MATH-449 - Biostatistics
EPFL, Spring 2025
Problem Set 6

1. A statistics student did an internship with a power company, where she was hired to analyze the time it took before cracks developed in the company's new turbine prototype. 41 turbines were observed in a testing facility for two months, where engineers had carefully recorded the time it took before noticeable cracks were discovered. The power company had installed machines that could make turbines rotate to simulate the rotation they could experience in a real-world environment.

   The following outcomes were recorded:

   - Some of the turbines developed cracks before the two months were over.
   - Some turbines were observed for the whole two months without any cracks.
   - For a significant number of turbines, the machines that enforced the rotation stopped working during the study. The power company didn't have the resources to repair these machines, making the turbines they rotated unobserved.

   The power company was interested in the time it took before cracks developed if the rotation enforcing machines did not stop. The turbines whose machines stopped were considered censored. The time it took for the machines to fail was thought to be unrelated to the time it took before cracks developed so that the observed (non-censored) turbines were representative of all turbines.

   a) Classify the above outcomes using the variables $\tilde{T}_i$ and D_i from the lectures.

      Being familiar with survival analysis, the student calculated the Kaplan-Meier curve, which estimates the survival probability as a function of t. The estimate along with approximate 95% confidence intervals (for each fixed t), $\hat{S}(t) \pm 1.96\hat{\sigma}(t)$, is plotted in Figure 1.

   b) Based on the plot, find the probability of a turbine being crack-free after 30 days, with a 95% confidence interval. Use the plot to estimate the 10th percentile of the survival times, along with a 95% confidence interval.

      The Kaplan-Meier estimator estimates the true survival probability $P(T > t)$ as long as the censoring is independent.

   c) Given the information provided thus far in this example, argue that the censoring is independent.

      Challenge: After talking with some of the engineers, the student learned that the machines provided different rotational speeds to the turbines. She reasoned that the machines that provided higher rotational speed: 1) could make cracks appear faster due to increased stress on the turbines, and 2) were more likely to stop working during the study (thus leading to censoring) due to increased stress on the machines. She learned that the machines could broadly be categorised into two groups; those that provided fast rotational speed and those that provided slow rotational speed.
      After thinking about the problem for a bit, she realised that the result from b) could provide a misleading picture of the survival probability. However, she also found that she could use the extra information about the machines' rotation speeds to improve her statistical analysis.

   d) Can you guess what she did?[2]

   Solution

[2] Hint: what do we know about the censoring?

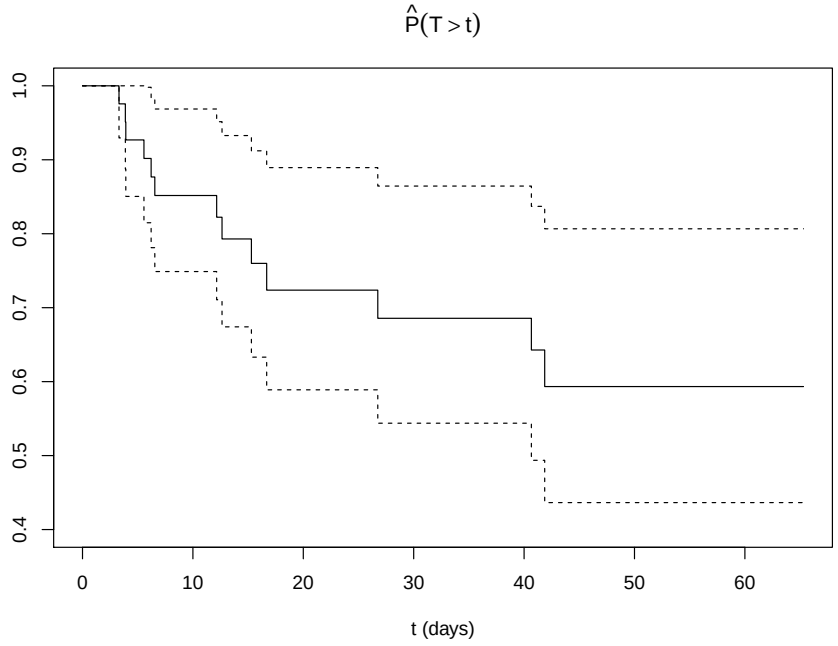

Figure 1:

a) $D_i = 1$ if turbine i is observed with cracks during the study, in which case $\tilde{T}_i = T_i$. $D_i = 0$ if turbine i is censored (i.e. if the machine rotation turbine i stops or no cracks appear before the end of the study), in which case $\tilde{T}_i < T_i$.

