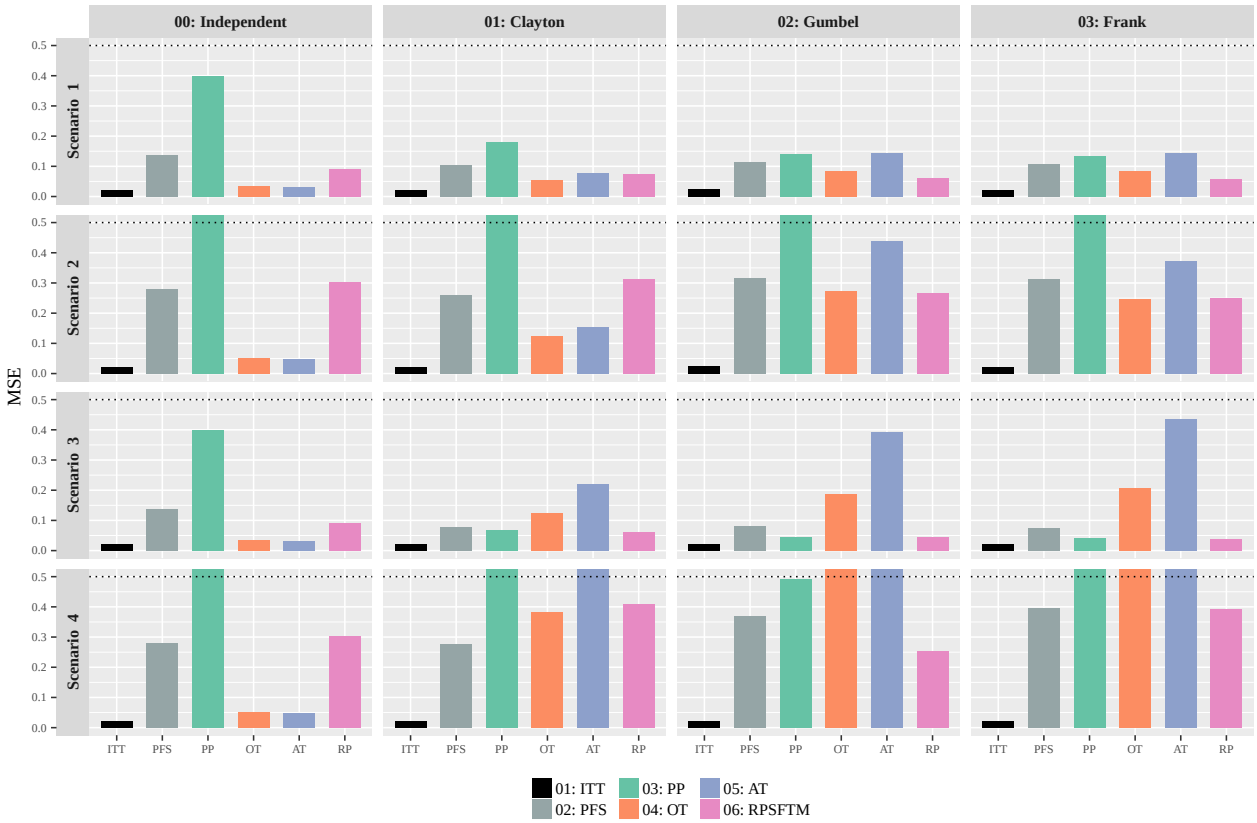
**title:** Supplemental material to:

Performance of randomization-based causal methods with and without integrating external data sources for adjusting overall survival in case of extensive treatment switches in placebo-controlled randomized oncology phase 3 trials

