd\log(\lambda_1) - x\lambda_1 t \\ & + (1-x)d\log(\lambda_0) - (1-x)\lambda_0 t \\ & + (1-d)\log(f_U(t)) + d\log(1-F_U(t)).\end{aligned} \tag{3.2}$$

The parameter-independent part $(1-d)\log(f_U(t)) + d\log(1-F_U(t))$ is irrelevant for estimation. For the sake of readability, it will be omitted in the following likelihoods. Suppose n realisations $(x_m, t_m, d_m), m = 1, \ldots, n$ of (X, T, D). Let $k = \sum_{m=1}^{n} x_m$ be the number of responders, $l_1 = \sum_{m=1}^{n} x_m \cdot d_m$ the number of uncensored responder survival times and $l_0 = \sum_{m=1}^{n}(1-x_m)\cdot d_m$ the number of uncensored non-responder survival times. The observations are arranged such that $x_1 = \cdots = x_k = 1$ and $t_1, \ldots, t_{l_1}$ are the uncensored responder survival times. Analogously, $x_{k+1} = \cdots = x_n = 0$ and $t_{k+1}, \ldots, t_{k+l_0}$ are the uncensored non-responder survival times. The likelihood of the n realisations then is

$$\begin{aligned}\prod_{m=1}^{n} L_{(x_m,t_m,d_m)}(p, \lambda_1, \lambda_0) = p^k \cdot \prod_{m=1}^{l_1} \lambda_1 \exp(-\lambda_1 t_m) \cdot \prod_{m=l_1+1}^{k} \exp(-\lambda_1 t_m) \\ \cdot \prod_{m=k+1}^{k+l_0} \lambda_0 \exp(-\lambda_0 t_m) \cdot \prod_{m=k+l_0+1}^{n} \exp(-\lambda_0 t_m).\end{aligned}$$

Hence, the log likelihood of the three-dimensional parameter $(p, \lambda_1, \lambda_0)$ is:

$$\begin{aligned}\log\left(\prod_{m=1}^{n} L_{(x_m,t_m,d_m)}(p, \lambda_1, \lambda_0)\right) = {} & k \cdot \log(p) + (n-k)\cdot\log(1-p) \\ & + l_1 \log(\lambda_1) - \lambda_1 \cdot \sum_{m=1}^{k} t_m \\ & + l_0 \log(\lambda_0) - \lambda_0 \cdot \sum_{m=k+1}^{n} t_m\end{aligned}$$

Therefore, the argument of the maximum can be found within the three separate summands. Finding the roots of the derivatives of the summands yields the following Maximum Likelihood estimators (MLEs):

$$\begin{aligned}\hat{p} &= \frac{k}{n}, \\ \hat{\lambda}_1 &= \frac{1}{\frac{1}{l_1}\sum\limits_{m=1}^{k} t_m}, \\ \hat{\lambda}_0 &= \frac{1}{\frac{1}{l_0}\sum\limits_{m=k+1}^{n} t_m}.\end{aligned}$$

Note that the estimation of $\hat{\lambda}_1$ is only well-defined if $l_1 \neq 0$. In the case $l_1 = 0$ there is no unique MLE for λ_1. The same is true for $\hat{\lambda}_0$ if $l_0 = 0$. Thus, technically, the MLE of the three-dimensional parameter $(p, \lambda_1, \lambda_0)$ only exists if $0 < k < n$. This won't matter in practice since studies will be designed such that the cases "all responders" or "no responders" have low probability.

3.2.1 Asymptotic distribution of MLEs

In this section, the asymptotic distribution of the MLEs is derived. Let $\vartheta := (p, \lambda_1, \lambda_0)$ be the three-dimensional model parameter. It is known that the asymptotic distribution of a multi-parameter MLE is multivariate normal (L 1983, p. 429-430):

$$\sqrt{n}(\hat{\vartheta} - \vartheta) \overset{D}{\to} N(0, I(\vartheta)^{-1})$$

$I(\vartheta)$ is the Fisher information matrix and is obtained by taking the negative expectation of the second derivation of the log likelihood:

$$I(\vartheta)_{i,j} = -\mathrm{E}\left[\frac{d^2}{d\vartheta_i d\vartheta_j} \log L_{(X,T,D)}(\vartheta)\right]$$

The log likelihood was derived in equation (3.2). The derivations are

$$\begin{aligned}
\frac{d^2}{dp^2} \log L_{(X,T,D)}(\vartheta) &= -\frac{X}{p^2} + \frac{1-X}{(1-p)^2} \\
\frac{d^2}{d\lambda_1^2} \log L_{(X,T,D)}(\vartheta) &= -\frac{XD}{\lambda_1^2} \\
\frac{d^2}{d\lambda_0^2} \log L_{(X,T,D)}(\vartheta) &= -\frac{(1-X)D}{\lambda_0^2}.
\end{aligned}$$

All "mixed" derivations $\frac{d^2}{d\vartheta_i d\vartheta_j} \log L_{(X,T,D)}(\vartheta)$ with $i \neq j$ are zero. It is $\mathrm{E}[X] = p$ and thus the Fisher information matrix is:

$$I(\vartheta) = \begin{pmatrix} -\frac{1}{p(1-p)} & 0 & 0 \\ 0 & -\frac{\mathrm{E}[XD]}{\lambda_1^2} & 0 \\ 0 & 0 & -\frac{\mathrm{E}[(1-X)D]}{\lambda_0^2} \end{pmatrix}$$

This yields

$$\begin{aligned}
\mathrm{Var}(\hat{p}) &\approx \frac{p \cdot (1-p)}{n}, \\
\mathrm{Var}(\hat{\lambda}_1) &\approx \frac{\lambda_1^2}{n \cdot \mathrm{E}[XD]}, \\
\mathrm{Var}(\hat{\lambda}_0) &\approx \frac{\lambda_0^2}{n \cdot \mathrm{E}[(1-X)D]}, \\
\mathrm{Cov}(\hat{p}, \hat{\lambda}_1) &\approx 0, \\
\mathrm{Cov}(\hat{p}, \hat{\lambda}_0) &\approx 0, \\
\text{and } \mathrm{Cov}(\hat{\lambda}_1, \hat{\lambda}_0) &\approx 0.
\end{aligned} \tag{3.3}$$

In particular, the MLEs of the different parameters are asymptotically uncorrelated. Note that $\mathrm{E}[XD] = p \cdot \mathrm{P}(T_S|_{X=1} \leq U)$ and $\mathrm{E}[(1-X)D] = (1-p) \cdot \mathrm{P}(T_S|_{X=0} \leq U)$. When estimating the variance, the MLEs for the true parameter are inserted and $\mathrm{E}[XD]$ is estimated by the mean $\frac{1}{n}\sum_{m=1}^{n} x_m d_m = l_1/n$. Analogously, $\mathrm{E}[(1-X)D]$ is estimated by l_0/n. This yields the variance estimators

$$\begin{aligned} \hat{\sigma}^2_{\hat{p}} &:= \frac{\hat{p} \cdot (1-\hat{p})}{n}, \\ \hat{\sigma}^2_{\hat{\lambda}_1} &:= \frac{\hat{\lambda}_1^2}{l_1}, \\ \text{and } \hat{\sigma}^2_{\hat{\lambda}_0} &:= \frac{\hat{\lambda}_0^2}{l_0}. \end{aligned} \tag{3.4}$$

3.2.2 Alternative parameterization of survival parameters

In this section, the MLEs and their asymptotic distribution of transformed model parameters is given. Two further parameterizations of the survival parameters are reasonable: Using $\theta_j := \log(\lambda_j)$ or $\eta_j := 1/\lambda_j$. The respective MLEs are equivalent, meaning $\hat{\theta}_j = \log(\hat{\lambda}_j)$ and $\hat{\eta}_j = 1/\hat{\lambda}_j$. However, the distribution of the MLEs differ. For the construction of approximate confidence intervals or hypothesis tests based on approximate normality, the choice of parameterization might affect the goodness of normal approximation. For that reason, the asymptotic distribution of $\hat{\theta}_j$ and $\hat{\eta}_j$ is also derived and compared to $\hat{\lambda}_j$. Proceeding as in Section 3.2.1, the asymptotic variances of the alternative MLEs are:

$$\begin{aligned} \mathrm{Var}(\hat{\theta}_1) &\approx \frac{1}{n\exp(\theta_1) \cdot \mathrm{E}[XT]}, \\ \mathrm{Var}(\hat{\theta}_0) &\approx \frac{1}{n\exp(\theta_0) \cdot \mathrm{E}[(1-X)T]}, \\ \mathrm{Var}(\hat{\eta}_1) &\approx \frac{\eta_1^3}{n \cdot (2 \cdot \mathrm{E}[XT] - \eta_1 \cdot \mathrm{E}[XD])} \\ \mathrm{Var}(\hat{\eta}_0) &\approx \frac{\eta_0^3}{n \cdot (2 \cdot \mathrm{E}[(1-X)T] - \eta_1 \cdot \mathrm{E}[(1-X)D])} \end{aligned} \tag{3.5}$$

When estimating the variance, the MLEs for the true parameters are inserted, $\mathrm{E}[XT]$ is estimated by the mean $\frac{1}{n}\sum_{m=1}^{k} t_m = l_1/n \cdot \hat{\eta}_1$, and $\mathrm{E}[(1-X)T]$ is estimated by the mean $l_0/n \cdot \hat{\eta}_0$. This yields

$$\begin{aligned} \hat{\sigma}^2_{\hat{\theta}_1} &:= \frac{1}{l_1}, \\ \hat{\sigma}^2_{\hat{\theta}_0} &:= \frac{1}{l_0}, \\ \hat{\sigma}^2_{\hat{\eta}_1} &:= \frac{\hat{\eta}_1^2}{l_1}, \\ \text{and } \hat{\sigma}^2_{\hat{\eta}_0} &:= \frac{\hat{\eta}_0^2}{l_0}. \end{aligned} \tag{3.6}$$

3.2.3 Approximate normality of MLEs

In this section, the empirical distribution of the MLEs of the survival parameters is assessed and compared. The distribution of $\hat{p}$ is well known. It is an approximately normally distributed and unbiased estimator of p. Furthermore, it has a root mean squared error (RMSE) of $\sqrt{p(1-p)/n}$. This section focusses on the MLEs of the survival parameters $\hat{\lambda}_1, \hat{\lambda}_0, \hat{\theta}_1, \hat{\theta}_0, \hat{\eta}_1$, and $\hat{\eta}_0$. The empirical distribution of the MLEs is investigated by the first simulation study with 28 scenarios and four different censoring distributions described in Section 2.2.3.

Figure 5 shows quantile-quantile plots of standardized MLEs. The Wald-type standardization is calculated by $(\hat{u}-u)/\hat{\sigma}_{\hat{u}}$, with parameter $u = \lambda_1, \eta_1, \theta_1$ and $\hat{\sigma}_{\hat{u}}$ being the square root of the variance estimator derived in sections 3.2.1 and 3.2.2. Results are similar in all scenarios. For clarity, only the scenarios with $\lambda_1 = \lambda_0 = 0.037, p = 0.5, n = 50, 70, 100$, and all four censoring distributions are shown. Further scenarios are shown in Appendix A.3 in Figure 20. It can be seen that the line for θ_1 is closest to standard normal, meaning that the normal approximation of $\hat{\theta}_1$ is the best. Hence, the construction of approximate confidence intervals and approximate tests will be based on the approximate normality of $\hat{\theta}_1$. For most of the following considerations, the parameterization (p, θ_1, θ_0) of the RSES model will be used.

Figure 6 shows bias and RMSE of $\hat{\theta}_1$ for $p = 0.5, \theta_1 = \log(0.037) \approx -3.3$ and $n = 50, \ldots, 500$. Interestingly, the bias is the greatest in the absence of censoring. However, the bias is generally low. For comparison, a bias of 0.02 in the estimation of the true value $\theta_1 = -3.3$ corresponds to a bias of approximately 2% of the median survival. The RMSE is greater for greater censoring probabilities.

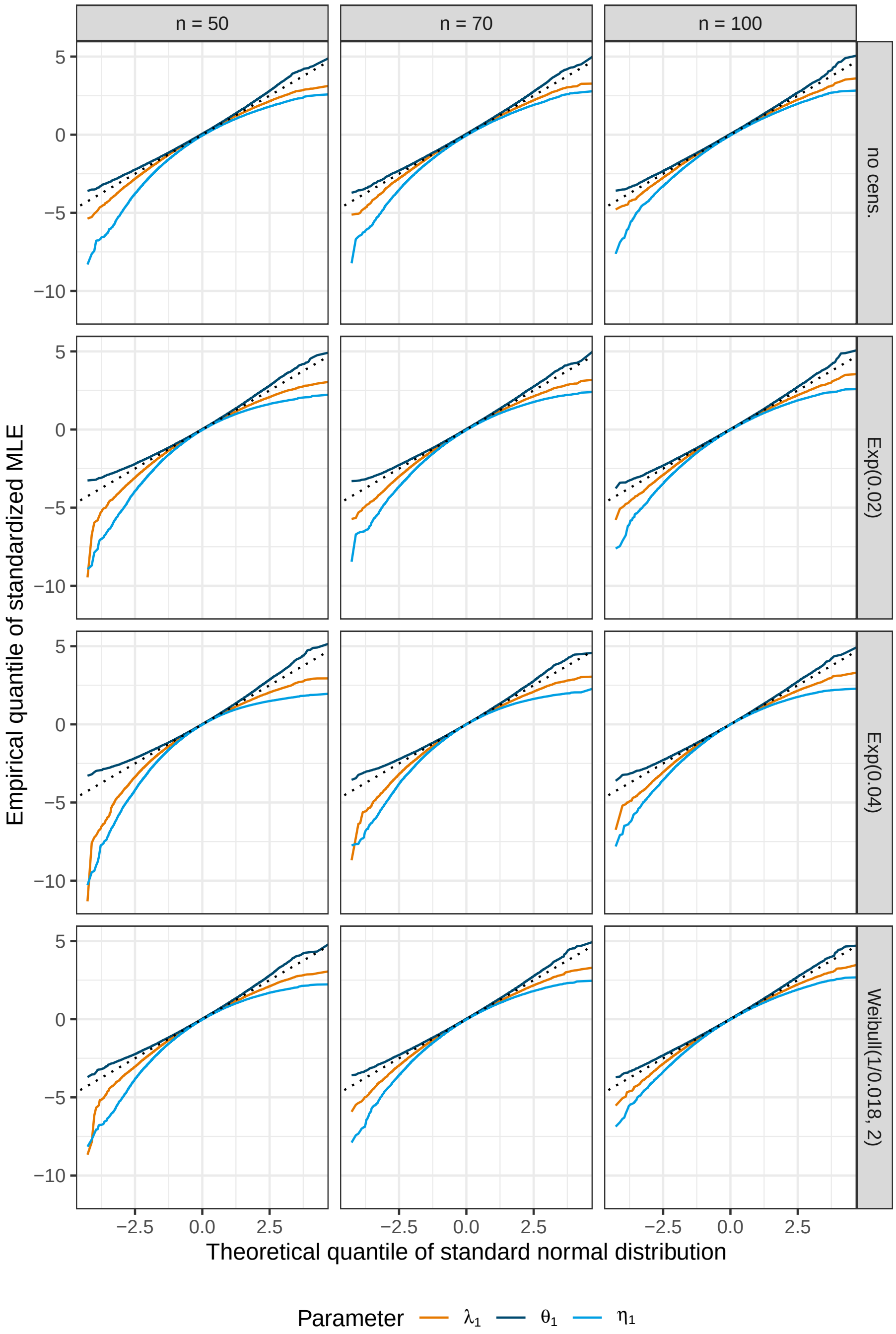


Figure 5: Quantile-quantile plots of standardized MLEs for $\lambda_1 = 0.037, p = 0.5, n = 50, 70, 100$, and four censoring distributions. Dotted black line indicates perfect agreement of empirical quantiles with standard normal distribution.

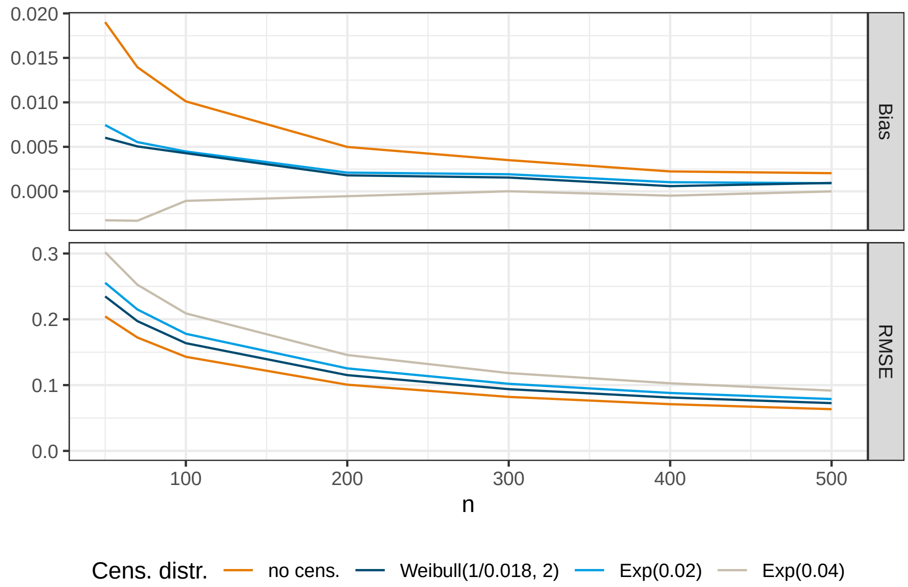


Figure 6: Bias and RMSE of $\hat{\theta}_1$ for different sample sizes and censoring distributions.

3.2.4 Correlation of MLEs

In this section, the pairwise correlation between the MLEs is investigated. Figure 7 shows estimated pairwise correlation coefficients with confidence intervals of $\hat{p}, \hat{\theta}_1$ and $\hat{\theta}_0$ for $p = 0.2$ and $\lambda_1 = \lambda_0 = 0.037$. Consistent with the findings in Section 3.2.1 regarding the asymptotic joint distribution of the MLEs, all estimators are approximately uncorrelated, especially for large sample sizes. Only the correlation of $\hat{p}$ and $\hat{\theta}_1$ is noticeable for small sample sizes.

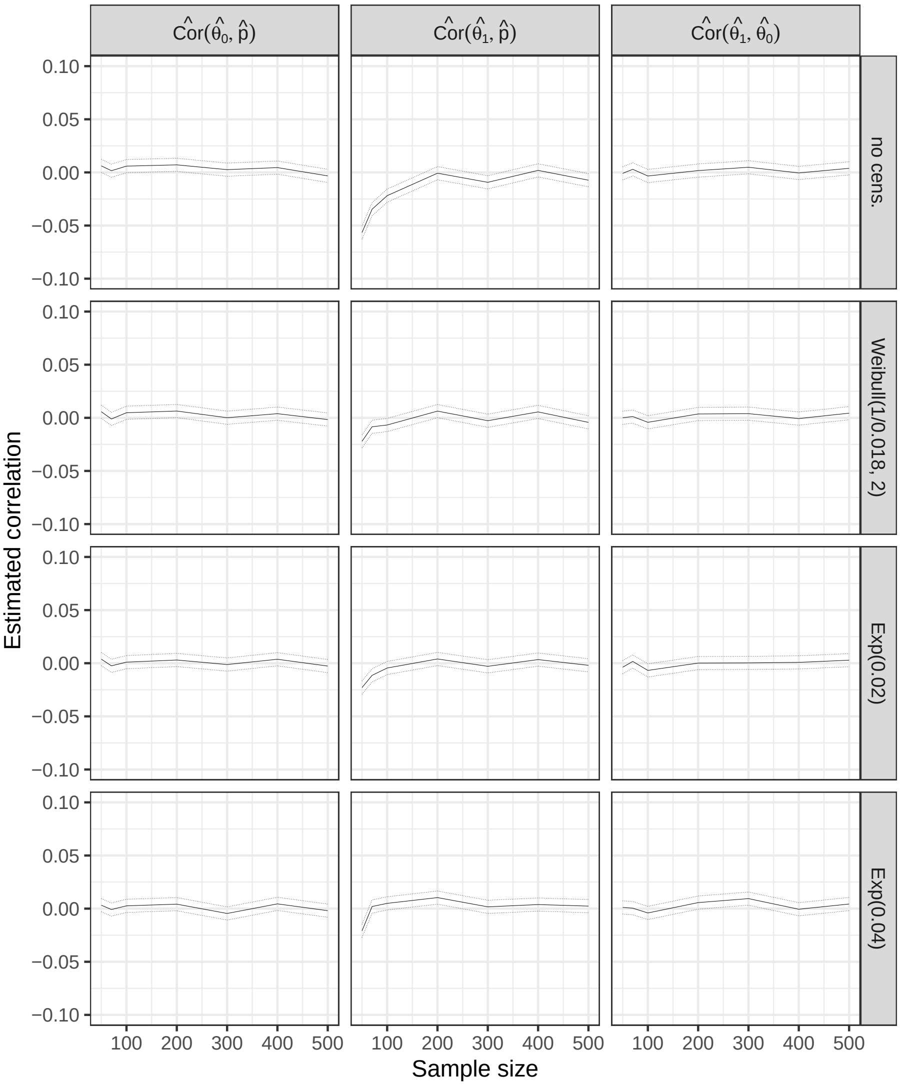


Figure 7: Estimated pairwise correlations (solid lines) and 95% confidence intervals (dotted lines) of $\hat{p}, \hat{\theta}_1$ and $\hat{\theta}_0$ for $p = 0.2, \lambda_1 = \lambda_0 = 0.037$ and four censoring distributions.

3.2.5 Approximate confidence intervals for model parameters

In this section, approximate confidence intervals for the model parameters are constructed by using the asymptotic normality of the MLEs. If $\hat{u}$ is an estimator for u and $\hat{\sigma}^2_{\hat{u}}$ an estimator for $\text{Var}(\hat{u})$ such that $(\hat{u} - u)/\hat{\sigma}_{\hat{u}}$ is approximately normally distributed, an approximate $1 - \alpha$ confidence interval is given by $\hat{u} \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{u}}$. For $\hat{p}$, this results in the well-known confidence interval for a binomial probability:

$$\text{CI}_p(\hat{p}) = \hat{p} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

The coverage probability of CI_p is dependent on the true response probability p_0 and can be calculated exactly by

$$\text{P}_{p_0}(\text{CI}_p(\hat{p}) \ni p_0) = \sum_{k \in \{0,\dots,n\}} 1_{\left\{|p_0 - k/n| \le z_{1-\alpha/2} \cdot \sqrt{k/n \cdot (1-k/n)/n}\right\}} \cdot f_{n,p_0}(k),$$

where f_{n,p_0} is the binomial density. Figure 8 shows the exact coverage probabilities for $p = 0.1, 0.2, 0.3, 0.4, 0.5$ and $n = 50, 70, 100, 200, 300, 400, 500$ and reveals the well-known weakness of the normal approximation for extreme values of p_0.

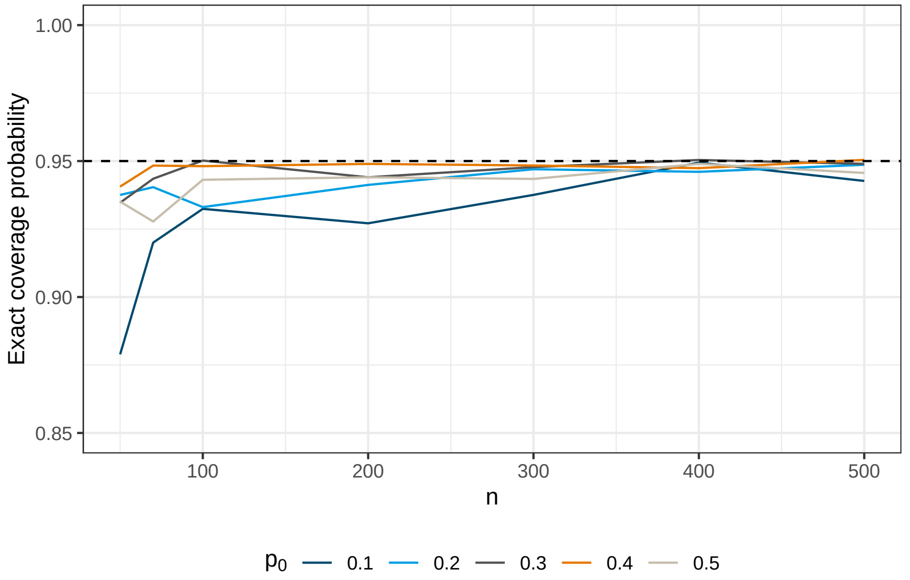


Figure 8: Exact coverage probability of approximate confidence intervals for the response probability p for different values of the true response probability p_0.

It was seen in Section 3.2.3 that $(\hat{\theta}_j - \theta_j)/\hat{\sigma}_{\hat{\theta}_j}$ with $\hat{\sigma}^2_{\hat{\theta}_j} = 1/l_j$ is approximately normal. Thus, an approximate $1 - \alpha$ confidence interval for θ_j is given by

$$\text{CI}_{\theta_j}(\hat{\theta}_j, l_j) = \hat{\theta}_j \pm z_{1-\alpha/2} \cdot \sqrt{1/l_j}.$$

Figure 9 shows estimated coverage probabilities of CI_{θ_1} for all scenarios and censoring distributions of the first simulation study. The coverage probability is calculated conditional on the existence of $\hat{\theta}_1$, i.e. conditional on $l_1 > 0$. All coverage probabilities are satisfying and coverage increases for greater censoring. The results for the coverage probability of CI_{θ_0} are similar and shown in Appendix A.3 in Figure 21.

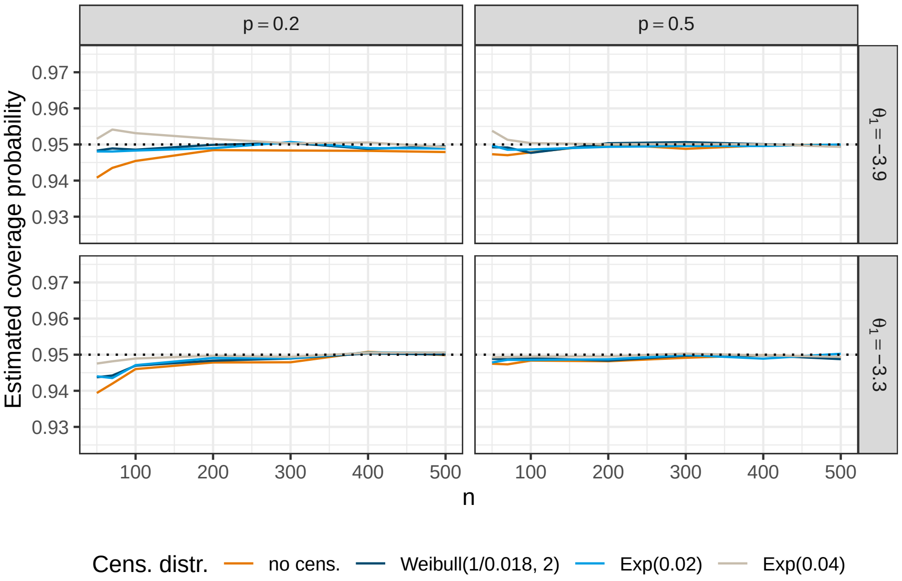


Figure 9: Estimated coverage probability of approximate confidence intervals for θ_1 with respect to sample size n, response probability p and true responder survival parameter θ_1. The true value of the non-responder survival parameter is $\theta_0 \approx -3.3$. The desired confidence level 0.95 is indicated by the dotted black line.

3.2.6 Exact distribution of MLEs

In this section, the exact distribution of the MLEs $\hat{p}, \hat{\theta}_1$ and $\hat{\theta}_0$ is investigated. The exact distribution is used to derive exact formulas for the expected value and variance of the MLEs. Furthermore, exact formulas for the pairwise covariances and correlation coefficients are derived. Additionally, the asymptotic behavior of the derived terms is investigated. Lastly, exact formulas for the approximate confidence intervals given in Section 3.2.5 are derived. For the survival parameters $\hat{\theta}_1$ and $\hat{\theta}_0$, only the cases of no censoring or exponential censoring are considered. For all calculations, assume $0 < p < 1$. The exact distribution of $\hat{p}$ is well known. $n \cdot \hat{p} = k$ follows a binomial distribution with parameters n and p. This section focusses on $\hat{\theta}_1$ and $\hat{\theta}_0$.

