Supplementary Material for Optimizing Primary Analyses in Randomized Controlled Trials with Multiple Endpoints: A Simulation Study with Application to Kidney Transplantation

Felix Herkner[1,2], Martin Posch[2], Gregor Bond[1], Franz König[2,*]

[1] Division of Nephrology and Dialysis, Department of Medicine III, Medical University of Vienna, Vienna, Austria

[2] Center for Medical Data Science, Medical University of Vienna, Vienna, Austria

[*] franz.koenig@meduniwien.ac.at

2025-09-11

A: Details on the data generating mechanism and performance measures

Details on the data generating mechanism

Three independent exponentially distributed variables are generated using base R's "rexp" function [6], representing latent time to death (D_{orig}), time to graft loss (G_{orig}) and time to first infection ($I_{1,orig}$). The distributions use parameters λ_d, λ_g and λ_i, respectively, i.e. $D_{orig} \sim Exp(\lambda_d)$, $G_{orig} \sim Exp(\lambda_g)$ and $I_{1,orig} \sim Exp(\lambda_i)$.

Participants are observed for a fixed follow-up time, s. The time-to-event (TTE) composite endpoint is then defined as time to the first of the events to be experienced by a participant, if it is observed within follow-up: $TTE_{composite} = min[D_{orig}, G_{orig}, I_{1,orig}, s]$. If participants experience the event within their follow-up (i.e. $TTE_{composite} < s$) they are marked as event and censored otherwise. A composite binary endpoint is defined by

$$Bin_{composite} = \begin{cases} 1 & \text{if } TTE_{composite} < s \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Note that no dropouts are simulated.

The time-to-event for the endpoint death is leaving D_{orig} unchanged (administrative censoring still applies; $D = min[D_{orig}, s]$). As people cannot loose their graft or be infected after dying, the two other endpoints are defined $G = min[G_{orig}, D_{orig}, s]$ and $I = min[I_{1,orig}, D_{orig}, s]$. The status variable indicating events for G is then defined

$$Status_G = \begin{cases} 1 & \text{if } G = G_{orig} \\ 2 & \text{if } G = D_{orig} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

and likewise for $Status_I$. The distinction of events allows for flexible handling of deaths occurring before one of the other events (e.g. censoring at time of death, or applying other competing risk methodology).

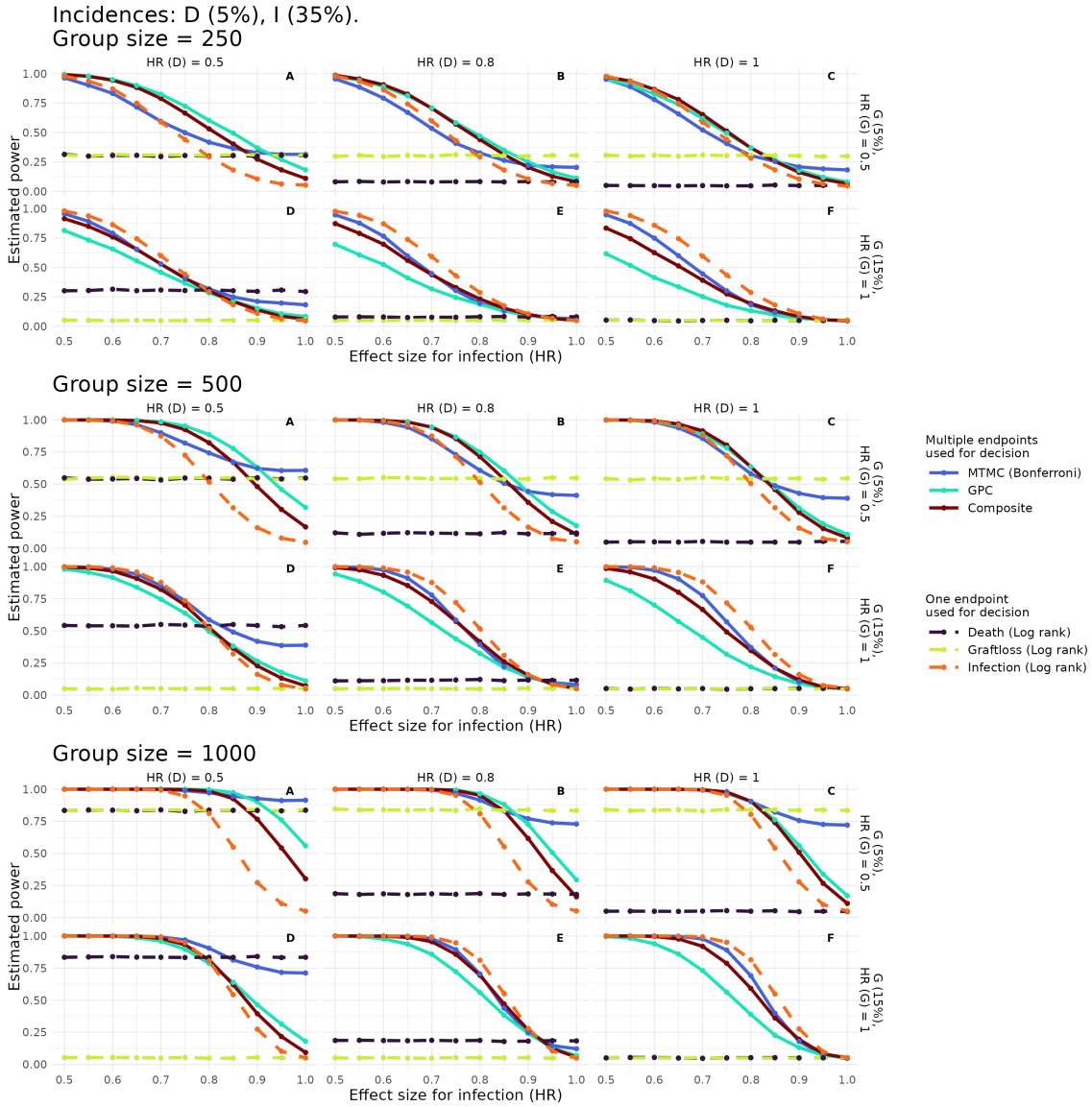
Times-to-rejection in Section "Motivating example revisited: Examples of two simulated studies" are generated in the same way as the other endpoints ($R_{orig} \sim Exp(\lambda_r)$ and accordingly included in the composite endpoint definitions. Observable time-to-event data for rejections take into account that rejections are both censored by death and graft loss: $R = min[R_{orig}, G_{orig}, D_{orig}, s]$ and the status variable extends Equation (2) to "1" marking rejections, "2" graft losses and "3" deaths, "0" otherwise.