b) See Figure 2. We draw a vertical line through 30 days to read off the point estimate along with 95% confidence intervals at $t = 30$; see the left panel. The estimate is seen to be approximately 0.68 with 95% confidence interval $(0.55, 0.86)$. The 10th percentile of the survival times is the t such that $P(T > t) = 0.9$. Thus, drawing a horizontal line at 0.9, seeing where that line intersects the survival curve and its confidence intervals (vertical lines) gives us an estimate of the 10th percentile. See the right panel. Reading off the plot we see that the 10th percentile is estimated to be approximately 6.9 with 95% confidence interval $(3.9, 16.7)$.

c) The phrase "The time it took for the machines to fail was thought to be unrelated to the time it took before cracks developed so that the observed (non-censored) turbines were representative of all turbines." indicate (although it is not stated precisely) that censoring is independent.

d) The text before d) indicate a classic scenario where independent censoring fails; the distribution of the survival time T and the distribution of the censoring time $T^\star$ are both dependent on some variable L (in this case, rotation speed), making the censoring dependent. However, if one has measured L, one can hope the censoring is independent *given* L. If this holds, she can estimate the survival function for each value l of L. Since L has two values in the given example: "high speed" and "low speed", she can calculate the Kaplan-Meier estimate in each of those groups (since independent censoring is assumed to hold in each of the groups), and obtain an estimate of the survival function by 'averaging' the results. In more detail, she can calculate the Kaplan-Meier estimate among the turbines with high rotational speed, $\hat{S}(t|\text{"high speed"})$, and the Kaplan-Meier estimate among the turbines with low rotational speed, $\hat{S}(t|\text{"low speed"})$. Using the law

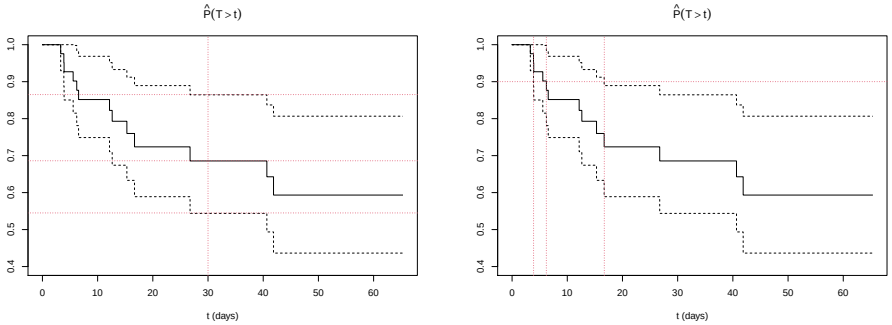

Figure 2:

of total probability she can obtain an estimator

$$\hat{\hat{S}}(t) = \hat{S}(t|\text{"high speed"}) \cdot \hat{P}(\text{"high speed"}) + \hat{S}(t|\text{"low speed"}) \cdot \hat{P}(\text{"low speed"})$$

where $\hat{P}$("high speed") is the average number of turbines with "high speed" in the start of the study, and $\hat{P}$("low speed") is the average number of turbines with "low speed" at the start of the study.

2. Let T^1 and T^2 be two independent survival times with respective hazards α^1 and α^2.

   a) Show that $S = \min(T^1, T^2)$ has hazard $\alpha^1 + \alpha^2$.

   b) **Challenge:** Show that $P(T^1 < T^2 | S = t) = \frac{\alpha^1(t)}{\alpha^1(t)+\alpha^2(t)}$. [1]

Solution

a) We have that the events $\{\min(T^1, T^2) > t\}$ and $\{T^1 > t, T^2 > t\}$ are equal. This leads to $P(\min(T^1, T^2) > t) = P(T^1 > t, T^2 > t) = P(T^1 > t)P(T^2 > t)$ where we used that T^1 and T^2 are independent to obtain in the last equality. Now, $P(T^i > t) = e^{-\int_0^t \alpha^i(s)ds}$ for $i \in \{1, 2\}$ by definition, which leads to $P(S > t) = P(T^1 > t)P(T^2 > t) = e^{-\int_0^t \alpha^1(s)+\alpha^2(s)ds}$. This shows that the survival function of $\min(T^1, T^2)$ is $e^{-\int_0^t \alpha^1(s)+\alpha^2(s)ds}$. From the relationship between survival functions and hazards we see that $\min(T^1, T^2)$ has hazard $\alpha^1 + \alpha^2$.

