SUPPLEMENTARY MATERIALS

Supplementary Methods

Generalized Pairwise Comparisons

A full description of generalized pairwise comparisons has been published previously [1]. We restrict our discussion to the analysis of survival data from randomized trials comparing an experimental to a control group. Pairwise comparisons are carried out on all possible pairs of patients, one from the experimental group (group T) and the other from the control group (group C). Let xi and yj be the survival times of a patient i from group T ($i = 1, \dots, nT$) and of a patient *j* from group C ($j = 1, \dots, nC$). A pair is classified as a "Win" if the survival of patient *i* is larger than that of patient *j* by at least m months, i.e., $xi - yj > m$; as a "Loss" in the opposite situation, i.e., $xi - yj < -m$; and 'neutral' in all other cases, i.e. $-m < xi - yj < m$, or one or both of the survival times is censored such that the pair cannot be classified as favorable or unfavorable. A pairwise score, noted $p_{ij}(m)$, takes the value 1, -1, or 0 in each of these respective cases. The net benefit, called "proportion in favor of treatment" in the original publication [1], is defined as

$$\Delta(m) = P[xi > yj + m] - P[yj > xi + m]$$

where $P[xi > yj + m]$ and $P[yj > xi + m]$ are, respectively, the probabilities for a random pair to be a Win or a Loss.

It can be calculated as the sum of the pairwise scores over all the pairs that can be formed between one patient from the treatment group and one patient from the control group:

$$\hat{\Delta}(m) = \frac{\sum_{i=1}^{nT} \sum_{j=1}^{nC} p_{ij}(m)}{nT.nC}$$

A confidence interval for $\hat{\Delta}(m)$, and a test of statistical significance, can be computed using a randomization test. When $m = 0$, this test is exactly equivalent to the Gehan-Wilcoxon test statistic. When $m > 0$, this test can be viewed as a generalization of the Gehan-Wilcoxon test statistic. The Gehan's modification of the Mann-Whitney test has been shown to be biased in presence of censored observations [2]. Further adjustments of the test statistic $\Delta(m)$ assign a pairwise score $p_{ij}(m)$ to pairs classified uninformative because of right censoring, in order to achieve better power under proportional hazards or when the treatment effect is delayed [3]. The adjusted pairwise score $p_{ij}(m)$ can take any value between -1 and 1. This adjusted procedure has been used in this manuscript. When $m = 0$ and under proportional hazards, there is a simple relationship between the net benefit, the hazard ratio, and the proportion of informative pairs f [4]:

$$\Delta = f.\frac{1 - HR}{1 + HR}$$

Details on the simulation parameters

In the simulations, survival times have been assumed to follow a negative exponential distribution. The negative exponential distribution has a rate parameter denoted λ. In all scenarios, survival times in the control group followed an exponential distribution with $\lambda_C = 0.1$.

