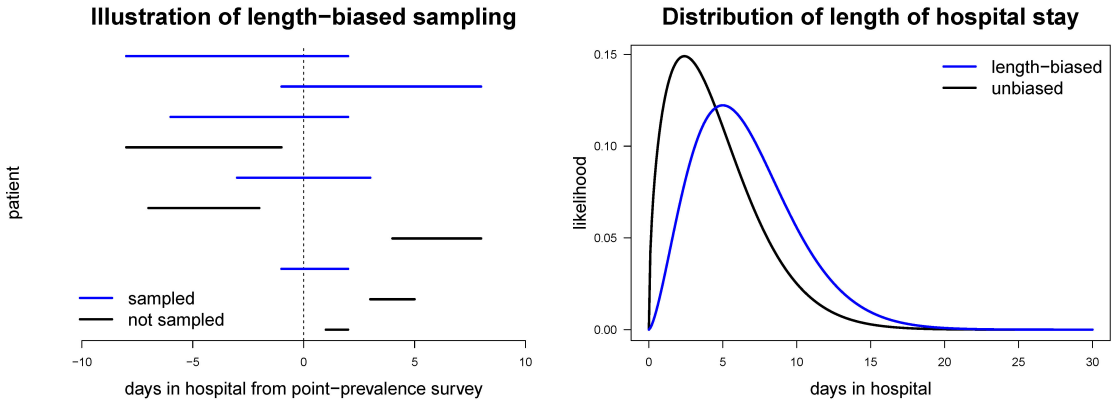
Simulation - weighting

Objective

In this simulation study we illustrate how to correct the length-bias in our models by weighting patients according to their length of stay.

Length bias & weighting

When sampling a hospital population in a one-day prevalence survey, a patient with a length of stay of 10 days will have twice the probability of being sampled compared to a patient with a length of stay of 5 days. The sampled distribution will therefore have length-biased hospital stays compared to the full underlying hospital population.



Under the **steady-state assumption**, the probability of being sampled is proportional to the length of stay. Therefore, the length bias can be corrected by weighting patients inversely to the length of stay. This way, overrepresented long-stayers are downweighted, and underrepresented short-stayers are upweighted.

Data generation

We simulate 1000 hospital cohort populations of ~20,000 patients using Weibull densities from which we take cross-sectional PPS samples of ~625 patients. The shape parameter is the same for all densities to ensure proportional hazards. Generating the cohort population includes a binary covariate „McCabe morbidity score“ with an influence on the infection hazard.

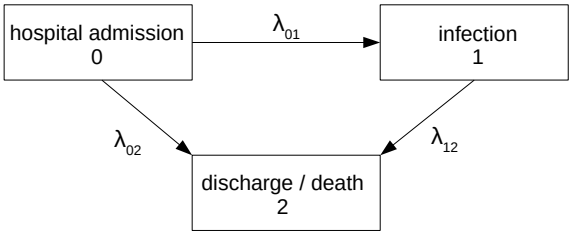
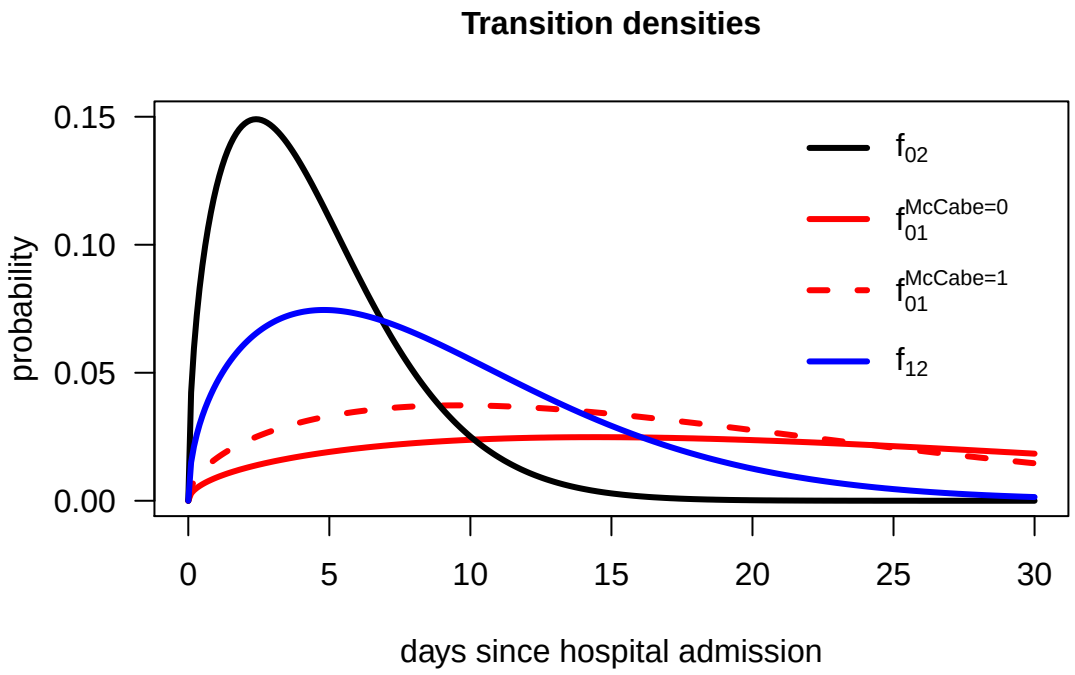