Performance measures

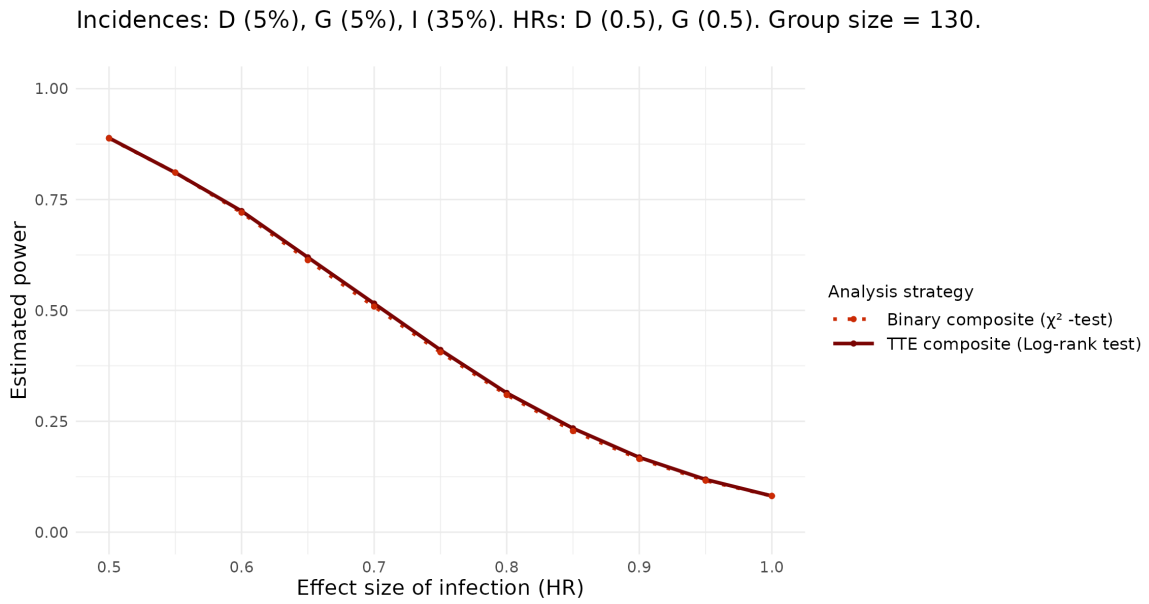
The main goal of the study is to estimate the probability of rightfully rejecting the global null hypothesis if there is a treatment effect on at least one of the endpoints. If treatment does not affect any of the endpoints, the rejection probability corresponds to the type-1-error rate. This quantity for each of the strategies in each scenario is estimated by the proportion, $\hat{p}$, of simulation repetitions in which the overall null hypothesis is rejected and therefore the decision is made to declare a significant difference between the study arms. Estimates from simulation studies are subject to uncertainty despite the great number of repetitions, and suitable measures of uncertainty should thus be reported [4]. This is commonly done by computing the Monte Carlo Standard Error (MCSE), that is in this case the standard error of the estimate of statistical power over all simulation repetitions in one scenario. Using a normal approximation of a binomial distribution, the MCSE of the power estimates are calculated as $MCSE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where n is the number of simulation repetitions. 10000 simulation repetitions result in standard errors of 0.005 and 0.004 for $\hat{p} = 0.5$ (where the standard error is maximal) and $\hat{p} = 0.8$, respectively. In the following presentation of results, corresponding confidence intervals in the graphs are approximately the same as the line width and are left out for reasons of clearer display.

B: Supplementary figures

An extension of Fig. 4 to larger group sizes is given in Supplementary Fig. 1. Additional results showing the comparison of power of a χ^2-test of a binary composite endpoint to the log-rank test of a time-to-event composite is shown in Supplementary Fig. 2.



Supplementary Figure 1: Estimated power of the approaches using time-to-event endpoint definitions in various scenarios including different groups sizes. The plot grids under each headline are similar to Fig. 4, only the group sizes are as indicated in the respective headline. From top to bottom, the grid of plots show scenarios with a group size of 250, 500 and 1000 in each study arm, respectively. In the first row of each plot grid, the expected proportion of graft losses in the control group amounts to 5% and a marked treatment effect on graft loss (hazard ratio (HR) = 0.5) is present. In the second row of each plot grid, the expected proportion of graft losses is 15% and there is no treatment effect on graft loss (HR = 1). Parameters fixed in all of the shown scenarios are the expected proportions of deaths and infections in the control group (5% and 35%, respectively). The x-axis of plots depicts the HR of infections (0.5, large difference, to 1, i.e., equal hazards for infection in both groups). On the y-axis the estimated power is plotted. The HR of death increases within a row from left to right taking values 0.5, 0.8, and 1. Solid lines identify procedures taking all endpoints into account; Bonferroni correction (blue), Composite endpoint (darkred), GPC (turquoise). Dashed lines indicate that tests are performed on one single endpoint without multiplicity correction; log-rank tests for differences of hazards of graft loss (light green), infections (orange), and death (black).



Supplementary Figure 2: Estimated power of two composite endpoint definitions in a scenario only varying effect sizes of infection. The expected proportion of deaths, graft losses and infections in the control group is fixed to 5%, 5% and 35%, respectively. A marked treatment effect on graft loss and death (hazard ratio (HR) = 0.5) is present. The x-axis of plots depicts the HR of infections (0.5, a strong treatment effect, to 1, i.e., no effect, left to right). On the y-axis the estimated power is plotted. The solid line identifies the Composite endpoint defined as time-to-first-event (tested via log-rank-tests, dark red), the dashed line the binary composite endpoint (tested via χ^2-tests, red).

C: Details on the hypothetical scenarios

Details of the scenario parameters used in Section "Motivating example revisited: Examples of two simulated studies" are provided. In the first hypothetical example, the expected proportions of death (D), graft loss (G) and rejections (R) in the control group are set to 5% and infections (I) 35%, the (time-constant) hazard ratios (HR) between the treatment groups are chosen to be 1 (no difference between groups) for D, and 0.5 for G, R and I (corresponding to a fairly strong treatment effect between groups).

For the second example, the expected proportion of D in the control group is set to be 15%, while the respective proportions of G, R and I are the same as previously (5%, 5% and 35%, respectively). The HRs are now set to 1 (D), 0.7 (G), 1 (R) and 0.5 (I).

Note that while the follow-up time for each participant is set to $s = 10$ time units in the simulation study, a follow-up time of $s = 9$ is used in the examples. This only influences the choice of the parameters of exponential distributions because these are find to reach the desired incidences (see Supplement Section A).

D: Simulation parameters

All combinations of parameters given in Supplementary Table 1 were investigated except hazard ratios (HR), where only HRs 0.5, 0.7, 0.9, and 1 were considered for combination with all other parameters, resulting in 512 scenarios. Additional 259 scenarios with HRs for death and infections in between, i.e., 0.55, ..., 0.65, 0.75, ..., 0.85, 0.95 were run to provide for more granularity in plots.