References

1 Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-

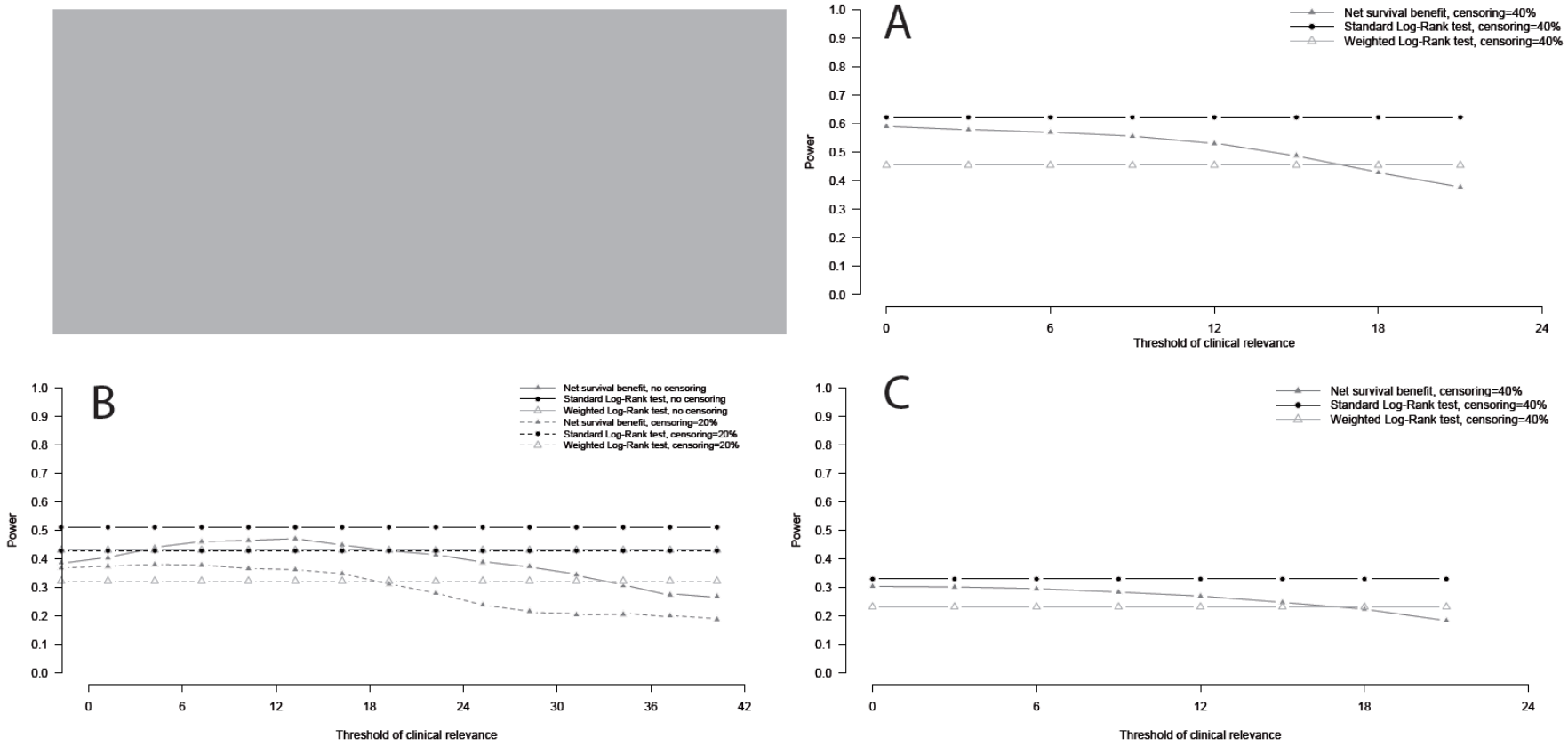
sample problem. *Stat Med* 2010; **29**: 3245–57.

2 Efron B. The two sample problem with censored data. Univ. of C. Vol. 4, Fifth Berkeley