Table 1: Parameter values of the Weibull densities

~	Shape parameter	Scale parameter
$\lambda_{01}^{\text{McCabe=0}}$	1.5	30
$\lambda_{01}^{\text{McCabe=1}}$	1.5	20
λ_{02}	1.5	5
λ_{12}	1.5	10



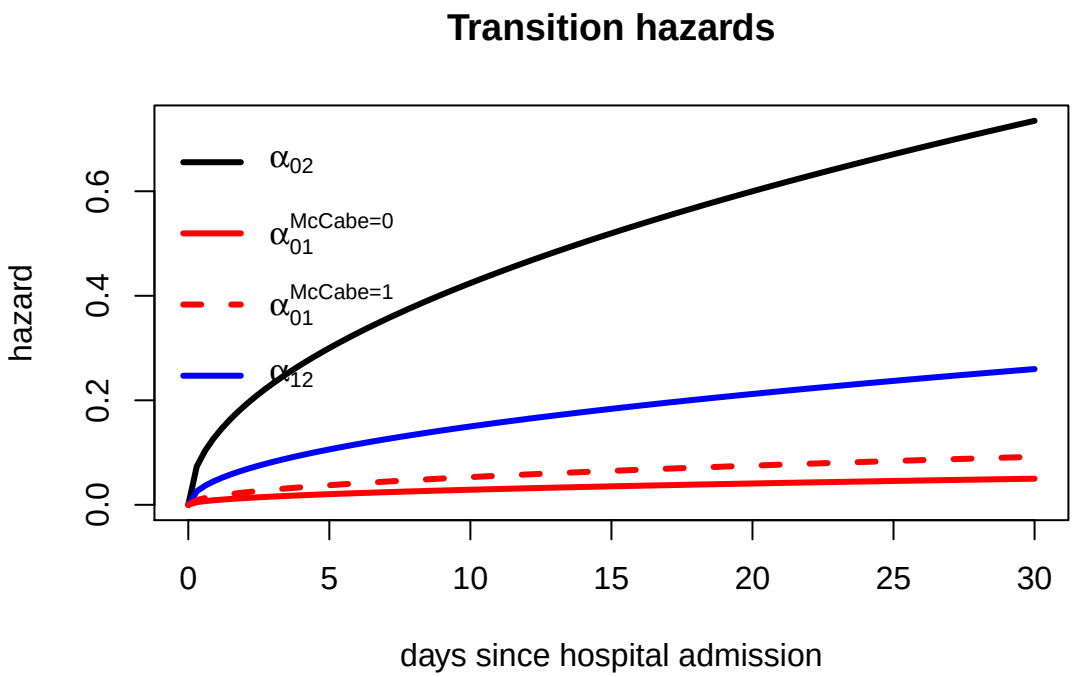
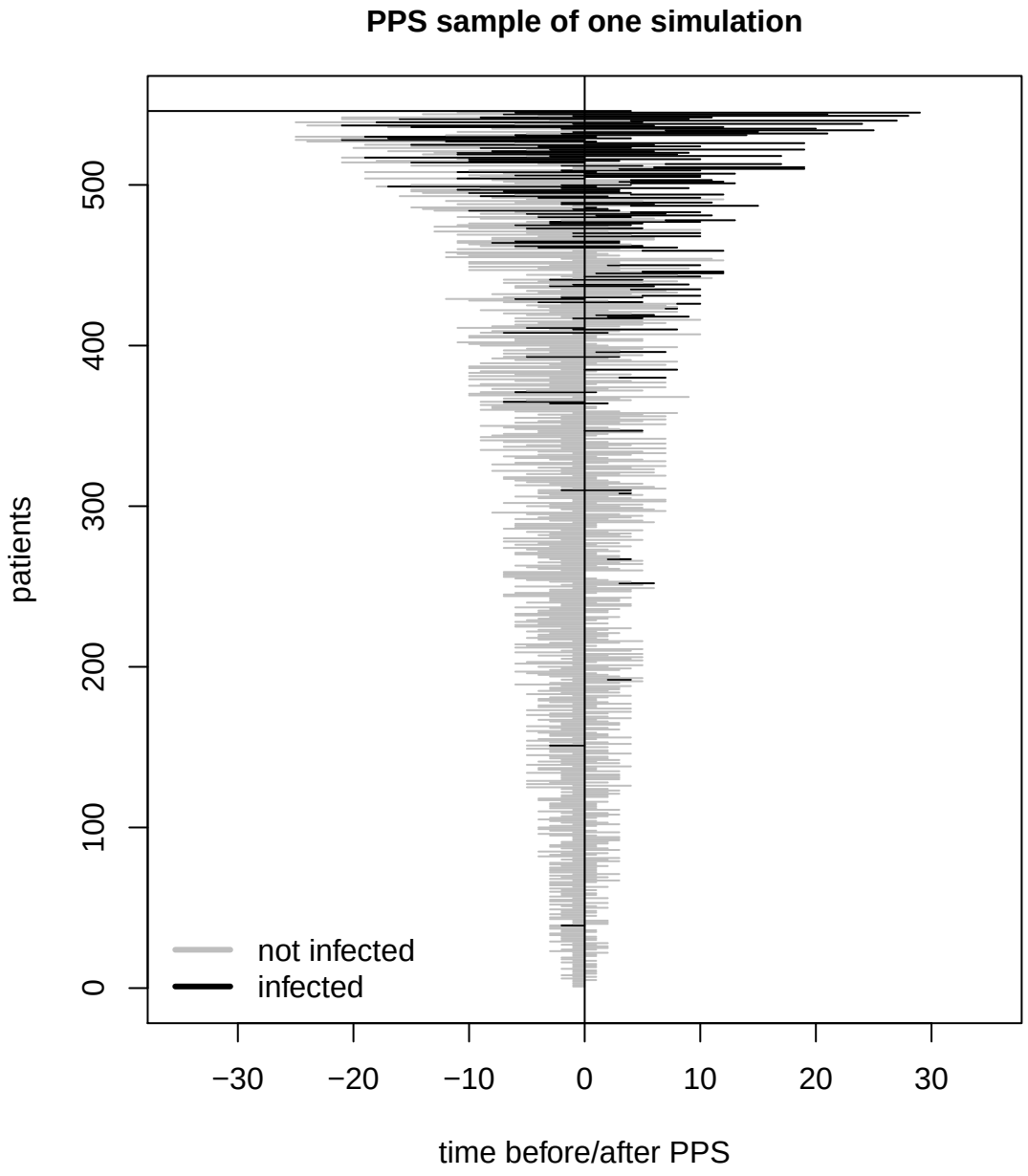