# A Simulation study I: MSE and forest plots

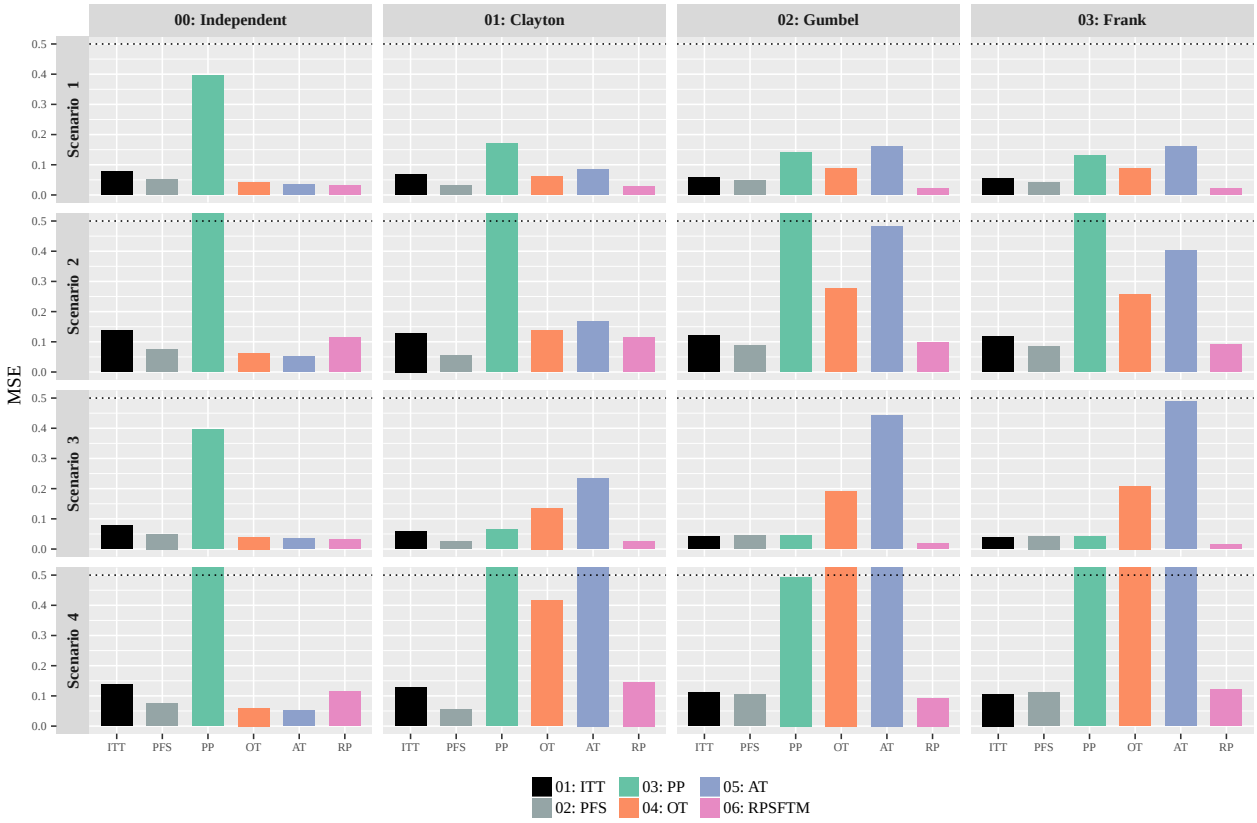
## A.1 MSE under the null




*Bar exceeding the dotted line is omitted.

Figure A1: MSE of HR in simulation study I (under the null)

## A.2 MSE under the alternative




*Bar exceeding the dotted line is omitted.

Figure A2: MSE of HR in simulation study I (under the alternative)

## A.3 Forest plot under the null

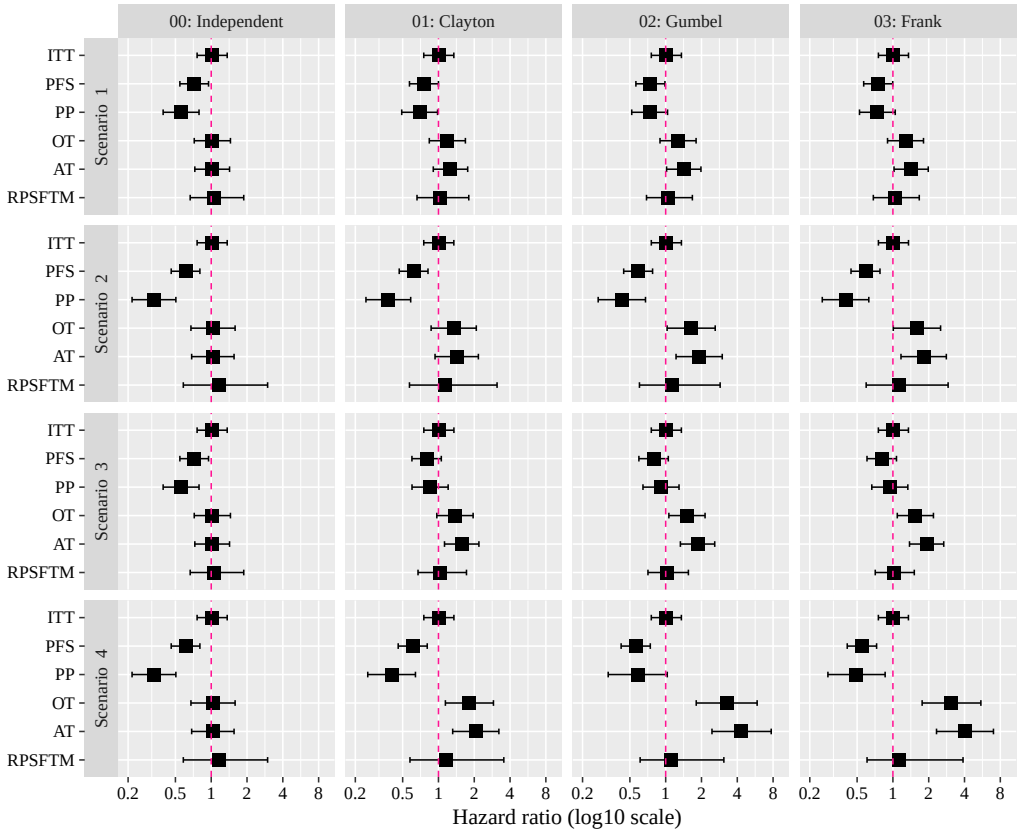



Figure A3: Forest plot of HR in simulation study I (under the null)