Furthermore, 1113 scenarios for group sizes of 250, 500 and 1000 were investigated (see Supplementary Fig. 5-8).

The weighted Bonferroni correction is simulated separately and every combination of weights applied to the endpoints in selected scenarios. Selected scenarios are expected proportions of death and graft loss of 5% and infections 35% in the control group (which is the setting expected to closest reflect real study conditions in kidney transplant immunosuppression studies following up patients in the first year after transplantation) and every combination HRs 0.5, 0.7 and 1 for each of the endpoints (resulting in 27 scenarios). An additional 16 scenarios were again run to provide smoother plots. For each of these scenarios, every possible combination of weights is applied as specified in Supplementary Table 1, see the Plot grids in Supplementary Fig. 4. This must take into account that the weights must sum to one. This yields 66 possible combinations of weights in every single scenario. In the additional scenarios, not all weights were applied.

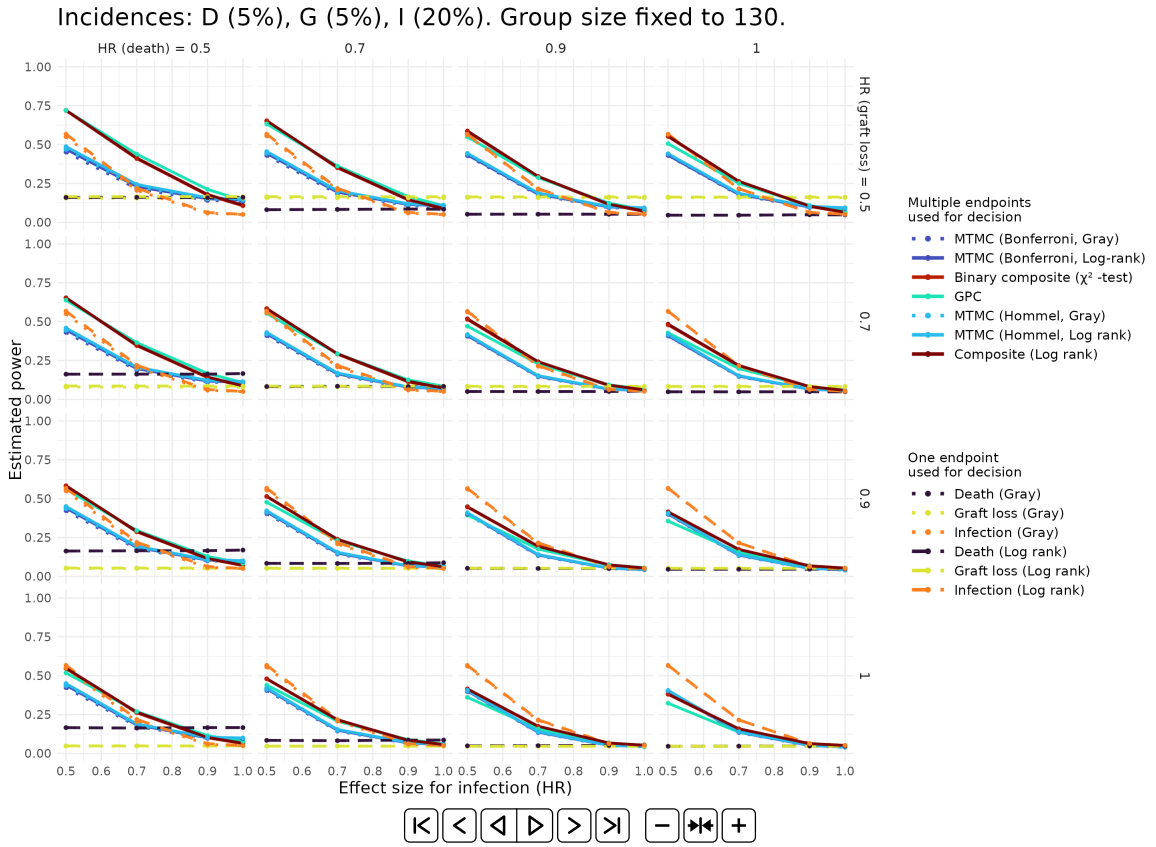
Supplementary Table 1: Important parameters of the simulation and their specification. Following the terminology introduced by Benda et al. [1] and Friede and colleagues [2], disease specific features may be estimable from preliminary data (like incidences of endpoints) or must be assumed entirely (like underlying patterns of missing data). DSF (e) indicates Disease Specific Features which are estimable, and DSF (a) ones that have to be assumed. Design choices (DC) might be constrained (DC (c)) due to, e.g., external factors. MCSE = Monte Carlo standard error, CE = Composite endpoint, MTMC = Multiple testing and multiplicity correction, GPC = Generalized pairwise comparisons.

Parameter	Type	Investigated values	Description
Expected proportion of death within follow-up	DSF (e)	5%, 15% (control group)	Expected proportion of deaths in the control group within observation period
Expected proportion of graft loss within follow-up	DSF (e)	5%, 15% (control group)	Expected proportion of graft losses in the control group within observation period taking into account censoring by death
Expected proportion of infections within follow-up	DSF (e)	20%, 35% (control group)	Expected proportion of infections in the control group within observation period taking into account censoring by death
Hazard ratios (HR)	DSF (e)	0.5, 0.55, ...,0.95, 1	Ranging from treatment being considered strongly favourable (HR = 0.5) to no effect (HR = 1) for each endpoint separately
Bonferroni weights	DC	0, 0.1, 0.2, ..., 0.9, 1	Weights to be used for each endpoint, if weights are used to split the alpha level
Group size	DC (c)	130, 250, 500, 1000	Per group
Individual follow-up time	DC	10 months	Fixed follow-up per participant
Simulation repetitions	DC	10.000	Number of simulation repetitions. Calculated to meet a reasonable MCSE for power estimates of ¡ 0.005 or 0.5%
Statistical significance	DC	5%	Two-sided tests
Type of endpoints	DC	Time-to-event, binary	Binary endpoints only reported for the composite endpoint
Testing strategies and endpoint definitions	DC	(i) CE	Time-to-event (log rank test) and binary (χ^2 test)
		(ii) MTMC	Time-to-event (log rank tests, Gray's test)
		(iii) GPC	Time-to-event
Hypothesis tests	DC	Log rank test	Implemented in the *survival* R package [7]
		Gray's test	Implemented in the *cmprsk* R package [3].
		GPC	Implemented in the *BuyseTest* R package [5]