b) We have

$$\begin{aligned} P(T^1 < T^2 | S = t) &= \lim_{h \to 0+} P(T^1 < T^2 | t \leq S < t + h) \\ &= \lim_{h \to 0+} \frac{P(T^1 < T^2, t \leq S < t + h)}{P(t \leq S < t + h)} \\ &= \lim_{h \to 0+} \frac{\frac{1}{h} P(T^1 < T^2, t \leq S < t + h)}{\frac{1}{h} P(t \leq S < t + h)} \\ &= \frac{\lim_{h \to 0+} \frac{1}{h} P(T^1 < T^2, t \leq S < t + h)}{\lim_{h \to 0+} \frac{1}{h} P(t \leq S < t + h)} \end{aligned} \tag{1}$$

where we assumed that the limits in the numerator and the denominator exists. Let $f_{T^i}(t) = \alpha^i(t) e^{-\int_0^t \alpha^i(s)ds}$ denote the marginal density function of T^i. Focusing on the

[1] Hint: consider the limit $\lim_{h \to 0+} P(T^1 < T^2 | t \leq S < t + h)$ and use the result from a).

numerator first, we have by the law of total probability that

$$\begin{aligned}P(T^1 < T^2, t \le S < t+h) &= \int_0^\infty P(T^1 < T^2, t \le S < t+h | T^2 = u) f_{T^2}(u) du \\ &= \int_t^\infty P(T^1 < T^2, t \le S < t+h | T^2 = u) f_{T^2}(u) du.\end{aligned}$$

To obtain the second line, we used that $P(T^1 < T^2, t \le S < t+h | T^2 = u) = 0$ for $u < t$. Furthermore, by the law of total probability,

$$\begin{aligned}&\int_t^\infty P(T^1 < T^2, t \le S < t+h | T^2 = u) f_{T^2}(u) du \\ &\qquad = \int_t^\infty \int_0^\infty P(T^1 < T^2, t \le S < t+h | T^2 = u, T^1 = v) f_{T^1|T^2}(v) dv f_{T^2}(u) du \\ &\qquad = \int_t^\infty \int_t^{\min(u,t+h)} P(T^1 < T^2, t \le S < t+h | T^2 = u, T^1 = v) f_{T^1}(v) dv f_{T^2}(u) du \end{aligned} \tag{2}$$

where we used that T^1 and T^2 are independent (hence the conditional density $f_{T^1|T^2}(v)$ is equal to the marginal density $f_{T^1}(v)$), and that $P(T^1 < T^2, t \le S < t+h | T^2 = u, T^1 = v) = 0$ unless $v \in [t, \min(u, t+h)]$. In fact, for $u > t$ and $v \in [t, \min(u, t+h)]$ we have that $P(T^1 < T^2, t \le S < t+h | T^2 = u, T^1 = v) = 1$. This gives that (2) reduces to

$$\int_t^\infty \int_t^{\min(u,t+h)} f_{T^1}(v) dv f_{T^2}(u) du = \int_t^\infty \Big(P(T^1 > t) - P(T^1 > \min(u, t+h)) \Big) f_{T^2}(u) du.$$

Thus, by multiplying with $1/h$ and taking the limit we get that the numerator in (1) is (moving the limit inside the integral)

$$\begin{aligned}&\lim_{h \to 0+} \frac{1}{h} \int_t^\infty \Big(P(T^1 > t) - P(T^1 > \min(u, t+h)) \Big) f_{T^2}(u) du \\ &\qquad = \int_t^\infty -\frac{d}{dt} P(T^1 > t) f_{T^2}(u) du \\ &\qquad = \alpha^1(t) P(T^1 > t) P(T^2 > t). \end{aligned} \tag{3}$$

Here we used the fact that, for any $u > t$,

$$\begin{aligned}\lim_{h \to 0+} \frac{1}{h} \Big(P(T^1 > \min(u, t+h)) - P(T^1 > t) \Big) &= \lim_{h \to 0+} \frac{1}{h} \Big(P(T^1 > t+h) - P(T^1 > t) \Big) \\ &= \frac{d}{dt} P(T^1 > t),\end{aligned}$$

and that $\frac{d}{dt} P(T^1 > t) = -\alpha^1(t) P(T^1 > t)$ by the relationship between hazard functions and survival functions. For the denominator in (1) we have

$$\begin{aligned}\lim_{h \to 0+} \frac{1}{h} P(t \le S < t+h) &= \lim_{h \to 0+} \frac{1}{h} P(t \le S < t+h | S \ge t) P(S \ge t) \\ &= \big(\alpha^1(t) + \alpha^2(t) \big) P(S \ge t) \\ &= \big(\alpha^1(t) + \alpha^2(t) \big) P(T^1 \ge t) P(T^2 \ge t), \end{aligned} \tag{4}$$

where we used the law of total probability in the first line, the result from a) and the definition of the hazard function in the second line, and the fact that T^1 and T^2 are independent in the last line. Thus, by dividing (3) by (4) we get the advertised result.

3. (Exercise 3.1 from ABG 2008) The data in the table are from Freireich et al. (1963) and show the result of a study where children with leukemia are treated with a drug (6-MP) to prevent relapse, and where this treatment is compared with placebo. The numbers in the table are remission lengths in weeks; a '*' indicates a censored observation.