Symp. on Math. Statist. and Prob. 1967. 831-853 p.

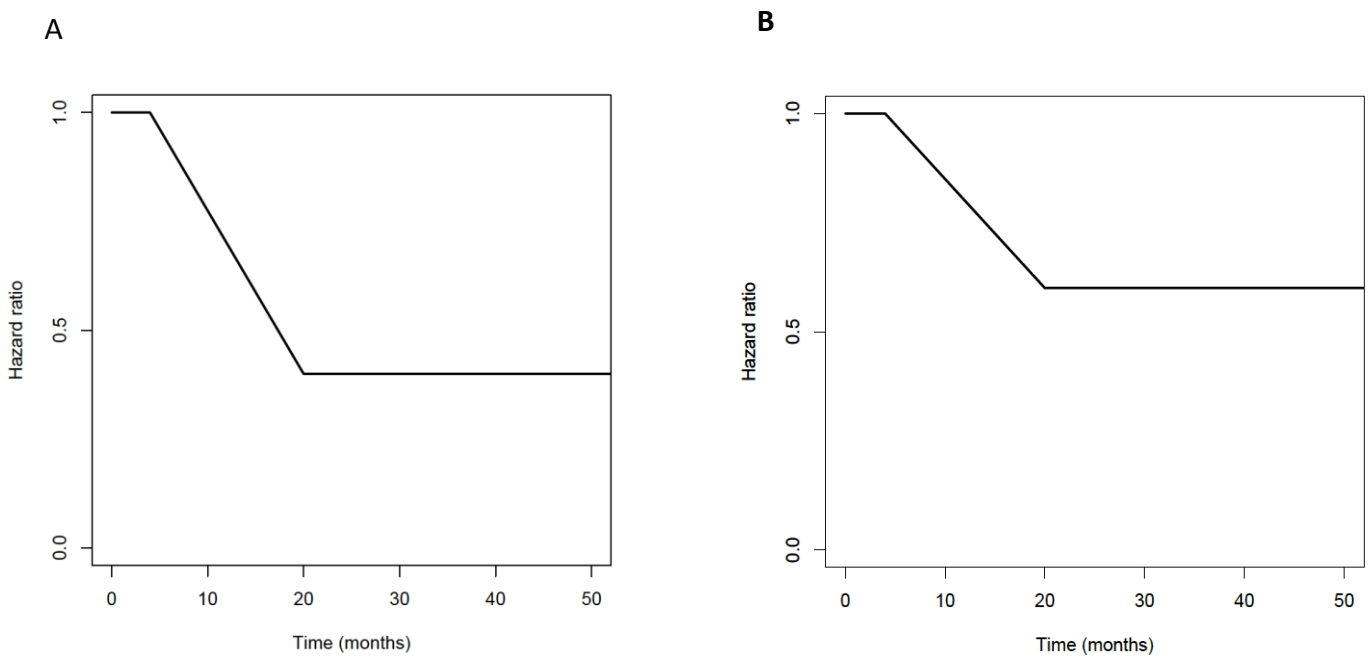
3 Péron J, Buyse M, Ozenne B, Roche L, Roy P. An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. *Stat Methods Med Res* 2016; : 96228021665832.