## A.4 Forest plot under the alternative

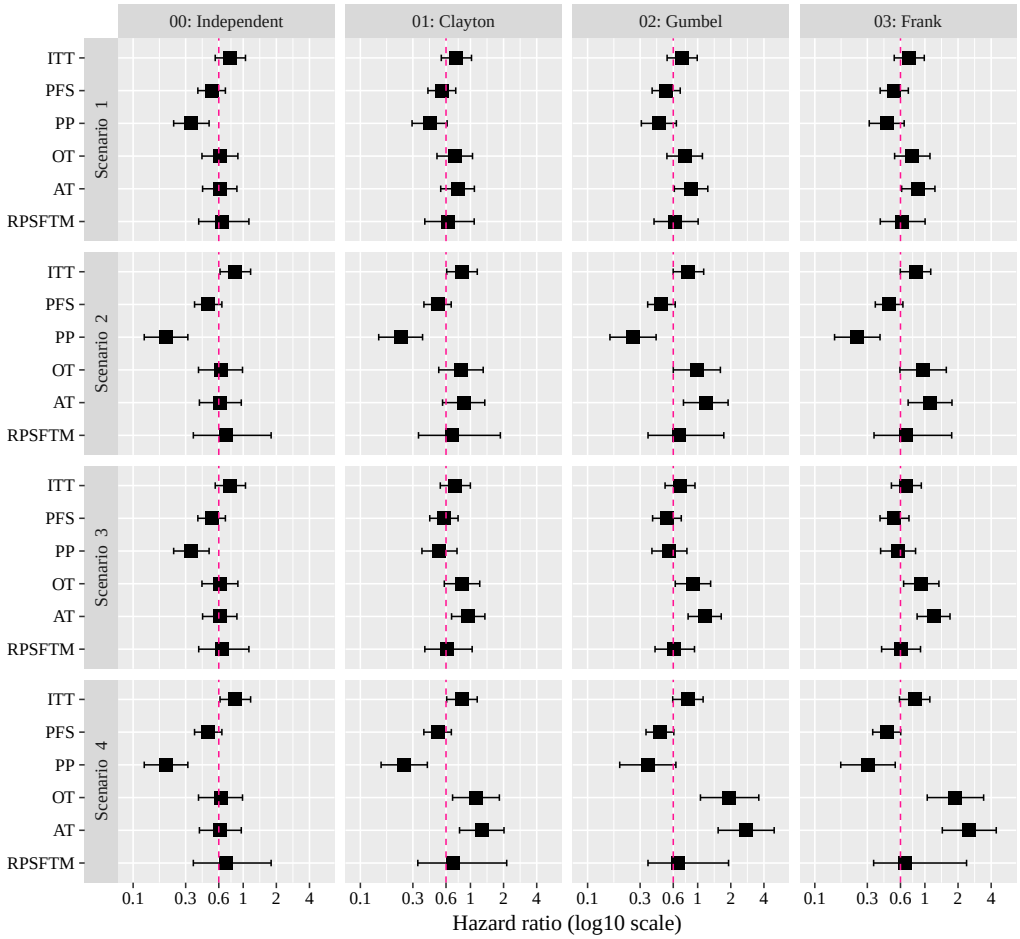
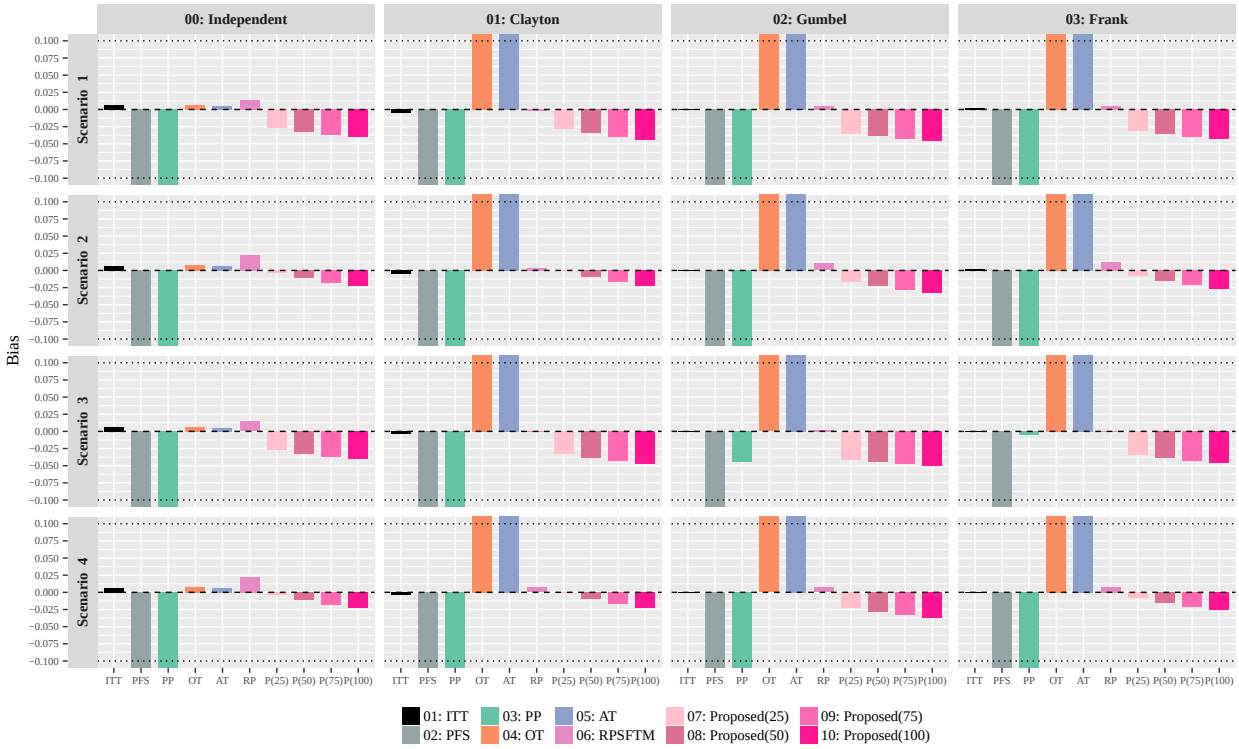



