Non-proportionality and hypothesis testing in survival analysis: a simulation experiment

Jef Moerman

2025-09-28

Introduction

This document was established to further reflect on an analysis performed for the National Cancer Registre. As a part of their recruitement procedure for a statistician, the candidates were given a data set to assess whether the incidence period (`incperiod`: 2010-2013, 2014-2017, 2018-2021) was a determining factor for the survival of newly diagnosed cancer patients. This data consisted of data corrected for factors:

- gender (`fld_sx`)
- age (`agegr`)
- some comorbidities (`comorb1`, `comorb2`, `comorb3`)
- symptoms at diagnosis (`who`)
- cancer stage (`cStage`)

The follow-up of patients was 3 years, after which they were administratively censored

- survival time was named as: `surv_yy3`
- event indicator was named as: `status3`

Because of data ownership, the presentation was confidential and could not be shared with third parties. This reason was the inspiration to make a simulation study to further question my inference made from the received data.

I reproduced a data set with similar dynamics as the received data, without using estimated parameters of the received data but rather selecting parameters more or less reflecting the dynamics of the received data.

On my presentation, I fitted a Cox proportional hazards model on data where the variable of interest did obey the proportional hazards assumption, but some variables did not. From this I want to adress two concerns with the analysis:

- how does non-proportionality with a known functional dependence present itself in the Schoenfeld residuals
- how reliable is the Wald test for the effect of the incidence period?

In the rest of this article, I discuss:

- the dynamics of the simulated data
- the Schoenfeld residuals of a cox model fitted on these data
- an iterative resampling experiment and its p-value distribution

Data simulation

Simulating covariates

My simulation experiment consists of 4002 patients equally distributed across the three incidence periods 2010-2013, 2014-2017 and 2018-2021 (1334 per period). I wanted to reproduce data in which some covariates were independent of the other variables and some were dependent. Following covariates were randomly sampled independent of other predictor variables:

- gender (`fld_sx`)
- age (`agegr`)
- comorbidity (`comorb1`) (`comorb2`, `comorb3` are not present in my simulation)

To make the data more complex, I decided to make `incperiod` associated with factors:

- cancer stage at diagnosis (`cStage`)
- who score at diagnosis (`who`)

To reflect improved screening management in my simulation, the incidence period 2018-2021 has another distribution of these two variables than incidence periods 2010-2013 and 2014-2017.

Simulating survival times

To simulate survival times, I used weighted sampling of survival times on a time grid of 500 time stamps. I chose time stamp 360 as the three year survival mark, allowing administrative censoring for sim-patients that survived longer than the follow-up time

My sampling technique consisted of the following steps:

1. $log(\lambda)_i$: calculate log-lambda for a patient with given covariates at every timestamp i
2. $\lambda_i = \exp(log(\lambda)_i)$: exponentiate this log-lambda to obtain risks of an event at every timestamp i
3. $r_i = 1 - \exp(-\lambda_i)$: correction in order to not obtain risks larger than 1
4. $f_i = \prod_{j=1}^{i}(1 - r_i)$: calculate the relative frequencies of a cohort of patients with the same covariates
5. $p_i = f_i \times r_i$: the probability of having an event on timestamp i is the risk of an event multiplied by the relative frequency of the remainder of the cohort of identical patients

Variables that obeyed proportionality (gender, age and comorbidity) were introduced in the survival time simulation as:

$$\log \lambda(var) \sim \log \lambda_0 + \beta_{var}$$

Non-proportionality for cancer stage and WHO score was introduced by generating log-lambda values altering at every time stamp. The effect of cancer stage at diagnosis was modeled as a linearly increasing effect of time:

$$\log \lambda(cStage) \sim \log \lambda_0 + \beta_{cStage} \times t$$

The effect of WHO score was modeled as an effect exponentially decaying with time. The chosen relaxation time was 1 year:

$$\log \lambda(who) \sim \log \lambda_0 + \beta_{who} \times \exp(-t/\,y)$$

Schoenfeld residuals

Having `who` and `cStage` as predictors with more than two categories, I wanted to disentangle the dynamics of the Schoenfeld residuals of a Cox model trained on a subset of the data containing only the baseline category and one of the other categories.

For this paragraph, two data sets were generated with survival times dependent on the other covariates and either `who` or `cStage`, suppressing the effect of one of the two on the survival times. This step was

undertaken because `who` and `cStage` are both dependent on `incperiod` and therefore correleted with each other. Suppressing the effect of one covariate leaves a possibility to clearly examine the effect of the other covariate

For each of the two data sets the following steps were undertaken:

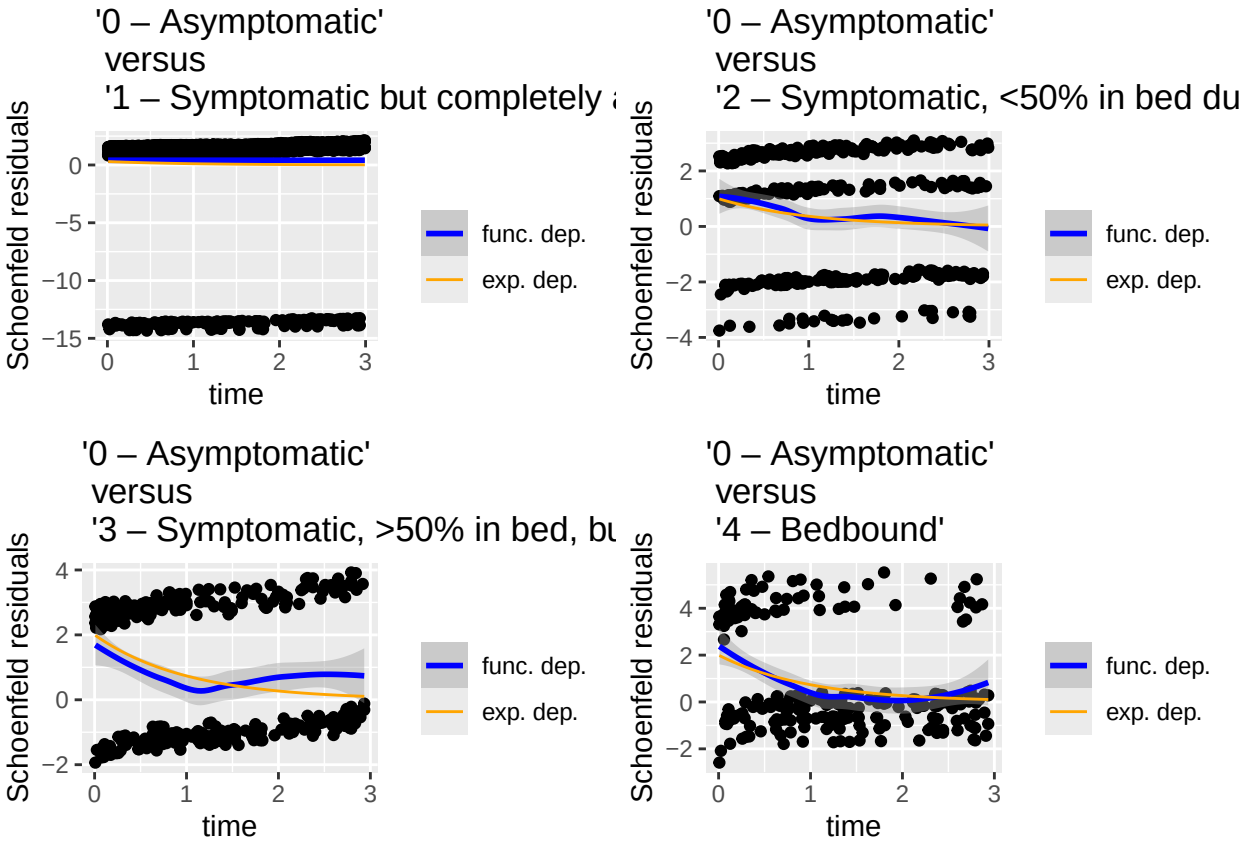
1. filter data set on baseline level ("0 – Asymptomatic" for `who` and "0" for `cStage`) and one of the other levels
2. train Cox proportional hazards model on covariates `incperiod`, `who` and `cStage`
3. calculate the expected functional form of the Schoenfeld residuals with respect to survival time
4. plot Schoenfeld residuals against time, add a locally weighted average and the expected functional dependence

By performing this investigation, a view is generated of how the Schoenfeld residuals and their functional time dependence present of one two-level covariate, without confounding of the other levels of the variable or by other covariates.