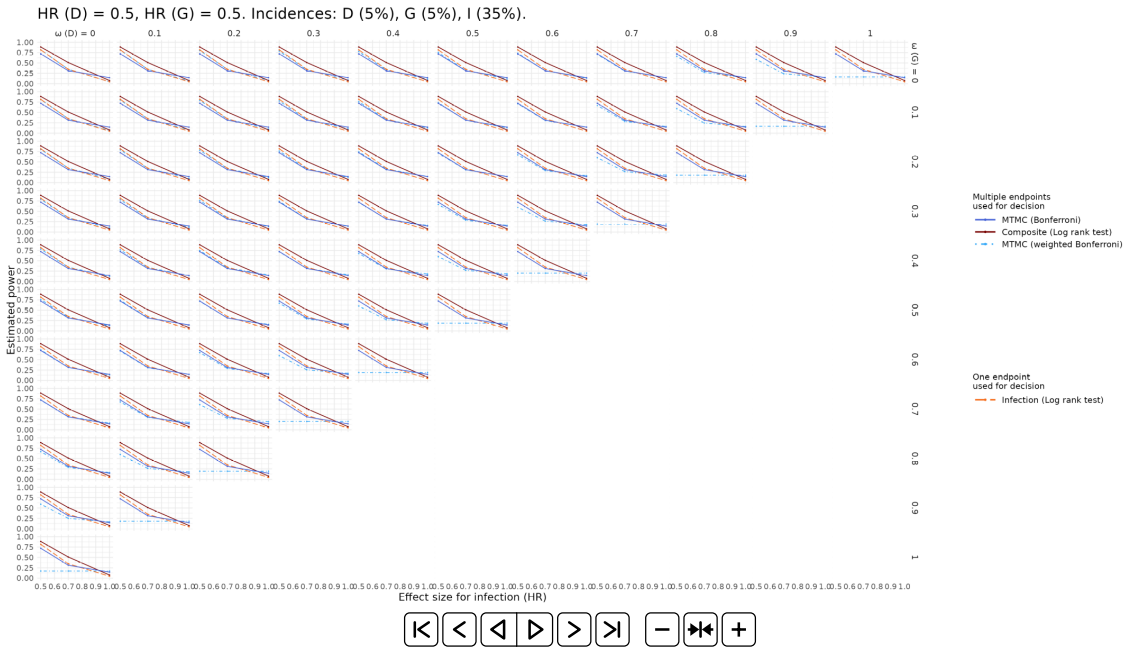
E: Plots of all scenarios evaluated

The following slide shows show plots for all the scenarios evaluated. For some of the plots, additional HRs of infections were simulated to provide for more granulated plots. These are not included here but would be seen as more granular measurements in the plots shown. The comparisons of strategies on time-to-event endpoint definitions is shown in Supplementary Fig. 3. Contrasting weighted and unweighted Bonferroni multiplicity correction in all investigated scenarios is shown in Supplementary Fig. 4. In these scenarios, the group size is always fixed to 130 participants per group. The results of simulations with larger group sizes are shown in Supplementary Fig. 5-S8.

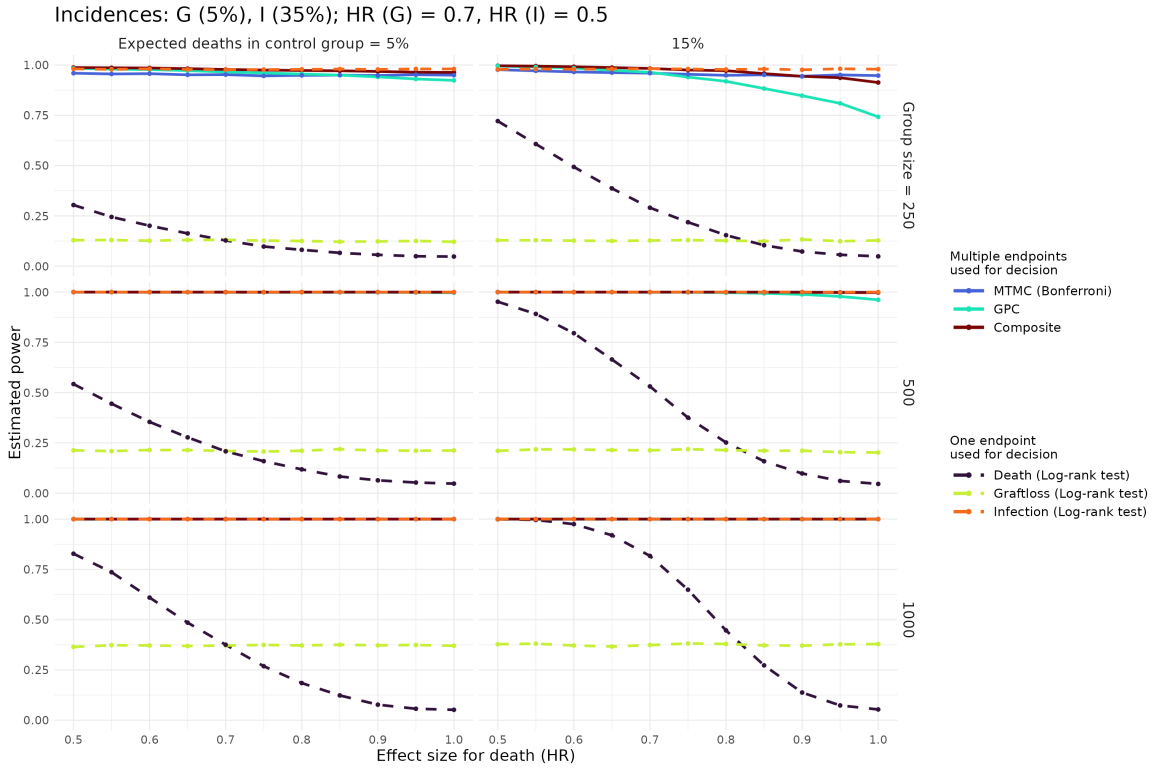
If the slide shows do not work, please try to switch to another PDF reader (tested for Adobe Acrobat).