Figure A4: Forest plot of HR in simulation study I (under the alternative)

# B Simulation study II: bias and coverage probability

## B.1 Bias under the null




*Bar exceeding the dotted line is omitted.

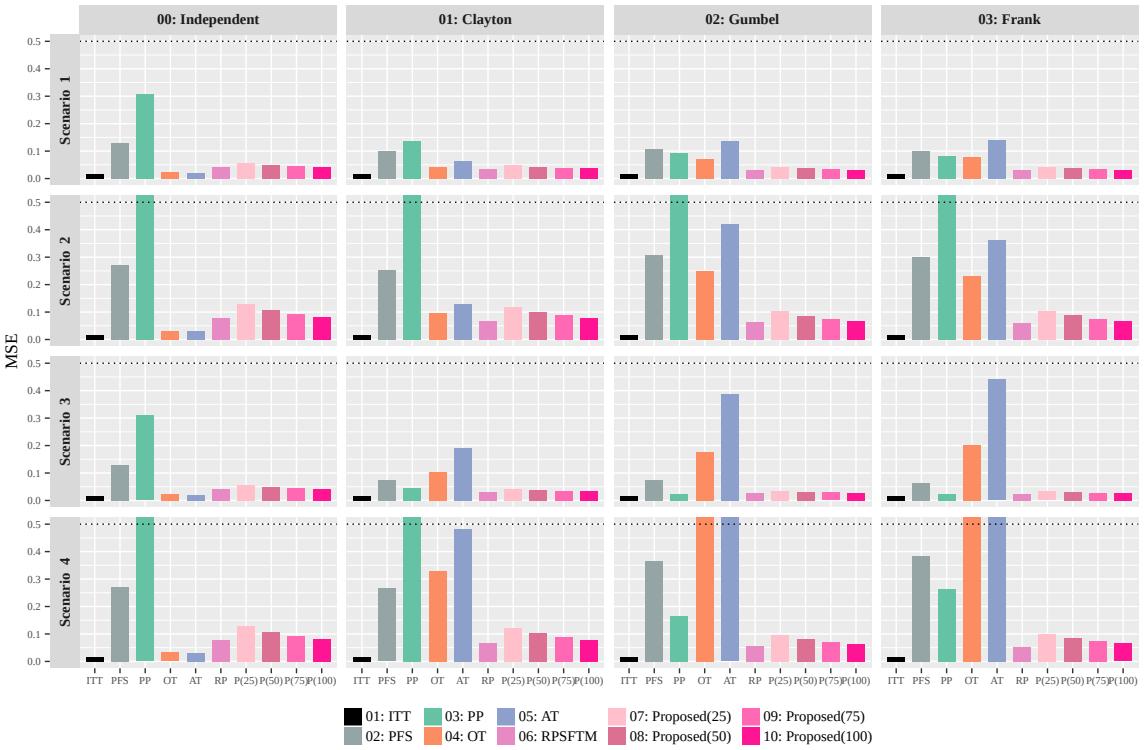
Figure B5: Bias of HR in simulation study II (under the null)

## B.2 Bias under the alternative


*Bar exceeding the dotted line is omitted.

Figure B6: Bias of HR in simulation study II (under the alternative)