Firstly, the case of no censoring is considered. Conditional on $k > 0$, $\sum\limits_{m=1}^{k} t_m$ is the sum of k independent exponentially distributed random variables and thus follows a $\Gamma(k, \lambda_1)$-distribution (with shape/rate parameterization). After scaling with factor $1/k$,

$$\exp(-\hat{\theta}_1) = \frac{1}{k} \cdot \sum_{m=1}^{k} t_m \sim \Gamma(k, k\lambda_1).$$

Analogously,

$$\exp(-\hat{\theta}_0) \sim \Gamma(n-k, (n-k)\lambda_0).$$

Now, the exact expectation, variance and covariance and their asymptotic behavior is derived without using MLE properties. As $-\hat{\theta}_1$ is the logarithm of a Γ-distributed random variable, the expectation of $\hat{\theta}_1$ conditional on k with $k > 0$ is known to be

$$\mathrm{E}\Big[\hat{\theta}_1 \;\Big|\; k\Big] = \theta_1 + \log(k) - \psi(k), \tag{3.7}$$

where ψ is the digamma function (Abramowitz et al. 1972, p. 258). The digamma function is defined as the derivative of the logarithm of the gamma function:

$$\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log(\Gamma(x))$$

Since $\hat{\theta}_1$ does not exist for $k = 0$, interest lies in the expectation of $\hat{\theta}_1$ conditional on the existence of $\hat{\theta}_1$, i.e. conditional on $k > 0$. The random variable $k|_{k>0}$ has the distribution function

$$F_{k|_{k>0}}(z) = \frac{f_{n,p}(z) - f_{n,p}(0)}{1 - f_{n,p}(0)}, \tag{3.8}$$

where $f_{n,p}$ is the binomial distribution function. By the law of total expectation (Ross 2010, p. 333),

$$\begin{aligned} \mathrm{E}\Big[\hat{\theta}_1 \;\Big|\; k > 0\Big] &= \mathrm{E}\Big[\mathrm{E}[\hat{\theta}_1 \mid k] \;\Big|\; k > 0\Big] \\ &= \theta_1 + \mathrm{E}[\log(k) - \psi(k) \mid k > 0]. \end{aligned} \tag{3.9}$$

Analogously,

$$\mathrm{E}\Big[\hat{\theta}_0 \;\Big|\; k < n\Big] = \theta_0 + \mathrm{E}[\log(n-k) - \psi(n-k) \mid k < n].$$

The digamma function has the known property

$$\tfrac{1}{2k} \leq \log(k) - \psi(k) \leq \tfrac{1}{k} \text{ for } k > 0. \tag{3.10}$$

The random variable k converges in probability to infinity for $n \to \infty$, meaning $\mathrm{P}(k > M) \overset{n \to \infty}{\to} 1$ for arbitrarily large M. Thus, by the continuous mapping theorem (CMT, Mann and Wald (1943)),

$$\log(k) - \psi(k) \overset{P,\ n \to \infty}{\to} 0.$$

By the Portmanteau theorem (Klenke 2014, p. 254),

$$\mathrm{E}[\log(k) - \psi(k)] \overset{n \to \infty}{\to} 0,$$

since $\log(k) - \psi(k)$ is bounded. Hence,

$$\mathrm{E}\Big[\hat{\theta}_1 \;\Big|\; k > 0\Big] \overset{n \to \infty}{\to} \theta_1.$$

Analogously, $\mathrm{E}\Big[\hat{\theta}_0 \;\Big|\; k < n\Big] \overset{n\to\infty}{\to} \theta_0$. Note, however, that $\hat{\theta}_1$ and $\hat{\theta}_0$ are positively biased due to inequality (3.10). A bias-corrected version of the MLE could be obtained by $\hat{\theta}_1 - \log(k) + \psi(k)$ and is unbiased in the absence of censoring. However, in the presence of censoring, the correction turned out to be too large and produced worse estimates than the MLE. Thus, it is not further evaluated.

Now, $\mathrm{Var}\Big(\hat{\theta}_1 \;\Big|\; k > 0\Big)$ is assessed. For $k > 0$ and since $-\hat{\theta}_1$ is the logarithm of a Γ-distributed random variable, it is

$$\mathrm{Var}\Big(\hat{\theta}_1 \;\Big|\; k\Big) = \psi^{(1)}(k),$$

where $\psi^{(1)}$ is the polygamma function of order 1 (Abramowitz et al. 1972, p. 260). For $k > 0$, it has the property

$$\frac{1}{2k^2} \leq \psi^{(1)}(k) - \frac{1}{k} \leq \frac{1}{k^2}. \tag{3.11}$$

By the law of total variance (Ross 2010, p. 348) and using equation (3.7),

$$\begin{aligned}
\mathrm{Var}\Big(\hat{\theta}_1 \;\Big|\; k > 0\Big) &= \mathrm{E}\Big[\mathrm{Var}\Big(\hat{\theta}_1 \;\Big|\; k\Big) \;\Big|\; k > 0\Big] + \mathrm{Var}\Big(\mathrm{E}\Big[\hat{\theta}_1 \;\Big|\; k\Big] \;\Big|\; k > 0\Big) \\
&= \mathrm{E}\Big[\psi^{(1)}(k) \;\Big|\; k > 0\Big] + \mathrm{Var}(\theta_1 + \log(k) - \psi(k) \mid k > 0) \\
&= \mathrm{E}\Big[\psi^{(1)}(k) \;\Big|\; k > 0\Big] + \mathrm{Var}(\log(k) - \psi(k) \mid k > 0).
\end{aligned}$$

It is clear that $\mathrm{Var}\Big(\hat{\theta}_1 \;\Big|\; k > 0\Big) \overset{n\to\infty}{\to} 0$. The asymptotic variance of $\hat{\theta}_1$ is derived in formula (3.5) as

$$\frac{1}{n\exp(\theta_1) \cdot \mathrm{E}[XT]}.$$

In the case of no censoring, $T|_{X=1} \sim \mathrm{Exp}(\lambda_1)$ and thus

$$\begin{aligned}
\mathrm{E}[XT] &= \mathrm{E}[\mathrm{E}[XT \mid X]] \\
&= \mathrm{E}[X/\lambda_1] \\
&= p/\lambda_1 \\
&= p/\exp(\theta_1).
\end{aligned}$$

Hence, the asymptotic variance formula simplifies to $\frac{1}{np}$. To prove the asymptotic correctness of this formula, it has to be shown that

$$\mathrm{Var}\Big(\hat{\theta}_1 \;\Big|\; k > 0\Big)/(1/(np)) \overset{n\to\infty}{\to} 1.$$

It is known that $k/n \overset{D}{\to} p$ and $f_{n,p}(0) \overset{n\to\infty}{\to} 0$. Thus, due to equation (3.8), $k|_{k>0}/n \overset{D}{\to} p$. Since p is a constant, this implies $k|_{k>0}/n \overset{P}{\to} p$. By CMT, it follows $n/k|_{k>0} \overset{P}{\to} 1/p$ and $np/k^2|_{k>0} = n^2/k^2|_{k>0} \cdot p/n \overset{P}{\to} 0$. It cannot be concluded $\mathrm{E}[|p \cdot n/k|_{k>0} - 1|] \overset{n\to\infty}{\to} 0$ directly since $p \cdot n/k - 1$ is not bounded. However, if δ with $0 < \delta < p$ is chosen, it can be written

$$n/k|_{k>0} = n/k|_{k>0} \cdot 1_{k/n\geq\delta} + n/k|_{k>0} \cdot 1_{k/n<\delta}. \tag{3.12}$$

The first term is bounded above by $1/\delta$. Due to $1_{k/n\geq\delta} \overset{P}{\to} 1$ and the Portmanteau theorem,

$$\mathrm{E}\big[n/k|_{k>0} \cdot 1_{k/n\geq\delta}\big] \overset{n\to\infty}{\to} 1/p.$$

For the second term it is

$$\mathrm{E}\big[n/k|_{k>0} \cdot 1_{k/n<\delta}\big] \leq n \cdot \mathrm{P}(k/n < \delta).$$

By Hoeffding's inequality (Hoeffding 1994),

$$\mathrm{P}(k/n < \delta) \leq \exp(-2n(p-\delta)^2).$$

Hence,

$$\mathrm{E}\big[n/k|_{k>0} \cdot 1_{k/n<\delta}\big] \overset{n\to\infty}{\to} 0.$$

It follows $\mathrm{E}[n/k|_{k>0}] \overset{n\to\infty}{\to} 1/p$ and

$$\mathrm{E}[1 - p \cdot n/k \mid k > 0] \overset{n\to\infty}{\to} 0.$$

Analogously, $\mathrm{E}\big[np/k^2 \bigm| k > 0\big] \overset{n\to\infty}{\to} 0$. Using this and inequalities (3.11) and (3.10) leads to:

$$\begin{aligned}
&\left|\frac{\mathrm{Var}\Big(\hat{\theta}_1 \Bigm| k > 0\Big)}{1/(np)} - 1\right| \\
&= \Big|np \cdot \mathrm{E}\Big[\psi^{(1)}(k) - 1/(np) \Bigm| k > 0\Big] + np \cdot \mathrm{Var}(\log(k) - \psi(k) \mid k > 0)\Big| \\
&\leq \mathrm{E}\Big[np \cdot \Big|\psi^{(1)}(k) - 1/k\Big| + |p \cdot n/k - 1| \Bigm| k > 0\Big] \\
&\qquad + np \cdot \mathrm{E}\big[(\log(k) - \psi(k))^2 \bigm| k > 0\big] \\
&\leq \mathrm{E}\big[np/k^2 \bigm| k > 0\big] + \mathrm{E}[1 - p \cdot n/k \mid k > 0] + \mathrm{E}\big[np/k^2 \bigm| k > 0\big] \\
&\overset{n\to\infty}{\to} 0
\end{aligned}$$

Hence,

$$\mathrm{Var}\Big(\hat{\theta}_1 \Bigm| k > 0\Big)/(1/(np)) \overset{n\to\infty}{\to} 1. \tag{3.13}$$

Analogously, $\mathrm{Var}\Big(\hat{\theta}_0 \Bigm| k < n\Big)/(1/(n(1-p))) \overset{n\to\infty}{\to} 1$. Note that the above calculations hold also true if it is conditioned on $0 < k < n$, thus

$$\mathrm{Var}\Big(\hat{\theta}_1 \Bigm| 0 < k < n\Big)/(1/(np)) \overset{n\to\infty}{\to} 1$$

and

$$\mathrm{Var}\Big(\hat{\theta}_0 \Bigm| 0 < k < n\Big)/(1/(n(1-p))) \overset{n\to\infty}{\to} 1.$$

This will be used in the following.

Now the correlation of $\hat{\theta}_1$ and $\hat{\theta}_0$ is considered. Conditional on k, $\hat{\theta}_1$ and $\hat{\theta}_0$ are independent as functions of distinct independent random variables. Thus, by the law of total covariance (Ross 2010, p. 381) and using equation (3.7),

$$\begin{aligned}
&\operatorname{Cov}\Big(\hat{\theta}_1, \hat{\theta}_0 \Bigm| 0 < k < n\Big)\\
&= \mathrm{E}\Big[\operatorname{Cov}\Big(\hat{\theta}_1, \hat{\theta}_0 \Bigm| k\Big) \Bigm| 0 < k < n)\Big]\\
&\qquad + \operatorname{Cov}\Big(\mathrm{E}\Big[\hat{\theta}_1 \Bigm| k\Big], \mathrm{E}\Big[\hat{\theta}_0 \Bigm| k\Big] \Bigm| 0 < k < n\Big)\\
&= \operatorname{Cov}(\theta_1 + \log(k) - \psi(k), \theta_0 + \log(n-k) - \psi(n-k) \mid 0 < k < n)\\
&= \operatorname{Cov}(\log(k) - \psi(k), \log(n-k) - \psi(n-k) \mid 0 < k < n).
\end{aligned}$$

Using inequality (3.10), it is

$$\begin{aligned}
\Big|\operatorname{Cov}\Big(\hat{\theta}_1, \hat{\theta}_0 \Bigm| 0 < k < n\Big)\Big| \leq\ & \mathrm{E}[1/k \cdot 1/(n-k) \mid 0 < k < n]\\
&+ \mathrm{E}[1/k \mid 0 < k < n] \cdot \mathrm{E}[1/(n-k) \mid 0 < k < n]
\end{aligned}$$

and

$$\begin{aligned}
n \cdot \Big|\operatorname{Cov}\Big(\hat{\theta}_1, \hat{\theta}_0 \Bigm| 0 < k < n\Big)\Big| \leq\ & \mathrm{E}[n/k \cdot 1/(n-k) \mid 0 < k < n] +\\
& \mathrm{E}[n/k \mid 0 < k < n] \cdot \mathrm{E}[1/(n-k) \mid 0 < k < n].
\end{aligned}$$

As previous, $n/k|_{0<k<n} \overset{P}{\to} 1/p$ and $1/(n - k|_{0<k<n}) \overset{P}{\to} 0$. Using the same trick as in (3.12), it is

$$\mathrm{E}[n/k \cdot 1/(n-k) \mid 0 < k < n] \overset{n\to\infty}{\to} 0,$$

$$\mathrm{E}[n/k \mid 0 < k < n] \overset{n\to\infty}{\to} 1/p,$$

and

$$\mathrm{E}[1/(n-k) \mid 0 < k < n] \overset{n\to\infty}{\to} 0.$$

In total, it is

$$\begin{aligned}
&\operatorname{Cor}(\hat{\theta}_1, \hat{\theta}_0 | 0 < k < n)\\
&= \frac{\operatorname{Cov}\Big(\hat{\theta}_1, \hat{\theta}_0 \Bigm| 0 < k < n\Big)}{\sqrt{\operatorname{Var}\Big(\hat{\theta}_1 \Bigm| 0 < k < n\Big) \cdot \operatorname{Var}\Big(\hat{\theta}_0 \Bigm| 0 < k < n\Big)}}\\
&= \frac{n \cdot \operatorname{Cov}\Big(\hat{\theta}_1, \hat{\theta}_0 \Bigm| 0 < k < n\Big)}{\sqrt{\operatorname{Var}\Big(\hat{\theta}_1 \Bigm| 0 < k < n\Big) \cdot np \cdot \operatorname{Var}\Big(\hat{\theta}_0 \Bigm| 0 < k < n\Big) \cdot n(1-p)}} \cdot \sqrt{p(1-p)}\\
&\overset{n\to\infty}{\to} 0,
\end{aligned}$$

since the numerator converges to 0 and the denominator was shown in (3.13) to converge to 1.

Finally, the correlation of $\hat{p}$ and $\hat{\theta}_1$ is considered. By the law of total covariance and since $\hat{p}$ is constant for fixed k,

$$\begin{aligned}
\operatorname{Cov}\Big(\hat{p}, \hat{\theta}_1 \;\Big|\; k > 0\Big) &= \operatorname{E}\Big[\operatorname{Cov}\Big(\hat{p}, \hat{\theta}_1 \;\Big|\; k\Big) \;\Big|\; k > 0\Big] \\
&\qquad + \operatorname{Cov}\Big(\operatorname{E}[\hat{p} \mid k], \operatorname{E}\Big[\hat{\theta}_1 \;\Big|\; k\Big] \;\Big|\; k > 0\Big) \\
&= \operatorname{Cov}(\hat{p}, \theta_1 + \log(k) - \psi(k) \mid k > 0) \\
&= \operatorname{Cov}(\hat{p}, \log(k) - \psi(k) \mid k > 0) \\
&= \operatorname{E}[\hat{p} \cdot (\log(k) - \psi(k)) \mid k > 0] \\
&\qquad - \operatorname{E}[\hat{p} \mid k > 0] \cdot \operatorname{E}[\log(k) - \psi(k) \mid k > 0] \\
&= \operatorname{E}[(\hat{p} - \operatorname{E}[\hat{p} \mid k > 0]) \cdot (\log(k) - \psi(k)) \mid k > 0]
\end{aligned}$$

Thus,

$$\begin{aligned}
\left| n \cdot \operatorname{Cov}\Big(\hat{p}, \hat{\theta}_1 \;\Big|\; k > 0\Big) \right| &\leq \operatorname{E}[|\hat{p} - \operatorname{E}[\hat{p} \mid k > 0]| \cdot n/k \mid k > 0] \\
&\overset{n\to\infty}{\to} 0,
\end{aligned}$$

since $\hat{p}|_{k>0} - \operatorname{E}[\hat{p} \mid k > 0] \overset{p}{\to} 0$, $n/k|_{k>0} \overset{p}{\to} 1/p$, and by using the trick in (3.12). Hence, as before,

$$\begin{aligned}
\operatorname{Cor}\Big(\hat{p}, \hat{\theta}_1 \;\Big|\; k > 0\Big) &= \frac{\operatorname{Cov}\Big(\hat{p}, \hat{\theta}_1 \;\Big|\; k > 0\Big)}{\sqrt{\operatorname{Var}(\hat{p} \mid k > 0) \cdot \operatorname{Var}\Big(\hat{\theta}_1 \;\Big|\; k > 0\Big)}} \\
&= \frac{n \cdot \operatorname{Cov}\Big(\hat{p}, \hat{\theta}_0 \;\Big|\; k > 0\Big)}{\sqrt{p(1-p) \cdot \operatorname{Var}(\hat{\theta}_1) \cdot np}} \cdot \sqrt{p} \\
&\overset{n\to\infty}{\to} 0.
\end{aligned}$$

Analogously, $\operatorname{Cor}\Big(\hat{p}, \hat{\theta}_0 \;\Big|\; k < n\Big) \overset{n\to\infty}{\to} 0$.

Now consider the case of exponential censoring. Let $U \sim \operatorname{Exp}(\lambda_U)$. Then, the responder survival times $t_1, \ldots, t_k$ are realisations of $T \sim \operatorname{Exp}(\lambda_1 + \lambda_U)$. As seen above, conditional on k and l_1 with $l_1 > 0$ (note that $l_1 > 0$ implies $k > 0$), it is

$$\frac{1}{l_1} \cdot \sum_{m=1}^{k} t_m \sim \Gamma(k, l_1 \cdot (\lambda_1 + \lambda_U)).$$

Thus,

$$\operatorname{E}\Big[\hat{\theta}_1 \;\Big|\; k, l_1\Big] = \log(\lambda_1 + \lambda_U) + \log(l_1) - \psi(k). \tag{3.14}$$

This yields

$$\operatorname{E}\Big[\hat{\theta}_1 \;\Big|\; l_1 > 0\Big] = \log(\lambda_1 + \lambda_U) + \operatorname{E}[\log(l_1) - \psi(k) \mid l_1 > 0]. \tag{3.15}$$

Note that l_1 is the number of uncensored responder survival times. Hence, l_1 is binomially distributed with size n and probability $\mathrm{P}(X = 1 \text{ and } T_S < U) = p \cdot \lambda_1/(\lambda_U + \lambda_1)$. Thus,

$$l_1/n \xrightarrow{P} p \cdot \lambda_1/(\lambda_U + \lambda_1)$$

and therefore

$$l_1/k = l_1/k \cdot n/k \xrightarrow{P} \lambda_1/(\lambda_U + \lambda_1).$$

So asymptotically, it is non-surprisingly seen that

$$\begin{aligned}\mathrm{E}\Big[\hat{\theta}_1 \;\Big|\; l_1 > 0\Big] &= \log(\lambda_1 + \lambda_U) + \mathrm{E}[\log(l_1) - \psi(k) \mid l_1 > 0] \\ &= \log(\lambda_1 + \lambda_U) + \mathrm{E}[\log(l_1/k) \mid l_1 > 0] + \mathrm{E}[\log(k) - \psi(k) \mid l_1 > 0] \\ &\overset{n\to\infty}{\to} \log(\lambda_1 + \lambda_U) + \log(\lambda_1/(\lambda_U + \lambda_1)) + 0 \\ &= \theta_1.\end{aligned}$$

Analogously, $\mathrm{E}\Big[\hat{\theta}_0 \;\Big|\; l_0 > 0\Big] = \log(\lambda_0 + \lambda_U) + \mathrm{E}[\log(l_0) - \psi(n-k) \mid l_0 > 0]$.

For the variance, it can be proceeded analogously to the case without censoring and thus

$$\mathrm{Var}\Big(\hat{\theta}_1 \;\Big|\; l_1 > 0\Big) = \mathrm{E}\Big[\psi^{(1)}(k) \;\Big|\; l_1 > 0\Big] + \mathrm{Var}(\log(l_1) - \psi(k) \mid l_1 > 0)$$

and

$$\mathrm{Var}\Big(\hat{\theta}_0 \;\Big|\; l_0 > 0\Big) = \mathrm{E}\Big[\psi^{(1)}(n-k) \;\Big|\; l_0 > 0\Big] + \mathrm{Var}(\log(l_0) - \psi(n-k) \mid l_0 > 0).$$

For the covariance of $\hat{\theta}_1$ and $\hat{\theta}_0$, it is

$$\begin{aligned}&\mathrm{Cov}\Big(\hat{\theta}_1, \hat{\theta}_0 \;\Big|\; l_1 > 0, l_0 > 0\Big) \\ &\quad = \mathrm{Cov}(\log(l_1) - \psi(k), \log(l_0) - \psi(n-k) \mid l_1 > 0, l_0 > 0).\end{aligned}$$

For the covariance of $\hat{p}$ and $\hat{\theta}_1$, it is

$$\mathrm{Cov}\Big(\hat{p}, \hat{\theta}_1 \;\Big|\; l_1 > 0\Big) = \mathrm{E}[(\hat{p} - \mathrm{E}[\hat{p} \mid l_1 > 0]) \cdot (\log(l_1) - \psi(k)) \mid l_1 > 0].$$

For the covariance of $\hat{p}$ and $\hat{\theta}_0$, it is

$$\mathrm{Cov}\Big(\hat{p}, \hat{\theta}_0 \;\Big|\; l_0 > 0\Big) = \mathrm{E}[(\hat{p} - \mathrm{E}[\hat{p} \mid l_0 > 0]) \cdot (\log(l_0) - \psi(n-k)) \mid l_0 > 0].$$

The asymptotic behavior of these terms can be investigated analogously to the case of no censoring and confirms the asymptotic distribution derived in Section 3.2.1.

Knowing the exact distribution of $\hat{\theta}_j$ in the case of no censoring or exponential censoring allows the exact calculation of coverage probabilities of the asymptotic confidence intervals. The coverage probability depends on the true response probability p_0. The upper limit of $\mathrm{CI}_{\theta_1}(\hat{\theta}_1, l_1)$ lies above the true value θ_1 if

$$\theta_1 \leq \hat{\theta}_1 + \sqrt{1/l_1}.$$

This is equivalent to

$$\lambda_1 \exp(-\hat{\theta}_1) \leq \exp\left(\sqrt{1/l_1}\right). \tag{3.16}$$

The analogous statement is true for the lower limit. In the case of no censoring, the distribution of $\lambda_1 \exp(-\hat{\theta}_1)$ conditional on $l_1 = k$ was shown to be $\Gamma(k, k)$. Thus, the coverage probability of CI_{θ_1} conditional on k is independent of θ_1 and is given by

$$\begin{aligned}\mathrm{P}\Big(\mathrm{CI}_{\theta_1}(\hat{\theta}_1, k) \ni \theta_1 \;\Big|\; k = k\Big) = {} & F_{\Gamma(k,k)}\left(\exp\left(z_{1-\alpha/2}\sqrt{1/k}\right)\right) \\ & - F_{\Gamma(k,k)}\left(\exp\left(-z_{1-\alpha/2}\sqrt{1/k}\right)\right),\end{aligned}$$

where $F_{\Gamma(k,k)}$ is the distribution function of the $\Gamma(k, k)$-distribution. The unconditional coverage probability with true response probability p_0 then is

$$\mathrm{P}_{p_0}(\mathrm{CI}_{\theta_1}(\hat{\theta}_1, k) \ni \theta_1 | k > 0) = \frac{\sum\limits_{k\in\{1,\ldots,n\}} f_{n,p_0}(k) \cdot \mathrm{P}\Big(\mathrm{CI}_{\theta_1}(\hat{\theta}_1, k) \ni \theta_1 \;\Big|\; k = k\Big)}{1 - f_{n,p_0}(0)}. \tag{3.17}$$

Analogously, the coverage probability of CI_{θ_0} is

$$\begin{aligned}& P_{p_0}(\mathrm{CI}_{\theta_0}(\hat{\theta}_0, n-k) \ni \theta_0 | k < n) \\ & \quad = \frac{\sum\limits_{k\in\{0,\ldots,n-1\}} f_{n,p_0}(k) \cdot \mathrm{P}\Big(\mathrm{CI}_{\theta_0}(\hat{\theta}_0, n-k) \ni \theta_0 \;\Big|\; k = k\Big)}{1 - f_{n,p_0}(n)}\end{aligned} \tag{3.18}$$

with

$$\begin{aligned}\mathrm{P}\Big(\mathrm{CI}_{\theta_0}(\hat{\theta}_0, n-k) \ni \theta_0 \;\Big|\; k = k\Big) = {} & F_{\Gamma(n-k,n-k)}\left(\exp\left(z_{1-\alpha/2}\sqrt{1/(n-k)}\right)\right) \\ & - F_{\Gamma(n-k,n-k)}\left(\exp\left(-z_{1-\alpha/2}\sqrt{1/(n-k)}\right)\right).\end{aligned}$$

In the case of an exponential censoring distribution $U \sim \mathrm{Exp}(\lambda_U)$, the distribution of $(\lambda_1 + \lambda_U) \cdot \exp(-\hat{\theta}_1)$ conditional on k and l_1 was shown to be $\Gamma(k, l_1)$. Since equation (3.16) is equivalent to

$$(\lambda_1 + \lambda_U) \cdot \exp(-\hat{\theta}_1) \leq \exp\left(\sqrt{1/l_1}\right)/q_1$$

with $q_1 := \lambda_1/(\lambda_U + \lambda_1)$, the coverage probability of CI_{θ_1} conditional on k and l_1 is given by

$$\begin{aligned}\mathrm{P}\Big(\mathrm{CI}_{\theta_1}(\hat{\theta}_1, k) \ni \theta_1 \;\Big|\; k = k, l_1 = l_1\Big) = {} & F_{\Gamma(k,l_1)}\left(\exp\left(z_{1-\alpha/2}\sqrt{1/l_1}/q_1\right)\right) \\ & - F_{\Gamma(k,l_1)}\left(\exp\left(-z_{1-\alpha/2}\sqrt{1/l_1}/q_1\right)\right).\end{aligned}$$

Conditional on k, l_1 is binomially distributed with size k and probability q_1. The unconditional distribution of l_1 is binomial with size n and probability $p_0 \cdot q_1$. Thus, the common density of (k, l_1) conditional on $l_1 > 0$ is

$$f|_{l_1>0}(k, l_1) = \frac{f_{n,p}(k) \cdot f_{k,q_1}(l_1)}{1 - f_{n,p_0\cdot q_1}(0)}.$$

The unconditional coverage probability with true response probability p_0 then is

$$\begin{aligned} &P_{p_0}(\mathrm{CI}_{\theta_1}(\hat{\theta}_1, l_1) \ni \theta_1 | l_1 > 0) \\ &= \frac{\sum_{k=1}^{n} \sum_{l_1=1}^{k} f_{n,p}(k) \cdot f_{k,q_1}(l_1) \cdot \mathrm{P}\Big(\mathrm{CI}_{\theta_1}(\hat{\theta}_1, k) \ni \theta_1 \;\Big|\; k = k, l_1 = l_1\Big)}{1 - f_{n,p_0 \cdot q_1}(0)}. \end{aligned} \tag{3.19}$$

Analogous formulas apply to the coverage probability of CI_{θ_0}. The derived formulas (3.17), (3.18), (3.19) for the exact coverage probability are validated by the first simulation study. See Figure 23 in Appendix A.4 for a comparison of exactly calculated and estimated coverage probabilities.

3.3 Hypothesis testing

In this section, tests for the comparison of an experimental group E and a control group C with respective parameter triples $(p_E, \theta_{1,E}, \theta_{0,E})$ and $(p_C, \theta_{1,C}, \theta_{0,C})$ are derived. The subsections comprise the development of an approximate and an exact test, the derivation of formulas for the exact calculation of rejection probabilities of these tests, and the construction of approximate confidence intervals for the differences of model parameters. The global null hypothesis H_0 is an intersection of three local null hypotheses, as described in Section 2.3:

$$\begin{aligned} H_{p,0}&\colon\ p_C = p_E \\ H_{\theta_1,0}&\colon\ \theta_{1,C} = \theta_{1,E} \\ H_{\theta_0,0}&\colon\ \theta_{0,C} = \theta_{0,E} \end{aligned}$$

3.3.1 Approximate RSES test

In this section, an approximate test of H_0 is derived. Let X_i be the response status, $T_{S,i}$ the survival time, U_i the censoring time, $T_i = \min(T_{S,i}, U_i)$ the observed event-free time, and $D_i = 1_{T_{S,i} \le U_i}$ the event indicator in each of the two groups $i = E, C$. In group i, n_i realisations $(x_{i,m}, t_{i,m}, d_{i,m})$ of (X_i, T_i, D_i) with $i = E, C$ and $m = 1, \ldots, n_i$ are observed. Let $k_i = \sum_{m=1}^{n_i} x_{i,m}$ be the number of responders, $l_{1,i} = \sum_{m=1}^{n_i} x_{i,m} \cdot d_{i,m}$ the number of uncensored responder survival times and $l_{0,i} = \sum_{m=1}^{n_i} (1 - x_{i,m}) \cdot d_{i,m}$ the number of uncensored non-responder survival times in group i. The observations are arranged such that $x_{i,1} = \cdots = x_{i,k_i} = 1$ and $t_{i,1}, \ldots, t_{l_{1,i}}$ are the uncensored responder survival times in group i. Analogously, $x_{i,k_i+1} = \cdots = x_{n_i} = 0$ and $t_{i,k_i+1}, \ldots, t_{k_i+l_{0,i}}$ are the uncensored non-responder survival times in group i.

For the local hypotheses $H_{p,0}, H_{\theta_1,0}, H_{\theta_0,0}$, test statistics similar to Wald test statistics (Lehmann and Romano 2010, p. 508) are constructed by standardizing the difference between the MLEs of both groups. The difference is divided by an estimator of the standard deviation of the difference. Here, the formulas for the asymptotic variance in (3.3) and (3.5) are used and the unknown true parameters are replaced by their MLEs under the local null hypothesis.