Table 2: Data averages over 1000 simulations. *The length of stays of infected/non-infected are are time-dependent biased.

~	Full cohort	PPS
n	20000	559
incidence proportion	0.083	
prevalence proportion		0.163
length of stay	5.58	8.94
length of stay (infected)*	14.42	17.68
length of stay (non-infected)*	4.78	6.56
time to infection (infected)	4.91	5.52
McCabe (mean)	0.20	0.21



Methods

We fit three logistic regression/Cox models:

- using the full cohort population
- using the PPS sample
- using the PPS sample and weighting by length of stay

We also compare the Rhame & Sudderth formula.

$$\text{Incidence proportion} = \text{Prevalence proportion} \times \frac{\text{length of stay (all patients)}}{\text{time from infection to discharge (infected patients)}}$$

- using data from only from the PPS
- using data from the full cohort

Bias is reported as the difference to the estimate from the full cohort ($\hat{\beta}_{\cdot} - \hat{\beta}_{\text{full cohort}}$). The distribution of the bias over the 1000 studies are summarized using boxplots.

Logistic regression

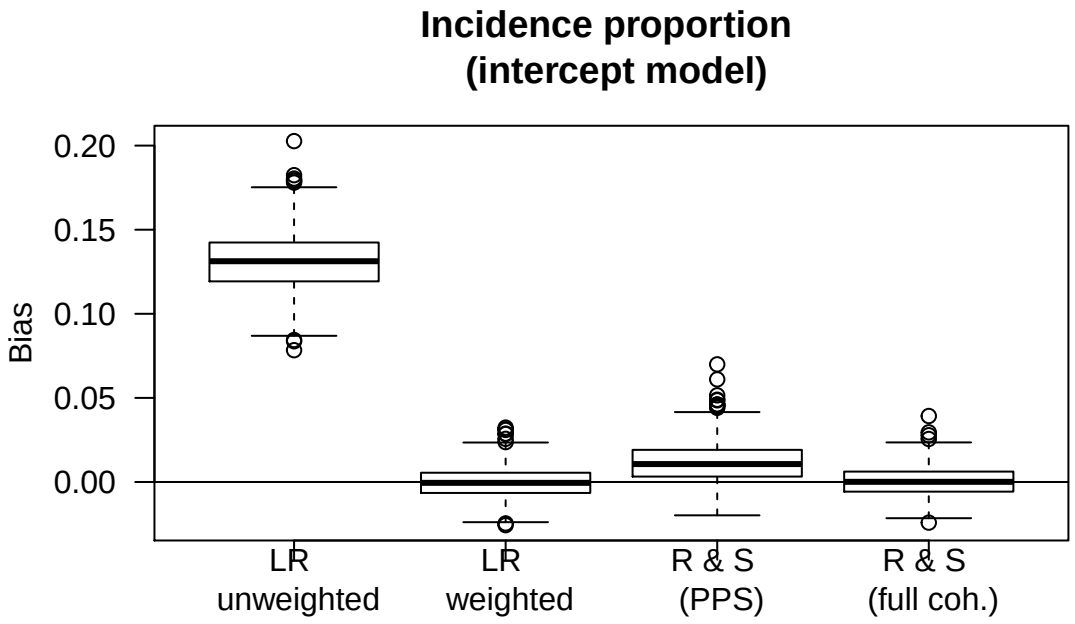
Logistic regression models are fit to the outcome "infected" (1) vs "not infected" (0) by the end of stay.