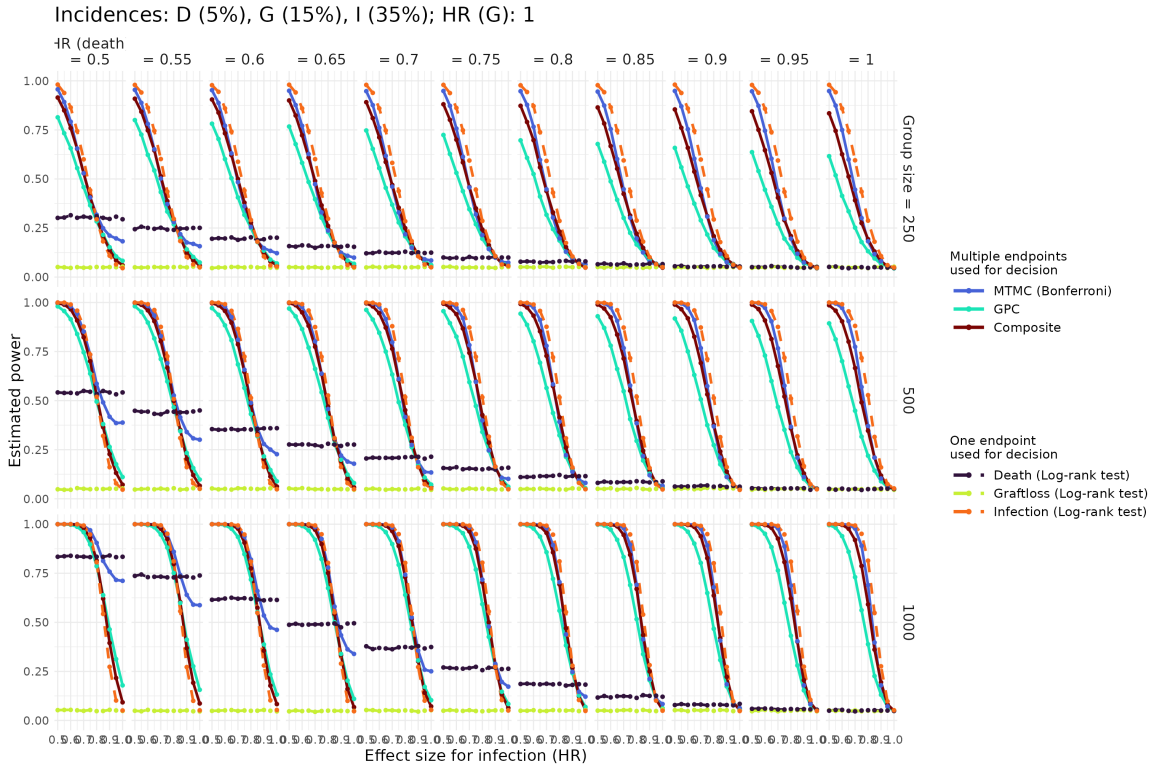
Supplementary Figure 3: Estimated power of the approaches using time-to-event endpoint definitions in various scenarios. The titles indicate the parameters not explicitly shown in the plots. These are the expected proportions of deaths (D), graft loss (G) and infections (I) in the control group. The group size is fixed to 130 participants. The x-axis of plots depicts the HR of infections (0.5, large difference, to 1, i.e., equal hazards for infection in both groups). On the y-axis the estimated power is plotted. The facet labels of each column give the hazard ratio (HR) of death, the rows the HR of graft loss. The HR of death increases within a row from left to right taking values 0.5, 0.7, 0.9, and 1 and likewise for graft loss in columns. Solid lines identify procedures taking all endpoints into account; Bonferroni correction (blue), Composite endpoint (darkred), GPC (turquoise). Dashed lines indicate that tests are performed on one single endpoint without multiplicity correction; log-rank tests for differences of hazards of graft loss (light green), infections (orange), and death (black). Dotted lines indicate that Gray's test was used instead of log-rank tests.



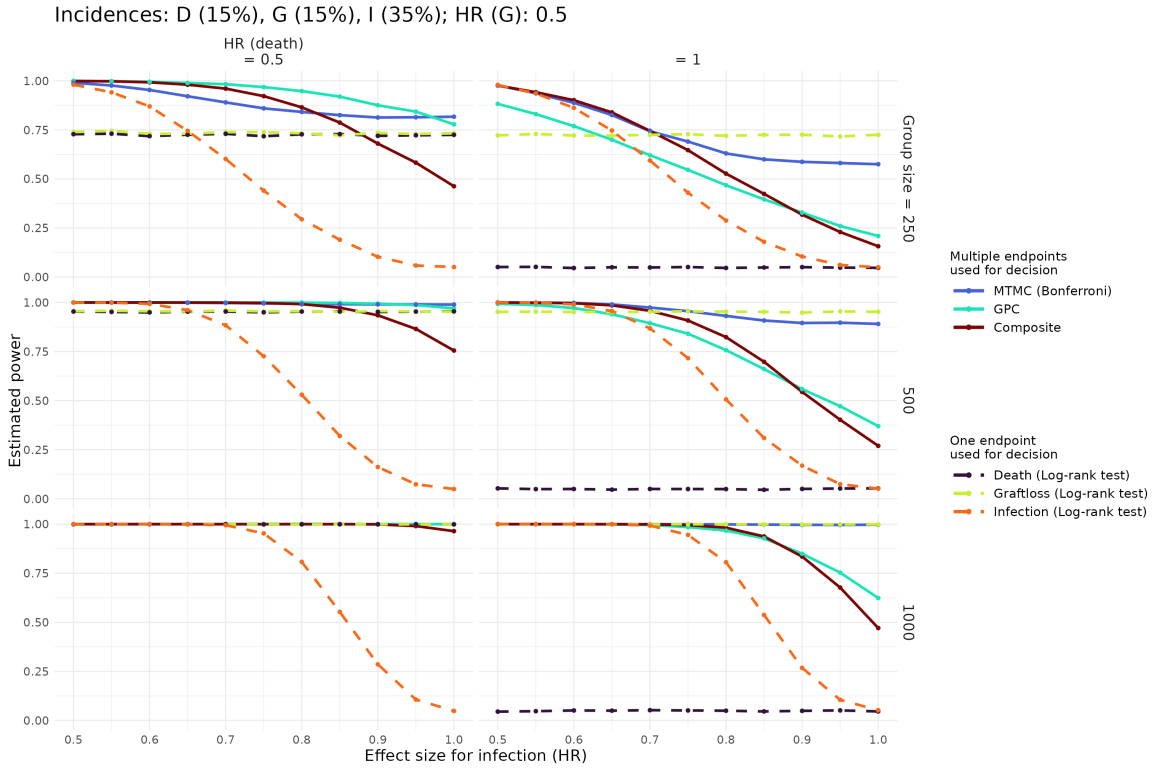
Supplementary Figure 4: Comparison of estimated power of a weighted Bonferroni procedure and aforementioned approaches in various scenarios. Time-to-event endpoint definitions are again used and differences between groups tested by log-rank tests. The expected proportion of deaths (D), graft losses (G) and infections (I) in the control group is fixed to 5%, 5% and 35%, respectively, in all shown plots. Hazard ratios (HR) of D and G vary in each plot and are given in the title of each figure. The x-axis of plots depicts the HR of infections (0.5, a strong treatment effect, to 1, i.e., no effect). On the y-axis the estimated power is plotted. The weights, ω, applied to adjust the tests p-values varies and is given in the column (weight of the test of D) and row (weight of the test of G) facet labels . Solid lines identify procedures taking all endpoints into account; Multiple testing and multiplicity correction (MTMC) using log-rank tests and Bonferroni correction (blue), Composite endpoint (darkred). Dashed lines indicate log-rank tests of infections (orange). Dashed-dotted lines are the weighted version of the Bonferroni correction (blue). Note that as the weights must sum to one, the weight of the tests of I follow from the other two weights. For the same reason, combinations of weights in the right-lower triangle do not exist and the plots remain empty.