The MLE of the response probability $p_E = p_C$ under $H_{p,0}$ is $\tilde{p} = \frac{n_E \hat{p}_E + n_C \hat{p}_C}{n_E + n_C}$. For the test of $H_{p,0}$, this yields the well-known two proportion z-test with test statistic

$$T_p = \frac{\hat{p}_E - \hat{p}_C}{\sqrt{\tilde{p}(1-\tilde{p})(\frac{1}{n_E} + \frac{1}{n_C})}}.$$

If $\tilde{p} \in \{0, 1\}$ (which means that everyone or no one is a responder), T_p is set to zero because $H_{p,0}$ should not be rejected in this case.

For the variance estimation of $\hat{\theta}_{1,E}$ and $\hat{\theta}_{1,C}$, θ_1 and $\mathrm{E}[XT]$ have to be estimated under $H_{\theta_1,0}$. The MLE of θ_1 under $H_{\theta_1,0}$ is

$$-\log\left(\frac{1}{l_{1,E} + l_{1,C}}\left(\sum_{i=1}^{k_E} t_{E,i} + \sum_{i=1}^{k_C} t_{C,i}\right)\right).$$

$\mathrm{E}[XT]$ is estimated by the mean $1/(n_E + n_C) \cdot \left(\sum\limits_{m=1}^{k_E} t_{E,m} + \sum\limits_{m=1}^{k_C} t_{C,m}\right)$. Hence, $\mathrm{Var}(\hat{\theta}_{1,E})$ is estimated by $\frac{1}{n_E} \cdot \frac{n_E + n_C}{l_{1,E} + l_{1,C}}$ and $\mathrm{Var}(\hat{\theta}_{1,C})$ by $\frac{1}{n_C} \cdot \frac{n_E + n_C}{l_{1,E} + l_{1,C}}$. The test statistic for the local test of $H_{\theta_1,0}$ then is

$$T_{\theta_1} = \frac{\hat{\theta}_{1,E} - \hat{\theta}_{1,C}}{\sqrt{\frac{n_E + n_C}{l_{1,E} + l_{1,C}}\left(\frac{1}{n_E} + \frac{1}{n_C}\right)}}.$$

T_{θ_1} cannot be calculated if $l_{1,E} = 0$ or $l_{1,C} = 0$. In this case, T_{θ_1} is set to zero to accept the null hypothesis due to insufficient information. However, this case should be rare in practice.

Analogously, the test statistic for the local test of $H_{\theta_0,0}$ is

$$T_{\theta_0} = \frac{\hat{\theta}_{0,E} - \hat{\theta}_{0,C}}{\sqrt{\frac{n_E + n_C}{l_{0,E} + l_{0,C}}\left(\frac{1}{n_E} + \frac{1}{n_C}\right)}}.$$

Again, if $l_{0,E} = 0$ or $l_{0,C} = 0$, T_{θ_0} is set to zero.

The three test statistics are asymptotically standard normally distributed under their respective null hypothesis. Since the MLEs are asymptotically uncorrelated, the three test statistics are also asymptotically uncorrelated. Thus, the intersection of the three local hypotheses is tested by assuming independence of the test statistics and testing every hypothesis at the local level $\tilde{\alpha} = 1 - \sqrt[3]{1-\alpha}$. The local level is chosen such that $(1-\tilde{\alpha})^3 = 1 - \alpha$. The local test procedure consists in calculating the test statistic and comparing its absolute value to the respective quantile $z_{1-\tilde{\alpha}/2}$ of the normal distribution. The global hypothesis H_0 can be rejected if at least one of the local hypotheses can be rejected. This will asymptotically control the Type I error rate at the level α. This test procedure is called *approximate RSES test* throughout this dissertation. The level $\tilde{\alpha}$ is based on an equal split of α. However, the procedure can easily be adapted to other allocation methods.

3.3.2 Exact calculation of rejection probability

In this section, exact formulas for the rejection probability of the approximate test are derived for the case of no censoring or exponential censoring. Let $U \sim$

$\operatorname{Exp}(\lambda_U)$ be the censoring time and $p_E, \lambda_{1,E}, \lambda_{0,E}, p_C, \lambda_{1,C}, \lambda_{0,C}$ the distribution parameters in experimental and control group. For the case of no censoring, plug $\lambda_U = 0$ into the following calculations. It is seen in Section 3.2.6 that conditional on the number of responders k_E and the number of uncensored responder survival times $l_{1,E}$,

$$\frac{1}{l_{1,E}} \cdot \sum_{m=1}^{k_E} t_{E,m} \sim \Gamma(k_E, l_{1,E} \cdot (\lambda_{1,E} + \lambda_U)).$$

Analogously, conditional on k_C and $l_{1,C}$,

$$\frac{1}{l_{1,C}} \cdot \sum_{m=1}^{k_C} t_{C,m} \sim \Gamma(k_C, l_{1,C} \cdot (\lambda_{1,C} + \lambda_U)).$$

Thus, conditional on $k_E, k_C, l_{1,E}$ and $l_{1,C}$,

$$\begin{aligned} \exp\left(T_{\theta_1} \cdot \sqrt{\frac{n_E + n_C}{l_{1,E} + l_{1,C}} \left(\frac{1}{n_E} + \frac{1}{n_C}\right)}\right) &= \frac{\frac{1}{l_{1,E}} \cdot \sum\limits_{m=1}^{k_E} t_{E,m}}{\frac{1}{l_{1,C}} \cdot \sum\limits_{m=1}^{k_C} t_{C,m}} \\ &\sim \beta'\left(k_C, k_E, \frac{l_{1,E} \cdot (\lambda_{1,E} + \lambda_U)}{l_{1,C} \cdot (\lambda_{1,C} + \lambda_U)}\right), \end{aligned} \tag{3.20}$$

where $\beta'(\alpha, \beta, q)$ is the beta prime distribution, also known as beta distribution of the second kind (Johnson et al. 1995, p. 51-52), scaled with factor q. It can be defined by its density

$$f_{\beta'(\alpha,\beta,q)}(x) = \frac{\left(\frac{x}{q}\right)^{\alpha-1} \left(1 + \left(\frac{x}{q}\right)\right)^{-\alpha-\beta}}{q \cdot B(\alpha, \beta)} \qquad \text{for } x > 0,$$

where B is the Beta function. Conditional on $k_E, k_C, l_{1,E}$ and $l_{1,C}$ with $l_{1,E}, l_{1,C} > 0$, the probability to accept $H_{\theta_1,0}$ can be calculated by

$$\begin{aligned} \mathrm{P}(\text{accept } H_{\theta_1,0} \mid k_E, k_C, l_{1,E}, l_{1,C}) &= \mathrm{P}\big(T_{\theta_1} \le z_{1-\tilde{\alpha}/2} \bigm| k_E, k_C, l_{1,E}, l_{1,C}\big) \\ &\quad - \mathrm{P}\big(T_{\theta_1} \le -z_{1-\tilde{\alpha}/2} \bigm| k_E, k_C, l_{1,E}, l_{1,C}\big) \\ &= F_{\beta'\left(k_C, k_E, \frac{l_{1,E} \cdot (\lambda_{1,E}+\lambda_U)}{l_{1,C} \cdot (\lambda_{1,C}+\lambda_U)}\right)}(u(l_{1,E}, l_{1,C})) \\ &\quad - F_{\beta'\left(k_C, k_E, \frac{l_{1,E} \cdot (\lambda_{1,E}+\lambda_U)}{l_{1,C} \cdot (\lambda_{1,C}+\lambda_U)}\right)}(1/u(l_{1,E}, l_{1,C}))\,, \end{aligned}$$

with local level $\tilde{\alpha}$ and

$$u(l_{1,E}, l_{1,C}) = \exp\left(z_{1-\tilde{\alpha}/2} \cdot \sqrt{\frac{n_E + n_C}{l_{1,E} + l_{1,C}} \cdot \left(\frac{1}{n_E} + \frac{1}{n_C}\right)}\right).$$

For $l_{1,E} = 0$ or $l_{1,C} = 0$, the MLE doesn't exist and thus we have to accept $H_{\theta_1,0}$: $\mathrm{P}(\text{accept } H_{\theta_1,0} \mid k_E, k_C, l_{1,E}, l_{1,C}) = 1$. Let $q_{j,i} = \lambda_{j,i}/(\lambda_{j,i} + \lambda_U)$ be the probability of observing an event for one specific patient in response stratum j and group i. The probability to accept $H_{\theta_1,0}$ conditional only on k_E and k_C then is

$$\begin{aligned}
&\mathrm{P}(\text{accept } H_{\theta_1,0} \mid k_E, k_C) \\
&\quad = \sum_{l_{1,E}=0}^{k_E} \sum_{l_{1,C}=0}^{k_C} f_{k_E,q_{1,E}}(l_{1,E}) \cdot f_{k_C,q_{1,C}}(l_{1,C}) \cdot \mathrm{P}(\text{accept } H_{\theta_1,0} \mid k_E, k_C, l_{1,E}, l_{1,C}).
\end{aligned}$$

Analogously, conditional on $k_E, k_C, l_{0,E}$ and $l_{0,C}$ with $l_{0,E}, l_{0,C} > 0$, the probability to accept $H_{\theta_0,0}$ can be calculated by

$$\begin{aligned}
&\mathrm{P}(\text{accept } H_{\theta_0,0} \mid k_E, k_C, l_{0,E}, l_{0,C}) \\
&\quad = \mathrm{P}\big(T_{\theta_0} \leq z_{1-\tilde{\alpha}/2} \bigm| k_E, k_C, l_{0,E}, l_{0,C}\big) - \\
&\qquad \mathrm{P}\big(T_{\theta_0} \leq -z_{1-\tilde{\alpha}/2} \bigm| k_E, k_C, l_{0,E}, l_{0,C}\big) \\
&\quad = F_{\beta'\left(n_C-k_C, n_E-k_E, \frac{l_{0,E}\cdot(\lambda_{0,E}+\lambda_U)}{l_{0,C}\cdot(\lambda_{0,C}+\lambda_U)}\right)}\left(u(l_{0,E}, l_{0,C})\right) - \\
&\qquad F_{\beta'\left(n_C-k_C, n_E-k_E, \frac{l_{0,E}\cdot(\lambda_{0,E}+\lambda_U)}{l_{0,C}\cdot(\lambda_{0,C}+\lambda_U)}\right)}\left(1/u(l_{0,E}, l_{0,C})\right).
\end{aligned}$$

For $l_{0,E} = 0$ or $l_{0,C} = 0$, $\mathrm{P}(\text{accept } H_{\theta_0,0} \mid k_E, k_C, l_{0,E}, l_{0,C}) = 1$. The probability to accept $H_{\theta_0,0}$ conditional only on k_E and k_C is

$$\begin{aligned}
&\mathrm{P}(\text{accept } H_{\theta_0,0} \mid k_E, k_C) \\
&\quad = \sum_{l_{0,E}=0}^{n_E-k_E} \sum_{l_{0,C}=0}^{n_C-k_C} f_{n_E-k_E,q_{0,E}}(l_{0,E}) \cdot f_{n_C-k_C,q_{0,C}}(l_{0,C}) \cdot \\
&\qquad\qquad \mathrm{P}(\text{accept } H_{\theta_0,0} \mid k_E, k_C, l_{0,E}, l_{0,C}).
\end{aligned}$$

Let $I_p(k_E, k_C)$ be the indicator of accepting $H_{p,0}$. $I_p(k_E, k_C) = 1$ if and only if $|T_p(k_E, k_C)| \leq z_{1-\tilde{\alpha}/2}$. The unconditional probability to accept the global null hypothesis H_0 then is

$$\begin{aligned}
&\mathrm{P}(\text{Accept } H_0) \\
&\quad = \sum_{k_E,k_C} f_{n_E,p_E}(k_E) \cdot f_{n_C,p_C}(k_C) \cdot I_p(k_E, k_C) \\
&\qquad\qquad \cdot \mathrm{P}(\text{Accept } H_{\theta_1,0} \mid k_E, k_C) \cdot \mathrm{P}(\text{Accept } H_{\theta_0,0} \mid k_E, k_C).
\end{aligned} \tag{3.21}$$

The formulas for the exact calculation of rejection probabilities are validated by the second simulation study to assess test characteristics described in Section 2.3. See Figure 24 in Appendix A.4 for a comparison of estimated and exactly calculated rejection probabilities.

3.3.3 Approximate confidence intervals for parameter differences

In this section, approximate confidence intervals for the parameter differences are derived. The canonical effect measures for the local tests of $H_{p,0}, H_{\theta_1,0}, H_{\theta_0,0}$ are the parameter differences corresponding to the numerators of the test statistics, as described in Section 2.3.3. Let $\hat{\Delta}_p = \hat{p}_E - \hat{p}_C, \hat{\Delta}_{\theta_1} = \hat{\theta}_{1,E} - \hat{\theta}_{1,C}$ and $\hat{\Delta}_{\theta_0} = \hat{\theta}_{0,E} - \hat{\theta}_{0,C}$ be the estimators of these effect measures. Using the asymptotic variances derived in (3.4) and (3.6) and the independence of the estimators, the variances of the effect estimators can be estimated by

$$\begin{aligned} \hat{\sigma}^2_{\hat{\Delta}_p} &:= \frac{\hat{p}_E \cdot (1-\hat{p}_E)}{n_E} + \frac{\hat{p}_C \cdot (1-\hat{p}_C)}{n_C}, \\ \hat{\sigma}^2_{\hat{\Delta}_{\theta_1}} &:= \frac{1}{l_{1,E}} + \frac{1}{l_{1,C}}, \\ \text{and } \hat{\sigma}^2_{\hat{\Delta}_{\theta_0}} &:= \frac{1}{l_{0,E}} + \frac{1}{l_{0,C}}. \end{aligned} \tag{3.22}$$

Then, two-sided $1-\alpha$ confidence intervals for the effect measures are given by $\hat{\Delta}_u \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{\Delta}_u}$, where u is one of the parameters p, θ_1 or θ_0. Note that these confidence intervals are only asymptotically equivalent to the test decisions of the local tests. The reason for this is that for the local tests, the variance of the parameter differences is estimated under the local null hypothesis, providing a more accurate variance estimation under the local null hypothesis.

3.3.4 Exact RSES test

In this section, an exact test of H_0 is developed in the case of no censoring. Furthermore, an exact formula for calculating the rejection probability of the exact test is derived. To construct an exact test, $H_{\theta_1,0}$ and $H_{\theta_0,0}$ are tested conditionally on k_E and k_C. By doing so, the test statistic T_{θ_1} becomes a monotone transformation of the simplified test statistic

$$\tilde{T}_{\theta_1} := \hat{\theta}_{1,E} - \hat{\theta}_{1,C}.$$

It is

$$\exp(\tilde{T}_{\theta_1}) = \frac{\exp(-\hat{\theta}_{1,C})}{\exp(-\hat{\theta}_{1,E})}$$

and in Section 3.2.6 it was shown $\exp(-\hat{\theta}_{1,E}) \sim \Gamma(k_E, k_E\lambda_{1,E})$ and $\exp(-\hat{\theta}_{1,C}) \sim \Gamma(k_C, k_C\lambda_{1,C})$. Thus, as described in Section 3.3.2,

$$\exp(\tilde{T}_{\theta_1}) \sim \beta'\left(k_C, k_E, \frac{k_E\lambda_{1,E}}{k_C\lambda_{1,C}}\right).$$

The beta prime distribution is also described in Section 3.3.2. Under the null hypothesis $\lambda_{1,E} = \lambda_{1,C}$, the exact distribution of $\tilde{T}_{\theta_1}$ is independent of $\lambda_{1,E}$ and $\lambda_{1,C}$. Conditional on k_E, k_C, the exact p-value of the test of $H_{\theta_1,0}$ can then be calculated by

$$p_{\theta_1}\left(\tilde{T}_{\theta_1} \mid k_E, k_C\right) = 1 - F_{\beta'\left(k_C,k_E,\frac{k_E}{k_C}\right)}\left(\exp(\tilde{T}_{\theta_1})\right) + F_{\beta'\left(k_C,k_E,\frac{k_E}{k_C}\right)}\left(\exp(-\tilde{T}_{\theta_1})\right).$$

$\tilde{T}_{\theta_0}$ and $p_{\theta_0}\big(\tilde{T}_{\theta_0} \mid k_E, k_C\big)$ are defined analogously.

$H_{p,0}$ can be tested exactly by using any exact test for the comparison of two binomial proportions, e.g. Boschloo's test (Boschloo 1970). For consistency, an exact test based on the test statistic T_p is used in this thesis. This test was investigated and named *Z-Pooled Exact Unconditional Test* in an article about exact tests for the difference of two binomial proportions (Mehrotra et al. 2003). Let k_E and k_C be the observed number of responders in the treatment groups. The exact two-sided p-value depends on the true response probability p_0 under H_0 and can be calculated by

$$p_{p,p_0}(T_p(k_E,k_C)) = \sum_{m_E=0}^{n_E}\sum_{m_C=0}^{n_C} f_{n_E,p_0}(m_E) f_{n_C,p_0}(m_C) \cdot 1_{|T_p(m_E,m_C)| \geq |T_p(k_E,k_C)|}. \tag{3.23}$$

Since the true response probability p_0 under H_0 is not known, the test decision of the exact test of $H_{p,0}$ is made with the maximum p-value

$$p_p(T_p) = \max_{p_0 \in [0,1]} p_{p,p_0}(T_p).$$

Hence, the exact test procedure consists in computing the exact local p-value p_p for the test of $H_{p,0}$ and the exact local p-values p_{θ_1} and p_{θ_0} conditionally on k_E and k_C. If one of the p-values is smaller than the local level $\tilde{\alpha}$, the global hypothesis H_0 is rejected. Let A_p be the rejection region of the exact test of $H_{p,0}$. Then the Type I error rate is:

$$\begin{aligned}
&\mathrm{P}(\text{Reject } H_0)\\
&= \mathrm{P}(\text{Reject } H_{p,0}) + \mathrm{P}\left(\text{Accept } H_p \text{ and } (\text{Reject } H_{\theta_1,0} \text{ or Reject } H_{\theta_0,0})\right)\\
&= \mathrm{P}(\text{Reject } H_{p,0})\\
&\qquad + \sum_{(k_E,k_C)\notin A_p} f(k_E,k_C) \cdot \mathrm{P}(\text{Reject } H_{\theta_1,0} \text{ or Reject } H_{\theta_0,0} \mid k_E,k_C)\\
&= \mathrm{P}(\text{Reject } H_{p,0})\\
&\quad + \sum_{(k_E,k_C)\notin A_p} f(k_E,k_C) \cdot (1 - \mathrm{P}(\text{Accept } H_{\theta_1,0} \text{ and Accept } H_{\theta_0,0} \mid k_E,k_C))\\
&= \mathrm{P}(\text{Reject } H_{p,0})\\
&\quad + \sum_{(k_E,k_C)\notin A_p} f(k_E,k_C)\\
&\qquad\quad \cdot (1 - \mathrm{P}(\text{Accept } H_{\theta_1,0} \mid k_E,k_C) \cdot \mathrm{P}(\text{Accept } H_{\theta_0,0} \mid k_E,k_C))\\
&\leq \mathrm{P}(\text{Reject } H_{p,0}) + \sum_{(k_E,k_C)\notin A_p} f(k_E,k_C) \cdot (1 - (1-\tilde{\alpha}) \cdot (1-\tilde{\alpha}))\\
&= \mathrm{P}(\text{Reject } H_{p,0}) + (1 - \mathrm{P}(\text{Reject } H_{p,0})) \cdot (1 - (1-\tilde{\alpha})^2)\\
&= 1 - (1 - \mathrm{P}(\text{Reject } H_{p,0})) \cdot (1-\tilde{\alpha})^2\\
&\leq 1 - (1-\tilde{\alpha})^3\\
&= \alpha
\end{aligned} \tag{3.24}$$

The fourth equality holds because conditional on k_E and k_C, T_{θ_1} and T_{θ_0} are independent. The inequality after that is actually an equality because T_{θ_1} and

T_{θ_0} have a continuous distribution and hence their exact tests exploit the local level. Thus, the procedure yields an exact test for H_0 controlling the Type I error rate at the level α.

The rejection probability of this exact testing procedure can be calculated exactly. Conditional on k_E, k_C, the local tests of $H_{\theta_j,0}$ reject the null hypothesis if $p_{\theta_j}(\tilde{T}_{\theta_j}) \leq \tilde{\alpha}$. Let $c_{\theta_j}(k_E, k_C, \tilde{\alpha}) > 0$ be the critical value defined by the equation

$$p_{\theta_j}\big(c_{\theta_j}(k_E, k_C, \tilde{\alpha})\big) = \tilde{\alpha}.$$

c_{θ_j} can be calculated numerically by solving the equation. Then, the acceptance probability of $H_{\theta_1,0}$ conditional on k_E, k_C can be calculated by

$$\begin{aligned} \mathrm{P}(\text{Accept } H_{\theta_1,0} \mid k_E, k_C) = F_{\beta'\left(k_C, k_E, \frac{k_E \lambda_{1,E}}{k_C \lambda_{1,C}}\right)}&(\exp(c_{\theta_1}(k_E, k_C, \tilde{\alpha}))) \\ - F_{\beta'\left(k_C, k_E, \frac{k_E \lambda_{1,E}}{k_C \lambda_{1,C}}\right)}&(\exp(-c_{\theta_1}(k_E, k_C, \tilde{\alpha}))) \,. \end{aligned}$$

Note that if $k_E = 0$ or $k_C = 0$, $H_{\theta_1,0}$ cannot be rejected. So in this case it is $\mathrm{P}(\text{Accept } H_{\theta_1,0} \mid k_E, k_C) = 1$. The acceptance probability of $H_{\theta_0,0}$ can be calculated analogously.

The critical value of the exact test of $H_{p,0}$ is the largest value $c_p(\tilde{\alpha})$ such that

$$p_p\big(c_p(\tilde{\alpha})\big) \leq \tilde{\alpha}. \tag{3.25}$$

Note that the function p_p is not continuous since it includes the function p_{p,p_0} defined in equation (3.23) which is the sum of indicator functions. Hence, the solution of equation (3.25) cannot be found by standard numerical methods. However, $T_p(k_E, k_C)$ can only take a finite number of values, since $k_i \in \{0, \ldots, n_i\}$. Thus, T_p can be calculated for every pair of (k_E, k_C). Then, the exact p-value p_p can be calculated for every possible value of T_p. For this, the maximum of p_{p,p_0} as defined in equation (3.23) can be found by performing a grid search over the interval $[0, 1]$. The critical value is then the largest of the exact p-values fulfilling equation (3.25).

With these considerations, the exact acceptance probability of the exact RSES test can be calculated by

$$\begin{aligned} \mathrm{P}(\text{Accept } H_0) = \sum_{\substack{k_E, k_C \\ T_p(k_E, k_C) \leq c_p(\tilde{\alpha})}} & f_{n_E, p_E}(k_E) \cdot f_{n_C, p_C}(k_C) \\ & \cdot \mathrm{P}(\text{Accept } H_{\theta_1,0} \mid k_E, k_C) \\ & \cdot \mathrm{P}(\text{Accept } H_{\theta_0,0} \mid k_E, k_C). \end{aligned} \tag{3.26}$$

Formula (3.26) is used to calculate Type I error rate and power of the exact test in the scenarios of the second simulation study described in Section 2.4. The results of these calculations are shown in Appendix A.1.

3.4 Assessment of test characteristics

In this section, Type I error rate and power of the approximate test are analysed and compared to the logrank test and the stratified logrank test. Furthermore, coverage probability of the approximate confidence intervals for parameter difference and correlation of the local test decisions is assessed. Probabilities and correlation coefficients are estimated by the second simulation study described in Section 2.4.1.

3.4.1 Assessment of Type I error rate

In this section, Type I error rates are evaluated in six scenarios with four different censoring distributions. In scenarios TOE1.1, TOE1.2, and TOE1.3, the survival of responders and non-responders is equal. In scenarios TOE2.1, TOE2.2, and TOE2.3, a hazard ratio of 0.4 between responders and non-responders is assumed. The response probability varies over the three values $p = 0.13, 0.26, 0.52$. See Section 2.4 for a detailed description of the scenarios.

Figure 10 shows Type I error rate of the approximate RSES test, the logrank test, and the stratified logrank test in all scenarios. The two logrank tests adhere to the nominal level pretty well. The approximate RSES test exceeds the nominal level for small sample sizes but performs similar to the logrank tests for larger sample sizes.

To further investigate the Type I error rate of the approximate RSES test, Figure 11 shows the estimated rejection probabilities of the local tests compared to the local level $\tilde{\alpha} = 1 - \sqrt[3]{1-\alpha}$. The rejection probabilities of the local tests are not denoted Type I rate to prevent confusion with the Type I error rates of the tests of survival difference. Three things are noticeable:

1. It is seen that for small response probabilities (scenarios TOE1.1 and TOE2.1), the rejection probability of $H_{\theta_1,0}$ considerably exceeds $\tilde{\alpha}$. This is due to the poor normal approximation of the small responder stratum. For a sample size of 50 per group and a response probability of 0.13, the expected number of responders is only 6.5 per group.
2. In the left bottom panel (scenario TOE2.1), it is seen that rejection probability of $H_{\theta_1,0}$ for the strongest censoring distribution Exp(0.04) is not monotonous decreasing with sample size. In this scenario, the survival of responders in the control group is Exp(0.0284). This means that only 42% of responders are expected to have an event. Hence, with a group sample size of 50 and a response probability of 0.13, the case of no events in the responder stratum in the control group occurs relatively often with a probability of 6 %. Then, the MLE $\hat{\theta}_{1,C}$ doesn't exist and $H_{\theta_1,0}$ cannot be rejected.
3. The downward spikes in the rejection probability of $H_{p,0}$ in the right panels (scenarios TOE1.3 and TOE2.3) are due to the discreteness of the number of responders.

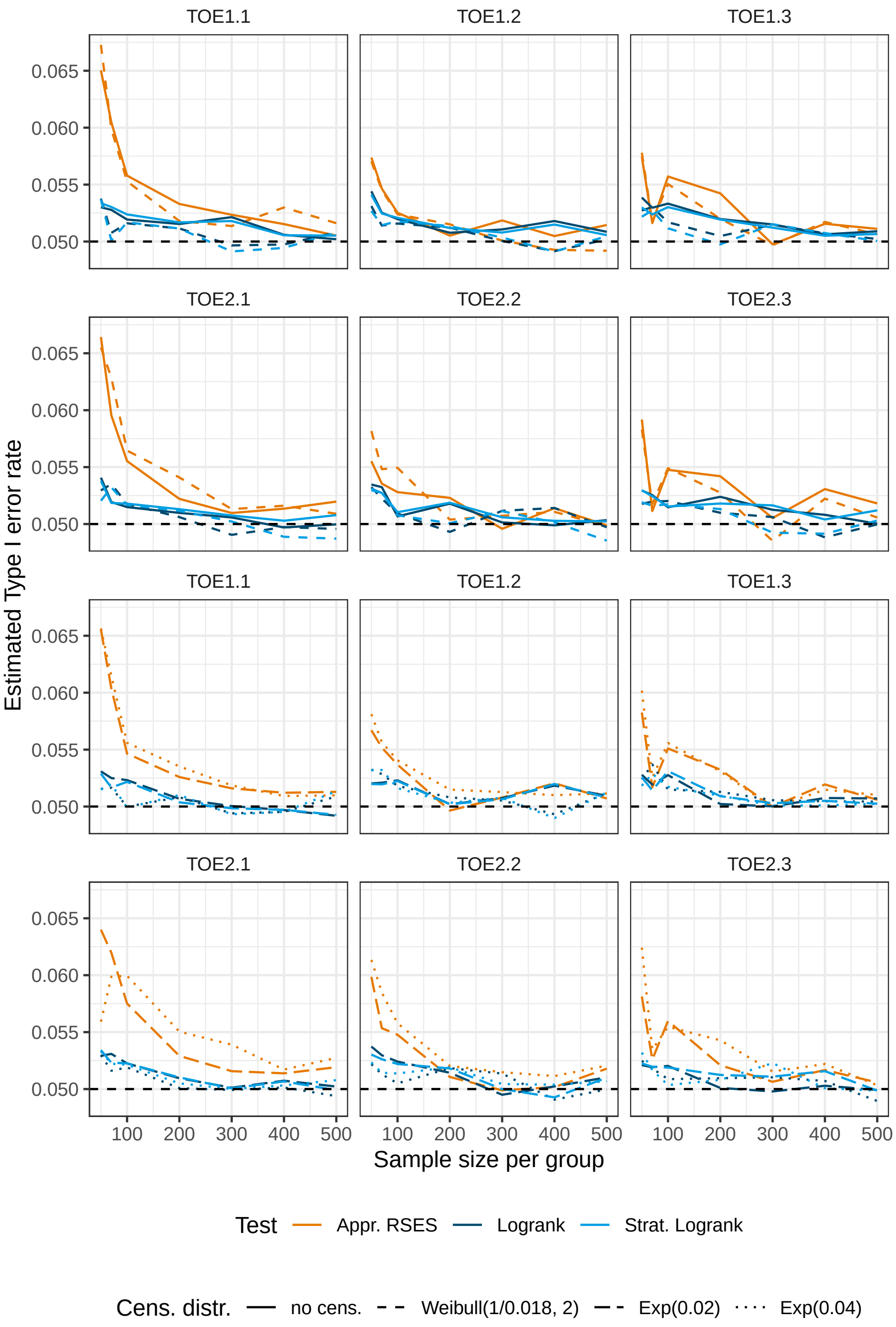


Figure 10: Estimated Type I error rate for three tests, four censoring distributions and sample sizes 50, . . . , 500 in six scenarios TOE1.1, . . . , TOE2.3. For better visibility, only two censoring distributions per panel are shown, resulting in 12 panels.