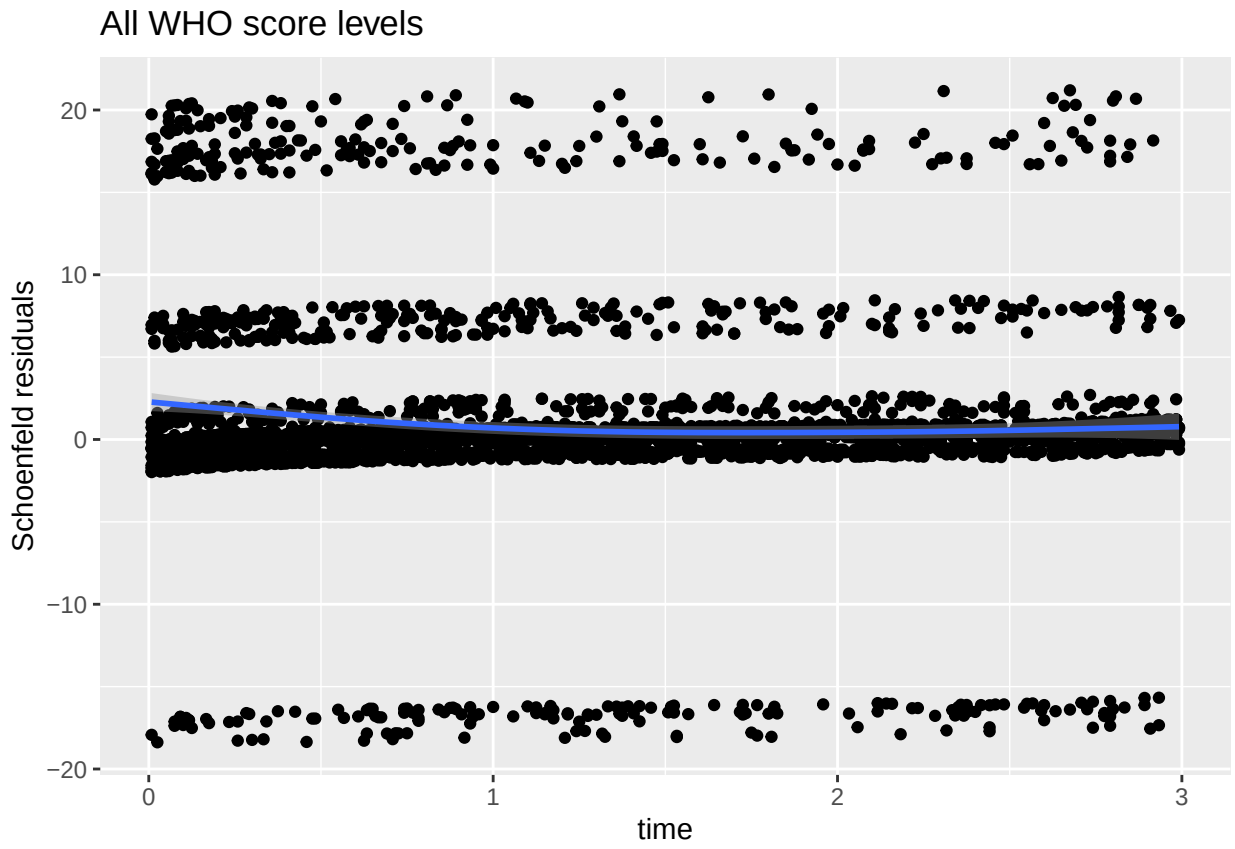
Isolated effects of `who`

Schoenfeld residuals of models trained on the different subsets of `who` and a model trained on the whole synthetic data set.