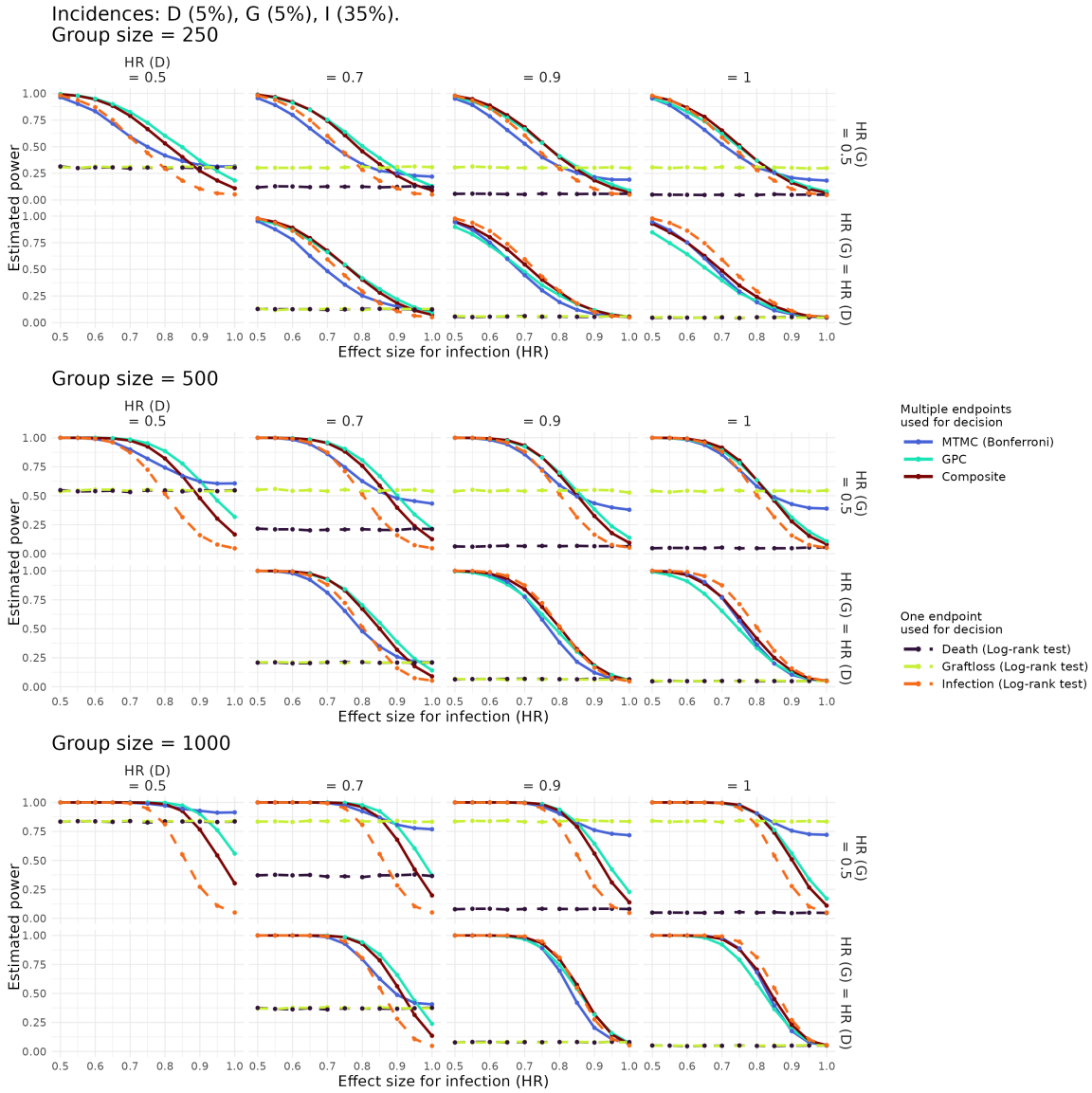
Supplementary Figure 5: Estimated power of the approaches using time-to-event endpoint definitions in various scenarios including different group sizes. The titles list the parameters not explicitly shown in the plots. Incidences in the title mean the expected proportion of observed deaths (D), graft loss (G) and infections (I) in the control group. The x-axis depicts the hazard ratio (HR) of death (0.5, large difference, to 1, i.e., equal hazards for infection in both groups). On the y-axis the estimated power is plotted. The left column of plots are scenarios with incidences of death of 5%, in the right column the incidence of death is 15%. The group sizes increase by rows from 250 to 500 and 1000 participants per group. Solid lines identify procedures taking all endpoints into account; Bonferroni correction (blue), Composite endpoint (darkred), GPC (turquoise). Dashed lines indicate that tests are performed on one single endpoint without multiplicity correction; log-rank tests for differences of hazards of graft loss (light green), infections (orange), and death (black).



Supplementary Figure 6: Estimated power of the approaches using time-to-event endpoint definitions in various scenarios including different group sizes. The titles list the parameters not explicitly shown in the plots. Incidences in the title mean the expected proportion of observed deaths (D), graft loss (G) and infections (I) in the control group. The x-axis depicts the hazard ratio (HR) of infections (0.5, large difference, to 1, i.e., equal hazards for infection in both groups). On the y-axis the estimated power is plotted. Hazard ratios of death vary by column and group sizes by row. Solid lines identify procedures taking all endpoints into account; Bonferroni correction (blue), Composite endpoint (darkred), GPC (turquoise). Dashed lines indicate that tests are performed on one single endpoint without multiplicity correction; log-rank tests for differences of hazards of graft loss (light green), infections (orange), and death (black).



Supplementary Figure 7: Estimated power of the approaches using time-to-event endpoint definitions in various scenarios including different group sizes. The titles list the parameters not explicitly shown in the plots. Incidences in the title mean the expected proportion of observed deaths (D), graft loss (G) and infections (I) in the control group. The x-axis depicts the hazard ratio (HR) of infections (0.5, large difference, to 1, i.e., equal hazards for infection in both groups). On the y-axis the estimated power is plotted. Hazard ratios of death vary by column (left column 0.5, right column 1) and group sizes by row (250, 500 and 1000 participants per group). Solid lines identify procedures taking all endpoints into account; Bonferroni correction (blue), Composite endpoint (darkred), GPC (turquoise). Dashed lines indicate that tests are performed on one single endpoint without multiplicity correction; log-rank tests for differences of hazards of graft loss (light green), infections (orange), and death (black).