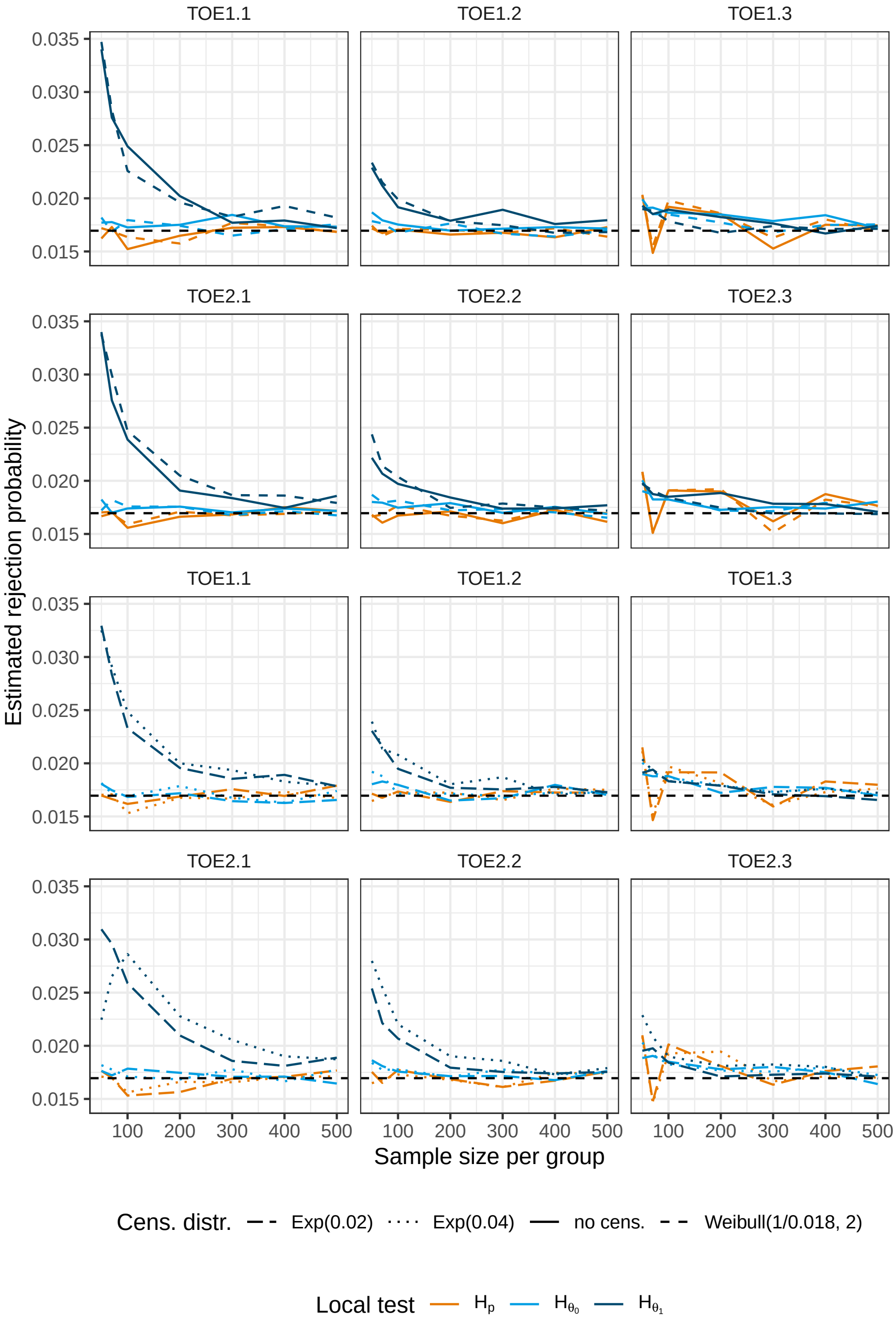


Figure 11: Estimated rejection probabilities of local tests for various censoring distributions and sample sizes 50, . . . , 500 in six scenarios TOE1.1, . . . , TOE2.3. For better visibility, only two censoring distributions per panel are shown, resulting in 12 panels.

3.4.2 Assessment of Power

In this section, power is evaluated in five scenarios. In all scenarios, responder survival is better than non-responder survival. In the first two scenarios (+resp), survival benefit of the experimental group is solely due to a higher response probability. In scenarios 3 and 4 (+resp +surv), there is additionally a survival benefit of the experimental group within the strata. In Scenario 5 (+surv), the response probabilities are equal in both groups. The survival benefit of the experimental group is solely due to a better survival of both responders and non-responders. See Section 2.4 for a detailed description of the scenarios.

Figure 12 shows power for sample sizes per group from 50 to 500. When survival benefit is solely due to response benefit (+resp), the approximate RSES test is much more powerful than the logrank test. Since the stratified logrank test only considers survival differences within the response strata, its power equals the significance level. The higher the survival benefit within the strata compared to the response benefit, the better perform the logrank tests compared to the approximate RSES test. This is because they don't spend significance level to test response difference. When there is no response difference at all (+surv), the logrank tests are more powerful.

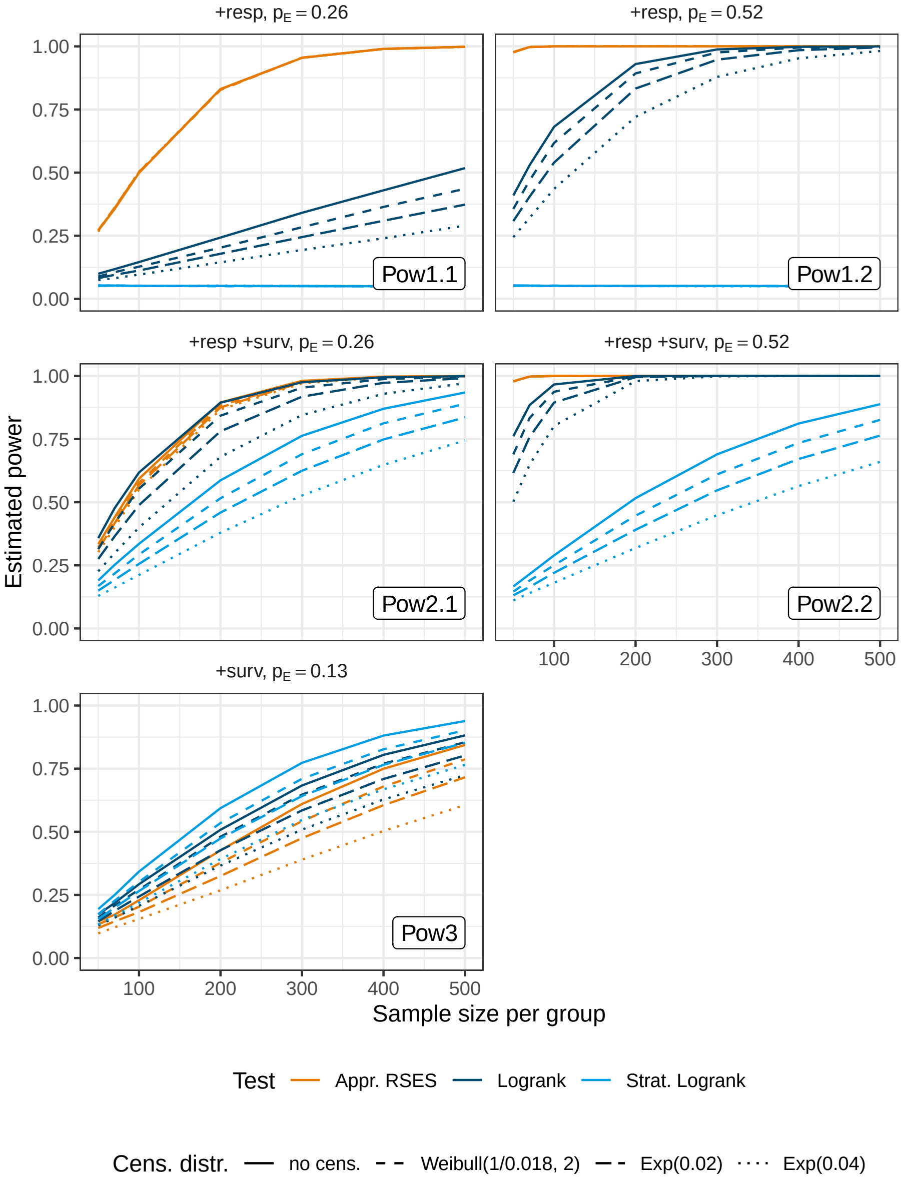


Figure 12: Estimated power for sample sizes 50, . . . , 500, three tests, four censoring distributions, and different treatment effect scenarios Pow1.1, . . . , Pow3. p_E denotes the response probability in the experimental group. +resp means that the experimental group has a response benefit. +surv means that both response strata in the experimental group have better survival than in the control group. +resp +surv means that both effects are present.

3.4.3 Assessment of coverage probability of approximate confidence intervals for parameter difference

In this section, the coverage probability of the approximate confidence intervals derived in Section 3.3.3 is investigated. Figure 13 shows estimated coverage probabilities. In most cases, the coverage probability is very close to the desired value of 95%. In all cases, the coverage probability is within 95% ± 2%. It is noticeable that for small response probabilities and small sample sizes (scenarios TOE1.1, TOE2.1, Pow1.1, Pow2.1, Pow3), the coverage probability for the parameter difference regarding θ_1 deviates from 95%. Interestingly, the size and direction of this deviation depends on the censoring distribution. While coverage probabilities are too small in the case of no censoring, they are too large in the case of the strongest censoring.

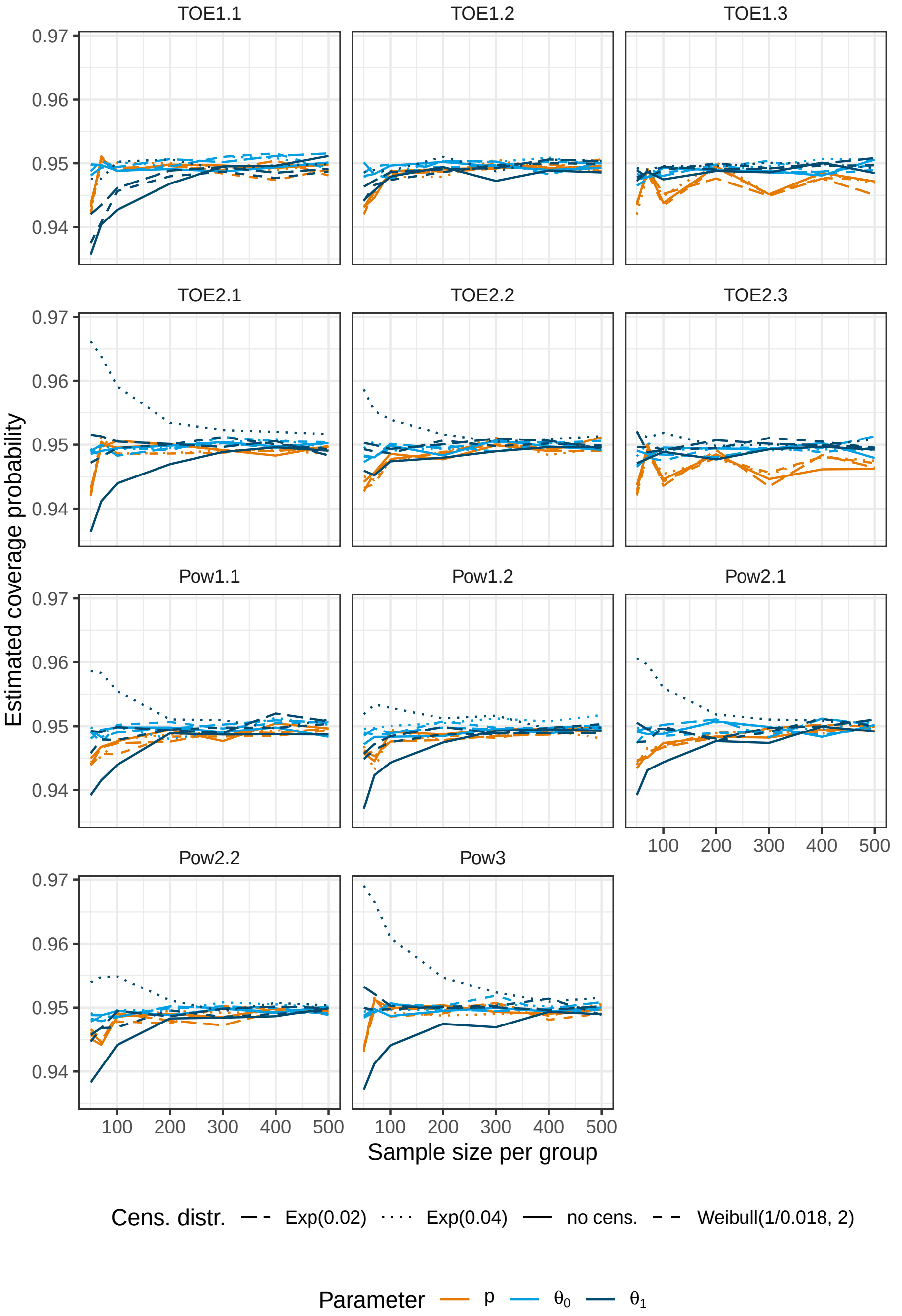


Figure 13: Estimated coverage probability of approximate confidence intervals for parameter differences between treatment groups. Estimated coverage probabilities are shown for various sample sizes, the three model parameters p, θ_1, θ_2, different scenarios TOE1.1, . . . , Pow3, and censoring disributions.

3.4.4 Assessment of independence of local test decisions

In Section 3.3.1, it is argued that the three local test statistics are asymptotically independent. Based on that, the significance level α is split multiplicatively into the local levels $\tilde{\alpha} = 1 - \sqrt[3]{1-\alpha}$ such that $(1-\tilde{\alpha})^3 = 1-\alpha$. In this section, the pairwise independence of the rejections of the local test statistics is investigated as described in Section 2.4.5. Let R_p, R_{θ_1}, and R_{θ_0} be the binary random variables indicating the rejection of the respective local test. Figure 14 shows estimated pairwise Pearson correlation coefficients between R_p, R_{θ_1}, and R_{θ_0} for some scenarios of the second simulation study. Further scenarios are shown in Figure 22 in Appendix A.3, but results are similar. It is seen that correlation is generally small and approaches zero for larger sample sizes. Only in scenarios with small response probabilities and sample sizes, there is a small positive correlation of R_p and R_{θ_1} with estimated correlation coefficients between 0.03 and 0.1. This is consistent with the correlation of $\hat{p}$ and $\hat{\theta}_1$ that is seen in Section 3.2.4 for these scenarios.

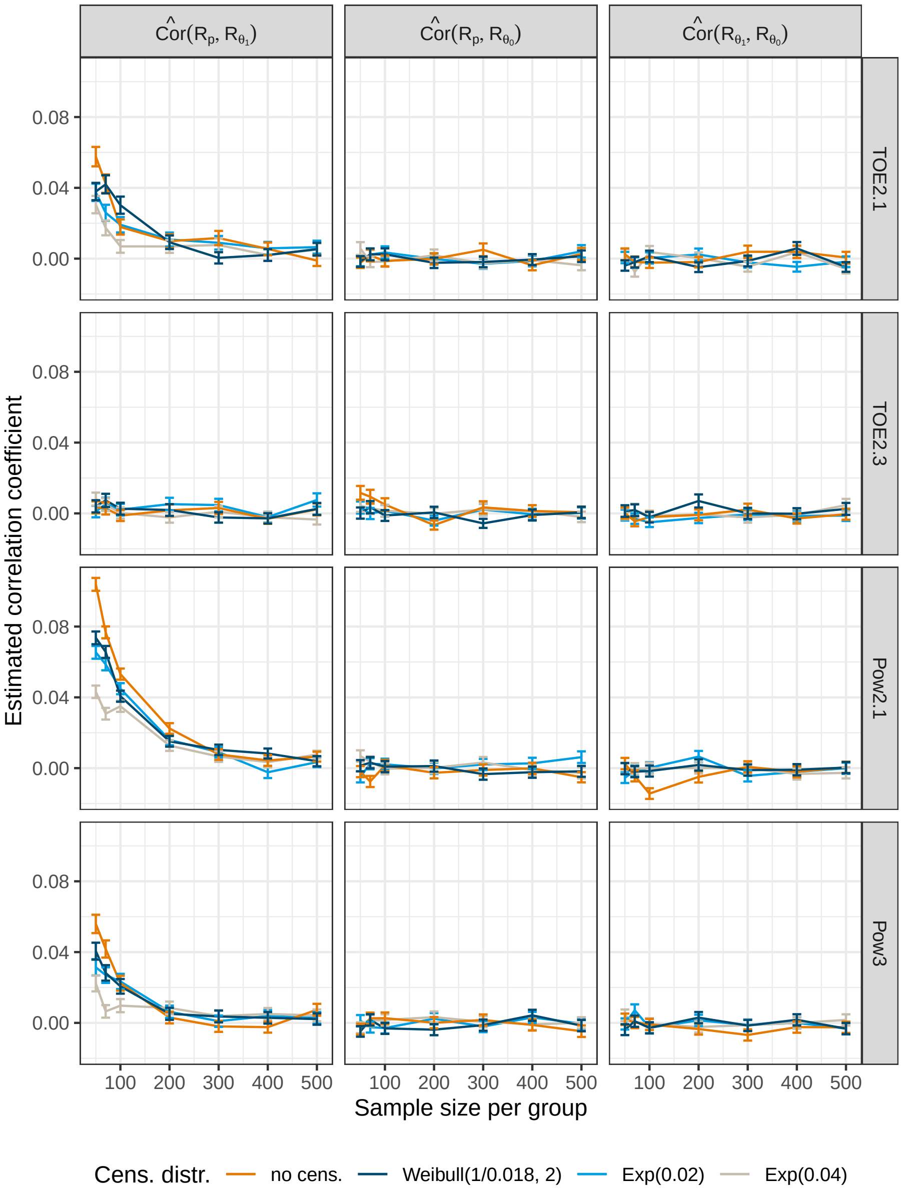


Figure 14: Estimated correlation coefficients of pairs of local test rejections $R_p, R_{\theta_1}, R_{\theta_0}$ in four scenarios TOE2.1, TOE2.3, Pow2.1, and Pow3 for different censoring distributions. Error bars indicate ± standard error.

3.5 Sample size calculation

In this section, an approximate and an exact sample size calculation method are derived for the approximate RSES test. Firstly, approximate formulas for the probabilities to falsely accept the local null hypotheses $H_{p,0}, H_{\theta_1,0}, H_{\theta_0,0}$ are derived. This is done using the approximate normality of the test statistics. Let n_E, n_C be the sample sizes and $p'_i, \theta'_{j,i}$ the specified parameter values under the assumed alternative hypothesis for group $i = E, C$ and response stratum $j = 0, 1$. The test statistic for testing $H_{p,0}$ is

$$T_p = \frac{\hat{p}_E - \hat{p}_C}{\sqrt{\tilde{p}(1-\tilde{p})(\frac{1}{n_E} + \frac{1}{n_C})}}$$

with $\tilde{p} = \frac{n_E \hat{p}_E + n_C \hat{p}_C}{n_E + n_C}$. To estimate the expectation of T_p under the alternative hypothesis, the expectation of the MLEs of the response probabilities $\mathrm{E}[\hat{p}_i] = p'_i$ are plugged in. Let $p' = \frac{n_E p'_E + n_C p'_C}{n_E + n_C}$ be the expectation of $\tilde{p}$. Then,

$$\mathrm{E}[T_p] \approx \frac{p'_E - p'_C}{\sqrt{p'(1-p')(\frac{1}{n_E} + \frac{1}{n_C})}}.$$

The variance of $\hat{p}_E - \hat{p}_C$ is given by

$${\sigma'_p}^2 := \frac{p'_E(1-p'_E)}{n_E} + \frac{p'_C(1-p'_C)}{n_C}.$$

Thus, the variance of T_p is approximately

$$\mathrm{Var}(T_p) \approx \frac{{\sigma'_p}^2}{p'(1-p')(\frac{1}{n_E} + \frac{1}{n_C})}$$

Hence, under the alternative it is

$$T_p \overset{\text{appr}}{\sim} N\left(\frac{p'_E - p'_C}{\sqrt{p'(1-p')(\frac{1}{n_E} + \frac{1}{n_C})}}, \frac{{\sigma'_p}^2}{p'(1-p')(\frac{1}{n_E} + \frac{1}{n_C})}\right).$$

The two-sided local test of $H_{p,0}$ rejects the null hypothesis at level $\tilde{\alpha}$ if $T_p > z_{1-\tilde{\alpha}/2}$ or $T_p < -z_{1-\tilde{\alpha}/2}$. Thus, the rejection probability of $H_{p,0}$ under the specified alternative is approximately

$$\begin{aligned} 1 - \beta_p = 1 - \Phi &\left(\frac{z_{1-\frac{\tilde{\alpha}}{2}} \cdot \sqrt{p'(1-p')(\frac{1}{n_E} + \frac{1}{n_C})} - |p'_E - p'_C|}{\sigma'_p}\right) \\ &+ \Phi\left(\frac{-z_{1-\frac{\tilde{\alpha}}{2}} \cdot \sqrt{p'(1-p')(\frac{1}{n_E} + \frac{1}{n_C})} - |p'_E - p'_C|}{\sigma'_p}\right). \end{aligned}$$

The test statistic for testing $H_{\theta_1,0}$ is

$$T_{\theta_1} = \frac{\hat{\theta}_{1,E} - \hat{\theta}_{1,C}}{\sqrt{\frac{n_E + n_C}{l_{1,E} + l_{1,C}}\left(\frac{1}{n_E} + \frac{1}{n_C}\right)}}.$$

To take censoring into account, an assumption of the probability that a patient is not censored has to be made for each group and response stratum. Let $q_{j,i}$ be the probability that a patient in group i and response stratum j is not censored. If the censoring distribution is exponential with parameter λ_U, the probability $q_{j,i}$ can be calculated by $\lambda'_{j,i}/(\lambda'_{j,i} + \lambda_U)$, where $\lambda'_{j,i} = \exp(\theta'_{j,i})$. The expected number of events in each group and stratum then is $\mathrm{E}[l_{j,i}] = n_i \cdot p'_i \cdot q_{j,i}$. Plugging in these estimates yields

$$\mathrm{E}[T_{\theta_1}] \approx \frac{\theta'_{1,E} - \theta'_{1,C}}{\sqrt{\frac{n_E+n_C}{n_E \cdot p'_E \cdot q_{1,E} + n_C \cdot p'_C \cdot q_{1,C}} \left(\frac{1}{n_E} + \frac{1}{n_C}\right)}}$$

The asymptotic variance of $\hat{\theta}_{1,i}$ is given in equation (3.5):

$$\mathrm{Var}(\hat{\theta}_{1,i}) \approx \frac{1}{n_i \exp(\theta_{1,i}) \cdot \mathrm{E}[X_i T_i]}$$

X_i is the response indicator and T_i the observed event-free time of a patient in group i. It is $\mathrm{E}[X_i T_i] = \mathrm{E}[T_i \mid X_i = 1] \cdot \mathrm{P}(X_i = 1)$. Conditional on $X_i = 1$, T_i is the minimum of the responder survival time in group i and the censoring time. Hence, as shown in appendix A.5, $\mathrm{E}[T_i \mid X_i = 1] = \frac{1}{\lambda_{j,i}} \cdot q_{1,i}$. Since $\mathrm{P}(X_i = 1) = p'_i$ and $\exp(\theta_{1,i}) = \lambda_{1,i}$, it follows

$$\mathrm{Var}(\hat{\theta}_{1,i}) \approx \frac{1}{n_i p'_i q_{1,i}}.$$

Therefore, the variance of $\hat{\theta}_{1,E} - \hat{\theta}_{1,C}$ is approximately

$$\sigma'_{\theta_1}{}^2 := \frac{1}{n_E p'_E q_{1,E}} + \frac{1}{n_C p'_C q_{1,C}}.$$

Hence, under the alternative,

$$T_{\theta_1} \overset{\text{appr}}{\sim} N\left(\frac{\theta'_{1,E} - \theta'_{1,C}}{\sqrt{\frac{n_E+n_C}{n_E \cdot p'_E \cdot q_{1,E} + n_C \cdot p'_C \cdot q_{1,C}} \left(\frac{1}{n_E} + \frac{1}{n_C}\right)}}, \frac{\sigma_{\theta_1}{}'}{\frac{n_E+n_C}{n_E \cdot p'_E \cdot q_{1,E} + n_C \cdot p'_C \cdot q_{1,C}} \left(\frac{1}{n_E} + \frac{1}{n_C}\right)} \right).$$

Thus, the rejection probability of $H_{\theta_1,0}$ at level $\tilde{\alpha}$ under the specified alternative is approximately

$$\begin{aligned} 1 - \beta_{\theta_1} =& 1 - \Phi\left(\frac{z_{1-\frac{\tilde{\alpha}}{2}} \cdot \sqrt{\frac{n_E+n_C}{n_E \cdot p'_E \cdot q_{1,E} + n_C \cdot p'_C \cdot q_{1,C}} \left(\frac{1}{n_E} + \frac{1}{n_C}\right)} - |\theta'_{1,C} - \theta'_{1,E}|}{\sigma_{\theta_1}{}'} \right) \\ &+ \Phi\left(- \frac{z_{1-\frac{\tilde{\alpha}}{2}} \cdot \sqrt{\frac{n_E+n_C}{n_E \cdot p'_E \cdot q_{1,E} + n_C \cdot p'_C \cdot q_{1,C}} \left(\frac{1}{n_E} + \frac{1}{n_C}\right)} - |\theta'_{1,C} - \theta'_{1,E}|}{\sigma_{\theta_1}{}'} \right). \end{aligned}$$

Analogously, the rejection probability of $H_{\theta_0,0}$ at level $\tilde{\alpha}$ under the specified alternative is approximately

$$1-\beta_{\theta_0} = 1-\Phi\left(\frac{z_{1-\frac{\tilde{\alpha}}{2}}\cdot\sqrt{\frac{n_E+n_C}{n_E\cdot(1-p'_E)\cdot q_{0,E}+n_C\cdot(1-p'_C)\cdot q_{0,C}}\left(\frac{1}{n_E}+\frac{1}{n_C}\right)}-|\theta'_{0,C}-\theta'_{0,E}|}{{\sigma_{\theta_0}}'}\right)$$

$$+\,\Phi\left(-\frac{z_{1-\frac{\tilde{\alpha}}{2}}\cdot\sqrt{\frac{n_E+n_C}{n_E\cdot(1-p'_E)\cdot q_{0,E}+n_C\cdot(1-p'_C)\cdot q_{0,C}}\left(\frac{1}{n_E}+\frac{1}{n_C}\right)}-|\theta'_{0,C}-\theta'_{0,E}|}{{\sigma_{\theta_0}}'}\right).$$

with

$${\sigma'_{\theta_0}}^2 := \frac{1}{n_E(1-p'_E)q_{0,E}}+\frac{1}{n_C(1-p'_C)q_{0,C}}.$$

Due to the asymptotic independence of the three test statistics, the probability to not reject H_0, i.e. to accept all three local null hypotheses simultaneously, is approximately equal to the product of the acceptance probabilities of the three local null hypotheses.

Let $r = n_E/n_C$ be the desired sample size ratio. Specify power $1-\beta$, significance level α, and all distribution parameters, and set the local level to $\tilde{\alpha} = 1-\sqrt[3]{1-\alpha}$. Then, due to $n_E = r\cdot n_C$, the acceptance probabilities of the three local tests can be viewed as functions of n_C. Thus, the required control group sample size n_C is the solution of the equation

$$\beta_p(n_C)\cdot\beta_{\theta_1}(n_C)\cdot\beta_{\theta_0}(n_C) = \beta,$$

which can be determined numerically. If it is desired to split Type I error rate or power differently to weight certain hypotheses, the calculation method can easily be adapted to such changes.

In the case of no censoring or exponential censoring, power can be calculated exactly as shown in Section 3.3.2. In this case, the required sample size can be calculated exactly by an iterative method:

1. Start with the approximate sample size.
2. Calculate exact power $1-\mathrm{P}(\text{Accept } H_0)$ with formula (3.21).
3. Increase sample size if power is too low, decrease sample size if power is too high.
4. Iterate steps 2 and 3.

The approximate sample sizes are calculated for 29 scenarios and 4 censoring distributions to obtain a power of 0.8 at a significance level of 0.05. The 29 scenarios can be divided into six constellations that represent different combinations of survival parameters in treatment groups and response strata. In each constellation, 5 different response probabilities p_E in the experimental group are considered. The constellations are:

- Constellation 1: Equal survival in all strata
- Constellation 2: Better survival of responders in experimental group (1)
- Constellation 3: Better survival of responders in experimental group (2)

- Constellation 4: Better survival of responders and non-responders in experimental group
- Constellation 5: Better survival of non-responders in experimental group, even better survival of responders in experimental group
- Constellation 6: Better survival of responders in control group, better survival of non-responders in experimental group, even better survival of responders in experimental group

The constellations are described in more detail in Section 2.5. Power with the calculated sample sizes is estimated in the third simulation study described in Section 2.5.2. Figure 15 shows the estimated power in the 29 scenarios and 4 censoring distributions.

It is seen that the approximate sample size calculation method works pretty well for all scenarios and censoring distributions, with power values almost always between 80% and 85%. Power tends to be higher than the desired value of 80% for small sample sizes. Exact sample size calculation for the scenarios with small sample sizes ($n_E, n_C < 100$) shows that approximate sample sizes are never too small and exceed exact sample sizes by a maximum of 2 and mostly by only 0 or 1 (Figure 16).