We calculate two models

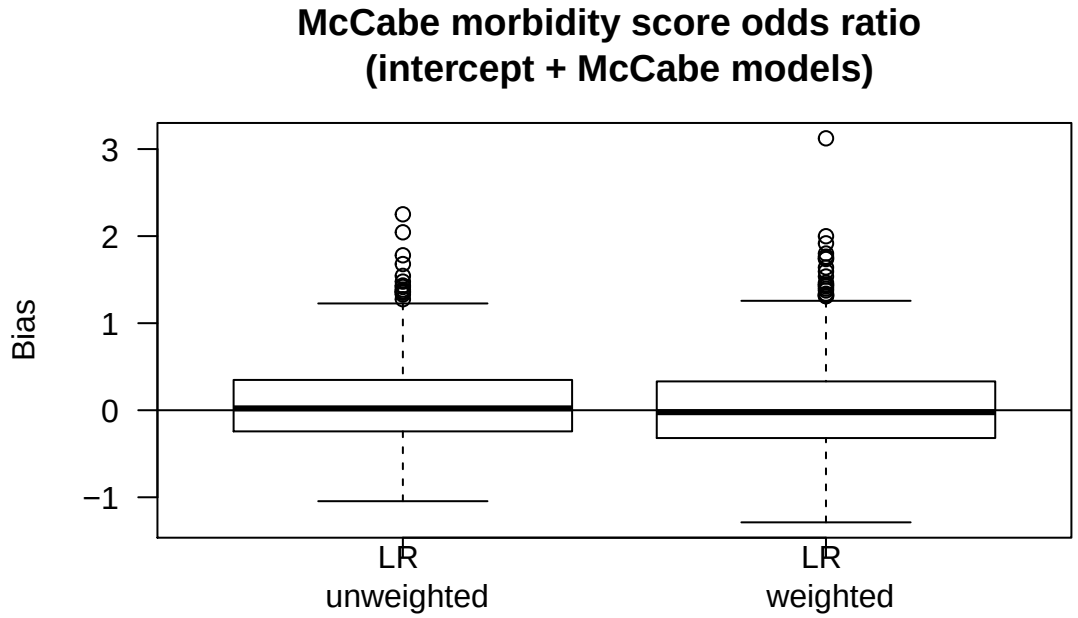
- intercept only
- intercept + McCabe score

The intercept is reported as a proportion using proportion $= \frac{\exp(\hat{\beta}_0)}{\exp(\hat{\beta}_0)+1}$.