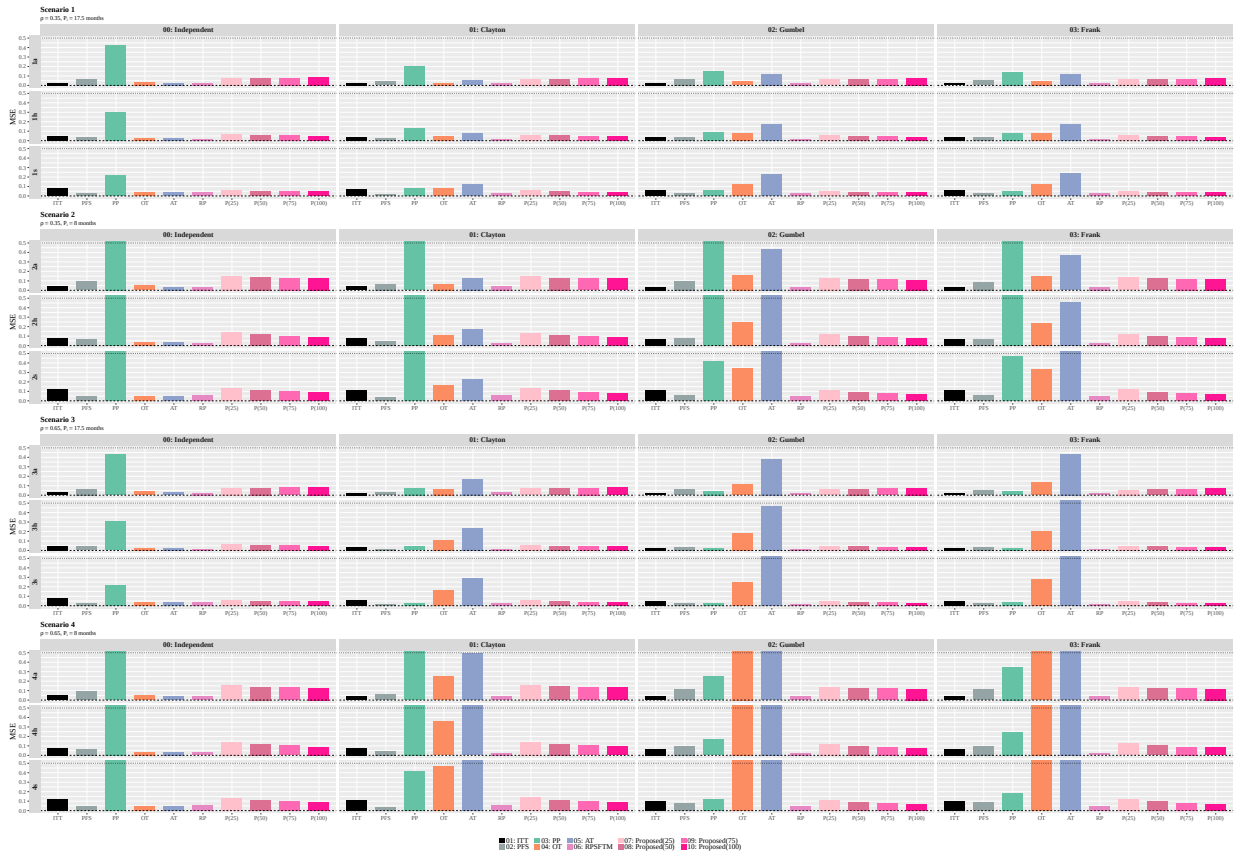
## B.3 MSE under the null




*Bar exceeding the dotted line is omitted.

Figure B7: MSE of HR in simulation study II (under the null)

## B.4 MSE under the alternative




*Bar exceeding the dotted line is omitted.

Figure B8: MSE of HR in simulation study II (under the alternative)