Exact power calculation is numerically extensive for large sample sizes. Thus, exact sample size calculation is best applied to finetuning small approximate sample sizes, since the approximation works very well for large approximate sample sizes.

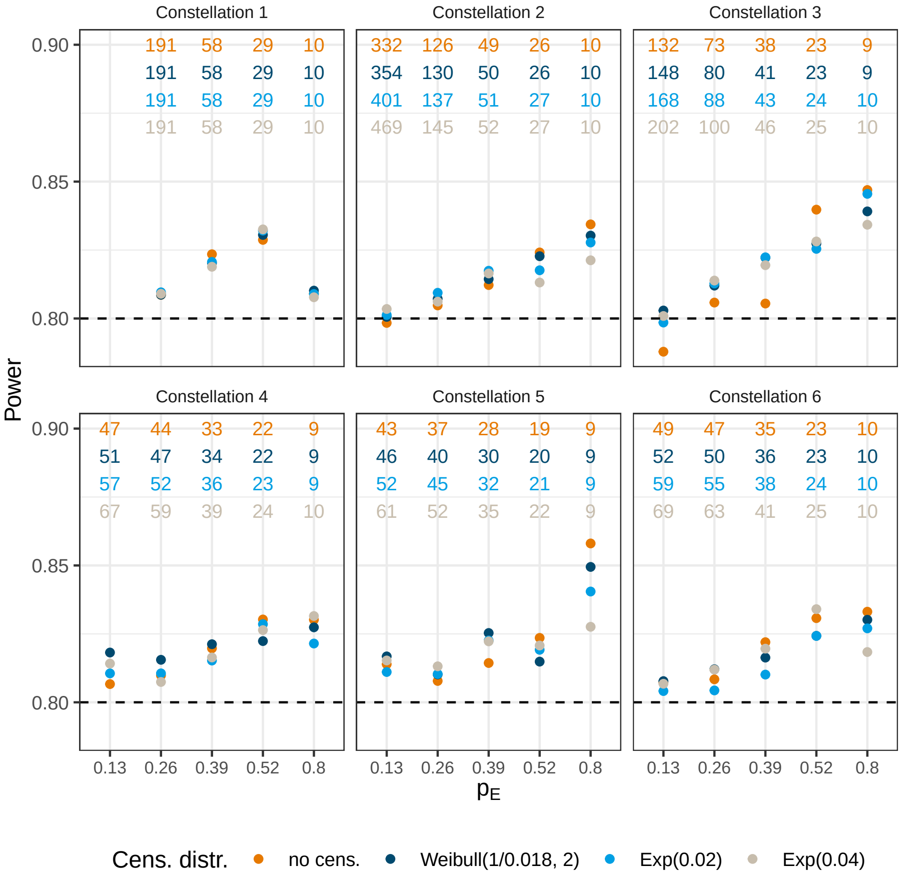


Figure 15: Estimated power for calculated sample sizes in various scenarios. The x axis contains different values of the response probability in the experimental group p_E. Response probability in the control group is set to $p_C = 0.13$ in every scenario. Calculated sample sizes per group are shown in the top row. Constellations $1, \ldots, 6$ comprise different constellations of the survival parameters $\lambda_{j,i}$ in stratum j and treatment group i.

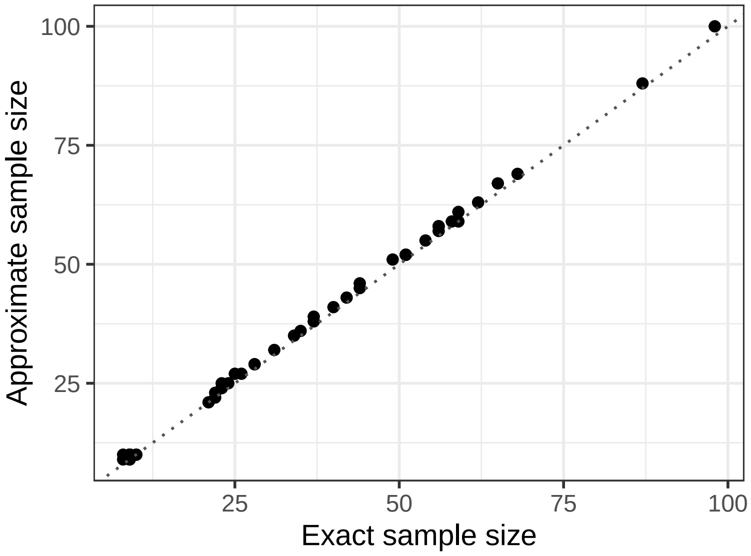


Figure 16: Approximate versus exact sample size for all scenarios where the approximate sample size is smaller than 100. Dotted line indicates equality of approximate and exact sample size.

3.6 Example

Huober et al. (2019) investigated the effect of Lapatinib (L), Trastuzumab (T), and a combination of both (L+T) on pathological complete response (pCR) and survival in patients with HER2-positive early breast cancer. They investigated differences in event-free survival and overall survival between the groups by Cox regression. The sample size calculation was based on the primary endpoint pCR and is described in Baselga et al. (2012). They reported group-wise response rates $p_L = 0.22, p_T = 0.28, p_{L+T} = 0.48$, and overall survival rates at 6 years $S_L(6) = 0.82, S_T(6) = 0.79$, and $S_{L+T}(6) = 0.85$. Furthermore, they estimated hazard ratios between responders and non-responders within the treatment groups: $\lambda_{1,L}/\lambda_{0,L} = 0.54, \lambda_{1,T}/\lambda_{0,T} = 0.45$, and $\lambda_{1,L+T}/\lambda_{0,L+T} = 0.28$. Under the assumption of exponentially distributed survival within the response strata, the RSES distribution parameters can be derived by formula (2.1):

$$S_i(t) = p_i \exp(\lambda_{1,i}t) + (1 - p_i)\exp(\lambda_{0,i}t)$$

Plugging in the values for $t = 6$ and solving for $\lambda_{1,i}$ and $\lambda_{0,i}$ yields the distribution parameters given in Figure 17 together with the survival functions. Under these assumptions, survival of responders is considerably better in all groups. Due to the highest response probability and the best survival for responders, the combination L+T has the best overall survival. Even though a treatment with T leads to a higher response probability compared to L, the non-responder survival in T is worse and the responder survival is almost equal. This results in a better survival of L compared to T. However, survival differences are small between all groups, with survival probabilities at 7 years after randomisation ranging from 0.76 to 0.83. Huober et al. (2019) did not report estimated survival curves within one stratum within one treatment group which is necessary for a detailed check of the distribution assumption of exponential survival within strata. However, they reported survival functions within treatment groups. The curves indicate that the hazard in the first two years after randomisation is considerably lower than for later times. This indicates that the RSES model distribution assumptions are not satisfied. Thus, the RSES model should not be applied here. However, it will be applied in the following for illustrative purposes.

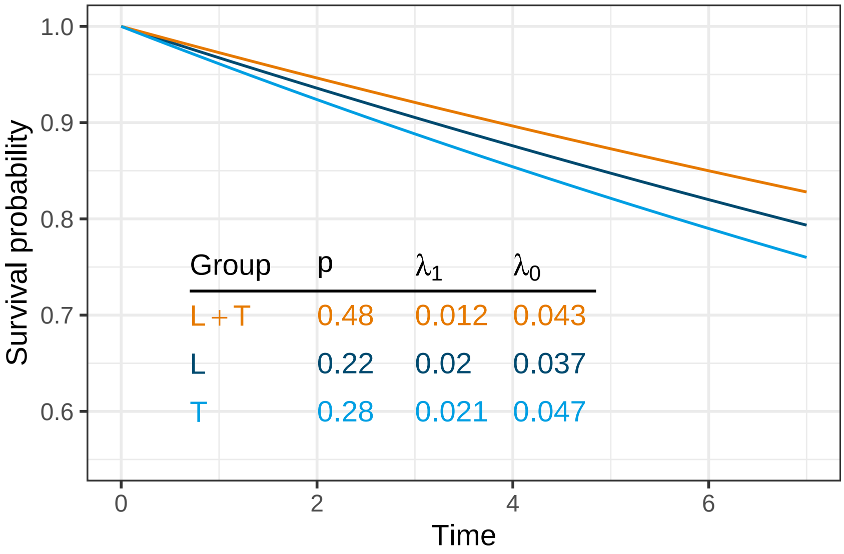
The example trial seemed to have administrative censoring at around 7 years after randomisation. The censoring distribution before that can be derived under the assumption of exponential censoring and independence of censoring and survival. Let $U \sim \text{Exp}(\lambda_U)$ be the censoring distribution. Then, the proportion still at risk at time t in group i is given by $S_i(t) \cdot \exp(-\lambda_U t)$. Plugging in the reported proportions at risk at 6 years of $0.48, 0.52$, and 0.57 in L, T, and L+T, respectively, yields the estimates of $0.09, 0.07$, and 0.07 for λ_U. For the following calculations and simulations, the mean $\lambda_U = 0.075$ of these estimates is taken. The probability to observe an event for a patient in group i and stratum j, assuming administrative censoring at 7 years after randomisation, can be calculated by

$$\begin{aligned} q_{j,i} &= \int_0^7 \lambda_{j,i} \exp(-\lambda_{j,i} t) \exp(-\lambda_u t))\mathrm{d}t \\ &= \frac{\lambda_{j,i}}{\lambda_{j,i} + \lambda_U} \Big(1 - \exp\big(-(\lambda_{j,i} + \lambda_U) \cdot 7\big)\Big). \end{aligned}$$

These assumptions and the method in Section 3.5 are used to calculate the approximate sample sizes for each of the three pairwise comparisons L vs. T, L vs. L+T, and T vs. L+T to obtain a power of 0.8 at a significance level of 0.05 and a randomisation ratio of $r = 1$. The global null hypothesis for one comparison, e.g. for L vs. T, is:

$$H_0 : p_L = p_T \text{ and } \theta_{1,L} = \theta_{1,T} \text{ and } \theta_{0,L} = \theta_{0,T}$$

Table 2 shows the calculated sample sizes for the approximate RSES test for each pairwise comparison. Furthermore, it shows corresponding rejection probabilities of the approximate RSES test, the logrank test, the stratified logrank test, and the three tests of the local hypotheses of the approximate RSES test. Rejection probabilities are estimated by the fourth simulation study described in Section 2.6. Power of the approximate RSES test is very close to the desired value in all scenarios. The required sample size is smallest for the comparison of L+T and L, since these groups differ substantially in response probability, responder survival, as well as non-responder survival. Both the logrank test and stratified logrank test have low power for all comparisons since the differences of marginal survival distributions are small. When comparing T and L+T, the considerable response rate advantage of L+T is the main effect of a better survival in L+T. Since the stratified logrank test deliberately ignores any survival benefit arising from a response benefit, it has considerably lower power than the logrank test for this comparison. When comparing L and T, T has a better response rate which has an positive effect on survival. However, non-responders in group T have worse survival than in group L. These effects partly compensate each other when compared by the logrank test. Thus, the logrank test has lower power than the stratified logrank test, since the stratified logrank test ignores the survival benefit arising from the response benefit. A closer look at the rejection probabilities of the local RSES tests reveals the reason for the striking power difference between approximate RSES test and the logrank tests: the rejection of $H_{p,0}$ is mainly responsible for the rejection of the global null hypothesis H_0.



Group	p	λ_1	λ_0
L+T	0.48	0.012	0.043
L	0.22	0.02	0.037
T	0.28	0.021	0.047

Figure 17: Estimated distribution parameters (response probability p, responder hazard λ_1, non-responder hazard λ_0) and survival functions of the three treatment groups L, T, and L+T in the example study.

Table 2: Estimated rejection probabilities for each of the three pairwise comparisons between the treatment groups L, T, and L+T of the example study. Probabilities are shown for the approximate RSES test, logrank test, stratified logrank test, and the three test of the local hypotheses of the approximate RSES test. Total sample sizes n are calculated by the derived approximate method.

Test	L vs. T $n = 1504$	L vs. L+T $n = 128$	T vs. L+T $n = 236$
Logrank	0.29	0.07	0.21
Strat. Logrank	0.42	0.06	0.09
Appr. RSES	0.80	0.80	0.81
Local test of $H_{p,0}$	0.72	0.79	0.79
Local test of $H_{\theta_1,0}$	0.03	0.01	0.06
Local test of $H_{\theta_0,0}$	0.28	0.03	0.03

Chapter 4

Discussion

Comparing treatments regarding their effect on response rates and survival is a common objective in oncological research (Huober et al. 2019; Bear et al. 2006). Furthermore, promising therapies can acquire preliminary approval based on response benefit which greatly accelerates the availability of new therapies for patients (Food and Drug Administration 2020; Food and Drug Administration 2022). When planning and analysing studies in this context, the correlation between the endpoints response and survival has to be considered. This can be done by the RSES model proposed by Xia et al. (2014).

This dissertation comprises the derivation of basic properties of the RSES model, the construction of estimators and hypothesis tests, and the development of sample size calculation methods. Furthermore, these methods are evaluated within simulation studies and applied to a clinical example. Additionally, the developed approximate RSES test is compared with the logrank test and the stratified logrank test, since both methods are widely used in the described context (Huober et al. 2019; Bear et al. 2006). It is found that when the RSES model distribution of two treatment groups is compared, parameter differences can translate to survival differences in three ways: equal survival, uniformly better survival in one of the groups, or crossing survival curves. Furthermore, estimating the survival parameters within the response strata parameterized as $\theta_j = \log(\lambda_j)$ yields the best approximate normal distribution, as compared to using λ_j or $\eta_j = 1/\lambda_j$. The resulting estimators $\hat{\theta}_j$ have low bias and low root mean squared error. The coverage probability of approximate confidence intervals for the model parameters is very close to the desired confidence level. The pairwise correlation of the estimators is very low and can be neglected for calculations. The developed approximate RSES test adheres well to the significance level. Exceedance of the significance level is only notable for extreme cases where there is a very small expected number of patients in one response stratum of one treatment group. Regarding power, the performance of the approximate RSES test compared to the logrank test and stratified logrank test depends on the constellation of model parameters. When survival benefit in the experimental group is mainly due to more responders, the approximate RSES test is considerably more powerful than the logrank tests. This advantage decreases and can be reversed if survival benefit in the experimental group is mainly due to better survival within the response strata. Applying the approximate RSES test to a clinical example also shows that it is considerably more powerful to detect survival differences mainly originating from a difference in response probabilities. Coverage probability of the derived approximate confidence intervals for model parameter differences is very close to the desired confidence level. The derived approximate sample size calculation method works very well. Approximate sample sizes deviate only very little from exact sample sizes, even for small sample sizes.

It should be mentioned that the assumptions of the RSES model are relatively strict as the survival distribution within response strata is assumed to be exponential (Xia et al. 2014). Thus, before applying the derived estimators or tests to real data, the actual distribution of survival in the response strata has to be evaluated. However, if the distribution assumptions are fulfilled, the derived estimators and confidence intervals perform very well. The results of the derived approximate RSES test have to be interpreted carefully, since the rejection of the null hypothesis H_0 indicates a difference between treatment groups only regarding the RSES model parameters. Although differences between model parameters almost always indicate a difference of survival distributions, as is seen in Section 3.1, such a survival difference does not necessarily indicate a uniform survival benefit. However, the same is true for the logrank test (Mantel 1966) which tests the null hypothesis that survival distributions are equal in both treatment groups. If the logrank test rejects the null hypothesis, it cannot be concluded that survival in one group is uniformly better than in the other group. Instead, the assessment of survival benefit is usually done by comparison of the survival distributions estimated by the Kaplan-Meier method (Kaplan and Meier 1958). Analogous to this approach, the result of the approximate RSES test has to be interpreted under consideration of the estimated parameters and the resulting survival distributions.

In practice, more flexible methods may be desired for estimating and testing survival distributions conditional on a binary response variable. For example, a completely non-parametric estimation approach could consist of estimating survival distributions within response strata by Kaplan-Meier estimators. These could be combined with the estimated response probabilities to estimate the survival distribution of all patients. Testing, on the other hand, could be based on an effect measure that indicates survival benefit. This has the advantage of objectively quantifying survival benefit in a way that is consistent with hypothesis testing. Also, the specification of a summary measure is required within the more and more commonly used Estimands framework (Pohl et al. 2021). Possible choices for such a summary measure are the average hazard ratio (Rauch et al. 2018; Brückner and Brannath 2017) or the difference in Restricted Mean Survival Times (RMST) between the groups (Royston and Parmar 2013). Combining a non-parametric survival estimation method that considers the response status with a meaningful effect measure like RMST could be a flexible way to analyse studies in the described context (Food and Drug Administration 2020). Brückner, Burger, et al. (2018) constructed weighted Kaplan-Meier estimators that consider response status. Furthermore, they give approximate tests for testing the RMST difference and the average hazard ratio. When planning such a study, more concrete assumptions have to be made. For this, the RSES model might be well suited. If, for example, the RMST shall be compared between two cancer therapies that are known to affect both response probability and survival, the effect of these therapies can be modeled by the RSES model for calculating the required sample size. The estimators derived in this dissertation can be used to estimate the model parameters from previous studies. This approach is similar to the Schoenfeld formula that is commonly used for sample size calculation for the logrank test (Schoenfeld 1983). For this formula, the hazard ratio between the treatment groups is assumed to be constant. Just like the RSES model, this assumption is convenient for sample size calculation but might be too restrictive for analysis. If it is desired to test the survival difference between two treatments by the logrank test in the setting of the RSES model, calculating the required sample size with the Schoenfeld formula is not possible due to the violation of the proportional hazards assumption. In this case, a flexible method proposed by Lakatos (1988) can be used for sample size calculation.

The developed exact RSES test might be less interesting for application in practice due to the requirement of no censoring. However, the approach to calculating exact p-values and critical values, and to use this for the development of an exact test, can be applied to other situations and other models.

The main limitation of the results in this dissertation is that the derived methods rely on the relative strict assumptions of the RSES model. No assessment of the robustness of these methods against model misspecification was done. Thus, the approximate RSES test has to be applied carefully in practice. The same is true for the sample size calculation methods as they are tailored to the approximate RSES test. However, the derived methods for point estimation and confidence interval estimation of parameters and parameter differences are useful to develop further methods and to derive assumptions for study planning, as described before. In addition, the findings about the approximate independence of parameter estimators and local tests might be helpful for developing new testing strategies and sample size calculation methods. Another limitation is that this thesis does not develop a method for planning and analysing a study where an early interim decision based on response rates is made, as described in the guideline of the Food and Drug Administration (2020). The derived approximate RSES test assesses parameter differences which, under the assumptions of the RSES model, translate to survival difference. Although a test of the difference between response rates is built into the approximate RSES test as it is one of the three local hypotheses, it is not really suitable for a combination of an interim decision regarding response and a final decision regarding survival. One reason for this is that a positive interim decision, i.e. a rejection of $H_{p,0}$, would already mean that the approximate RSES test rejects the null hypothesis regarding the final endpoint. If, as described before, the test for survival difference is based on an effect measure like RMST, the correlation between an interim decision based on response rates and the final assessment regarding RMST can be modeled by the RSES model. For this, the findings of this dissertation regarding the RSES model, parameter estimation, and correlation between parameters are an important foundation. Brückner, Burger, et al. (2018) proposed a testing strategy with an interim decision that controls the Type I error rate by using a combination test. Although they consider the response status for estimating the treatment effect and calculating the test statistic, the described interim decision is based on survival differences and not on differences in response rates. However, the testing strategy by Brückner, Burger, et al. (2018) may be adaptable to the situation where an interim decision is made based on response rates. Liu and Hu (2016) derived and compared different strategies to control the Type I error rate in the setting of an interim decision based on response rates.

A strength of this dissertation is its comprehensiveness. Every aspect of the RSES model is investigated in detail. For example, the translation of parameter differences to differences of survival distributions is described in Section 3.1 for all possible constellations. With regard to estimation, estimators for all eligible parameterizations of the RSES model are derived and their approximate normality is assessed in Section 3.2. Also, the exact distribution of the derived estimators is investigated in detail. Exact formulas are given for the calculation of coverage probabilities of approximate confidence intervals and of rejection probabilities of the tests. These formulas are validated by simulation studies. Furthermore, the complete R code for applying the derived methods in practice is given. Another strength of this thesis is the generalizability of the presented approach. The performed derivation of point estimators and confidence intervals can easily be applied to other parametric survival models. The same is true for the derived testing procedure and sample size calculation methods.

Further research is needed to develop distribution estimators and a test of survival difference with more flexible distribution assumptions. Then, the RSES model assumptions can be used to derive a sample size calculation method. Another aspect of further research is the planning of studies that combine preliminary and final approval, as described in the guideline of the Food and Drug Administration (2020). For considering the correlation of surrogate endpoint and survival in such a testing strategy, the RSES model and the derived methods in this thesis can be used.

To conclude, this thesis contains a comprehensive investigation of the RSES model. It provides point estimators and confidence interval estimators for the RSES model which are necessary for applying the RSES model in practice. The detailed description of the distribution of these estimators is a useful basis for developing further methods within the RSES model. The derived approximate RSES test and exact RSES test might not always be applicable in practice. However, they can be used for comparison with newly developed testing methods. Furthermore, the general approaches used in this dissertation regarding the derivation of estimators, confidence intervals, hypothesis tests, sample size calculation, and exact calculations can be applied to further models describing the relationship between a surrogate endpoint and a survival endpoint.

Chapter 5

Summary

5.1 English

The primary endpoint in oncology is usually overall survival, where differences between therapies may only be observable after many years. To avoid withholding of a promising therapy, preliminary approval based on a surrogate endpoint is possible. The approval can be confirmed later by assessing overall survival. When planning and analysing trials in this context, the correlation between surrogate endpoint and overall survival has to be taken into account. For the binary surrogate endpoint response, this relation can be modeled by means of the responder stratified exponential survival (RSES) model that was proposed elsewhere. The RSES model has three parameters: response probability p, the logarithmic hazard of responders θ_1, and the logarithmic hazard of non-responders θ_0. The aim of this dissertation is to investigate the RSES model and to develop and evaluate methods for parameter estimation, hypothesis testing, and sample size calculation within the RSES model.

Estimators for the parameters p, θ_1, θ_0 are derived by the Maximum Likelihood method. Approximate confidence intervals for the model parameters are constructed and are found to have very satisfying coverage probability. A hypothesis test for the difference of model parameters between two treatment groups, called approximate RSES test, is constructed. When it is compared with the logrank test and the stratified logrank test regarding power, results vary based on the scenario. When survival benefit in one group is mainly due to response benefit, the approximate RSES test is considerably more powerful than the other tests. Approximate confidence intervals for the parameter differences are derived and show very satisfying coverage probability. Where possible, exact formulas for the calculation of coverage probabilities and rejection probabilities are given. An approximate and an exact sample size calculation method for the approximate RSES test are developed. The sample size calculation method is applied to a clinical example and the power of the approximate RSES test, the logrank test, and the stratified logrank test is compared within this example. The approximate RSES test turns out to be considerably more powerful.

It is discussed that the assumptions of the RSES model are relatively strict. Also,

the results of the approximate RSES test have to be interpreted carefully, since a rejection of the null hypothesis does not necessarily translate to a uniform survival benefit. In practice, more flexible methods may be desired for estimating and testing survival distributions conditional on a binary response variable. Testing could be based on an effect measure that indicates survival benefit, like the Restricted Mean Survival Time (RMST). Combining a non-parametric survival estimation method that considers the response status with a meaningful effect measure like RMST could be a flexible way to analyse studies in the described context. When planning such a study with concrete assumptions, the RSES model can be applied. Also, it is pointed out that the approach presented in this thesis is applicable to other parametric survival models. Further research is needed to develop distribution estimators and a test of survival difference with more flexible distribution assumptions, as well as extending the methods to the situation of an early interim decision based on response rates.

It is concluded that this thesis contains a comprehensive investigation of the RSES model. It provides point estimators and confidence interval estimators for the RSES model which are necessary for applying the RSES model in practice. Furthermore, the general approaches used in this dissertation regarding the derivation of estimators, confidence intervals, hypothesis tests, sample size calculation, and exact calculations can be applied to further models describing the relationship between a surrogate endpoint and a survival endpoint.

5.2 Deutsch

Der primäre Endpunkt in der Onkologie ist in der Regel das Überleben, wobei Unterschiede zwischen den Therapien möglicherweise erst nach vielen Jahren erkennbar sind. Um eine vielversprechende Therapie nicht vorzuenthalten, ist eine vorläufige Zulassung aufgrund eines Surrogatendpunkts möglich. Die Zulassung kann später durch die Untersuchung des Überlebens bestätigt werden. Bei der Planung und Analyse von Studien in diesem Zusammenhang muss die Korrelation zwischen Surrogatendpunkt und Überleben berücksichtigt werden. Für den binären Surrogatendpunkt Response lässt sich diese Korrelation mit Hilfe des in der Literatur vorgeschlagenen Responder Stratified Exponential Survival (RSES) Modells beschreiben. Das RSES-Modell hat drei Parameter: die Response-Wahrscheinlichkeit p, das logarithmische Hazard der Responder θ_1 und das logarithmische Hazard der Non-Responder θ_0. Ziel dieser Dissertation ist es, das RSES-Modell zu untersuchen und Methoden zur Parameterschätzung, zum Hypothesentest und zur Berechnung der Fallzahl im Rahmen des RSES-Modells zu entwickeln und zu untersuchen.

Es werden Schätzer für die Parameter p, θ_1, θ_0 mit der Maximum-Likelihood-Methode abgeleitet. Approximative Konfidenzintervalle für die Modellparameter werden konstruiert und zeigen eine sehr zufriedenstellende Überdeckungswahrscheinlichkeit. Es wird ein Hypothesentest für die Differenz der Modellparameter zwischen zwei Behandlungsgruppen, der so genannte approximative RSES-Test, konstruiert. Beim Vergleich mit dem Logrank-Test und dem stratifizierten Logrank-Test hinsichtlich der Power variieren die Ergebnisse je nach Szenario. Wenn der Überlebensvorteil in einer Gruppe hauptsächlich auf einen Response-Vorteil zurückzuführen ist, hat der approximative RSES-Test deutlich höhere Power als die anderen beiden Tests. Approximative Konfidenzintervalle für die Parameterdifferenzen werden konstruiert und zeigen eine sehr zufriedenstellende Überdeckungswahrscheinlichkeit. Soweit möglich, werden exakte Formeln für die Berechnung der Überdeckungswahrscheinlichkeiten und der Ablehnungswahrscheinlichkeiten angegeben. Eine approximative und eine exakte Fallzahlberechnungsmethode für den approximativen RSES-Test werden entwickelt. Die Methode wird auf ein klinisches Beispiel angewandt, und die Power des approximativen RSES-Tests, des Logrank-Tests und des stratifizierten Logrank-Tests wird innerhalb dieses Beispiels verglichen. Der approximative RSES-Test zeigt wesentlich höhere Power.

Es wird diskutiert, dass die Annahmen des RSES-Modells relativ streng sind. Außerdem müssen die Ergebnisse des approximativen RSES-Tests mit Vorsicht interpretiert werden, da eine Ablehnung der Nullhypothese nicht unbedingt einen gleichmäßigen Überlebensvorteil anzeigt. In der Praxis werden möglicherweise flexiblere Methoden zur Schätzung und zum Vergleich von Überleben unter Berücksichtigung der Response gewünscht. Ein Gruppenvergleich könnte auf einem Effektmaß basieren, das einen Überlebensvorteil anzeigt, wie z.~B. die Restricted Mean Survival Time (RMST). Die Kombination einer nichtparametrischen Schätzmethode, die die Response berücksichtigt, mit einem aussagekräftigen Effektmaß wie der RMST könnte eine flexible Methode zur Analyse von Studien im beschriebenen Kontext darstellen. Wenn eine solche Studie mit konkreten Annahmen geplant wird, kann das RSES-Modell angewendet werden. Es wird weiterhin diskutiert, dass der in dieser Arbeit vorgestellte Ansatz auf andere parametrische Überlebensmodelle anwendbar ist. Weitere Forschung ist erforderlich, um Schätzer und Tests mit flexibleren Verteilungsannahmen zu entwickeln sowie um die Methoden auf die Situation einer frühen Zwischenauswertung auf Basis der Response auszuweiten.

Abschließend wird festgestellt, dass diese Dissertation eine umfassende Untersuchung des RSES-Modells enthält. Sie liefert Punktschätzer und Konfidenzintervalle für das RSES-Modell, die für die Anwendung des RSES-Modells in der Praxis notwendig sind. Darüber hinaus können die in dieser Dissertation verwendeten allgemeinen Ansätze zur Herleitung von Schätzern, Konfidenzintervallen, Hypothesentests, Fallzahlberechnungsmethoden und exakten Berechnungen auf weitere Modelle angewendet werden, die die Beziehung zwischen einem Surrogatendpunkt und einem Überlebensendpunkt beschreiben.

References

Abramowitz, M., Stegun, I. A., and Romer, R. H. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* 10th ed. Applied Mathematics Series no. 55. National Bureau of Standards.

Baselga, J., Bradbury, I., Eidtmann, H., Di Cosimo, S., de Azambuja, E., Aura, C., Gómez, H., Dinh, P., Fauria, K., Van Dooren, V., Aktan, G., Goldhirsch, A., Chang, T.-W., Horváth, Z., Coccia-Portugal, M., Domont, J., Tseng, L.-M., Kunz, G., Sohn, J. H., Semiglazov, V., Lerzo, G., Palacova, M., Probachai, V., Pusztai, L., Untch, M., Gelber, R. D., and Piccart-Gebhart, M. (2012). "Lapatinib with trastuzumab for HER2-positive early breast cancer (NeoALTTO): a randomised, open-label, multicentre, phase 3 trial". In: *The Lancet* 379.9816, pp. 633–640. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(11)61847-3. URL: https://www.sciencedirect.com/science/article/pii/S0140673611618473.