4 Buyse M. Reformulating the hazard ratio to enhance communication with clinical investigators. *Clin Trials* 2008; **5**: 641–2.

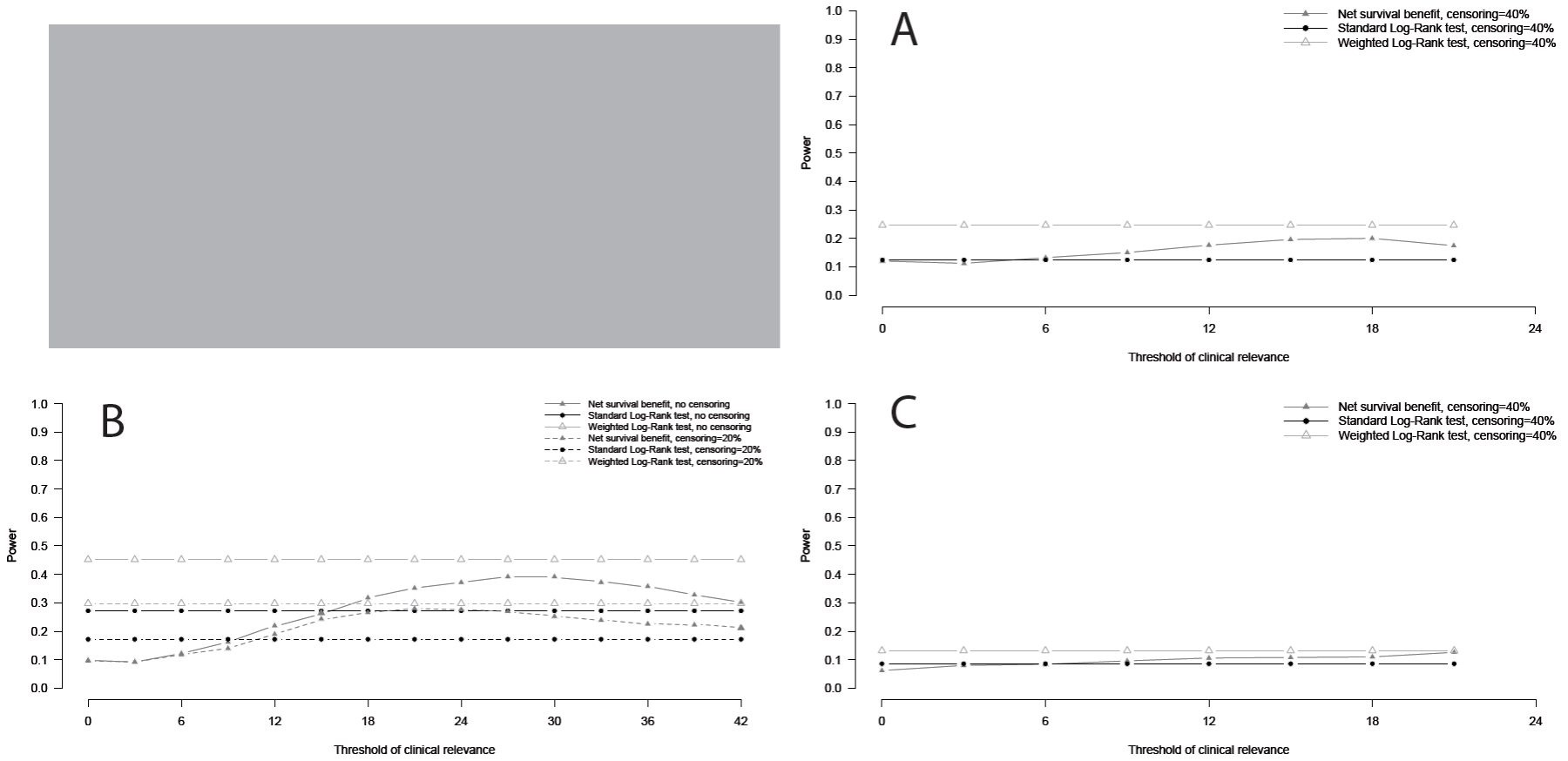
Supplementary Figures



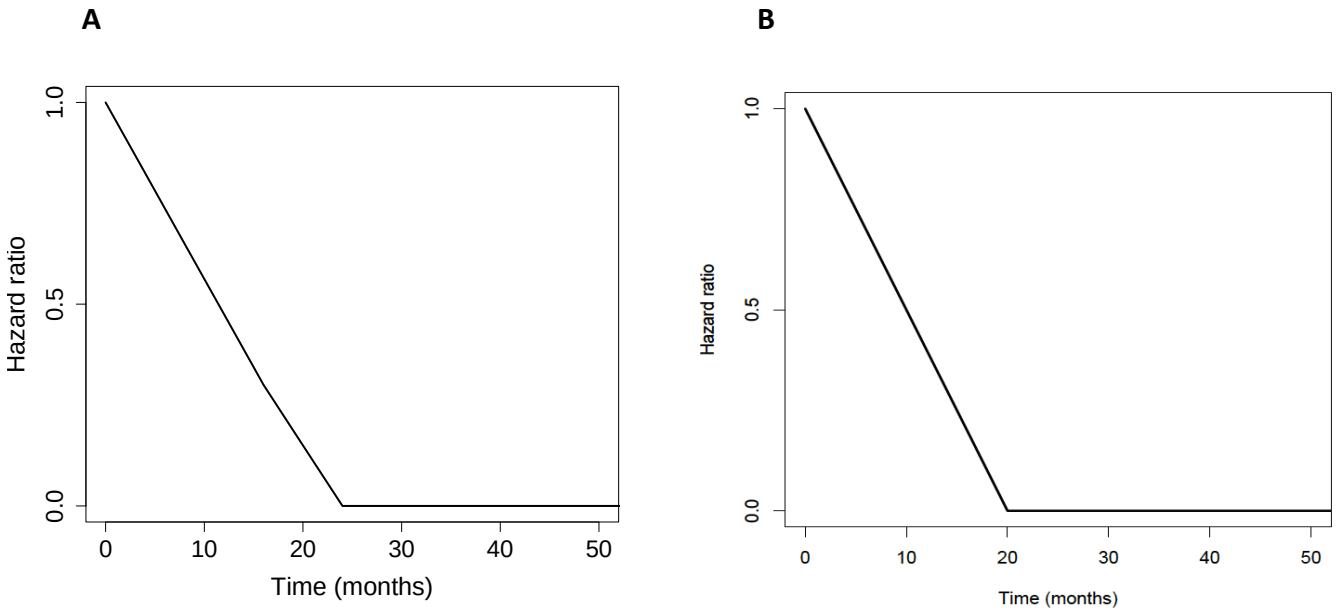
Supplementary Figure 1: **Power of the standard log-rank test, weighted log-rank test, and of the test of a net survival benefit of at least m months in a scenario of proportional hazards. A**) Hazard ratio = 0.65, and 40% of censored observations ; **B**) hazard ratio = 0.75, 0% and 20% of censored observations. **C**) hazard ratio = 0.75, 40 % of censored observations.