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



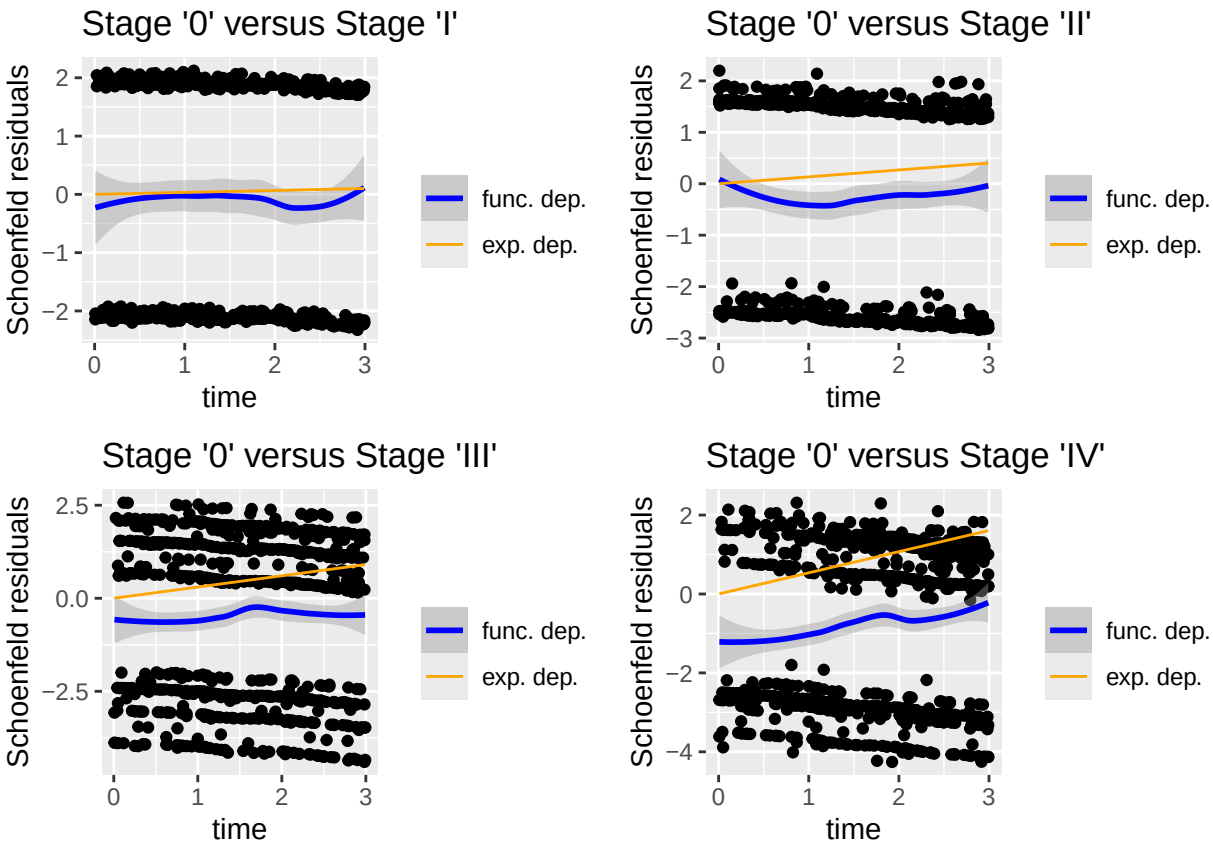

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



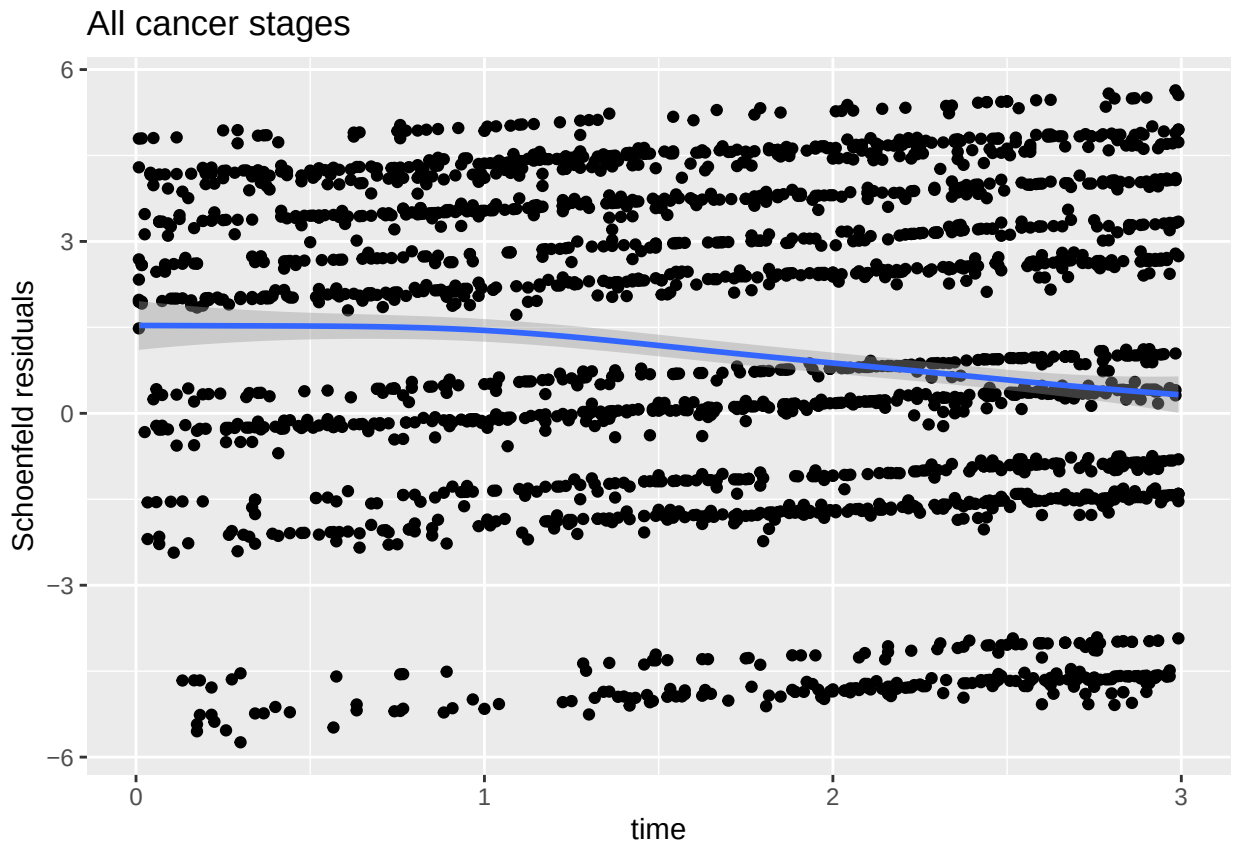
Isolated effects of `cStage`

Schoenfeld residuals of models trained on the different subsets of `cStage` and a model trained on the whole synthetic data set. A Simpson's paradox clearly appears in the all-level combined time dependence of the Schoenfeld residuals.

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Mixed-case Cox PH model (`who` and `cStage`)

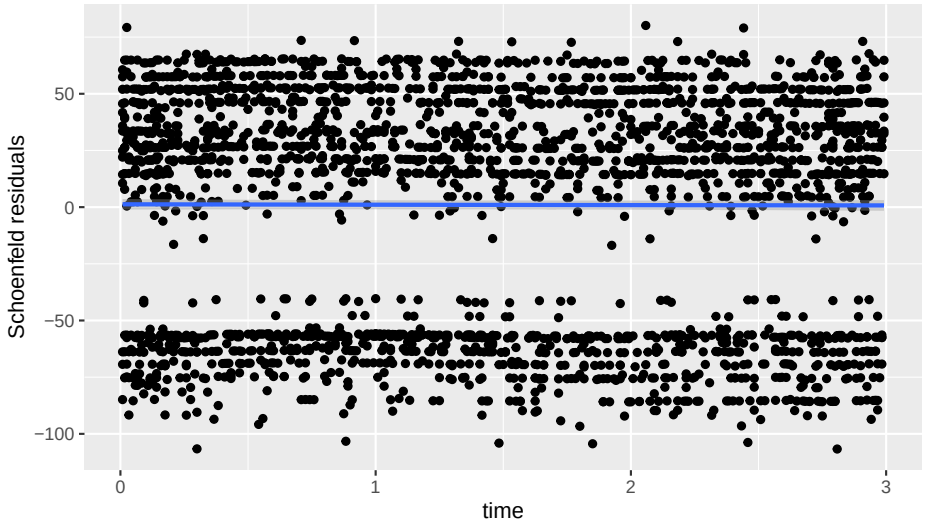
After an examination of the isolated Schoenfeld residuals of `who` and `cStage` we have more insight in how modeled effects present in them. I wanted to understand how the Cox model and the Schoenfeld residuals behave in a mixed-case data set.