Bear, H. D., Anderson, S., Smith, R. E., Geyer Jr, C. E., Mamounas, E. P., Fisher, B., Brown, A. M., Robidoux, A., Margolese, R., Kahlenberg, M. S., et al. (2006). "Sequential preoperative or postoperative docetaxel added to preoperative doxorubicin plus cyclophosphamide for operable breast cancer: National Surgical Adjuvant Breast and Bowel Project Protocol B-27". In: *J Clin Oncol* 24.13, pp. 2019–2027.

Boschloo, R. D. (1970). "Raised conditional level of significance for the 2 x 2-table when testing the equality of two probabilities". In: *Statistica Neerlandica* 24, pp. 1–9. DOI: 10.1111/j.1467-9574.1970.tb00104.x.

Brückner, M. and Brannath, W. (2017). "Sequential tests for non-proportional hazards data". In: *Lifetime data analysis* 23, pp. 339–352.

Brückner, M., Burger, H. U., and Brannath, W. (2018). "Nonparametric adaptive enrichment designs using categorical surrogate data". In: *Statistics in Medicine* 37.29, pp. 4507–4524.

Conforti, F., Pala, L., Sala, I., Oriecuia, C., De Pas, T., Specchia, C., Graffeo, R., Pagan, E., Queirolo, P., Pennacchioli, E., Colleoni, M., Viale, G., Bagnardi, V., and Gelber, R. D. (2021). "Evaluation of pathological complete response as surrogate endpoint in neoadjuvant randomised clinical trials of early stage breast cancer: systematic review and meta-analysis". In: *BMJ* 375. DOI: 10.1136/bmj-2021-066381. eprint: https://www.bmj.com/content/375/bmj-2021-066381.full.pdf. URL: https://www.bmj.com/content/375/bmj-2021-066381.

Cramér, H. (1974). *Mathematical methods of statistics.* Vol. 13. Princeton university press.

Dormuth, I., Liu, T., Xu, J., Yu, M., Pauly, M., and Ditzhaus, M. (2022). "Which test for crossing survival curves? A user's guideline". In: *BMC Medical Research Methodology* 22.1, p. 34.

Farrington, C. P. and Manning, G. (1990). "Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference of Non-Unity Relative Risk". In: *Statistics in Medicine* 9, pp. 1447–1454. DOI: 10.1002/sim.4780091208.

Food and Drug Administration (2020). *Guidance for industry: Pathologic complete response in neoadjuvant treatment of high-risk early-stage breast cancer: Use as an endpoint to support accelerated approval.* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/pathological-complete-response-neoadjuvant-treatment-high-risk-early-stage-breast-cancer-use.

– (2022). *CDER Drug and Biologic Accelerated Approvals Based on a Surrogate Endpoint As of December 31, 2021.* https://www.fda.gov/drugs/nda-and-bla-approvals/accelerated-approvals.

Hoeffding, W. (1994). "Probability inequalities for sums of bounded random variables". In: *The collected works of Wassily Hoeffding*, pp. 409–426.

Huober, J., Holmes, E., Baselga, J., Azambuja, E. de, Untch, M., Fumagalli, D., Sarp, S., Lang, I., Smith, I., Boyle, F., et al. (2019). "Survival outcomes of the NeoALTTO study (BIG 1–06): updated results of a randomised multicenter phase III neoadjuvant clinical trial in patients with HER2-positive primary breast cancer". In: *European journal of cancer* 118, pp. 169–177.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions, volume 2.* Vol. 289. John Wiley & Sons.

Kaplan, E. L. and Meier, P. (1958). "Nonparametric estimation from incomplete observations". In: *Journal of the American statistical association* 53.282, pp. 457–481.

Kesselheim, A. S., Wang, B., Franklin, J. M., and Darrow, J. J. (2015). "Trends in utilization of FDA expedited drug development and approval programs, 1987-2014: cohort study". In: *Bmj* 351.

Klenke, A. (2014). *Probability theory: a comprehensive course.* 2nd ed. Springer London.

L, L. (1983). *Theory of Point Estimation.* A Wiley publication in mathematical statistics. Wiley. URL: https://books.google.de/books?id=VcXdngEACAAJ.

Lakatos, E. (1988). "Sample sizes based on the log-rank statistic in complex clinical trials". In: *Biometrics*, pp. 229–241.

Lehmann, E. L. and Romano, J. P. (2010). *Testing statistical hypotheses.* 3rd ed. Springer.

Liu, Y. and Hu, M. (2016). "Testing multiple primary endpoints in clinical trials with sample size adaptation". In: *Pharmaceutical statistics* 15.1, pp. 37–45.

Mann, H. B. and Wald, A. (1943). "On stochastic limit and order relationships". In: *The Annals of Mathematical Statistics* 14.3, pp. 217–226.

Mantel, N. (1966). "Evaluation of survival data and two new rank order statistics arising in its consideration". In: *Cancer Chemother Rep* 50, pp. 163–170.

Mehrotra, D. V., Chan, I. S. F., and Berger, R. L. (2003). "A Cautionary Note on Exact Unconditional Inference for a Difference betwenn Two Independent Binomial Proportions". In: *Biometrics* 59, pp. 441–450. DOI: 10.1111/1541-0420.00051.

Pohl, M., Baumann, L., Behnisch, R., Kirchner, M., Krisam, J., and Sander, A. (2021). “Estimands—A Basic Element for Clinical Trials: Part 29 of a Series on Evaluation of Scientific Publications”. In: *Deutsches Ärzteblatt International* 118.51-52, p. 883.

R Core Team (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rauch, G., Brannath, W., Brückner, M., and Kieser, M. (2018). “The Average Hazard Ratio–A Good Effect Measure for Time-to-event Endpoints when the Proportional Hazard Assumption is Violated?” In: *Methods of information in medicine* 57.03, pp. 089–100.

Ross, S. (2010). *A first course in probability.* 8th ed. Pearson.

Royston, P. and Parmar, M. K. (2013). “Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome”. In: *BMC medical research methodology* 13.1, pp. 1–15.

Schoenfeld, D. A. (1983). “Sample-size formula for the proportional-hazards regression model”. In: *Biometrics*, pp. 499–503.

Wallach, J. D., Ross, J. S., and Naci, H. (2018). “The US Food and Drug Administration’s expedited approval programs: Evidentiary standards, regulatory trade-offs, and potential improvements”. In: *Clinical Trials* 15.3, pp. 219–229.

Xia, Y., Cui, L., and Yang, B. (2014). “A Note on Breast Cancer Trials With pCR-Based Accelerated Approval”. In: *Journal of Biopharmaceutical Statistics* 24:5, pp. 1102–1114. DOI: 10.1080/10543406.2014.931410.

Personal contribution to data acquisition / assessment and personal publications

No data were acquired for this thesis. The methods were applied to simulated data. Simulation of data and application of methods was done by myself.

Partial of this thesis have already been published in the following article:

Kilian, S., Krisam, J., and Kieser, M. (2022). *Analysis and sample size calculation within the responder stratified exponential survival model.* DOI: https://doi.org/10.48550/arXiv.2205.11599. arXiv: 2205.11599 `[stat.ME]`.

This article was written by myself and reviewed by the two coauthors J. Krisam and M. Kieser. In this article, the RSES model is introduced and possible constellations of survival distributions between two treatment groups within the RSES model are investigated, as it is done in Section 3.1 of this thesis. Furthermore, analogous to parts of Sections 3.2, 3.3, 3.4, and 3.5, methods for estimation, testing, and sample size calculation are derived and assessed, but only in the case of no censoring. These methods are then applied to the example described in Section 3.6.

Further personal publications:

Burgmaier, K., Kilian, S., Arbeiter, K., Atmis, B., Büscher, A., Derichs, U., Dursun, I., Duzova, A., Eid, L. A., Galiano, M., et al. (2021). “Early childhood height-adjusted total kidney volume as a risk marker of kidney survival in ARPKD”. In: *Scientific reports* 11.1, p. 21677.

Burgmaier, K., Kilian, S., Bammens, B., Benzing, T., Billing, H., Büscher, A., Galiano, M., Grundmann, F., Klaus, G., Mekahli, D., et al. (2019). “Clinical courses and complications of young adults with autosomal recessive polycystic kidney disease (ARPKD)”. In: *Scientific Reports* 9.1, p. 7919.

Dao Trong, P., Kilian, S., Jesser, J., Reuss, D., Aras, F. K., Von Deimling, A., Herold-Mende, C., Unterberg, A., and Jungk, C. (2023). “Risk Estimation in Non-Enhancing Glioma: Introducing a Clinical Score”. In: *Cancers* 15.9, p. 2503.

Dehne, S., Spang, V., Klotz, R., Kummer, L., Kilian, S., Hoffmann, K., Schneider, M. A., Hackert, T., Büchler, M. W., Weigand, M. A., et al. (2021). "Intraoperative Fractions of Inspiratory Oxygen Are Associated With Recurrence-Free Survival After Elective Cancer Surgery". In: *Frontiers in Medicine*, p. 2333.

Dikow, N., Moog, U., Karch, S., Sander, A., Kilian, S., Blank, R., and Reuner, G. (2019). "What do parents expect from a genetic diagnosis of their child with intellectual disability?" In: *Journal of Applied Research in Intellectual Disabilities* 32.5, pp. 1129–1137.

Fink, C., Alt, C., Schank, T. E., Sies, K., Kilian, S., and Schäkel, K. (2022). "Multiarm study comparing patient-reported and clinical outcome measures in patients undergoing antipsoriatic therapy with non-biological systemic agents in a real-world setting". In: *Journal of Dermatological Treatment* 33.7, pp. 2997–3004.

Fink, C., Kilian, S., Bertlich, I., Hoxha, E., Bardehle, F., Enk, A., and Haenssle, H. A. (2018). "Evaluation of capillary pathologies by nailfold capillaroscopy in patients with psoriasis vulgaris: study protocol for a prospective, controlled exploratory study". In: *BMJ open* 8.8, e021595.

Frese, C., Reissfelder, L.-S., Kilian, S., Felten, A., Laurisch, L., Schoilew, K., and Boutin, S. (2022). "Can the Acid-formation Potential of Saliva Detect a Caries-related Shift in the Oral Microbiome?" In: *Oral Health Prev Dent* 20, pp. 51–60.

Klotz, A.-L., Zajac, M., Ehret, J., Kilian, S., Rammelsberg, P., and Zenthöfer, A. (2020). "Short-term effects of a deterioration of general health on the oral health of nursing-home residents". In: *Clinical Interventions in Aging*, pp. 29–38.

– (2021). "Which factors influence the oral health of nursing-home residents with cognitive and motor impairments?" In: *Aging Clinical and Experimental Research* 33, pp. 85–93.

Körfer, D., Erhart, P., Wortmann, M., Dihlmann, S., Grond-Ginsbach, C., Kilian, S., Asatryan, A., Jung, G., Schmitz-Rixen, T., Böckler, D., et al. (2023). "Characteristics of patients with multiple arterial aneurysms". In: *Vasa*.

Mäder-Porombka, C., Homberg, A., Hörster, F., Brune, M., Kilian, S., Buckel, B., and Goldwasser, R. (2020). "Einfluss einer maßgeschneiderten Präanalytik-Fortbildung auf die Qualität der Laboruntersuchungen". In: *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 150, pp. 38–44.

Meisenbacher, K., Hagedorn, M., Skrypnik, D., Kilian, S., Böckler, D., Bischoff, M. S., and Peters, A. S. (2022). "Thoracic Endovascular Aortic Repair (TEVAR) First in Patients with Lower Limb Ischemia in Complicated Type B Aortic Dissection: Clinical Outcome and Morphology". In: *Journal of Clinical Medicine* 11.14, p. 4154.

Pilz, M., Kilian, S., and Kieser, M. (2022). "A note on the shape of sample size functions of optimal adaptive two-stage designs". In: *Communications in Statistics-Theory and Methods* 51.6, pp. 1911–1918.

Rammelsberg, P., Kilian, S., Büsch, C., and Kappel, S. (2020). "The effect of transcrestal sinus-floor elevation without graft on the long-term prognosis of maxillary implants". In: *Journal of Clinical Periodontology* 47.5, pp. 640–648.

Schmidt, S., Kunath, F., Coles, B., Draeger, D. L., Krabbe, L.-M., Dersch, R., Kilian, S., Jensen, K., Dahm, P., and Meerpohl, J. J. (2020a). "Intravesical bacillus Calmette-Guérin versus mitomycin C for Ta and T1 bladder cancer". In: *Cochrane Database of Systematic Reviews* 1.

– (2020b). “Intravesical Bacillus Calmette-Guérin versus mitomycin C for Ta and T1 bladder cancer: Abridged summary of the Cochrane Review”. In: *Investigative and clinical urology* 61.4, p. 349.

Schweizer, A., Fink, C., Bertlich, I., Toberer, F., Mitteldorf, C., Stolz, W., Enk, A., Kilian, S., and Haenssle, H. A. (2020). “Differentiation of combined nevi and melanomas: Case-control study with comparative analysis of dermoscopic features”. In: *JDDG: Journal der Deutschen Dermatologischen Gesellschaft* 18.2, pp. 111–118.

Sekundo, C., Langowski, E., Kilian, S., and Frese, C. (2020a). “Oral health and functional capacity of centenarians”. In: *Scientific Reports* 10.1, pp. 1–10.

– (2020b). “Periodontal and peri-implant diseases in centenarians”. In: *Journal of Clinical Periodontology* 47.10, pp. 1170–1179.

Sekundo, C., Langowski, E., Kilian, S., Wolff, D., Zenthöfer, A., and Frese, C. (2021). “Association of Dental and Prosthetic Status with Oral Health-Related Quality of Life in Centenarians”. In: *International Journal of Environmental Research and Public Health* 18.24, p. 13219.

Sonnenschein, S. K., Ciardo, A., Kilian, S., Ziegler, P., Ruetters, M., Splindler, M., and Kim, T.-S. (2021). “The impact of splinting timepoint of mobile mandibular incisors on the outcome of periodontal treatment—preliminary observations from a randomized clinical trial”. In: *Clinical Oral Investigations*, pp. 1–10.

Trong, P. D., Jesser, J., Deimling, A. von, Kilian, S., Herold-Mende, C., Unterberg, A., and Jungk, C. (2018). “NIMG-30. PREOPERATIVE PREDICTORS OF MALIGNANCY IN NON-ENHANCING GLIOMA IN THE ERA OF MOLECULAR CLASSIFICATION”. In: *Neuro-Oncology* 20.Suppl 6, p. vi182.

Zenthöfer, A., Ehret, J., Zajac, M., Kilian, S., Kostunov, J., Rammelsberg, P., and Klotz, A.-L. (2021). “How do changes in Oral health and chewing efficiency affect the changes of Oral-health-related quality of life of nursing-home residents in the short term?” In: *Clinical Interventions in Aging*, pp. 789–798.

Zenthöfer, A., Ehret, J., Zajac, M., Kilian, S., Rammelsberg, P., and Klotz, A.-L. (2020). “The effects of dental status and chewing efficiency on the oral-health-related quality of life of nursing-home residents”. In: *Clinical Interventions in Aging*, pp. 2155–2164.

Appendix A

Technical details and additional results

A.1 Test characteristics of exact test

Figure 18 shows exactly calculated Type I error rate of the exact RSES test in the case of no censoring for the Type I error scenarios of the second simulation study, but only for sample sizes $n_E = n_c = 50, 60, 70, 80, 90, 100$. The local exact tests of $H_{\theta_j,0}$ completely exploit the local level $\tilde{\alpha}$. However, due to the discreteness of k_E, k_C, the local exact test of $H_{p,0}$ is not able to do that. As a result, Type I error rate of the exact RSES test does not equal the significance level of $\alpha = 0.05$ but is slightly smaller. A further consequence is that the Type I error rate for scenarios TOE1.1, TOE1.2, and TOE1.3 is exactly the same as for scenarios TOE2.1, TOE2.2, and TOE2.3. Thus, only the latter three scenarios are shown.

Figure 19 shows exactly calculated power of the exact RSES test in the case of no censoring for the power scenarios of the second simulation study, and for sample sizes $n_E = n_c = 50, 60, 70, 80, 90, 100$. In scenarios Pow1.2 and Pow 2.2, power is almost equal and very large. This is due to the large response difference between treatment groups in these scenarios.

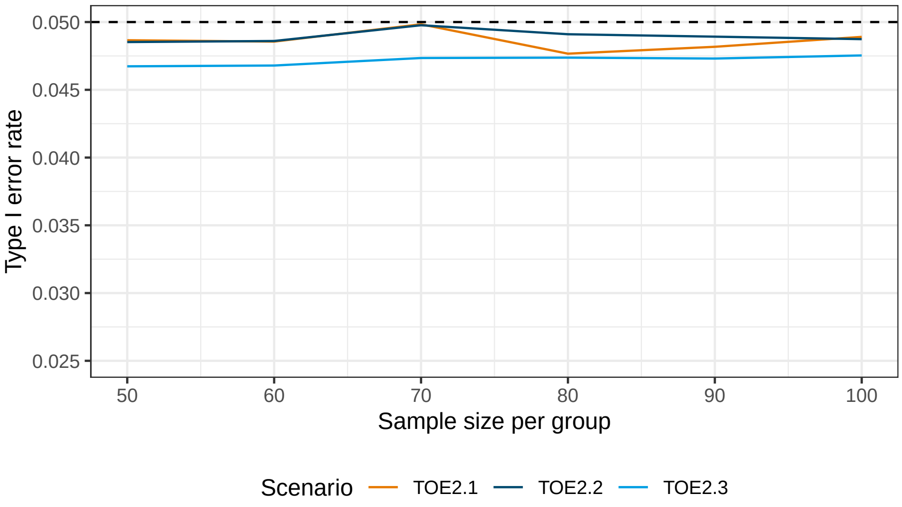


Figure 18: Exact Type I error rate of the exact RSES test for various sample sizes in three scenarios TOE2.1, TOE2.2, TOE2.3. Type I error rate in scenarios TOE1.1, TOE1.2, TOE1.3 is equal to the shown scenarios.

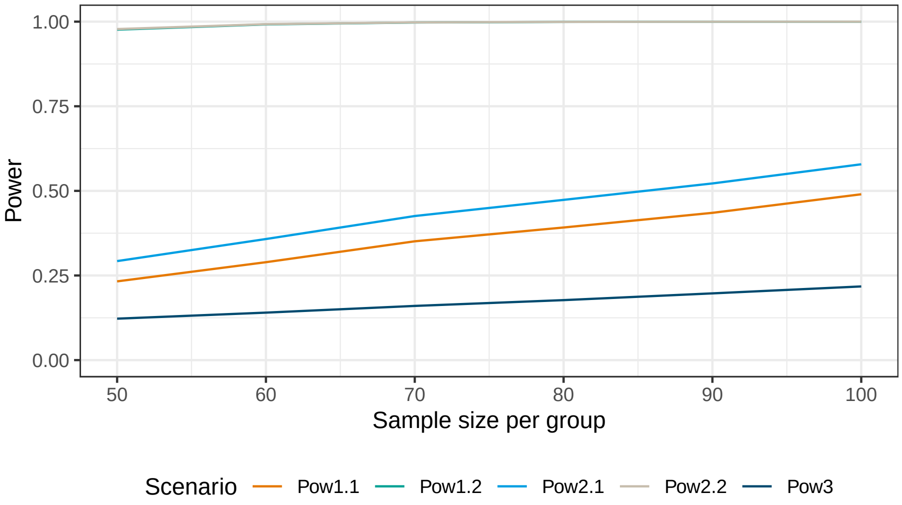


Figure 19: Exact power of the exact RSES test for various sample sizes in different treatment effect scenarios. In scenarios Pow1.2 and Pow 2.2, power is almost equal.

A.2 Logrank test statistics

The logrank test statistic introduced by Mantel (1966) with hypergeometric variance estimation is used in this dissertation. Let $t^{(i)}$ denote the event times. Let $Y^{(i)}$ be the total number at risk and $Y_E^{(i)}$ the number at risk in the experimental group immediately before $t^{(i)}$. Let $d^{(i)}$ be the total number of events and $d_E^{(i)}$ the number of events in the experimental group at $t^{(i)}$. Then

$$E^{(i)} = d \cdot \frac{Y_E^{(i)}}{Y^{(i)}}$$

is the expected number of events in the experimental group at $t^{(i)}$. The conditional variance of $d_E^{(i)}$ is derived from the hypergeometric distribution as

$$V^{(i)} = \frac{(Y^{(i)} - Y_E^{(i)}) \cdot Y_E^{(i)} \cdot (Y^{(i)} - d^{(i)}) \cdot d^{(i)}}{{Y^{(i)}}^2 \cdot (Y^{(i)} - 1)}.$$

The total number of observed and expected events are

$$O = \sum_i d_E^{(i)}$$

and

$$E = \sum_i E^{(i)}.$$

The approximate variance of $O - E$ is

$$V = \sum_i V^{(i)}.$$

The logrank test statistic then is

$$T_{\text{LR}} = \frac{O - E}{\sqrt{V}}.$$

For the stratified logrank test, the quantities O_j, E_j and V_j are calculated within each stratum j. The test statistic then is

$$T_{\text{sLR}} = \frac{\sum_j (O_j - E_j)}{\sqrt{\sum_j V_j}}.$$

A.3 Additional simulation results

Figure 20 shows additional results of the first simulation study described in Section 2.2.4 for assessing the approximate normality of $\hat{\lambda}_j, \hat{\theta}_j$ and $\hat{\eta}_j$.

Figure 21 shows additional results of the assessment of coverage probability of approximate confidence intervals for θ_j, as described in Section 2.2.6.

Figure 22 shows additional results for the assessment of the pairwise correlation between the local test decisions, as described in Section 2.4.5.

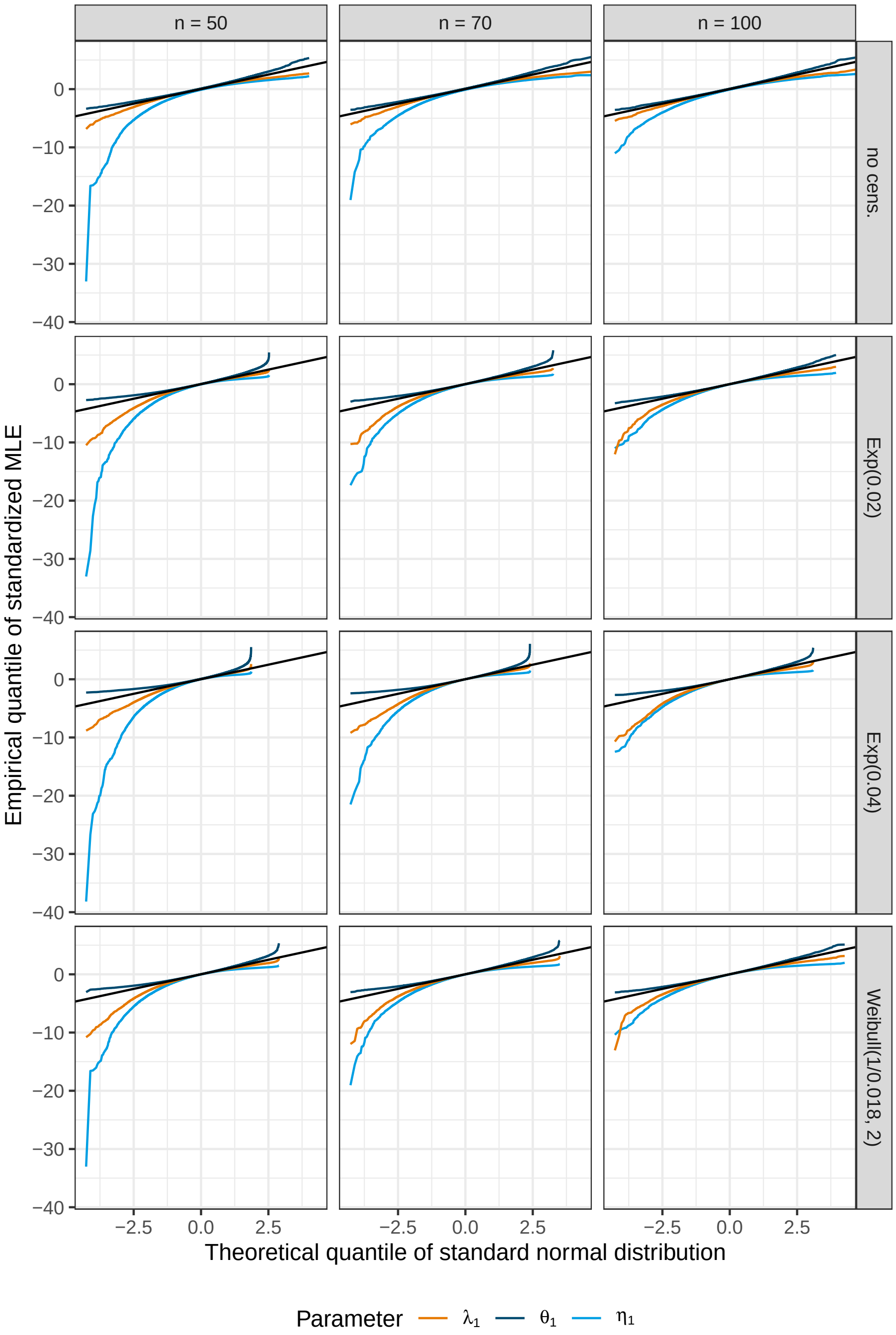


Figure 20: Quantile-quantile plots of standardized MLEs for $p = 0.2, \lambda_1 = 0.037, \lambda_0 = 0.02, n = 50, 70, 100$, and four censoring distributions. Dotted black line indicates perfect agreement of empirical quantiles with standard normal distribution.

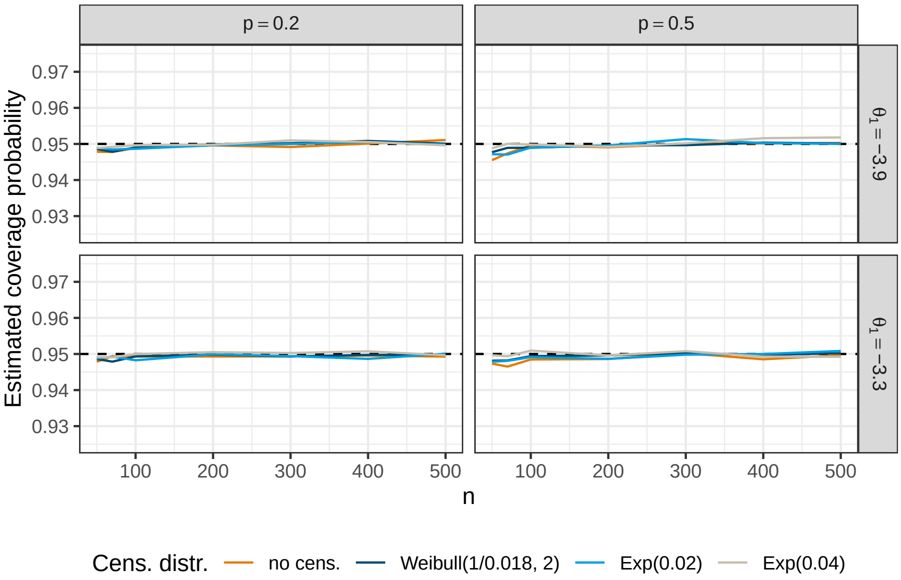


Figure 21: Estimated coverage probability of approximate confidence intervals for θ_0 with respect to sample size n, response probability p and responder survival parameter θ_1. The true value of the non-responder survival parameter is $\theta_0 \approx -3.3$.