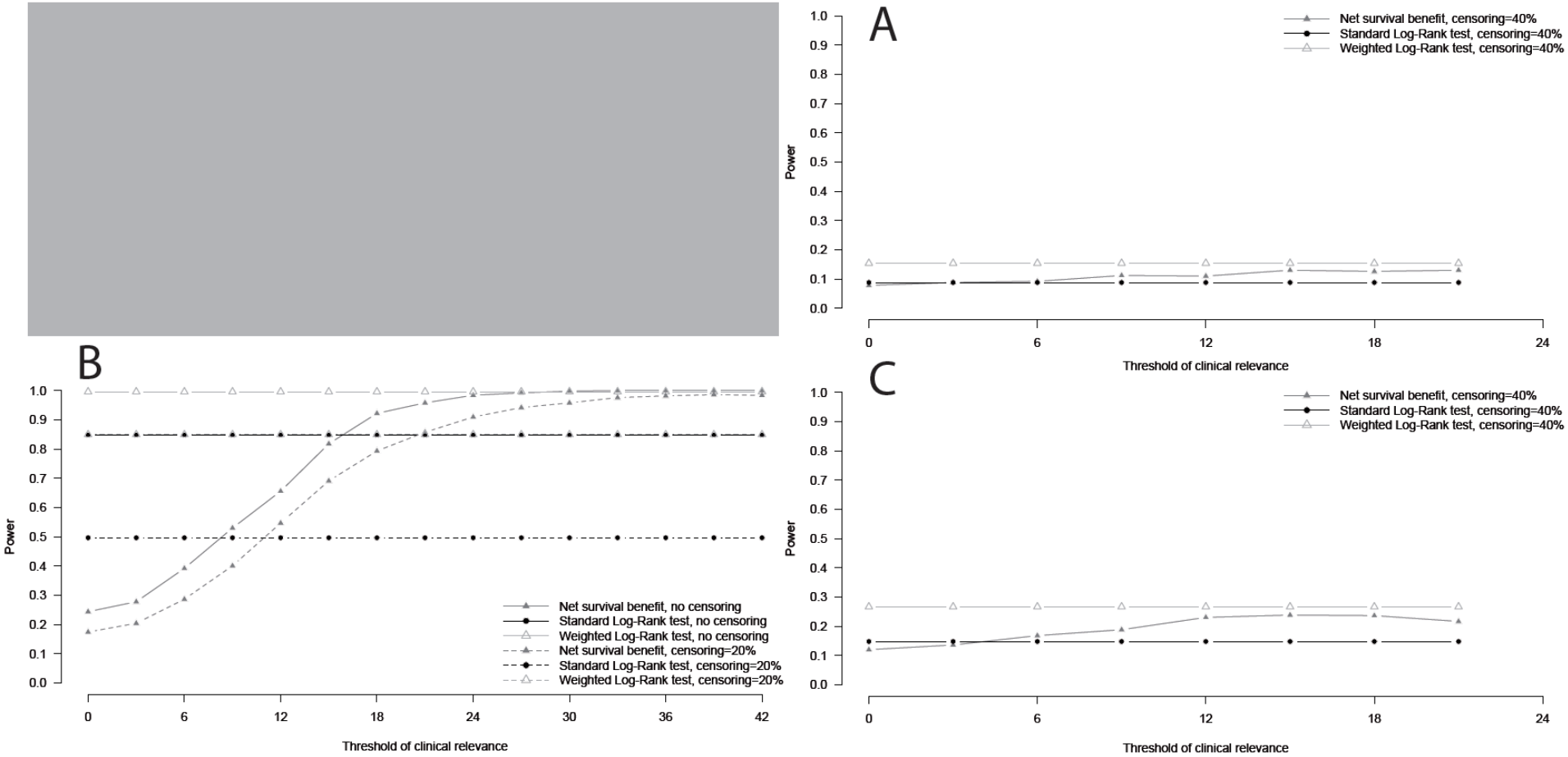
Supplementary Figure 2: Hazard ratio as a function of time, delayed survival differences.
A) Maximum hazard ratio=0.4 ; **B**) maximum hazard ratio=0.6.



Supplementary Figure 3: Power of the standard log-rank test, weighted log-rank test, and of the test of a net survival benefit of at least m months in a scenario of delayed treatment effect. A) Hazard ratio = 0.4, 40 % of censored observations; **B**) hazard ratio = 0.6, 0% and 20% of censored observations; **C)** hazard ratio = 0.6, 40% of censored observations.



Supplementary Figure 4: Hazard ratio as a function of time. A) Delayed cure rate ; **B)** quick cure rate.



Supplementary Figure 5: _Power of the standard log-rank test, weighted log-rank test, and of the test of a net survival benefit of at least m months in a scenario of: A) a delayed cure rate, 40 % of censored observations ; **B)** a quick cure rate, 0% and 20% of censored observations ; **C)** a quick cure rate, 40 % of censored observations.