I present in the regression coefficients for incidence period of a model with proportional main effects only, as has been the case for examining the residuals. Additionally, the results of a test for proportionality is included.

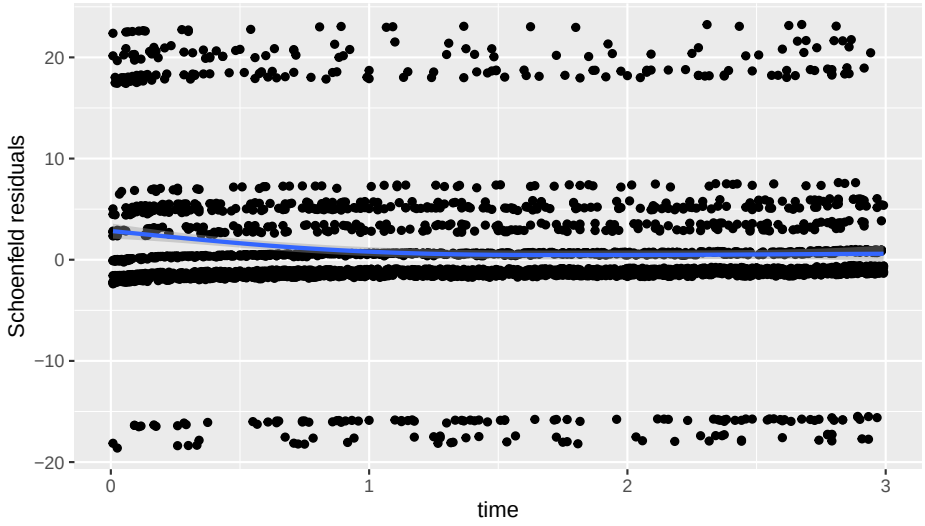
	chisq	df	p
incperiod	1.71	2	4.24e-01
who	28.60	5	2.74e-05
cStage	44.30	5	0.00e+00
GLOBAL	75.50	12	0.00e+00