Intercept model



Intercept + McCabe model

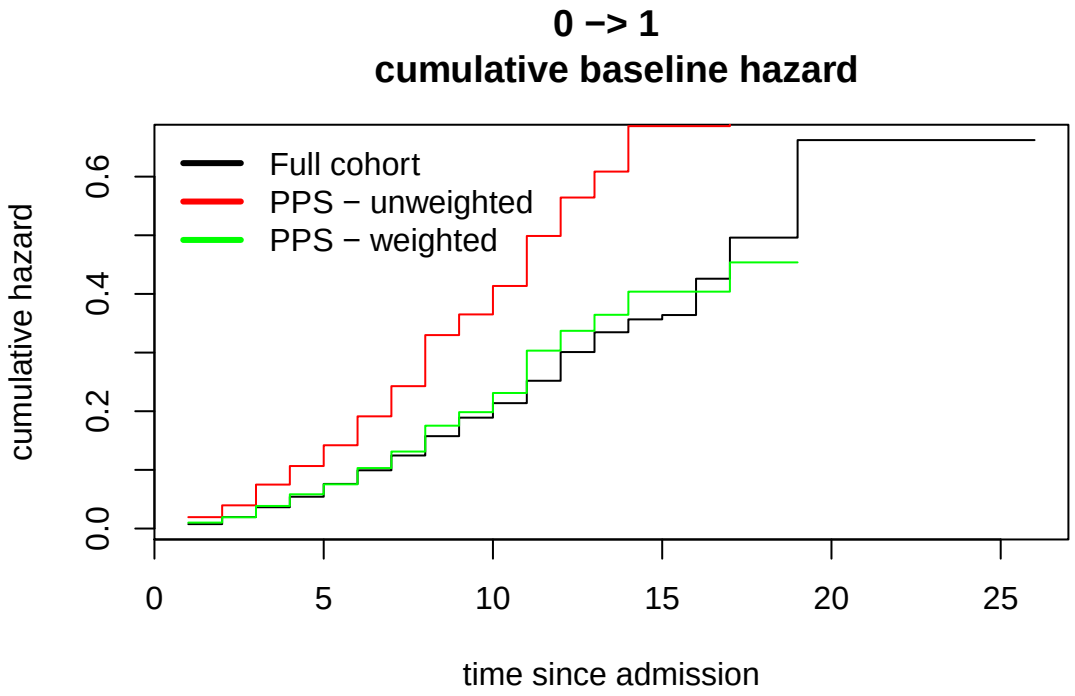


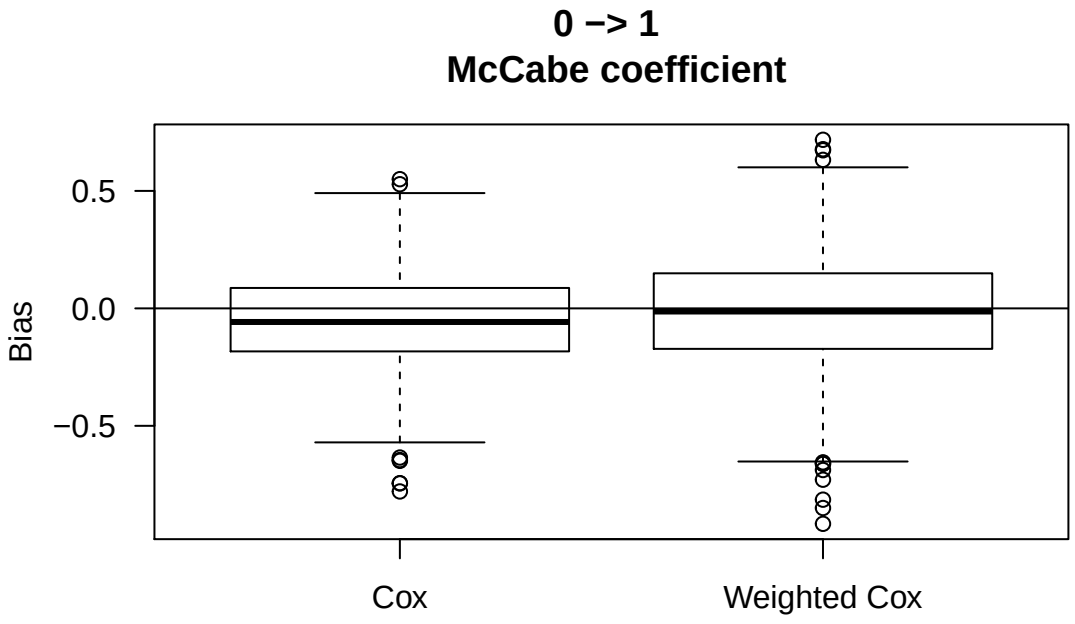
Cox regression

The plotted cumulative hazards are the Breslow estimator of the baseline hazard for the first simulated dataset.

The boxplot are the result of 1000 simulated datasets and show the bias ($\hat{\beta}_{\text{PPS sample}} - \hat{\beta}_{\text{full cohort}}$) of unweighted vs. weighted Cox regression for the log hazard ratio of McCabe score.

Infection hazard (0 -> 1)