## B.5 Coverage probability under the null

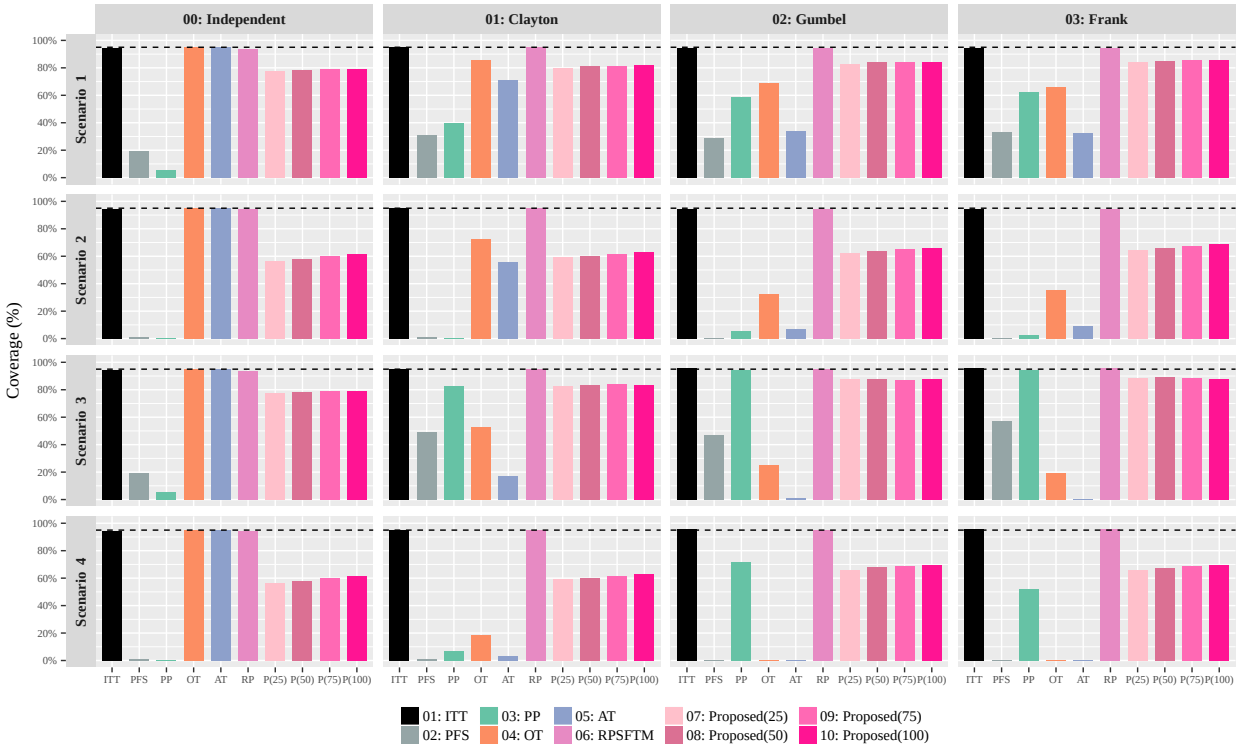



Figure B9: Coverage probability of HR 95% CI in simulation study II (under the null)

## B.6 Coverage probability under the alternative


Figure B10: Coverage probability of HR 95% CI in simulation study II (under the alternative)

# C Simulation study III: Effect of misspecification on type-I error rate

As explained in the Section 2.3, $\eta$ and $\xi$ of our proposed method were calibrated assuming that the true switching mechanisms and the true values of other design parameters (median $P_i$, median $U_i$, and $\rho$) were known. When the true scenario is misspecified, the type-I error rate may be inflated beyond the nominal level. In simulation study III, we assessed an impact of the following misspecification on type-I error rate control of our proposed method.

(a) Switching mechanisms: independent, Clayton, Gumbel, and Frank

(b) Median $P_i$ (proportion of crossover): 8 and 17.5 months

(c) $\rho$: 0.35 and 0.65

We first calibrated ($\eta^*$,$\xi^*$) in a specific situation via simulation. Data generation algorithm was completely the same as in the simulation study II. Next, we repeated generating phase 2 and 3 datasets changing the above (a)-(c) under the homogeneous null hypothesis, and calculated type-I error rate.

## C.1 Situation 1: switching mechanism was misspecified

We first assessed the impact of misspecifying the true switching mechanism. Note that median $P_i$ and $\rho$ were correctly specified. The results are shown in Table C1 and C2. The findings are as follows.

- When independent switching mechanism was mistakenly used for the calibration, type-I error rate was inflated across all scenarios. For example, in scenario 1h, type-I error rate was increased to 3-5% as shown in the top 4 rows in Table C1. Conversely, when dependent switching mechanism was mistakenly used for the calibration, a more stringent value of ($\eta^*$,$\xi^*$) was determined, which resulted in conservative results.

- When Clayton mechanism was mistakenly used for the calibration in the other dependent cases, type-I error rate was inflated.
- When Gumbel mechanism was mistakenly used for the calibration in Clayton mechanism, type-I error rate became smaller than the nominal level. On the other hand, in Frank mechanism, type-I error rate was inflated.
- When Frank mechanism was mistakenly used for the calibration in the other dependent cases, type-I error rate became smaller than the nominal level.
- All the above results were seen across all scenarios regardless of the value of $\alpha_0$.

### C.2 Situation 2: median $P_i$ was misspecified

Misspecification of median $P_i$ indicates that proportion of crossover was lower or higher than expected. The impact of type-I error control was shown in Table C3. Note that switching mechanism and $\rho$ were correctly specified. When ($\eta^*$,$\xi^*$) was calibrated in a dataset with infrequent crossover, type-I error rate became smaller than the nominal level. Conversely, the estimated proportion of crossover was higher than actually observed, and the type-I error rate was highly inflated. The same trends were seen across all scenarios regardless of the value of $\alpha_0$.

### C.3 Situation 3: $\rho$ was misspecified

The impact of misspecifying $\rho$ was shown in Table C4. Specification of $\rho$ does not affect the independent case, and hence, the results only in dependent cases were shown. Note that switching mechanism and median Pi were correctly specified. For Clayton and Gumbel mechanisms, the impact of misspecifying $\rho$ was little, whereas the impact was stronger (type-I error rate was inflated) for the Frank mechanism. The same trends were seen across all scenarios regardless of the value of $\alpha_0$.

Table C1: Type-I error when misspecifying switching mechanism under the homogeneous null (scenario 1h and 2h)