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

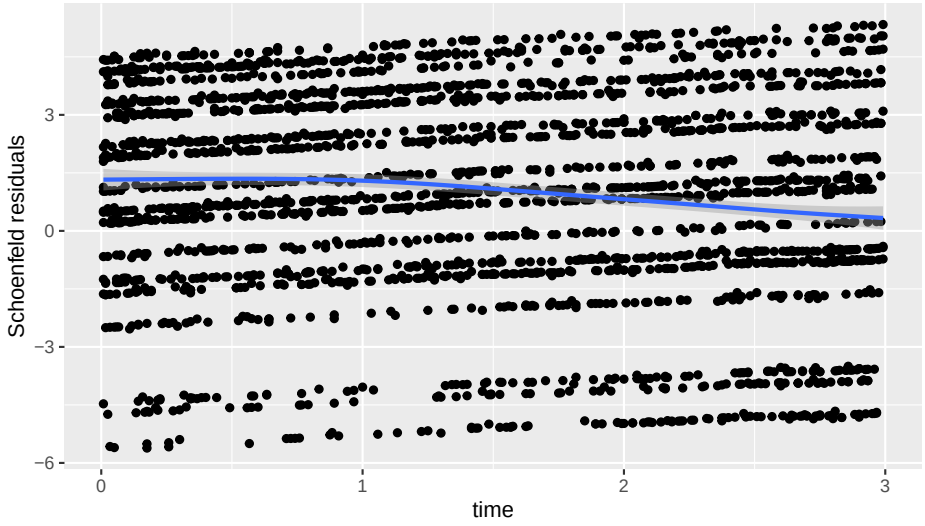
Incidence period



WHO score at diagnosis



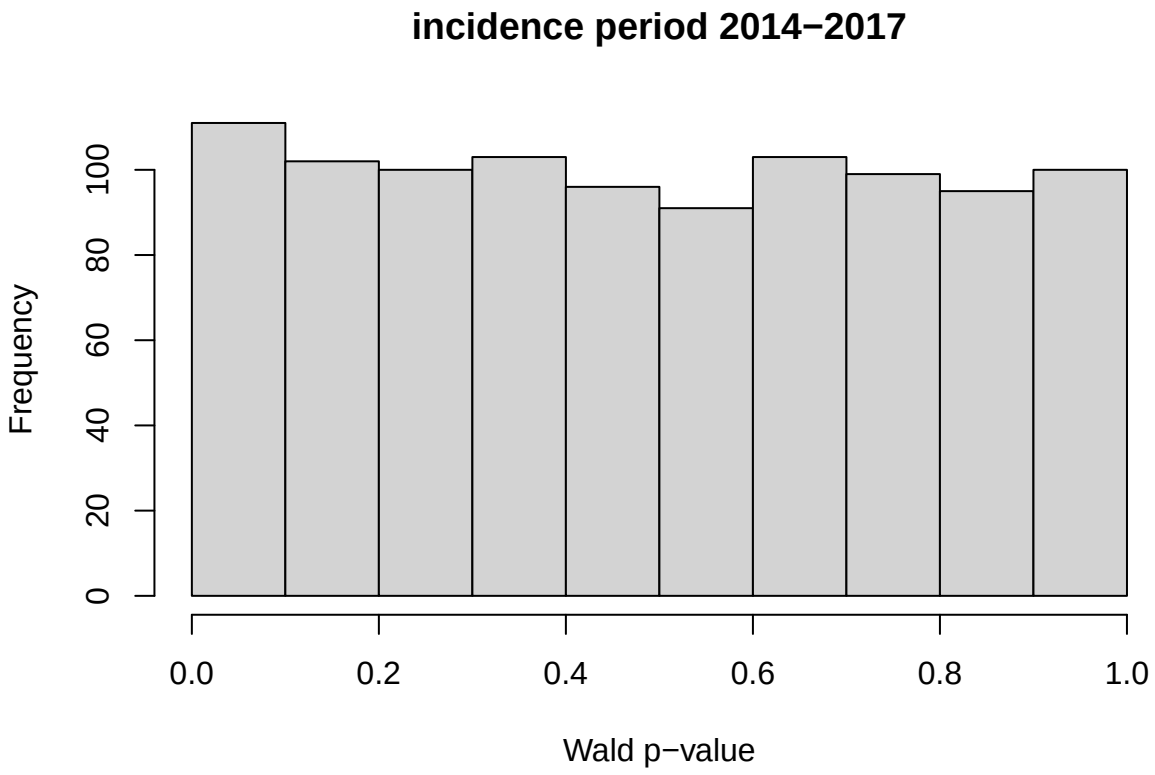
Cancer stage at diagnosis

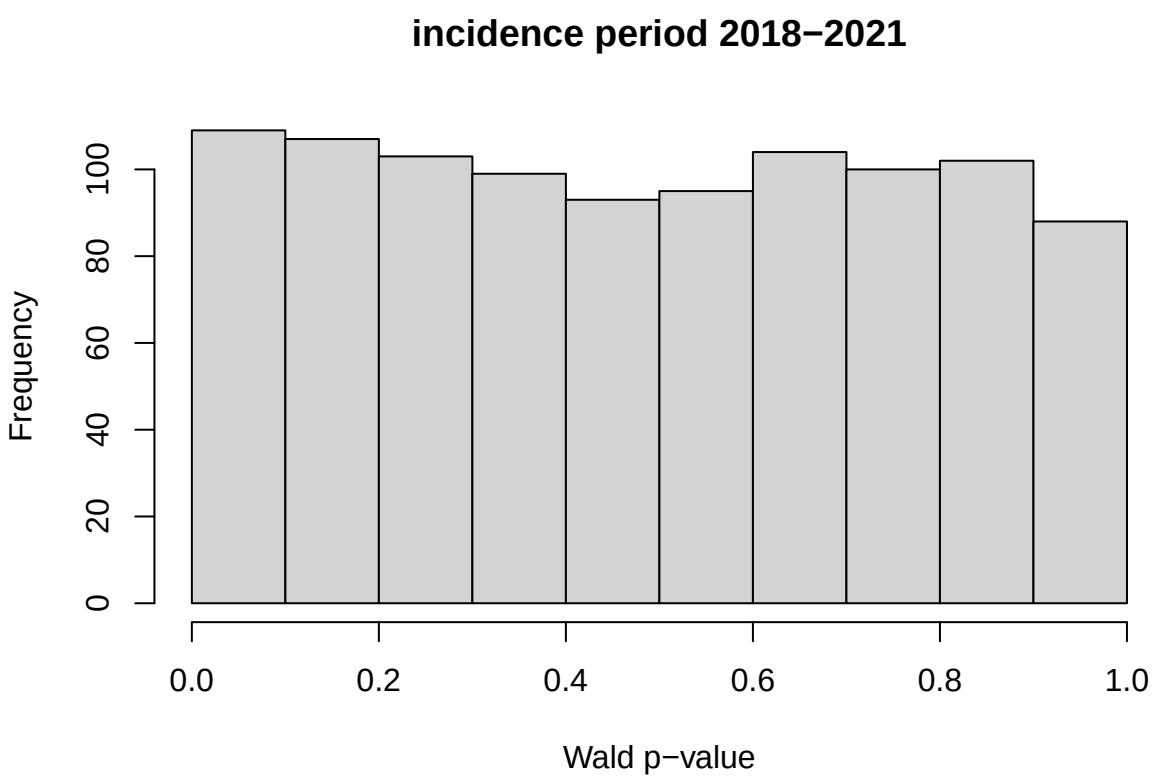


Distribution of the Wald p-values for `incperiod`

An exact test is a test that controls the type I error rate. This is equivalent to stating that the p-values of repeated sampling are uniformly distributed. For 1000 repeated experiments, a histogram was made of the Wald p-values to inspect the implications of the non-proportional confounding variables on hypothesis testing for the beta-coefficients of variable `incperiod` (incidence period).

```
## Warning in e$fun(obj, substitute(ex), parent.frame(), e$data): already
## exporting variable(s): effect
```