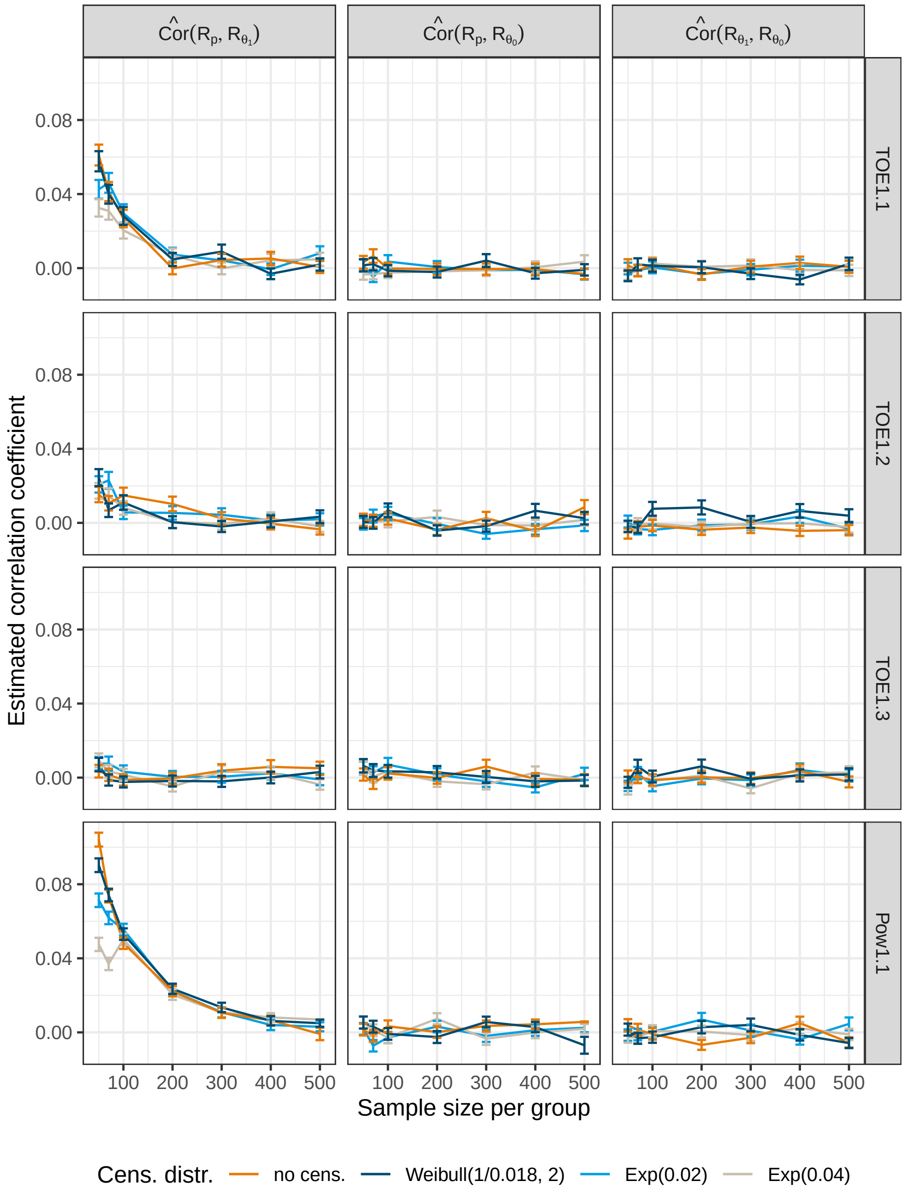


Figure 22: Estimated correlation coefficients of pairs of local test rejections $R_p, R_{\theta_1}, R_{\theta_0}$ in 4 scenarios TOE1.1, TOE1.2, TOE1.3, and Pow1.1 for different censoring distributions. Error bars indicate ± standard error.

A.4 Validation of exact calculations

Figure 23 shows a comparison of exactly calculated and estimated coverage probabilities of approximate confidence intervals for θ_1, as described in Section 2.2.7.

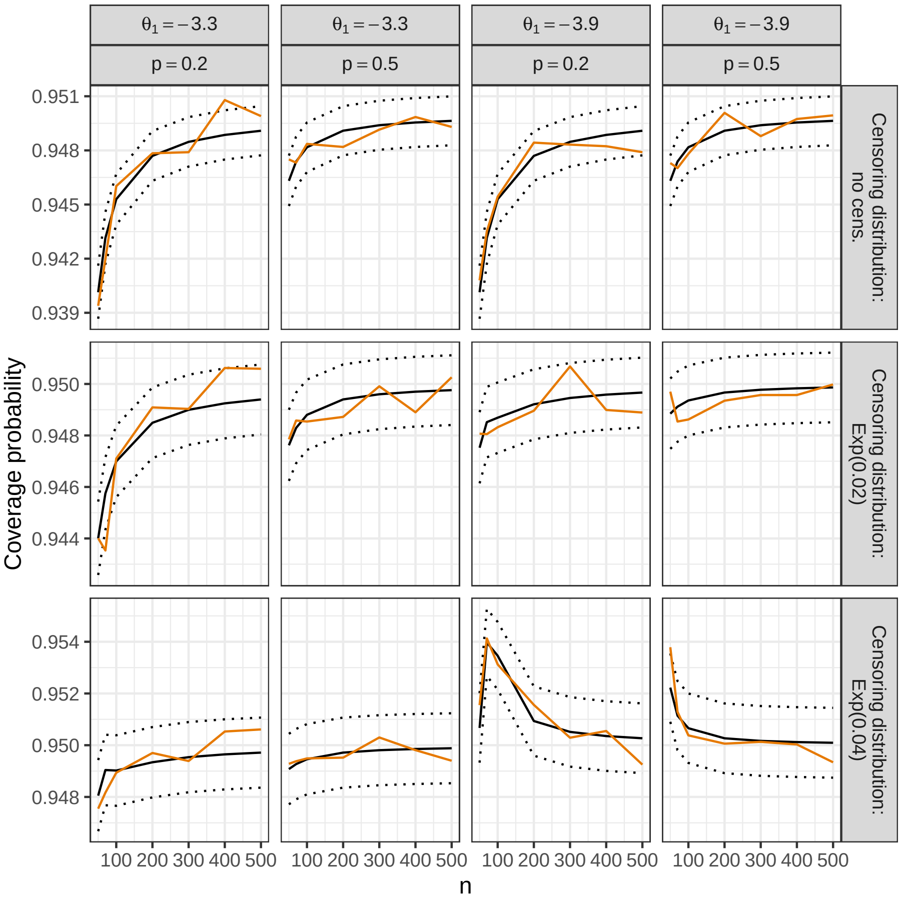


Figure 23: Coverage probability of approximate confidence intervals for θ_1 with respect to sample size n, response probability p, true responder survival parameter θ_1, and three censoring distributions. The true value of the non-responder survival parameter is $\theta_0 \approx -3.3$. The orange lines show estimated coverage probabilities. The solid black lines show exactly calculated coverage probabilities. The dotted black lines show 2.5% and 97.5% quantiles of the distribution of the estimated coverage probability. The estimated coverage probability is mostly within the quantiles which validates the exact calculation.

Figure 24 shows a comparison of exactly calculated and estimated rejection probabilities of the approximate RSES test, as described in Section 2.3.2.

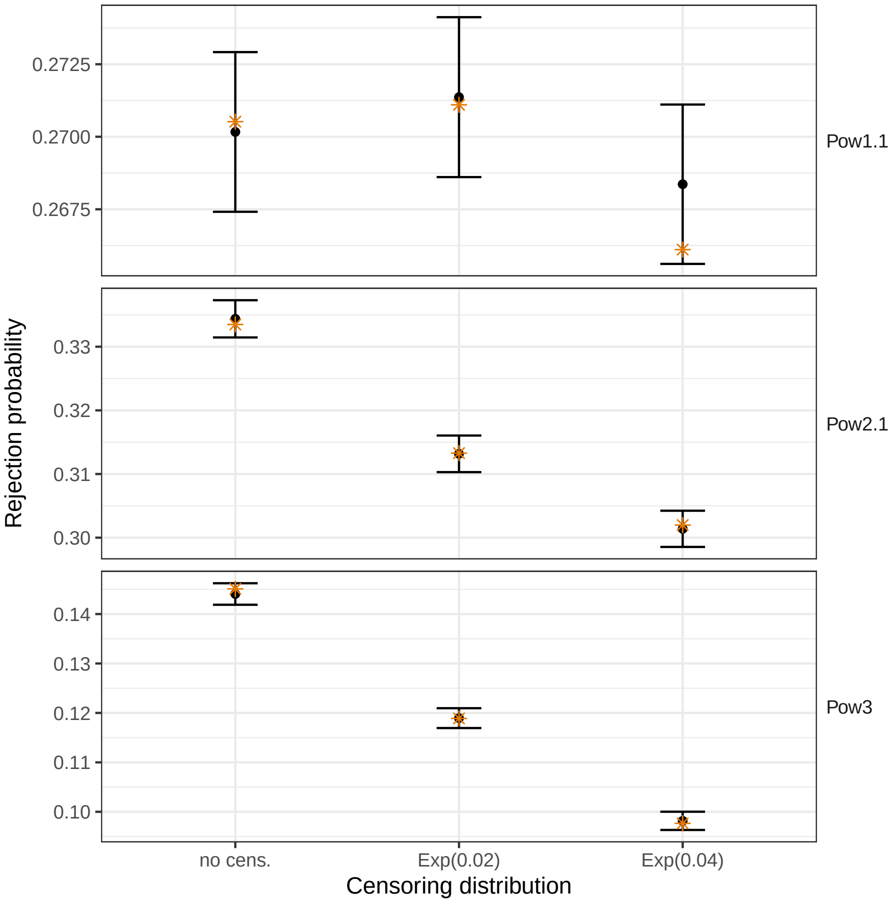


Figure 24: Rejection probability of the approximate RSES test for different scenarios Pow1.1, Pow2.1, Pow3 and three censoring distributions. Orange stars show estimated rejection probabilities by simulation. Black points show exactly calculated rejection probabilities. Black error bars span 2.5% and 97.5% quantiles of the distribution of the estimated rejection probability. The estimated coverage probability is always within the quantiles which validates the exact calculation.

A.5 Expectation of the observed event-free time

Let $T_S \sim \mathrm{Exp}(\lambda)$ be an exponentially distributed survival time and U an arbitrarily distributed censoring time. Let T_S and U be independent. The event-free time $T = \min(T_S, U)$ is observed. It will be shown that

$$\mathrm{E}[T] = \mathrm{E}[T_S] \cdot \mathrm{P}(T_S < U).$$

By the law of total expectation,

$$\begin{aligned}\mathrm{E}[T] &= \mathrm{E}[\mathrm{E}[T \mid U]] \\ &= \mathrm{E}[\mathrm{E}[T_S \cdot 1_{T_S<U} + U \cdot 1_{T_S \geq U} \mid U]].\end{aligned}$$

Since $1_{T_S<U} = 1 - 1_{T_S \geq U}$, it is $T_S \cdot 1_{T_S<U} = T_S - T_S \cdot 1_{T_S \geq U}$. Since T_S and U are independent, the distribution of T_S conditional on U is still $\mathrm{Exp}(\lambda)$. Thus, $\mathrm{E}[T_S \mid U] = \mathrm{E}[T_S]$ and $\mathrm{E}[T_S \cdot 1_{T_S \geq U} \mid U]$ can be calculated by

$$\begin{aligned}&\int\limits_U^\infty t \cdot \lambda \exp(-\lambda t)\mathrm{d}t \\ &= \int\limits_0^\infty (t' + U) \cdot \lambda \exp(-\lambda(t' + U))\mathrm{d}t' \\ &= \exp(-\lambda U) \cdot \left(\int\limits_0^\infty t' \cdot \lambda \exp(-\lambda t')\mathrm{d}t' + \int\limits_0^\infty U \cdot \lambda \exp(-\lambda t')\mathrm{d}t' \right) \\ &= \exp(-\lambda U) \cdot (\mathrm{E}[T_S] + U)\,.\end{aligned}$$

Note that $\exp(-\lambda U) = 1 - F_{T_S}(U) = \mathrm{E}[1_{T_S \geq U} \mid U]$. Thus, $\mathrm{E}[U \cdot 1_{T_S \geq U} \mid U] = U \cdot (1 - F_{T_S}(U))$. Hence, it is

$$\begin{aligned}\mathrm{E}[T] &= \mathrm{E}[\mathrm{E}[T \mid U]] \\ &= \mathrm{E}[\mathrm{E}[T_S \cdot 1_{T_S<U} + U \cdot 1_{T_S \geq U} \mid U]] \\ &= \mathrm{E}[\mathrm{E}[T_S] - (1 - F_{T_S}(U)) \cdot (\mathrm{E}[T_S] + U) + U \cdot (1 - F_{T_S}(U))] \\ &= \mathrm{E}[T_S] \cdot \mathrm{E}[(F_{T_S}(U))] \\ &= \mathrm{E}[T_S] \cdot \mathrm{P}(T_S < U).\end{aligned}$$