| scenario | Determination of $(\eta^*,\xi^*)$ | borrowing | True switching mechanism | | | |
|---|---|---|---|---|---|---|
| | | | Independent | Clayton | Gumbel | Frank |
| 1h | Independent | 0.25 | 2.5% | 3.3% | 5.0% | 4.1% |
| | | 0.5 | 2.5% | 3.7% | 4.9% | 4.5% |
| | | 0.75 | 2.5% | 3.4% | 4.5% | 4.2% |
| | | 1 | 2.5% | 3.1% | 4.6% | 4.3% |
| | Clayton | 0.25 | 2.0% | 2.5% | 3.9% | 3.1% |
| | | 0.5 | 1.8% | 2.5% | 3.7% | 3.3% |
| | | 0.75 | 1.9% | 2.4% | 3.6% | 3.3% |
| | | 1 | 2.0% | 2.5% | 4.3% | 3.8% |
| | Gumbel | 0.25 | 1.0% | 1.5% | 2.5% | 1.9% |
| | | 0.5 | 0.7% | 1.5% | 2.5% | 2.1% |
| | | 0.75 | 0.7% | 1.0% | 2.4% | 2.2% |
| | | 1 | 0.8% | 1.1% | 2.5% | 2.2% |
| | Frank | 0.25 | 1.5% | 1.9% | 3.3% | 2.5% |
| | | 0.5 | 1.3% | 1.8% | 2.8% | 2.5% |
| | | 0.75 | 1.2% | 1.7% | 2.7% | 2.4% |
| | | 1 | 1.2% | 1.5% | 2.7% | 2.5% |
| 2h | Independent | 0.25 | 2.2% | 3.2% | 4.0% | 4.0% |
| | | 0.5 | 2.2% | 3.3% | 4.5% | 4.1% |
| | | 0.75 | 2.4% | 3.2% | 4.6% | 4.0% |
| | | 1 | 2.5% | 3.5% | 4.7% | 4.1% |
| | Clayton | 0.25 | 1.8% | 2.5% | 3.2% | 3.2% |
| | | 0.5 | 1.5% | 2.5% | 3.4% | 3.1% |
| | | 0.75 | 1.8% | 2.4% | 3.8% | 3.5% |
| | | 1 | 1.9% | 2.4% | 4.4% | 3.6% |
| | Gumbel | 0.25 | 1.0% | 1.6% | 2.5% | 2.5% |
| | | 0.5 | 0.7% | 1.6% | 2.5% | 2.2% |
| | | 0.75 | 1.0% | 1.4% | 2.5% | 2.2% |
| | | 1 | 1.0% | 1.4% | 2.5% | 2.1% |
| | Frank | 0.25 | 1.4% | 1.9% | 2.5% | 2.5% |
| | | 0.5 | 1.1% | 1.7% | 2.9% | 2.5% |
| | | 0.75 | 1.3% | 1.8% | 2.7% | 2.5% |
| | | 1 | 1.3% | 1.7% | 3.0% | 2.5% |

Table C2: Type-I error when misspecifying switching mechanism under the homogeneous null (scenario 3h and 4h)

| scenario | Determination of $(\eta^*,\xi^*)$ | borrowing | True switching mechanism | | | |
|---|---|---|---|---|---|---|
| | | | Independent | Clayton | Gumbel | Frank |
| 3h | Independent | 0.25 | 2.5% | 4.0% | 5.1% | 6.5% |
| | | 0.5 | 2.5% | 3.8% | 4.9% | 7.0% |
| | | 0.75 | 2.5% | 3.6% | 4.4% | 6.4% |
| | | 1 | 2.5% | 3.6% | 4.6% | 6.1% |
| | Clayton | 0.25 | 1.6% | 2.5% | 3.2% | 3.9% |
| | | 0.5 | 1.5% | 2.5% | 3.1% | 4.5% |
| | | 0.75 | 1.7% | 2.5% | 3.2% | 4.5% |
| | | 1 | 1.7% | 2.4% | 3.5% | 4.3% |
| | Gumbel | 0.25 | 1.0% | 1.6% | 2.5% | 3.6% |
| | | 0.5 | 0.8% | 1.4% | 2.5% | 3.5% |
| | | 0.75 | 1.0% | 1.6% | 2.4% | 3.4% |
| | | 1 | 0.9% | 1.6% | 2.5% | 3.4% |
| | Frank | 0.25 | 0.6% | 1.0% | 1.4% | 2.5% |
| | | 0.5 | 0.5% | 0.8% | 1.7% | 2.5% |
| | | 0.75 | 0.6% | 0.9% | 1.4% | 2.5% |
| | | 1 | 0.6% | 0.9% | 1.4% | 2.5% |
| 4h | Independent | 0.25 | 2.2% | 3.7% | 4.9% | 5.1% |
| | | 0.5 | 2.2% | 3.3% | 4.7% | 4.8% |
| | | 0.75 | 2.4% | 3.2% | 4.6% | 5.1% |
| | | 1 | 2.5% | 3.5% | 4.4% | 5.1% |
| | Clayton | 0.25 | 1.4% | 2.4% | 3.4% | 3.7% |
| | | 0.5 | 1.4% | 2.2% | 3.3% | 3.5% |
| | | 0.75 | 1.6% | 2.5% | 3.5% | 3.6% |
| | | 1 | 2.0% | 2.4% | 3.5% | 3.9% |
| | Gumbel | 0.25 | 1.1% | 1.5% | 2.5% | 2.7% |
| | | 0.5 | 0.8% | 1.4% | 2.4% | 2.4% |
| | | 0.75 | 0.9% | 1.5% | 2.5% | 2.6% |
| | | 1 | 1.3% | 1.6% | 2.5% | 3.1% |
| | Frank | 0.25 | 0.6% | 1.7% | 2.3% | 2.5% |
| | | 0.5 | 0.6% | 1.1% | 2.3% | 2.5% |
| | | 0.75 | 0.7% | 1.0% | 2.3% | 2.5% |
| | | 1 | 0.8% | 1.1% | 2.2% | 2.5% |