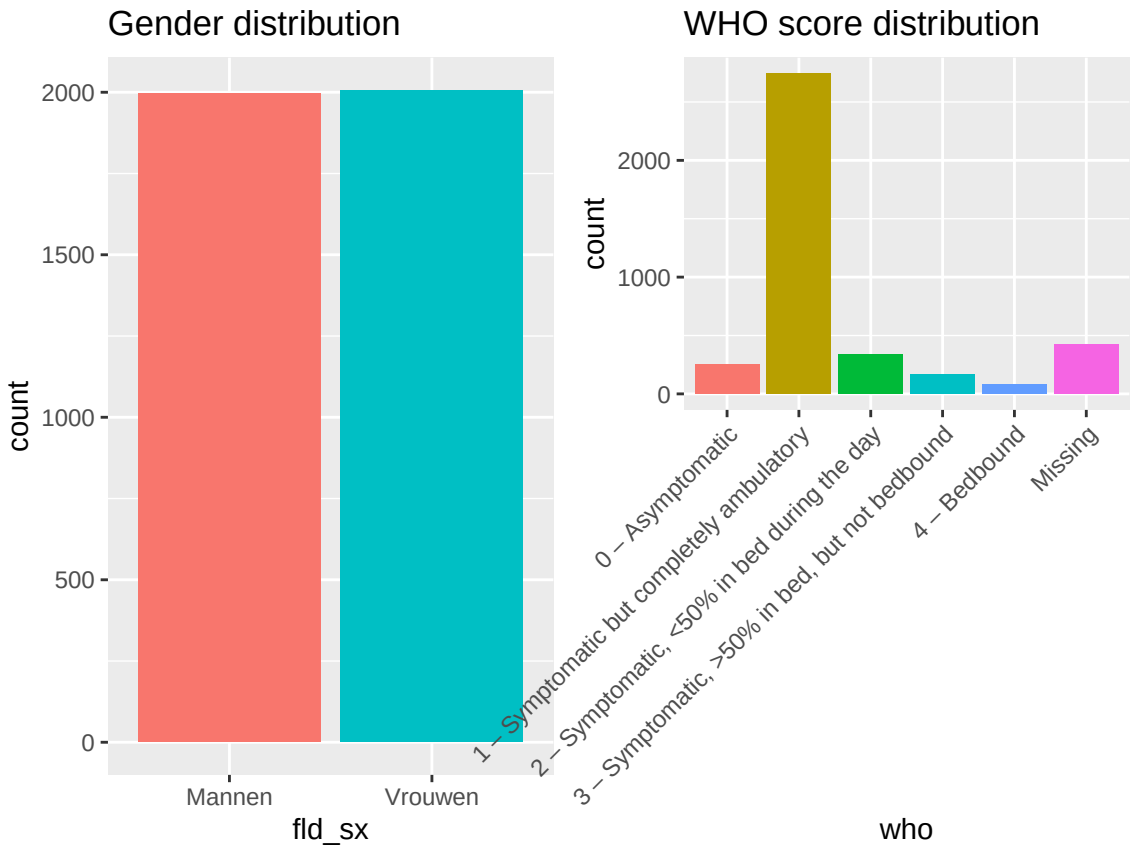
Conclusion

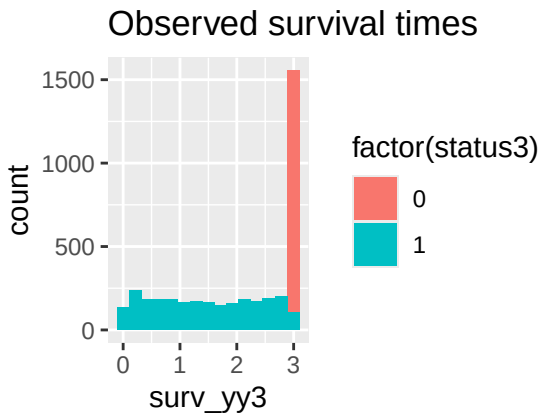
The functional form of the time-dependend log-hazard can be found in the functional form of the Schoenfeld residuals. This finding is applicable for two-level categorical predictors. The time dependence of covariates with more than two levels can be misleading when inspecting the Schoenfeld residuals of a Cox proportional hazards model. In the example given, cancer stage clearly showed a Simpson's paradox, giving a false impression that over time, the effect of cancer stage on survival declines, while in the simulation the effect increases.

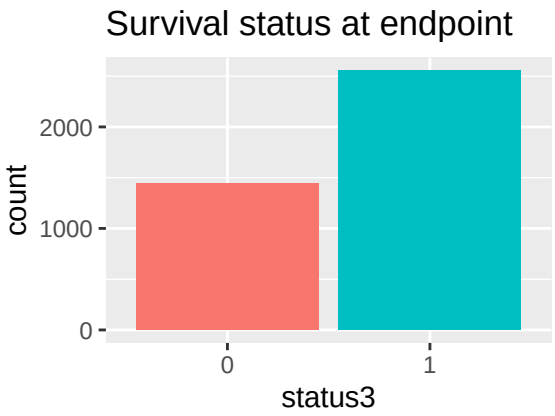
Despite strong deviations from proportionality, the Wald test remained exact for hypothesis testing of `incperiod`. Further research on this finding can be performed on:

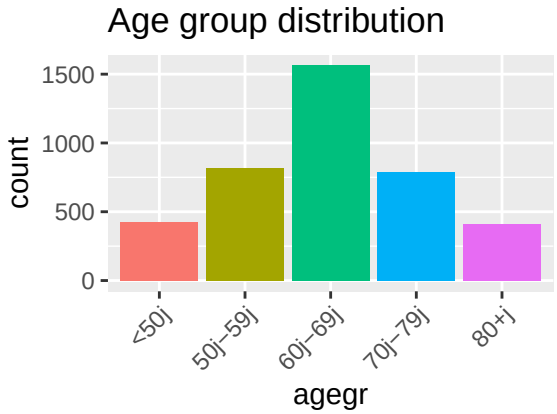
- the power of an alternative hypothesis, in which one of the incidence periods was given an inherent effect on survival
- further investigating the Wald test under the null hypothesis, but with stronger violations of the proportionality assumption

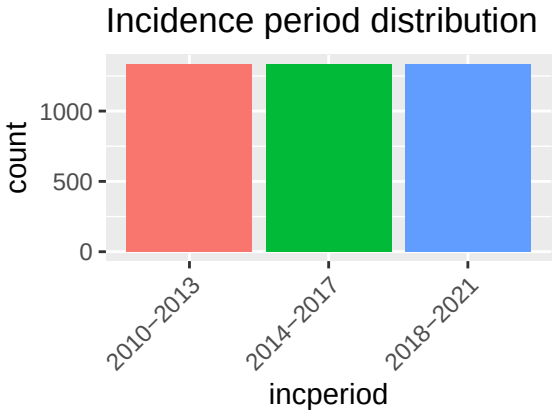
Appendix: observed variable distribution

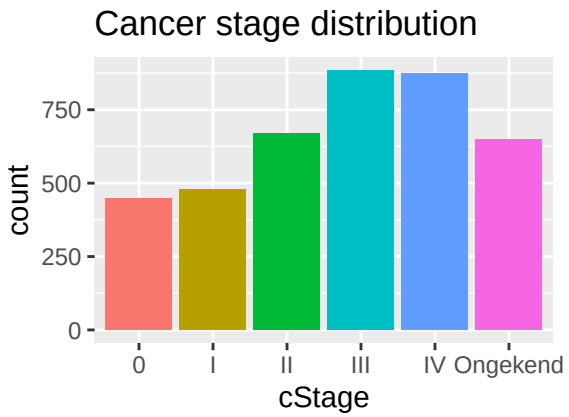


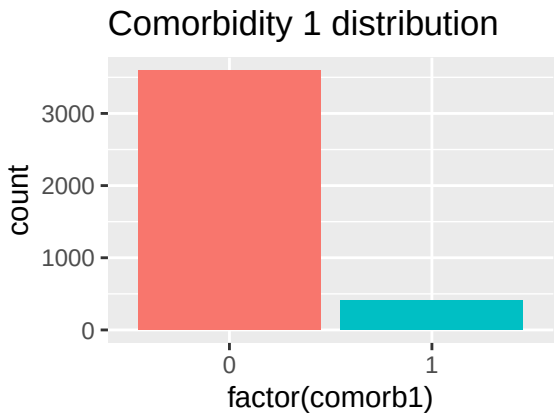

Observed survival times
1500
1000
500
0
count
0
1
2
3
surv_yy3
factor(status3)
0
1

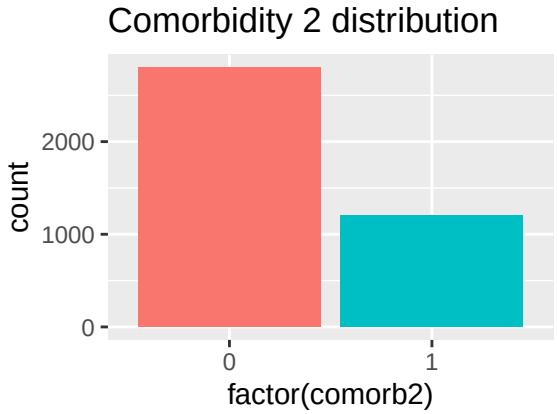

Survival status at endpoint
2000
1000
0
count
0
1
status3

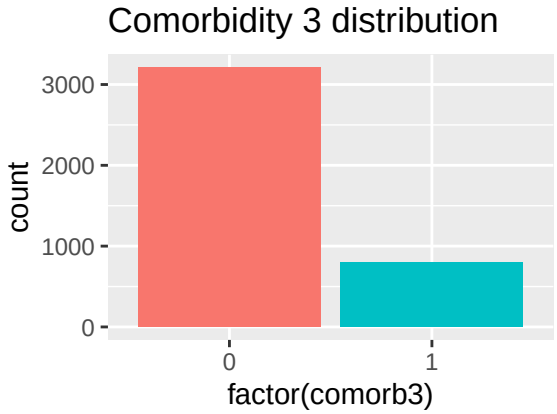

Age group distribution
1500
1000
500
0
count
<50j
50j–59j
60j–69j
70j–79j
80+j
agegr


Incidence period distribution
1000
500
0
count
2010–2013
2014–2017
2018–2021
incperiod


Cancer stage distribution
count
750
500
250
0
0
I
II
III
IV Ongekend
cStage


Comorbidity 1 distribution
count
3000
2000
1000
0
0
1
factor(comorb1)


Comorbidity 2 distribution
count
2000
1000
0
0
1
factor(comorb2)


Comorbidity 3 distribution
count
3000
2000
1000
0
0
1
factor(comorb3)