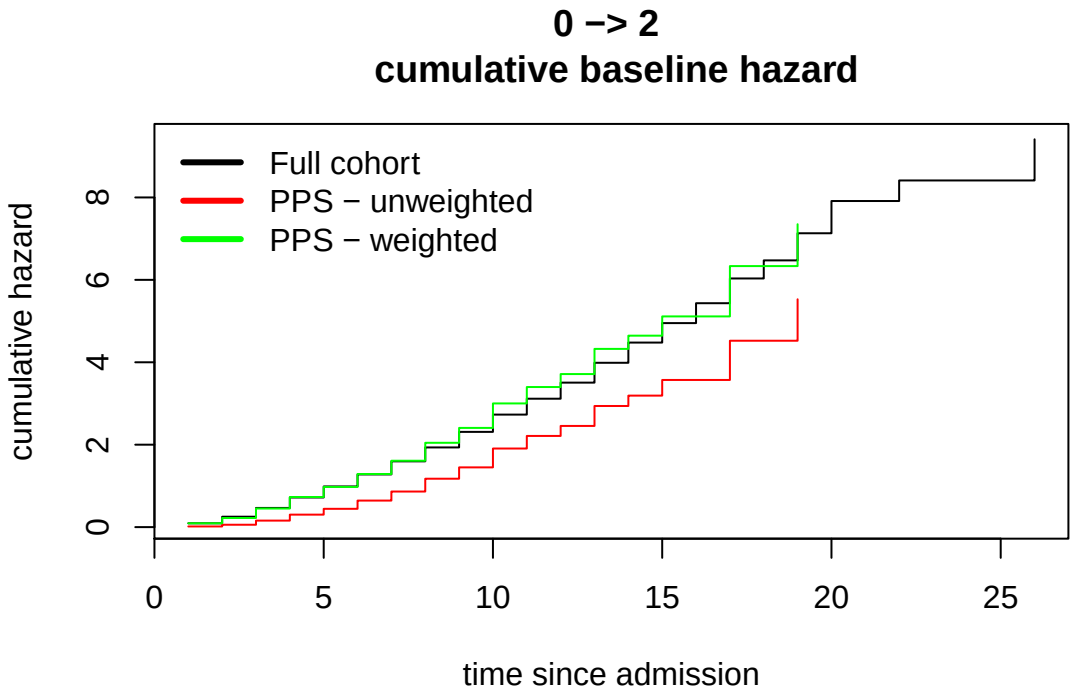


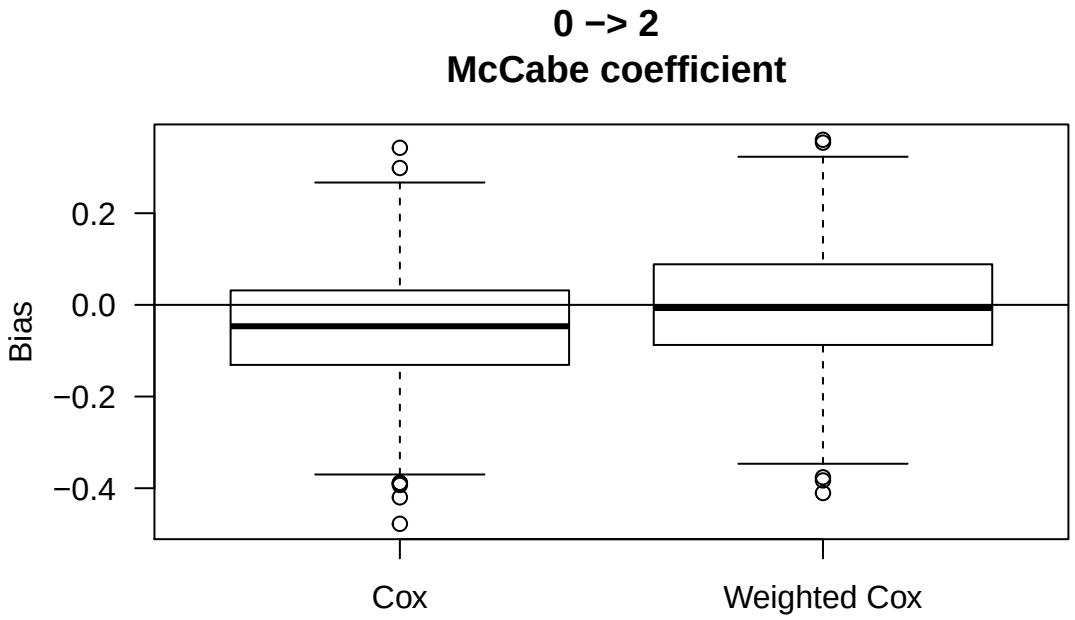
Average McCabe coefficient:

- full cohort population: 0.608

- PPS sample: 0.554
- PPS sample weighted: 0.595

Discharge hazard without infection (0 -> 2)