Table C3: Type-I error when misspecifying median $P_i$ under the homogeneous null

| Scenario (calibration) | True scenario | family | $\alpha_0 = 0.25$ | $\alpha_0 = 0.5$ | $\alpha_0 = 0.75$ | $\alpha_0 = 1$ |
|---|---|---|---|---|---|---|
| 1h (median $P_i$=17.5, $\rho = 0.35$) | 2h (median $P_i$=8, $\rho = 0.35$) | Independent | 0.3% | 0.3% | 0.3% | 0.3% |
| | | Clayton | 0.5% | 0.5% | 0.5% | 0.6% |
| | | Gumbel | 0.3% | 0.3% | 0.4% | 0.4% |
| | | Frank | 0.3% | 0.3% | 0.4% | 0.6% |
| 2h (median $P_i$=8, $\rho = 0.35$) | 1h (median $P_i$=17.5, $\rho = 0.35$) | Independent | 8.6% | 8.1% | 7.5% | 7.1% |
| | | Clayton | 8.6% | 8.8% | 8.3% | 7.3% |
| | | Gumbel | 10.3% | 9.7% | 9.2% | 9.0% |
| | | Frank | 8.1% | 8.5% | 8.2% | 7.9% |
| 3h (median $P_i$=17.5, $\rho = 0.65$) | 4h (median $P_i$=8, $\rho = 0.65$) | Independent | 0.3% | 0.3% | 0.3% | 0.3% |
| | | Clayton | 0.2% | 0.3% | 0.2% | 0.2% |
| | | Gumbel | 0.2% | 0.2% | 0.2% | 0.2% |
| | | Frank | 0.2% | 0.2% | 0.2% | 0.2% |
| 4h (median $P_i$=8, $\rho = 0.65$) | 3h (median $P_i$=17.5, $\rho = 0.65$) | Independent | 8.6% | 8.1% | 7.5% | 7.1% |
| | | Clayton | 9.0% | 8.2% | 7.9% | 7.2% |
| | | Gumbel | 8.6% | 8.9% | 8.1% | 7.9% |
| | | Frank | 10.1% | 10.4% | 10.2% | 9.5% |

Table C4: Type-I error when misspecifying $\rho$ under the homogeneous null

| Scenario (calibration) | True scenario | family | $\alpha_0 = 0.25$ | $\alpha_0 = 0.5$ | $\alpha_0 = 0.75$ | $\alpha_0 = 1$ |
|---|---|---|---|---|---|---|
| 1h (median $P_i$=17.5, $\rho$ = 0.35) | 3h (median $P_i$=17.5, $\rho$ = 0.65) | Clayton | 2.8% | 2.6% | 2.6% | 2.7% |
| | | Gumbel | 2.4% | 2.6% | 2.3% | 2.4% |
| | | Frank | 4.3% | 5.0% | 4.4% | 3.7% |
| 2h (median $P_i$=8, $\rho$ = 0.35) | 4h (median $P_i$=8, $\rho$ = 0.65) | Clayton | 2.9% | 2.5% | 2.4% | 2.4% |
| | | Gumbel | 3.0% | 2.7% | 2.5% | 2.4% |
| | | Frank | 3.4% | 3.3% | 3.0% | 3.1% |
| 3h (median $P_i$=17.5, $\rho$ = 0.65) | 1h (median $P_i$=17.5, $\rho$ = 0.35) | Clayton | 2.0% | 2.4% | 2.3% | 2.1% |
| | | Gumbel | 2.6% | 2.4% | 2.7% | 2.7% |
| | | Frank | 1.2% | 1.4% | 1.2% | 1.4% |
| 4h (median $P_i$=8, $\rho$ = 0.65) | 2h (median $P_i$=8, $\rho$ = 0.35) | Clayton | 2.0% | 2.2% | 2.5% | 2.4% |
| | | Gumbel | 1.9% | 2.2% | 2.5% | 2.7% |
| | | Frank | 1.9% | 1.6% | 2.1% | 1.9% |

## C.4 Concluding summary of study III

The single misspecification of factors (a)-(c) explained in an opening sentence of this section was largely related to type-I error rate control of our proposed method. Although misspecification of $\rho$ was problematic only in Frank switching mechanism, multiple misspecifications may lead to severe inflation or deflation of type-I error rate. Special cautions are required especially for the misspecification of median $P_i$ (proportion of crossover) because the impact was notable.

Note that the validity of factors (a)-(c) can be guessed using a dataset in placebo arm only. True crossover mechanism can be guessed by checking the longitudinal trend of crossover. For example, one can guess that crossover mechanism is close to the Clayton type if crossover occurred earlier in the study duration (see Section 5). Median $P_i$ and $\rho$ can be guessed from observed proportion of crossover. A plausible solution for the misspecification issue would be to predetermine various optimal sets of $(\eta^*,\xi^*)$ for various factors (a)-(c) at the planning stage, and select the value most fitted to the data during the study monitoring or just before primary statistical analysis. This process should be performed by independent statisticians using actual dataset of placebo arm only so that study integrity is ensured. The performance of this adaptive determination strategy is task to be tackled in the future when considering the application of our proposed method.