Supplementary Figure 8: Estimated power of the approaches using time-to-event endpoint definitions in various scenarios including different group sizes. The first title list the parameters not explicitly shown in the plots but constant in all shown scenarios. Incidences in the title mean the expected proportion of observed deaths (D), graft loss (G) and infections (I) in the control group. Three plot grids are shwon for Group sizes of 250 (top grid), 500 (middle grid) and 1000 (bottom grid) participants per group. The x-axis depicts the hazard ratio (HR) of infections (0.5, large difference, to 1, i.e., equal hazards for infection in both groups). On the y-axis the estimated power is plotted. Hazard ratios of death vary by column (again from 0.5, a strong treatment effect on deaths, to 1, treatment does not affect the event rate of death). The HR of graft loss varies by row; in the first row of plot grids, the HR of graft loss is 0.5. In the second row of plots, the HR of graft loss is equal to the HR of death in each plot (the first plot is empty, as the HRs of 0.5 are shwon in the first plot of the first row). Solid lines identify procedures taking all endpoints into account; Bonferroni correction (blue), Composite endpoint (darkred), GPC (turquoise). Dashed lines indicate that tests are performed on one single endpoint without multiplicity correction; log-rank tests for differences of hazards of graft loss (light green), infections (orange), and death (black).

F: Supplementary material for the "Hypothetical scenarios"

The cumulative incidence functions (CIF) plots of the individual endpoints in the two hypothetical examples in Section "Motivating example revisited: Examples of two simulated studies" are given in Supplementary Fig. 9.

Extending Table 2 in the paper, in the following Supplementary Table 2 also the p-values of testing a binary composite endpoint is reported. The rest of the table is the same as Table 2 in the paper.

Supplementary Table 2: Overview of the analysis of simulated studies chosen from scenarios 1 and 2, respectively, including a χ^2 test of a binary composite endpoint.

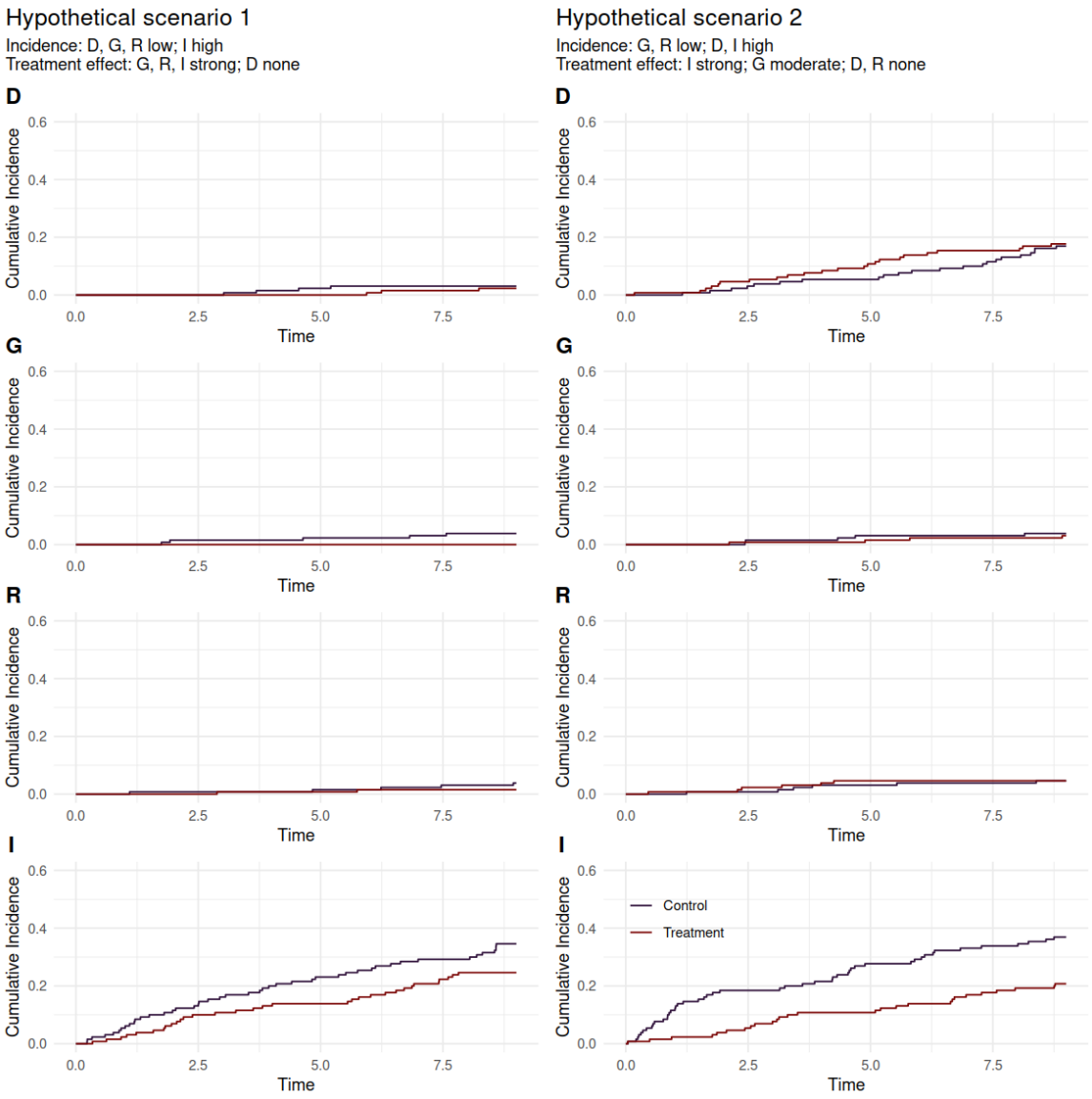
	Hypothetical scenario 1 *Incidence: D, G, R low; I high Treatment effect: G, R, I strong; D none*			**Hypothetical scenario 2** *Incidence: G, R low; D, I high Treatment effect: I strong; G moderate; D, R none*		
	Control	Treatment	p-value[3]	Control	Treatment	p-value[3]
Death *n (%)*	4 (3.1%)	3 (2.3%)	0.6891	22 (16.9%)	23 (17.7%)	0.7891
Graftloss *n (%)*	5 (3.8%)	0 (0%)	**0.0234**	5 (3.8%)	4 (3.1%)	0.7687
Rejection *n (%)*	5 (3.8%)	2 (1.5%)	0.2338	6 (4.6%)	6 (4.6%)	0.9729
Infection *n (%)*	45 (34.6%)	32 (24.6%)	0.0704	48 (36.9%)	27 (20.8%)	**0.0028**
Endpoints considered: D, G, R						
Composite *RMST*[1]	8.56	8.88	**0.0488**	8.07	7.83	0.9963
BIN CE *n (%)*	13 (10%)	5 (3.8%)	0.0506	32 (24.6%)	31 (23.8%)	0.8849
MTMC	-	-	0.0703	-	-	1
GPC	e% / f% / u%[2]		**0.0495**	e% / f% / u%[2]		0.9248
Death	100% / 3.1% / 2.2%			100% / 14.9% / 16.7%		
Graft loss	94.7% / 3.8% / 0%			68.4% / 3.1% / 2.5%		
Rejection	90.9% / 3% / 1.4%			62.8% / 3% / 2.4%		
Ties left	86.54%			57.41%		
Endpoints considered: D, G, I						
Composite *RMST*[1]	6.99	7.75	**0.0216**	6.24	7.21	**0.0114**
BIN CE *n (%)*	52 (40%)	35 (26.9%)	**0.0255**	69 (53.1%)	50 (38.5%)	**0.018**
MTMC	-	-	0.0703	-	-	**0.0085**
GPC	e% / f% / u%[2]		**0.0165**	e% / f% / u%[2]		0.0563
Death	100% / 3.1% / 2.2%			100% / 14.9% / 16.7%		
Graft loss	94.7% / 3.8% / 0%			68.4% / 3.1% / 2.5%		
Infection	90.9% / 28.4% / 18.6%			62.8% / 23.8% / 10.1%		
Ties left	43.85%			28.88%		
Endpoints considered: D, G, R, I						
Composite *RMST*[1]	6.89	7.68	**0.0163**	6.14	7.02	**0.0189**
BIN CE *n (%)*	55 (42.3%)	37 (28.5%)	**0.0196**	71 (54.6%)	53 (40.8%)	**0.0254**
MTMC	-	-	0.0937	-	-	**0.0113**
GPC[2]	e% / f% / u% [2]		**0.0111**	e% / f% / u%[2]		0.0806
Death	100% / 3.1% / 2.2%			100% / 14.9% / 16.7%		
Graft loss	94.7% / 3.8% / 0%			68.4% / 3.1% / 2.5%		
Rejection	90.9% / 3% / 1.4%			62.8% / 3% / 2.4%		
Infection	86.5% / 27.3% / 18%			57.4% / 21.4% / 9.2%		
Ties left	41.27%			26.88%		