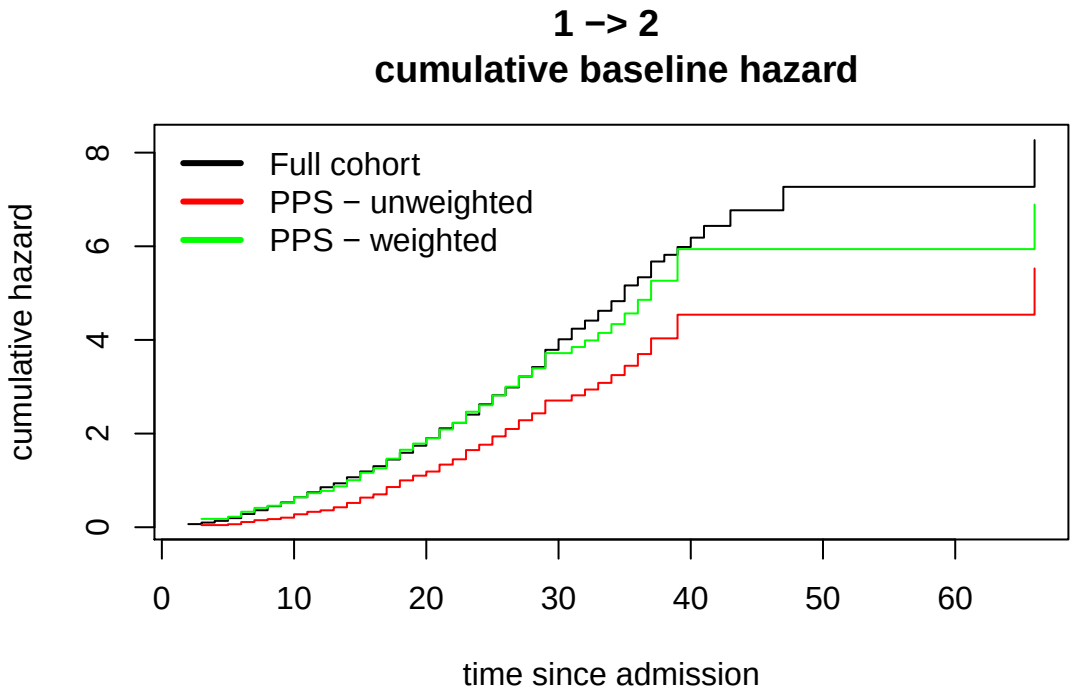


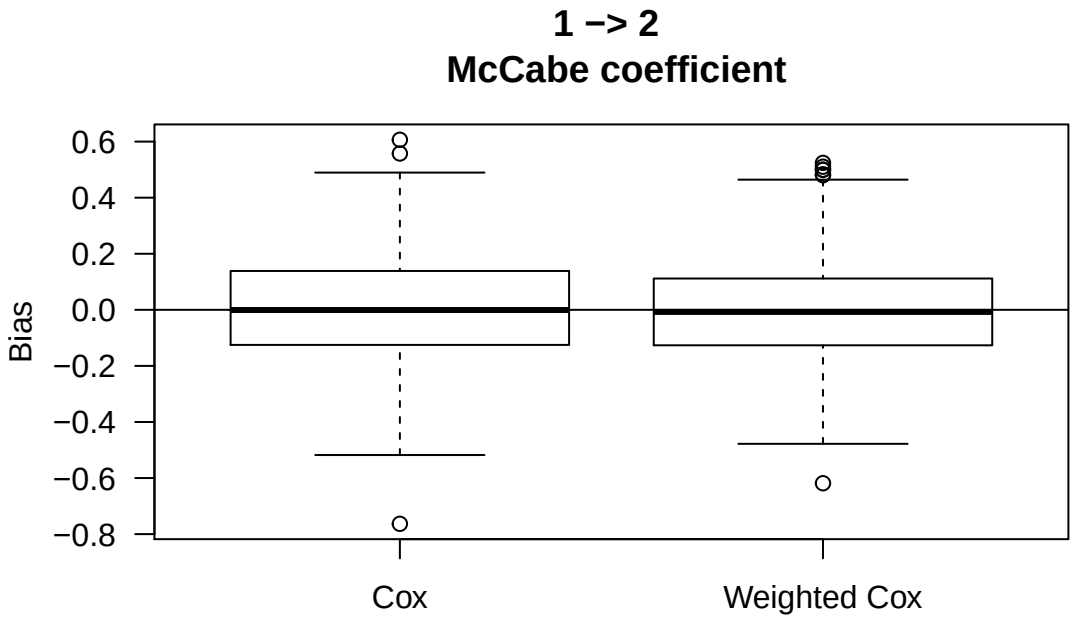


Average McCabe coefficient:

- full cohort population: -0.016
- PPS sample: -0.065
- PPS sample weighted: -0.019

Discharge hazard after infection (1 -> 2)





Average McCabe coefficient:

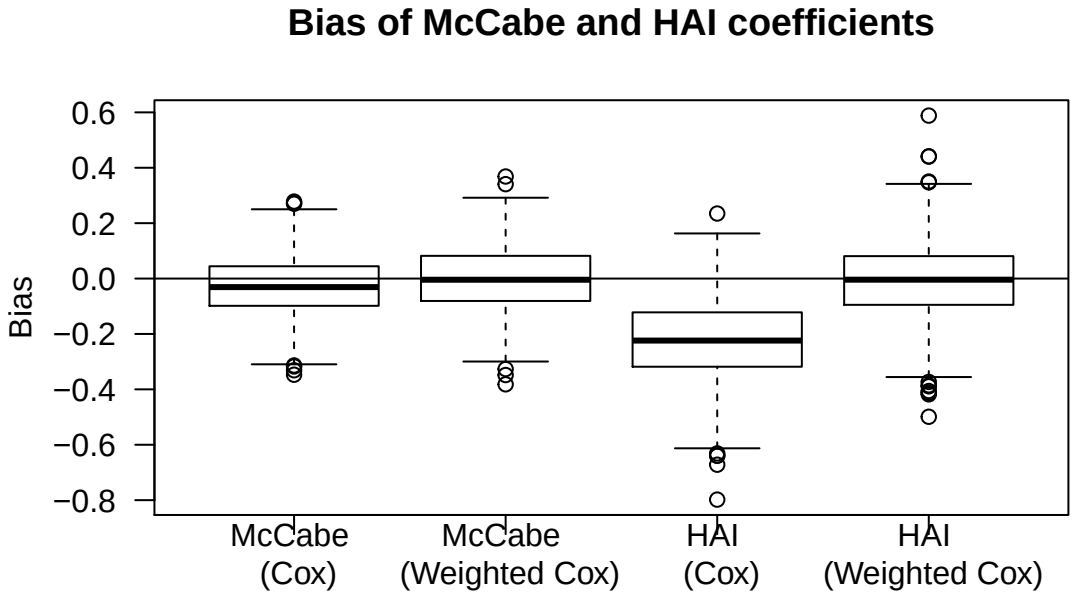
- full cohort population: 0.010

- PPS sample: 0.017
- PPS sample weighted: 0.009

HAI as time-dependent covariate

Data excerpt: one patient with 0 -> 2 transition and one patient with 0 -> 1 -> 2 transition:

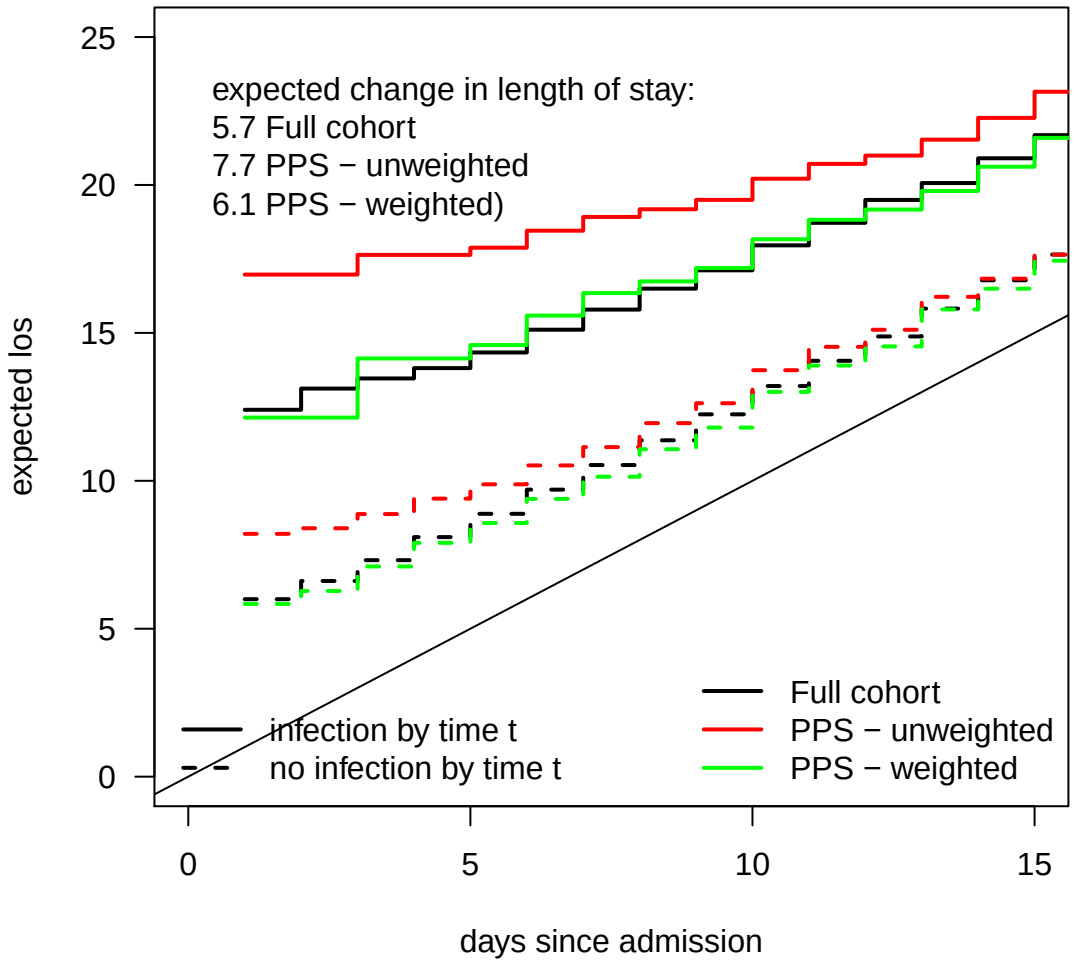
```
##    id from to entry exit mccabe hai
## 1   1    0  2     0    5      0   0
## 15 15    0  1     0    8      0   0
## 16 15    1  2     8   15      0   1
```



Change in length of stay

We determine the change in length of stay using the R package etm, again using the full cohort, the PPS sample, and a weighted PPS sample.

Here, the weighting is implemented by copying observations in the dataset according to their weight: patients with weight = 1 are copied once, patients with weight = $\frac{1}{2}$ are copied twice, etc.



Simulation - imputation

Objective

The purpose of this simulation is to verify that prospective follow-up information after PPS can be imputed for patients where this information is missing.

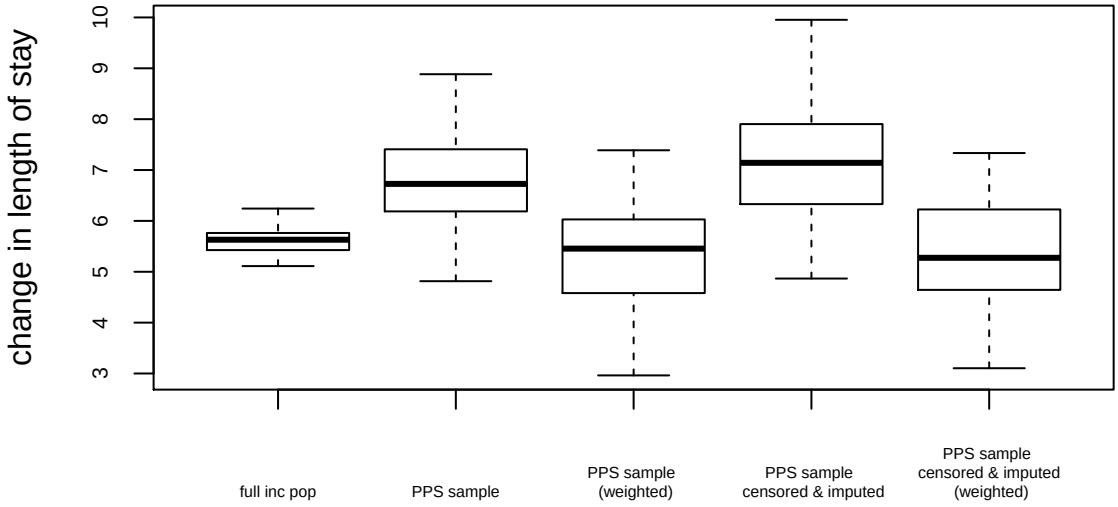
Data generation

We generate data with the same distribution as in the weighting simulation. For 100 datasets, we calculate the change in length of stay due to infection for the following datasets:

- Full incident cohort population
- The PPS sample
- The PPS sample, weighting patients by their length of stay
- The PPS sample with censored observations
  - Randomly, half of the patients are censored at the date of the PPS. We therefore impute their follow-up information from the non-missing half using Weibull densities (see Method section for details)
- The PPS sample with censored observations & imputation (weighted)

Results

The plot shows the change in length of stay for each dataset over 100 replications.



We see that the weighting works well in estimating the change in length of stay in the PPS sample. When we have censored observations, we see that the imputation recreates the same distribution as in the PPS sample. Therefore, we can also use weighting there to estimate the change in length of stay similarly to the full incidence population.

Dataset	Median change in length of stay
Full incidence cohort	5.6
PPS sample	6.7
PPS sample (weighted)	5.5
PPS sample censored & imputed	7.1
PPS sample censored & imputed (weighted)	5.3