Appendix B

R code

```
## Required Packages

require(tidyverse)
library(rootSolve)
library(extraDistr)

## Functions ####
## > RSES model ####

# Calculate survival function
Compute.Surr.Surv <- function(
    t,
    p,
    lambda.1,
    lambda.0
    ){
  # Input:
  #    t: time
  #    p: response probability
  #    lambda.1: hazard of responders
  #    lambda.1: hazard of non-responders
  # Output:
  #    value of survival function

  if (
    !all(
      c(
        length(t),
        length(p),
        length(lambda.0),
        length(lambda.1)
      ) == rep(length(t), 4)
    )
  ) {
    stop("All input vectors have to be the same length!")
  }

  return(
    sapply(
      1:length(t),
      function(i){
        (1-pi[i])*exp(-lambda.0[i]*t[i]) + pi[i]*exp(-lambda.1[i]*t[i])
      }
    )
  )
}

## > Confidence intervals for parameters ##########

# Calculates two-sided confidence interval for response probability
Calculate.approximate.CI.p <- function(
    p.hat,
    n,
    alpha
    ){
  # Input:
  #   p.hat: MLE for p
  #   n: sample size
  #   alpha: specifies confidence level 1 - alpha
  # Output:
  #   CI.ll: lower limit of confidence interval
  #   CI.ul: upper limit of confidence interval

  sd <- sqrt((p.hat*(1-p.hat))/n)
  CI.ll <- p.hat - qnorm(1-alpha/2)*sd
  CI.ul <- p.hat + qnorm(1-alpha/2)*sd
  return(
    list(
      CI.ll = CI.ll,
      CI.ul = CI.ul
    )
  )
}

# Calculates exact coverage probability of confidence interval for
# response probability
Calculate.cov.prob.CI.p <- function(
    p,
    n,
    alpha
    ){
  # Input:
  #   p: true response probability
  #   n: sample size
  #   alpha: specifies confidence level 1 - alpha
  # Output:
  #   exact coverage probability

  cov.prob <- sapply(
    1:length(p),
    function(i){
      k.vec <- 0:n[i]
      k <- k.vec[
        abs(p[i]-k.vec/n[i]) <=
          (qnorm(1-alpha/2) * sqrt(k.vec/n[i]*(1-k.vec/n[i])/n[i]))
      ]
      sum(dbinom(k, size = n[i], prob = p[i]))
    }
  )

  return(cov.prob)
}

# Calculates two-sided confidence interval for \theta_j
Calculate.approximate.CI.theta <- function(
    theta.hat,
    l,
    alpha
    ){
  # Input:
  #   theta.hat: MLE for \theta_j
  #   l: number of uncensored observations in stratum j
  #   alpha: specifies confidence level 1 - alpha
  # Output:
  #   CI.ll: lower limit of confidence interval
  #   CI.ul: upper limit of confidence interval

  sd <- sqrt(1/l)
  CI.ll <- theta.hat - qnorm(1-alpha/2)*sd
  CI.ul <- theta.hat + qnorm(1-alpha/2)*sd

  return(
    list(
      CI.ll = CI.ll,
      CI.ul = CI.ul
    )
  )
}

# Calculates exact coverage probability of confidence interval for
# \theta_j in the case of no censoring
Calculate.cov.prob.CI.theta.no.cens <- function(
    p,
    n,
    alpha
){
  # Input:
  #   p: true response probability
  #   n: sample size
  #   alpha: specifies confidence level 1 - alpha
  # Output:
  #   exact coverage probability

  cov.prob <- sapply(
    1:length(p),
    function(i){
      k <- 1:n[i]
      cp.cond.k <- pgamma(
        q = exp(qnorm(1-alpha/2)*sqrt(1/k)),
        shape = k,
        rate = k
      ) -
        pgamma(
          q = exp(-qnorm(1-alpha/2)*sqrt(1/k)),
          shape = k,
          rate = k
        )
      sum(
        dbinom(k, size = n[i], prob = p[i])*cp.cond.k)/
        (1-dbinom(0, size = n[i], prob = p[i]))
    }
  )

  return(cov.prob)
}

# Calculates exact coverage probability of confidence interval for
# \theta_j in the case of exponential censoring
Calculate.cov.prob.CI.theta.exp.cens <- function(
    p,
    n,
    alpha,
    lambda,
    lambda.cens
){
  # Input:
  #   p: true response probability
  #   n: sample size
  #   alpha: specifies confidence level 1 - alpha
  #   lambda: true value of \exp(\theta_j)
  #   lambda.cens: parameter of exponential censoring distribution
  # Output:
  #   exact coverage probability
  cov.prob <- sapply(
    1:length(p),
    function(i){
      # probability of not being censored
      p.C <- lambda[i]/(lambda[i]+lambda.cens[i])

      # possible combinations of k and l
      df.kl <- expand.grid(
        k = 1:n[i],
        l = 1:n[i]
      )
      df.kl <- df.kl[df.kl$l <= df.kl$k,]

      # coverage probability conditional on k and l
      cp.cond.kl <- pgamma(
        q = exp(qnorm(1-alpha/2)*sqrt(1/df.kl$l))/p.C,
        shape = df.kl$k,
        rate = df.kl$l
      ) -
        pgamma(
          q = exp(-qnorm(1-alpha/2)*sqrt(1/df.kl$l))/p.C,
          shape = df.kl$k,
          rate = df.kl$l
        )

      # expectation over conditional coverage probabilities
      sum(
        dbinom(df.kl$k, size = n[i], prob = p[i])*
          dbinom(df.kl$l, size = df.kl$k, prob = p.C)*cp.cond.kl
      )/
        (1-dbinom(0, size = n[i], prob = p[i]*p.C))
    }
  )

  return(cov.prob)
}

## > Approximate local test of H_{p, 0} ####

# Calculate approximate test of H_{p, 0}
Calculate.asymptotic.binomial.test <- function(
    k.E,
    k.C,
    n.E,
    n.C
){
  # Input:
  #   k.E: number of responders in experimental group
  #   k.C: number of responders in control group
  #   n.E: number of observations in experimental group
  #   n.C: number of observations in control group
  # Output:
  #   T.stat: test statistic
  #   p.value: p-value

  # Compute proportions and test statistic
  p.E <- k.E/n.E
  p.C <- k.C/n.C
  p <- (n.E*p.E + n.C*p.C)/(n.E + n.C)
  T.p <- ifelse(
    p %in% c(0,1),
    0,
    abs(p.E - p.C) / sqrt(p*(1-p)*(1/n.E + 1/n.C))
  )

  return(
    list(
      T.stat = T.p,
      p.value = 2*(1-pnorm(abs(T.p)))
    )
  )
}

# Calculate confidence interval of p_E - p_C
Calculate.asymptotic.binomial.CI <- function(
    k.E,
    k.C,
    n.E,
    n.C,
    alpha
    ){
  # Input:
  #   k.E: number of responders in experimental group
  #   k.C: number of responders in control group
  #   n.E: number of observations in experimental group
  #   n.C: number of observations in control group
  #   alpha: specifies confidence level 1 - alpha
  # Output:
  #   CI.ll: lower limit of confidence interval
  #   CI.ul: upper limit of confidence interval

  # Compute proportions and standard deviation
  p.E <- k.E/n.E
  p.C <- k.C/n.C
  diff <- p.E - p.C
  sd <- sqrt(p.E*(1-p.E)/n.E + p.C*(1-p.C)/n.C)

  return(
    list(
      CI.ll = diff-qnorm(1-alpha/2)*sd,
      CI.ul = diff+qnorm(1-alpha/2)*sd
    )
  )
}

# Calculate approximate sample size for test of H_{p, 0}
Calculate.asymptotic.binomial.sample.size <- function(
    p.E,
    p.C,
    r,
    alpha,
    power
    ){
  # Input:
  #   p.E: assumed response rate in experimental group
  #   p.C: assumed response rate in control group
  #   r: ratio of sample sizes n.E/n.C
  #   alpha: level two-sided test
  #   power: desired power
  # Output:
  #   n.C: required sample size in control group
  #   n.E: required sample size in experimental group

  n.C.ex <- (
    (
      qnorm(1-power)*
        sqrt(1/r*p.E*(1-p.E) + p.C*(1-p.C))
      - qnorm(1-alpha/2)*
        sqrt((r*p.E+p.C)/r*(1-(r*p.E+p.C)/(r+1)))
    )/
      (p.E-p.C))^2

  n.C <- ceiling(n.C.ex)
  n.E <- ceiling(r*n.C)

  list(
    n.C = n.C,
    n.E = n.E
  )
}

# Calculate approximate power for test of H_{p, 0}
Calculate.asymptotic.binomial.power <- function(
    n.E,
    n.C,
    p.E,
    p.C,
    alpha
    ){
```

```
  # Input:
  #   n.E: sample size in experimental group
  #   n.C: sample size in control group
  #   p.E: response rate in experimental group
  #   p.C: response rate in control group
  #   alpha: significance level
  # Output:
  #   approximate power

  # Total response rate
  p <- (n.E*p.E + n.C*p.C)/(n.E+n.C)
  # Joint standard deviation
  sigma.p <- sqrt(p.E*(1-p.E)/n.E + p.C*(1-p.C)/n.C)
  # Power
  power <- 1 - pnorm(
    (qnorm(1-alpha/2) * sqrt(
      p*(1-p)*(1/n.E + 1/n.C)
    ) -
      abs(p.E - p.C)) /
      sigma.p
  ) +
    pnorm(
      (-qnorm(1-alpha/2) * sqrt(
        p*(1-p)*(1/n.E + 1/n.C)
      ) -
        abs(p.E - p.C)) /
        sigma.p
    )
  return(power)
}

## > Approximate local test of H_{\theta_j, 0} ####

# Calculate approximate test of H_{\theta_j, 0}
Calculate.asymptotic.RSES.theta.test <- function(
    theta.E,
    theta.C,
    l.E,
    l.C,
    n.E,
    n.C
    ){
  # Input:
  #   theta.E: MLE of \theta_{j, E}
  #   theta.C: MLE of \theta_{j, C}
  #   l.E: number of uncensored observations in stratum j in experimental group
  #   l.C: number of uncensored observations in stratum j in control group
  #   n.E: number of observations in experimental group
  #   n.C: number of observations in control group
  # Output:
```

```
  #   T.stat: test statistic
  #   p.value: p-value

  T.theta <- ifelse(
    l.E == 0 | l.C == 0,
    0,
    abs(theta.E - theta.C)/sqrt((n.E+n.C)/(l.E+l.C)*(1/n.E + 1/n.C))
  )

  list(
    T.stat = T.theta,
    p.value = 2*(1-pnorm(T.theta))
  )
}

# Calculate approximate confidence interval for \theta_{j, E} - \theta_{j, C}
Calculate.asymptotic.RSES.theta.CI <- function(
    theta.E,
    theta.C,
    l.E,
    l.C,
    alpha
    ){
  # Input:
  #   theta.E: MLE of \theta_{j, E}
  #   theta.C: MLE of \theta_{j, C}
  #   l.E: number of uncensored observations in stratum j in experimental group
  #   l.C: number of uncensored observations in stratum j in control group
  #   n.E: number of observations in experimental group
  #   n.C: number of observations in control group
  #   alpha: specifies confidence level 1 - alpha
  # Output:
  #   CI.ll: lower limit of confidence interval
  #   CI.ul: upper limit of confidence interval

  diff <- theta.E - theta.C
  sd <- ifelse(
    l.E == 0 | l.C == 0,
    Inf,
    sqrt(1/l.E+1/l.C)
  )

  list(
    CI.ll = diff-qnorm(1-alpha/2)*sd,
    CI.ul = diff+qnorm(1-alpha/2)*sd
  )
}

# Calculate approximate sample size for test of H_{\theta_j, 0}
Calculate.asymptotic.RSES.theta.sample.size <- function(
    theta.E,
    theta.C,
    p.E,
    p.C,
    q.E,
    q.C,
    r,
    alpha,
    power
    ){
  # Input:
  #   theta.E: Assumed value for \theta_{j, E}
  #   theta.C: Assumed value for \theta_{j, C}
  #   p.E: assumed response rate in experimental group
  #   p.C: assumed response rate in control group
  #   q.E: assumed prop. of uncensored observations in stratum j in exp. group
  #   q.C: assumed prop. of uncensored observations in stratum j in contr. group
  #   r: sample size ratio
  #   alpha: significance level
  #   power: desired power
  # Output:
  #   n.E: required sample size in experimental group
  #   n.C: required sample size in control group

  n.C.ex <- (
    (
      qnorm(1-power)*sqrt(
        1/(r*p.E*q.E) + 1/(p.C*q.C)
      ) -
        qnorm(1-alpha/2)*
        sqrt(
          (1+r)^2/r /(r*p.E*q.E + p.C*q.C)
        )
    ) /
      (theta.E-theta.C)
  )^2
  n.C <- ceiling(n.C.ex)
  n.E <- ceiling(n.C*r)

  list(
    n.C = n.C,
    n.E = n.E
  )
}

# Calculate approximate power for test of H_{\theta_j, 0}
Calculate.asymptotic.RSES.theta.power <- function(
    n.E,
    n.C,
    theta.E,
    theta.C,
    p.E,
    p.C,
    q.E,
    q.C,
    alpha
  ){
  # Input:
  #   n.E: sample size in experimental group
  #   n.C: sample size in control group
  #   theta.E: Assumed value for \theta_{j, E}
  #   theta.C: Assumed value for \theta_{j, C}
  #   p.E: assumed response rate in experimental group
  #   p.C: assumed response rate in control group
  #   q.E: assumed prop. of uncensored observations in stratum j in exp. group
  #   q.C: assumed prop. of uncensored observations in stratum j in contr. group
  #   alpha: significance level
  # Output:
  #   approximate power

  power <- 1 - pnorm(
    (
      qnorm(1-alpha/2)*
        sqrt((n.E+n.C)/(n.E*p.E*q.E+n.C*p.C*q.C)*(1/n.E + 1/n.C))
      - abs(theta.C-theta.E)
    )/
      sqrt(1/(n.E*p.E*q.E)+1/(n.C*p.C*q.C))
  ) +
    pnorm(
      (
        -qnorm(1-alpha/2)*
          sqrt((n.E+n.C)/(n.E*p.E*q.E+n.C*p.C*q.C)*(1/n.E + 1/n.C))
        - abs(theta.C-theta.E)
      )/
        sqrt(1/(n.E*p.E*q.E)+1/(n.C*p.C*q.C))
    )

  return(power)
}

## > Approximate RSES test ####

# Calculate p-values of local tests
# The decision of the approximate RSES test is then made by
# comparing the minimum of the local p-values to the significance level
Calculate.all.local.p.values <- function(
    k.E,
    k.C,
    n.E,
    n.C,
  l.1E,
  l.1C,
  l.0E,
  l.0C,
  theta.1.diff,
  theta.0.diff
  ){
# Input:
#   k.E: number of responders in experimental group
#   k.C: number of responders in control group
#   n.E: number of observations in experimental group
#   n.C: number of observations in control group
#   l.1E: number of uncensored responders in experimental group
#   l.1C: number of uncensored responders in control group
#   l.0E: number of uncensored non-responders in experimental group
#   l.0C: number of uncensored non-responders in control group
#   theta.1.diff: difference of responder MLEs
#   theta.0.diff: difference of non-responder MLEs
# Output:
#   p.p: p-value of local test of H_{p, 0}
#   p.theta.1: p-value of local test of H_{\theta_1, 0}
#   p.theta.0: p-value of local test of H_{\theta_0, 0}

# Compute p-values
p.p <- Calculate.asymptotic.binomial.test(
  k.E = k.E,
  k.C = k.C,
  n.E = n.E,
  n.C = n.C
)$p.value

p.t1 <- Calculate.asymptotic.RSES.theta.test(
  theta.E = theta.1.diff,
  theta.C = 0,
  l.E = l.1E,
  l.C = l.1C,
  n.E = n.E,
  n.C = n.C
)$p.value

p.t0 <- Calculate.asymptotic.RSES.theta.test(
  theta.E = theta.0.diff,
  theta.C = 0,
  l.E = l.0E,
  l.C = l.0C,
  n.E = n.E,
  n.C = n.C
)$p.value

return(
    list(
      p.p = p.p,
      p.theta.1 = p.t1,
      p.theta.0 = p.t0
    )
  )
}

# Calculate approximate sample size for approximate RSES test
Calculate.asymptotic.RSES.sample.size <- function(
    theta.1.E,
    theta.1.C,
    theta.0.E,
    theta.0.C,
    p.E,
    p.C,
    q.1.E,
    q.1.C,
    q.0.E,
    q.0.C,
    r,
    alpha,
    power
    ){
  # Input:
  #   theta.1.E: Assumed value for \theta_{1, E}
  #   theta.1.C: Assumed value for \theta_{1, C}
  #   theta.0.E: Assumed value for \theta_{0, E}
  #   theta.0.C: Assumed value for \theta_{0, C}
  #   p.E: assumed response rate in experimental group
  #   p.C: assumed response rate in control group
  #   q.1.E: assumed proportion of uncensored responders in exp. group
  #   q.1.C: assumed proportion of uncensored responders in contr. group
  #   q.0.E: assumed proportion of uncensored non-responders in exp. group
  #   q.0.C: assumed proportion of uncensored non-responders in contr. group
  #   r: sample size ratio
  #   alpha: significance level
  #   power: desired power
  # Output:
  #   n.E: required sample size in experimental group
  #   n.C: required sample size in control group

  # logical vector to describe where are differences between parameters
  diff.vec <- c(
    p = p.E == p.C,
    theta.1 = theta.1.E == theta.1.C,
    theta.0 = theta.0.E == theta.0.C
  )
  if (all(diff.vec)) {
    stop("Not all pairs of parameters (E, C) can be equal.")
```

```
}

# local level
alpha.loc <- 1-(1-alpha)^(1/3)

# Find limits for sample size
beta.t <- ((1-power)/(1-alpha.loc)^sum(diff.vec))^(1/sum(!diff.vec))

n.p.t <- Calculate.asymptotic.binomial.sample.size(
  p.E = p.E,
  p.C = p.C,
  r = r,
  power = 1-beta.t,
  alpha = alpha.loc
)$n.C

n.theta.1.t <- Calculate.asymptotic.RSES.theta.sample.size(
  theta.E = theta.1.E,
  theta.C = theta.1.C,
  p.E = p.E,
  p.C = p.C,
  q.E = q.1.E,
  q.C = q.1.C,
  r = r,
  alpha = alpha.loc,
  power = 1-beta.t
)$n.C

n.theta.0.t <- Calculate.asymptotic.RSES.theta.sample.size(
  theta.E = theta.0.E,
  theta.C = theta.0.C,
  p.E = 1-p.E,
  p.C = 1-p.C,
  q.E = q.0.E,
  q.C = q.0.C,
  r = r,
  alpha = alpha.loc,
  power = 1-beta.t
)$n.C

# Define acceptance probability equation to solve
equation <- function(n.C){
  (1 - Calculate.asymptotic.binomial.power(
    n.E = r*n.C,
    n.C = n.C,
    p.E = p.E,
    p.C = p.C,
    alpha = alpha.loc
  )) *
    (1 - Calculate.asymptotic.RSES.theta.power(
          n.E = r*n.C,
          n.C = n.C,
          theta.E = theta.1.E,
          theta.C = theta.1.C,
          p.E = p.E,
          p.C = p.C,
          q.E = q.1.E,
          q.C = q.1.C,
          alpha = alpha.loc
        )) *
        (1- Calculate.asymptotic.RSES.theta.power(
          n.E = r*n.C,
          n.C = n.C,
          theta.E = theta.0.E,
          theta.C = theta.0.C,
          p.E = 1-p.E,
          p.C = 1-p.C,
          q.E = q.0.E,
          q.C = q.0.C,
          alpha = alpha.loc
        )) -
        (1 - power)
    }

    # Solve equation for n.C
    uniroot(
      f = equation,
      interval = c(
        min(c(n.p.t, n.theta.1.t, n.theta.0.t)[!diff.vec])-5,
        max(c(n.p.t, n.theta.1.t, n.theta.0.t)[!diff.vec])+5
      )
    )$root ->
      n.C.ex

    return(
      list(
        n.C = ceiling(n.C.ex),
        n.E = ceiling(r*n.C.ex)
      )
    )
  }

  # Calculate exact rejection probability of approximate RSES test
  # in the case of no censoring
  Calculate.exact.RSES.reject.prob.no.cens <- function(
      n.E,
      n.C,
      theta.1.E,
      theta.1.C,
      theta.0.E,
  theta.0.C,
  p.E,
  p.C,
  alpha
  ){
# Input:
#   n.E: sample size in experimental group
#   n.C: sample size in control group
#   theta.1.E: Assumed value for \theta_{1, E}
#   theta.1.C: Assumed value for \theta_{1, C}
#   theta.0.E: Assumed value for \theta_{0, E}
#   theta.0.C: Assumed value for \theta_{0, C}
#   p.E: assumed response rate in experimental group
#   p.C: assumed response rate in control group
#   alpha: significance level
# Output:
#   exact rejection probability

# local level
alpha.loc <- 1-(1-alpha)^(1/3)

# hazards in treatment groups and response strata
lam.E.1 <- exp(theta.1.E)
lam.C.1 <- exp(theta.1.C)
lam.E.0 <- exp(theta.0.E)
lam.C.0 <- exp(theta.0.C)

result <- 0

for (k.E in 0:n.E) {
  for (k.C in 0:n.C) {

    if (
      abs(
        Calculate.asymptotic.binomial.test(k.E, k.C, n.E, n.C)$T.stat
      ) > qnorm(1-alpha.loc/2)
    ) {
      add <- 0
    } else {
      # estimated rates
      p.E.e <- k.E/n.E
      p.C.e <- k.C/n.C

      # pooled rate
      p.e <- (n.E*p.E.e + n.C*p.C.e)/(n.E + n.C)

      if (k.E == 0 | k.C == 0) {
        u.1 <- 1
      } else {
        c.1 <- exp(qnorm(1-alpha.loc/2) * sqrt(1/p.e*(1/n.E + 1/n.C)))
```

```
          u.1 <- pbetapr(k.C*lam.C.1/(k.E*lam.E.1)*c.1, k.C, k.E) -
            pbetapr(k.C*lam.C.1/(k.E*lam.E.1) * 1/c.1, k.C, k.E)
        }

        if (k.E == n.E | k.C == n.C) {
          u.0 <- 1
        } else {
          c.0 <- exp(qnorm(1-alpha.loc/2) * sqrt(1/(1-p.e)*(1/n.E + 1/n.C)))
          u.0 <- pbetapr(
            (n.C-k.C)*lam.C.0/((n.E-k.E)*lam.E.0)*c.0,
            n.C-k.C,
            n.E-k.E
          ) - pbetapr(
            (n.C-k.C)*lam.C.0/((n.E-k.E)*lam.E.0) * 1/c.0,
            n.C-k.C,
            n.E-k.E
          )
        }

        add <- dbinom(k.E, n.E, p.E)*dbinom(k.C, n.C, p.C)*u.1*u.0
      }
      result <- result + add
    }
  }

  return(1-result)
}

# Calculate exact rejection probability of approximate RSES test
# in the case of exponential censoring
Calculate.exact.RSES.reject.prob.exp.cens <- function(
    n.E, n.C,
    theta.1.E,
    theta.1.C,
    theta.0.E,
    theta.0.C,
    lambda.cens,
    p.E,
    p.C,
    alpha
    ){
  # Input:
  #   n.E: sample size in experimental group
  #   n.C: sample size in control group
  #   theta.1.E: Assumed value for \theta_{1, E}
  #   theta.1.C: Assumed value for \theta_{1, C}
  #   theta.0.E: Assumed value for \theta_{0, E}
  #   theta.0.C: Assumed value for \theta_{0, C}
  #   p.E: assumed response rate in experimental group
  #   p.C: assumed response rate in control group
#   lambda.cens: parameter of exponential censoring distribution
#   r: sample size ratio
#   alpha: significance level
# Output:
#   exact rejection probability

# local level
alpha.loc <- 1-(1-alpha)^(1/3)

# hazards in treatment groups and response strata
lam.E.1 <- exp(theta.1.E)
lam.C.1 <- exp(theta.1.C)
lam.E.0 <- exp(theta.0.E)
lam.C.0 <- exp(theta.0.C)

# All combinations of l.i and k.i
df.l <- expand.grid(
  l.C = 1:n.C,
  l.E = 1:n.E,
  k.C = 1:n.C,
  k.E = 1:n.E
)

# All combinations of k.E, k.C, l.E, l.C where something has to be
# computed (0 < l.i < k.i)
ind <- df.l$l.E <= df.l$k.E & df.l$l.C <= df.l$k.C & df.l$l.E*df.l$l.C > 0

# Vector of values of k.i and l.i
k.E <- df.l$k.E[ind]
l.E <- df.l$l.E[ind]
k.C <- df.l$k.C[ind]
l.C <- df.l$l.C[ind]

# compute quantile for beta prime distribution function for every combination
c.1 <- exp(qnorm(1-alpha.loc/2) * sqrt((n.E+n.C)/(l.E+l.C)*(1/n.E + 1/n.C)))

# compute conditional acceptance probability of H_theta.1 and H_theta.0
# by beta prime distribution function for c.1 and 1/c.1 with true parameters
u.1 <- pbetapr(
  q = c.1,
  shape1 = k.C,
  shape2 = k.E,
  scale = l.E*(lam.E.1+lambda.cens)/l.C/(lam.C.1+lambda.cens)
) -
  pbetapr(
    q = 1/c.1,
    shape1 = k.C,
    shape2 = k.E,
    scale = l.E*(lam.E.1+lambda.cens)/l.C/(lam.C.1+lambda.cens)
  )
```

```
# u.0 is actually computed for wrong parameters (should be n.E-k.E and so on).
# Will be matched correctly later.
u.0 <- pbetapr(
  q = c.1,
  shape1 = k.C,
  shape2 = k.E,
  scale = l.E*(lam.E.0+lambda.cens)/l.C/(lam.C.0+lambda.cens)
) -
  pbetapr(
    q = 1/c.1,
    shape1 = k.C,
    shape2 = k.E,
    scale = l.E*(lam.E.0+lambda.cens)/l.C/(lam.C.0+lambda.cens)
  )

# all combinations of response combinations
k.comb <- expand.grid(
  k.C = 0:n.C,
  k.E = 0:n.E
)

# number of combinations of l.E and l.C for
# respective combination of k.E and k.C
n.l.comb <- k.comb$k.E*k.comb$k.C

# cumulative sum of these numbers for indexing the vectors correctly
n.l.comb.cum <- cumsum(c(0, n.l.comb))

# initiate acceptance probability vector by 0
accept.theta.1 <- rep(0, nrow(k.comb))

# update acceptance probability vector where there is a positive number
# of combinations of l.E and l.C
sapply(
  (1:nrow(k.comb))[n.l.comb > 0],
  function(i){
    # determine index range of combinations l.E and l.C belonging
    # to the specific combination k.E, k.C
    ind <- (n.l.comb.cum[i]+1):(n.l.comb.cum[i+1])

    # calculate sum of acceptance probabilities over this index range
    sum(
      u.1[ind]*
        dbinom(
          x = l.E[ind],
          size = k.comb$k.E[i],
          prob = lam.E.1/(lam.E.1+lambda.cens)
        )*
        dbinom(
          x = l.C[ind],
          size = k.comb$k.C[i],
          prob = lam.C.1/(lam.C.1+lambda.cens)
        )
    )
  }
) ->
  accept.theta.1[n.l.comb > 0]

# calculate probability of no events in one of the groups
# for every combination
theta.1.no.events <- dbinom(
  x = 0,
  size = k.comb$k.E,
  prob = lam.E.1/(lam.E.1+lambda.cens)
) +
  dbinom(
    x = 0,
    size = k.comb$k.C,
    prob = lam.C.1/(lam.C.1+lambda.cens)
  ) -
  dbinom(
    x = 0,
    size = k.comb$k.E,
    prob = lam.E.1/(lam.E.1+lambda.cens)
  ) *
  dbinom(
    x = 0,
    size = k.comb$k.C,
    prob = lam.C.1/(lam.C.1+lambda.cens)
  )

# repeat procedure for theta.0
accept.theta.0 <- rep(0, nrow(k.comb))
ind.n.l.comb.gr.0 <- (1:nrow(k.comb))[n.l.comb > 0]
sapply(
  ind.n.l.comb.gr.0,
  function(i){
    ind <- (n.l.comb.cum[i]+1):(n.l.comb.cum[i+1])
    sum(
      u.0[ind]*
        dbinom(
          x = l.E[ind],
          size = k.comb$k.E[i],
          prob = lam.E.0/(lam.E.0+lambda.cens)
        )*
        dbinom(
          x = l.C[ind],
          size = k.comb$k.C[i],
          prob = lam.C.0/(lam.C.0+lambda.cens)
          )
      )
    }
  ) ->
    accept.theta.0[nrow(k.comb)+1-ind.n.1.comb.gr.0]

  theta.0.no.events <- dbinom(
    x = 0,
    size = n.E-k.comb$k.E,
    prob = lam.E.0/(lam.E.0+lambda.cens)
  ) +
    dbinom(
      x = 0,
      size = n.C-k.comb$k.C,
      prob = lam.C.0/(lam.C.0+lambda.cens)
    ) -
    dbinom(
      x = 0,
      size = n.E-k.comb$k.E,
      prob = lam.E.0/(lam.E.0+lambda.cens)
    ) *
    dbinom(
      x = 0,
      size = n.C-k.comb$k.C,
      prob = lam.C.0/(lam.C.0+lambda.cens)
    )

  # Calculate sum over all combinations of k.E and k.C with respective
  # binomial probabilities
  sum(
    dbinom(k.comb$k.E, n.E, p.E)*dbinom(k.comb$k.C, n.C, p.C)*
      (Calculate.asymptotic.binomial.test(
        k.E = k.comb$k.E,
        k.C = k.comb$k.C,
        n.E = n.E,
        n.C = n.C
      )$p.value >= alpha.loc) *
      (accept.theta.1+theta.1.no.events)*
      (accept.theta.0+theta.0.no.events)
  ) ->
    result

  return(1-result)
}

Calculate.exact.RSES.sample.size <- function(
    theta.1.E,
    theta.1.C,
    theta.0.E,
    theta.0.C,
  p.E,
  p.C,
  lambda.cens,
  r,
  alpha,
  power
  ){
# Input:
#   theta.1.E: Assumed value for \theta_{1, E}
#   theta.1.C: Assumed value for \theta_{1, C}
#   theta.0.E: Assumed value for \theta_{0, E}
#   theta.0.C: Assumed value for \theta_{0, C}
#   p.E: assumed response rate in experimental group
#   p.C: assumed response rate in control group
#   lambda.cens: parameter of exponential censoring distribution
#   r: sample size ratio
#   alpha: significance level
#   power: desired power
# Output:
#   n.E: sample size in experimental group
#   n.C: sample size in control group
#   exact.power: exact power

# Probabilities of not being censored in treatment groups and
# response strata
q.1.E <- exp(theta.1.E)/(exp(theta.1.E)+lambda.cens)
q.1.C <- exp(theta.1.C)/(exp(theta.1.C)+lambda.cens)
q.0.E <- exp(theta.0.E)/(exp(theta.0.E)+lambda.cens)
q.0.C <- exp(theta.0.C)/(exp(theta.0.C)+lambda.cens)

# Starting value for iterative procedure
n.C <- Calculate.asymptotic.RSES.sample.size(
  theta.1.E = theta.1.E,
  theta.1.C = theta.1.C,
  theta.0.E = theta.0.E,
  theta.0.C = theta.0.C,
  p.E = p.E,
  p.C = p.C,
  q.1.E = q.1.E,
  q.1.C = q.1.C,
  q.0.E = q.0.E,
  q.0.C = q.0.C,
  r = r,
  alpha = alpha,
  power = power
)$n.C

# Exact power with starting value
power.ex <- Calculate.exact.RSES.reject.prob.exp.cens(
  theta.1.E = theta.1.E,
  theta.1.C = theta.1.C,
  theta.0.E = theta.0.E,
  theta.0.C = theta.0.C,
  lambda.cens = lambda.cens,
  p.E = p.E,
  p.C = p.C,
  alpha = alpha,
  n.C = n.C,
  n.E = ceiling(r*n.C)
)

# Increase sample size if exact power is lower than desired
if(power.ex < power){
  while(power.ex < power){
    n.C <- n.C+1

    power.ex <- Calculate.exact.RSES.reject.prob.exp.cens(
      theta.1.E = theta.1.E,
      theta.1.C = theta.1.C,
      theta.0.E = theta.0.E,
      theta.0.C = theta.0.C,
      lambda.cens = lambda.cens,
      p.E = p.E,
      p.C = p.C,
      alpha = alpha,
      n.C = n.C,
      n.E = ceiling(r*n.C)
    )
  }
} else {
  while(power.ex >= power){
    n.C <- n.C-1
    power.stored <- power.ex
    power.ex <- Calculate.exact.RSES.reject.prob.exp.cens(
      theta.1.E = theta.1.E,
      theta.1.C = theta.1.C,
      theta.0.E = theta.0.E,
      theta.0.C = theta.0.C,
      lambda.cens = lambda.cens,
      p.E = p.E,
      p.C = p.C,
      alpha = alpha,
      n.C = n.C,
      n.E = ceiling(r*n.C)
    )
  }
  n.C <- n.C+1
  power.ex <- power.stored
}
return(
    list(
      n.C = n.C,
      n.E = ceiling(r*n.C),
      exact.power = power.ex
    )
  )
}

# test function Calculate.exact.RSES.sample.size
if(F){
  Calculate.exact.RSES.sample.size (
    theta.1.E = -3.05,
    theta.1.C = -1.95,
    theta.0.E = -2.65,
    theta.0.C = -1.95,
    p.E = 0.8,
    p.C = 0.13,
    lambda.cens = 0.04,
    r = 1,
    alpha = 0.05,
    power = 0.8
  )
}

## > Logrank test and stratified logrank test ####

# Calculate logrank test
Calculate.logrank.test <- function(
    df
    ){
  # Input:
  #   df: a data frame with variables t (time), event (event indicator) and
  #       group ("E" or "C")
  # Output:
  #   O: number of observed events in E
  #   E: number of expected events in E
  #   V: approximate variance
  #   T.stat: test statistic

  # ordered vector of event times
  tau <- sort(unique(df$t[df$event]))

  if(length(tau) == 0){
    O <- NA
    E <- NA
    V <- NA
    T.LR <- NA
  } else {
    # number of events in E at each event time
    d1 <- sapply(
      tau,
      function(t) sum(df$group == "E" & df$t == t & df$event)
    )

    # number of total events at each event time
    d <- sapply(
      tau,
      function(t) sum(df$t == t & df$event)
    )

    # number at risk in E at each event time
    Y1 <- sapply(
      tau,
      function(t) sum(df$t >= t & df$group == "E")
    )

    # number at risk at each event time
    Y <- sapply(
      tau,
      function(t) sum(df$t >= t)
    )

    O <- d1
    E <- d * Y1/Y
    V <- ifelse(
      Y^2 * (Y - 1) == 0,
      0,
      ((Y - Y1) * Y1 * (Y - d)*d)/
        (Y^2 * (Y - 1))
    )
    if(sum(V) == 0) T.LR <- NA else T.LR <- sum(O - E)/sqrt(sum(V))
  }

  return(
    list(
      O = sum(O),
      E = sum(E),
      V = sum(V),
      T.stat = T.LR
    )
  )
}

# Calculate stratified logrank test
Calculate.strat.logrank.test <- function(
    df
    ){
  # Input:
  #   df: a data frame with variables t (time), event (event indicator),
  #       group ("E" or "C"), and resp (response status 0/1)
  # Output:
  #   T.stat: test statistic

  O.vec <- c()
  E.vec <- c()
  V.vec <- c()
  # Calculate number of observed and expected events and approximate variances
  # in each stratum
  for (stratum. in c(0, 1)) {
    Calculate.logrank.test(df[df$resp == stratum., ]) ->
      logrank.h

    if(is.na(logrank.h$T.LR)) next else {
      O.vec <- c(O.vec, logrank.h$O)
      E.vec <- c(E.vec, logrank.h$E)
      V.vec <- c(V.vec, logrank.h$V)
    }
  }

  Num <- sum(O.vec - E.vec)
  Denom <- sqrt(sum(V.vec))
  T.stat <- Num/Denom
  if(is.nan(T.stat)) T.stat <- NA

  return(
    list(
      T.stat = T.stat
    )
  )
}

## > Exact RSES test ####

# Calculate exact p-value for the exact test of H_{p, 0}
Exact.p.p <- function(
    k.E,
    k.C,
    n.E,
    n.C,
    p.vec
    ){
  # Input:
  #   k.E: number of responders in experimental group
  #   k.C: number of responders in control group
  #   n.E: sample size in experimental group
  #   n.C: sample size in control group
  #   p.vec: vector of values of true response probability
  # Output:
  #   vector of p-values corresponding to p.vec
  # Create data frame of all combinations (x.E, x.C) that are as or more extreme
  # than the observed combination
  expand.grid(
    x.E = 0:n.E,
    x.C = 0:n.C
  ) %>%
    mutate(
      T.p = Calculate.asymptotic.binomial.test(x.E, x.C, n.E, n.C)$T.stat
    ) %>%
    filter(
      T.p >= Calculate.asymptotic.binomial.test(k.E, k.C, n.E, n.C)$T.stat
    ) ->
    df.rej

  # For every true response probability in p.vec, calculate probability
  # of the data frame
  prob.vec <- NULL
  for (p in p.vec) {
    prob <- sum(dbinom(df.rej$x.E, n.E, p)*dbinom(df.rej$x.C, n.C, p))
    prob.vec <- c(prob.vec, prob)
  }

  return(prob.vec)
}

# Calculate critical value of test of H_{p, 0}
Exact.p.crit <- function(
    alpha.loc,
    n.E,
    n.C,
    size.acc = 3
    ){
  # Input:
  #   alpha.loc: local level
  #   n.E: sample size in experimental group
  #   n.C: sample size in control group
  #   size.acc: 10^(-size.acc) is the coarseness of the grid used to obtain
  #             the maximum p-value
  # Output:
  #   crit.val.lb: lower bound of critical value
  #   crit.val.mid: can be used as critical value
  #   crit.val.ub: upper bound of critical value
  #   max.size: maximum Type I error rate
  #   sizes: vector of Type I error rates dependent on true response probability

  # Create data frame of all test statistics ordered by test statistic
  expand.grid(
    x.E = 0:n.E,
    x.C = 0:n.C
  ) %>%
  mutate(
    stat = Calculate.asymptotic.binomial.test(x.E, x.C, n.E, n.C)$T.stat
  ) %>%
  arrange(desc(stat)) ->
  df.stat

# Extract stat, x.C and x.E as vector
stat <- df.stat$stat
x.C <- df.stat$x.C
x.E <- df.stat$x.E

# Find starting value for the search of critical value by taking the
# quantile of the normal distribution
start_value <- qnorm(1-alpha.loc/2)

# Find row number of df.stat corresponding to starting value
# <- row of df.stat where stat is maximal with stat <= start_value
# Special case with very small sample sizes can lead to stat > start_value
# for all rows. Then set i <- 1
i <- sum(stat>start_value)

# Define rough grid for p.C and p.E
acc <- 1
p.C <- seq(10^-acc, 1-10^-acc, by = 10^-acc)
p.E <- p.C

# Calculate exact Type I error rate for every pair (p.C, p.E)
sapply(
  1:length(p.C),
  function(j) dbinom(x.C[1:i], n.C, p.C[j])*dbinom(x.E[1:i], n.E, p.E[j])
) ->
  size.vec

# Increase index if maximal Type I error rate is too low
while (max(apply(size.vec, 2, sum)) <= alpha.loc) {
  i <- i+1

  # Compute new Type I error rates
  size.vec <- rbind(
    size.vec,
    dbinom(x.C[i], n.C, p.C)*dbinom(x.E[i], n.E, p.E)
  )
}

# Decrease index if maximal Type I error rate is too high and increase
# grid accuracy
for (acc in 1:size.acc) {

  # Define grid for p.C and p.E
  p.C <- seq(10^-acc, 1-10^-acc, by = 10^-acc)
```

```
    p.E <- p.C

    sapply(
      1:length(p.C),
      function(j) dbinom(x.C[1:i], n.C, p.C[j])*dbinom(x.E[1:i], n.E, p.E[j])
    ) ->
      size.vec

    # Decrease index if maximal Type I error rate is too high
    while (max(apply(size.vec, 2, sum)) > alpha.loc & i >= 1) {
      # Compute new Type I error rates
      size.vec <- size.vec[-i,]
      i <- i-1
    }
  }

  # Decrease index further as long as rows have the same test statistic value
  while (stat[i+1] == stat[i] & i >= 1) {
    size.vec <- size.vec[-i,]
    i <- i-1
  }

  # Critical value can now be chosen between stat[i+1] and stat[i]
  crit.val.mid <- (stat[i+1] + stat[i])/2

  # Return range of critical values and maximal Type I error rate
  return(
    list(
      crit.val.lb = stat[i],
      crit.val.mid = crit.val.mid,
      crit.val.ub = stat[i+1],
      max.size = max(apply(size.vec, 2, sum)),
      sizes = apply(size.vec, 2, sum)
    )
  )
}

# Calculate exact p-value for the exact test of H_{\theta_1, 0} conditional
# on k_E, k_C. For \theta_0, insert corresponding values of non-responders.
Cond.theta.p <- function(
    t.E = NULL,
    t.C = NULL,
    T.stat = NULL,
    k.E = NULL,
    k.C = NULL
    ){
  # Input:
  #   t.E: vector of responder survival times in experimental group
  #   t.E: vector of responder survival times in control group
  #   T.stat: monotone transformation of test statistic
  #   k.E: number of responders in experimental group
  #   k.C: number of responders in control group
  # Output:
  #   p-value

  # If vectors of survival times are given, calculate monotone transformation
  # of test statistic from these vectors
  if (!is.null(t.E) & !is.null(t.C)) {
    if (any(c(t.E, t.C) <= 0)) {
      stop("Non-positive survival times are not possible.")
    }
    k.E <- length(t.E)
    k.C <- length(t.C)
    T.stat <- log(1/k.C*sum(t.C)) - log(1/k.E*sum(t.E))
  }

  # Calculate exact p-value by bea prime distribution
  p.value <- ifelse(
    k.E*k.C == 0,
    1,
    1 - pbetapr(k.C/k.E*exp(abs(T.stat)), k.C, k.E) +
      pbetapr(k.C/k.E*exp(-abs(T.stat)), k.C, k.E)
  )

  return(p.value)
}

# Calculate critical value for the test of H_{\theta_1, 0} conditional on
# k_E, k_C. For \theta_0, insert corresponding values of non-responders.
Cond.theta.crit <- function(
    k.E,
    k.C,
    alpha.loc,
    upper = 100
    ){
  # Input:
  #   alpha.loc: local level
  #   k.E: number of responders in experimental group
  #   k.C: number of responders in control group
  #   upper: upper limit for the critical value needed for uniroot.all()
  # Output:
  #   critical value

  if (k.E*k.C <= 0) {
    stop("k.E and k.C must be greater 0.")
  }

  # Define equation for finding the root
  help.fun <- function(crit){
    1 - pbetapr(k.C/k.E*exp(crit), k.C, k.E) +
      pbetapr(k.C/k.E*exp(-crit), k.C, k.E) -
      alpha.loc
  }

  # Find the root
  rootSolve::uniroot.all(
    f = help.fun,
    interval = c(0, upper)
  ) ->
    crit.val

  return(crit.val)
}

# Calculate exact rejection probability of exact RSES test
Exact.rej.prob <- function(
    n.E,
    n.C,
    p.E,
    p.C,
    lambda.1.E,
    lambda.1.C,
    lambda.0.E,
    lambda.0.C,
    alpha.loc,
    crit.val
    ){
  # Input:
  #   n.E: sample size in experimental group
  #   n.C: sample size in control group
  #   p.E: vector of assumed response rates in experimental group
  #   p.C: vector of assumed response rates in control group
  #   lambda.1.E: Assumed value for \lambda_{1, E}
  #   lambda.1.C: Assumed value for \lambda_{1, C}
  #   lambda.0.E: Assumed value for \lambda_{0, E}
  #   lambda.0.C: Assumed value for \lambda_{0, C}
  #   alpha.loc: local significance level
  #   crit.val: critical value for test of H_{p, 0}
  # Output:
  #   vector of exact rejection probabilities corresponding to p.E and p.C

  prob.vec <- NULL

  for (i in 1:length(p.E)) {
    # Calculate test statistic of test of H_{p, 0}
    expand.grid(
      x.E = 0:n.E,
      x.C = 0:n.C
    ) %>%
      mutate(
    stat = Calculate.asymptotic.binomial.test(x.E, x.C, n.E, n.C)$T.stat,
    dens = dbinom(x.E, n.E, p.E[i])*dbinom(x.C, n.C, p.C[i]),
    prob = NA
  ) ->
  df

prob <- 0
for (j in 1:nrow(df)) {
  # If H_{p, 0} is already rejected, add probability of this response
  # outcome and go to next response outcome
  if(df$stat[j] >= crit.val){
    prob <- prob + 1*df$dens[j]
    df$prob[j] <- 1
    next()
  }

  # If H_{p, 0} is not rejected, calculate probability of rejection of the
  # other two local tests
  k.E <- df$x.E[j]
  k.C <- df$x.C[j]
  # If there is no responder and no non-responder comparison possible, skip
  if((k.E == 0 & k.C == n.C) | (k.E == n.E & k.C == 0)){
    df$prob[j] <- 0
    next()
  }

  # If non-responder comparison is possible, compute acceptance
  # probability of theta.0 test
  if((k.E != n.E & k.C != n.C)){
    crit.0 <- Cond.theta.crit(n.E-k.E, n.C-k.C, alpha.loc)[1]
    factor.0 <- (n.C-k.C)*lambda.0.C/((n.E-k.E)*lambda.0.E)
    ap.0 <- pbetapr(factor.0*exp(crit.0), n.C-k.C, n.E-k.E) -
      pbetapr(factor.0*exp(-crit.0), n.C-k.C, n.E-k.E)
  } else {
    ap.0 <- 1
  }

  # If responder comparison is possible, compute acceptance
  # probability of theta.1 test
  if((k.E != 0 & k.C != 0)){
    crit.1 <- Cond.theta.crit(k.E, k.C, alpha.loc)[1]
    factor.1 <- k.C*lambda.1.C/(k.E*lambda.1.E)
    ap.1 <- pbetapr(factor.1*exp(crit.1), k.C, k.E) -
      pbetapr(factor.1*exp(-crit.1), k.C, k.E)
  } else {
    ap.1 <- 1
  }

  # Add probability of response outcome multiplied with probability of
  # rejecting at least one test of H_{\theta_j, 0}
```

```
      df$prob[j] <- (1 - ap.0*ap.1)
      prob <- prob + df$dens[j] * (1 - ap.0*ap.1)
    }
    prob.vec <- c(prob.vec, prob)
  }

  return(prob.vec)
}
```

Danksagung

Ich danke meinem Doktorvater Prof. Dr. Meinhard Kieser für die Möglichkeit, bei ihm zu promovieren, den Vorschlag des Promotionsthemas, und dafür, dass er mir stets mit Rat und Tat zur Seite stand.

Weiterhin danke ich meinen Kolleginnen und Kollegen am IMBI, insbesondere Dr. Johannes Krisam und Dr. Marietta Kirchner, für den Austausch und den fachlichen Rat.

Außerdem danke ich meinem Vater für seine redaktionelle Hilfe bei der Erstellung meiner Dissertation.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema “Estimation, testing and sample size calculation within the responder stratified exponential survival model” handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Place, Date   doctoral candidate’s signature