[1] RMST = Restricted mean event-free survival time, restricted to 9 time units

[2] GPC pairs of the respective endpoint: e%: pairs evaluated / f%: favourable pairs / u%: unfavourable pairs. Percentages are fractions of ALL pairs.

[3] The first four p-values are calculated from log-rank tests of the individual endpoints. The composite time-to-event endpoint and the BIN CE are tested using a log-rank test and a χ^2 test, respectively. For MTMC, the smallest Bonferroni-adjusted p-value is reported. The overall p-value of the GPC is reported. Significant p-values (< 0.05) are bold.

BIN CE = Binary composite endpoint, D = Death, G = Graft loss, R = Rejections, I = Infections, MTMC = Multiple testing and multiplicity correction, GPC = Generalized pairwise comparisons



Supplementary Figure 9: Cumulative incidence functions (CIF) of individual endpoints in hypothetical scenario 1 (left column) and 2 (right column). The x-axis depicts the time on study, the y-axis the cumulative incidence. Note that the y-axis is limited to 0.6. The endpoints are death (D, first row), graft loss (G, second row), rejections (R) and infections (I). In blue, the CIF of the respective endpoint in the control group is given, in red the CIF of the treatment arm. The group sizes are 130 in both scenarios, the follow-up time is 9 time units. Scenario 1: underlying expected proportions in the control group for D, G and R are set to 5%, I 35%, and hazard ratios (HR) between groups set to 1 (D) and 0.5 (G, R and I). Scenario 2: expected proportions in the control group are set to 15%, 5%, 5% and 35%, and HRs set to 1, 0.7, 1, 0.5 for D, G, R and I, respectively.

G: Deriving simulation parameters

Expected proportions of events in the control group can be found using properties of the underlying distributions as follows. As death in the simplest setting is unaffected by the other endpoints, the expected proportion of death in the control group is

$$\mathbb{P}(D_{orig} < s) = 1 - e^{-\lambda_d s} \tag{3}$$

It is a well-known result that the minimum of n exponentially distributed variables with parameters $\lambda_1, ..., \lambda_n$ is again exponential with parameter $\lambda = \sum_{i=1}^{n} \lambda_i$. The expected proportion of graft loss in the control group within follow-up is then

$$\begin{aligned} \mathbb{P}(min(G_{orig}, D_{orig}, s) = G_{orig}) &= \mathbb{P}(min(G_{orig}, D_{orig}) \leq s, min(G_{orig}, D_{orig}) = G_{orig}) \\ &= \int_0^s f_G(x)\mathbb{P}(D_{orig} > x)dx \\ &= \lambda_g \int_0^s e^{-(\lambda_g+\lambda_d)x} dx \\ &= \frac{\lambda_g}{\lambda_g + \lambda_d}(1 - e^{-s(\lambda_g+\lambda_d)}) \end{aligned} \tag{4}$$

where we find λ_g in the last equation using a simple bisectional algorithm. In this way one can set λ_g so as to achieve a desired expected proportion of graft losses in the control group taking into account censoring by death.

In the same way, the parameters of the exponential distributions of infections and rejections can be calculated by setting the desired expected proportions in the control group.

References

[1] Norbert Benda, Michael Branson, Willi Maurer, and Tim Friede. Aspects of modernizing drug development using clinical scenario planning and evaluation. *Drug information journal: DIJ/Drug Information Association*, 44:299–315, 2010.

[2] Tim Friede, Richard Nicholas, Nigel Stallard, Susan Todd, Nicholas Parsons, Elsa Valdés-Márquez, and Jeremy Chataway. Refinement of the clinical scenario evaluation framework for assessment of competing development strategies with an application to multiple sclerosis. *Drug information journal: DIJ/Drug Information Association*, 44:713–718, 2010.

[3] Bob Gray. *cmprsk: Subdistribution Analysis of Competing Risks*, 2024. R package version 2.2-12.

[4] Tim P Morris, Ian R White, and Michael J Crowther. Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102, 2019.

[5] Brice Ozenne and Julien Peron. *BuyseTest: Implementation of the Generalized PairwiseComparisons*, 2025. R package version 3.2.0.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025.

[7] Terry M Therneau. *A Package for Survival Analysis in R*, 2024. R package version 3.8-3.