Placebo	1	1	2	2	3	4	4	5	5	8	8
	8	8	11	11	12	12	15	17	22	23	
6-MP	6	6	6	6*	7	9*	10	10*	11*	13	16
	17*	19*	20*	22	23	25*	32*	32*	34*	35*	

a) Compute the Nelson-Aalen estimates for the 6-MP group and for the placebo group. [2]

b) Plot both Nelson-Aalen estimates in the same figure. What can you learn from the plots?

c) Calculate the Kaplan-Meier estimate for both groups, and plot the results.

d) Estimate the probability of not having a remission the first ten weeks in both groups.

Solution We calculate the estimates here; plots can be found in the solution to problem 2.

We generate the following tables which contains the (possibly censored) times $\tilde{T}_i$, the indicator D_i which is 1 when there is an event and 0 when there is censoring. n.risk in a given row is the number of individuals at risk just before considering the individual who died/was censored in that row. $\hat{A}(\tilde{T}_i)$ is calculated using the hint in the footnote, and $\hat{S}(\tilde{T}_i) = \prod_{T_j \leq \tilde{T}_i}(1 - \Delta\hat{A}(T_j))$.

Placebo							6-MP						
i	$\tilde{T}_i$	D_i	n.risk	1/n.risk	$\hat{A}(\tilde{T}_i)$	$\hat{S}(\tilde{T}_i)$	i	$\tilde{T}_i$	D_i	n.risk	1/n.risk	$\hat{A}(\tilde{T}_i)$	$\hat{S}(\tilde{T}_i)$
1	1	1	21	0.0476	0.0976	0.902	1	6	1	21	0.0476	0.15	0.85
2	1	1	20	0.05	0.0976	0.902	2	6	1	20	0.05	0.15	0.85
3	2	1	19	0.0526	0.206	0.805	3	6	1	19	0.0526	0.15	0.85
4	2	1	18	0.0556	0.206	0.805	4	6	0	18	0.0556	0.15	0.85
5	3	1	17	0.0588	0.265	0.757	5	7	1	17	0.0588	0.209	0.8
6	4	1	16	0.0625	0.394	0.66	6	9	0	16	0.0625	0.209	0.8
7	4	1	15	0.0667	0.394	0.66	7	10	1	15	0.0667	0.276	0.746
8	5	1	14	0.0714	0.542	0.562	8	10	0	14	0.0714	0.276	0.746
9	5	1	13	0.0769	0.542	0.562	9	11	0	13	0.0769	0.276	0.746
10	8	1	12	0.0833	0.928	0.345	10	13	1	12	0.0833	0.359	0.684
11	8	1	11	0.0909	0.928	0.345	11	16	1	11	0.0909	0.45	0.622
12	8	1	10	0.1	0.928	0.345	12	17	0	10	0.1	0.45	0.622
13	8	1	9	0.111	0.928	0.345	13	19	0	9	0.111	0.45	0.622
14	11	1	8	0.125	1.2	0.253	14	20	0	8	0.125	0.45	0.622
15	11	1	7	0.143	1.2	0.253	15	22	1	7	0.143	0.593	0.533
16	12	1	6	0.167	1.56	0.16	16	23	1	6	0.167	0.76	0.444
17	12	1	5	0.2	1.56	0.16	17	25	0	5	0.2	0.76	0.444
18	15	1	4	0.25	1.81	0.12	18	32	0	4	0.25	0.76	0.444
19	17	1	3	0.333	2.15	0.08	19	32	0	3	0.333	0.76	0.444
20	22	1	2	0.5	2.65	0.04	20	34	0	2	0.5	0.76	0.444
21	23	1	1	1	3.65	0	21	35	0	1	1	0.76	0.444

For "What can you learn from the plots?" in b): From Figure 3 we see that the cumulative hazard estimate of the placebo group is higher than the cumulative hazard estimate of the 6-MP group, indicating that the placebo group has higher "cumulative risk" of relapse. It is not clear from the plot if the difference is statistically significant.

For problem d): From the table above and Figure 4 we see that the Kaplan-Meier estimate in the placebo group is 0.38 at $t = 10$, while the Kaplan-Meier estimate in the 6-MP group is 0.75 at $t = 10$. It is not clear from the plot if the difference is statistically significant.

[2] So far we have assumed absolutely continuous survival times, which implies that event times will not be tied. Note that some survival times are tied in the data set below. If T_j is an event times with d_j ties you may use the estimator $\Delta\hat{A}(T_j) = \sum_{i=0}^{d_j-1} \frac{1}{Z(T_j)-i}$, and set $\hat{A}(t) = \sum_{T_j \leq t} \Delta\hat{A}(T_j)$.

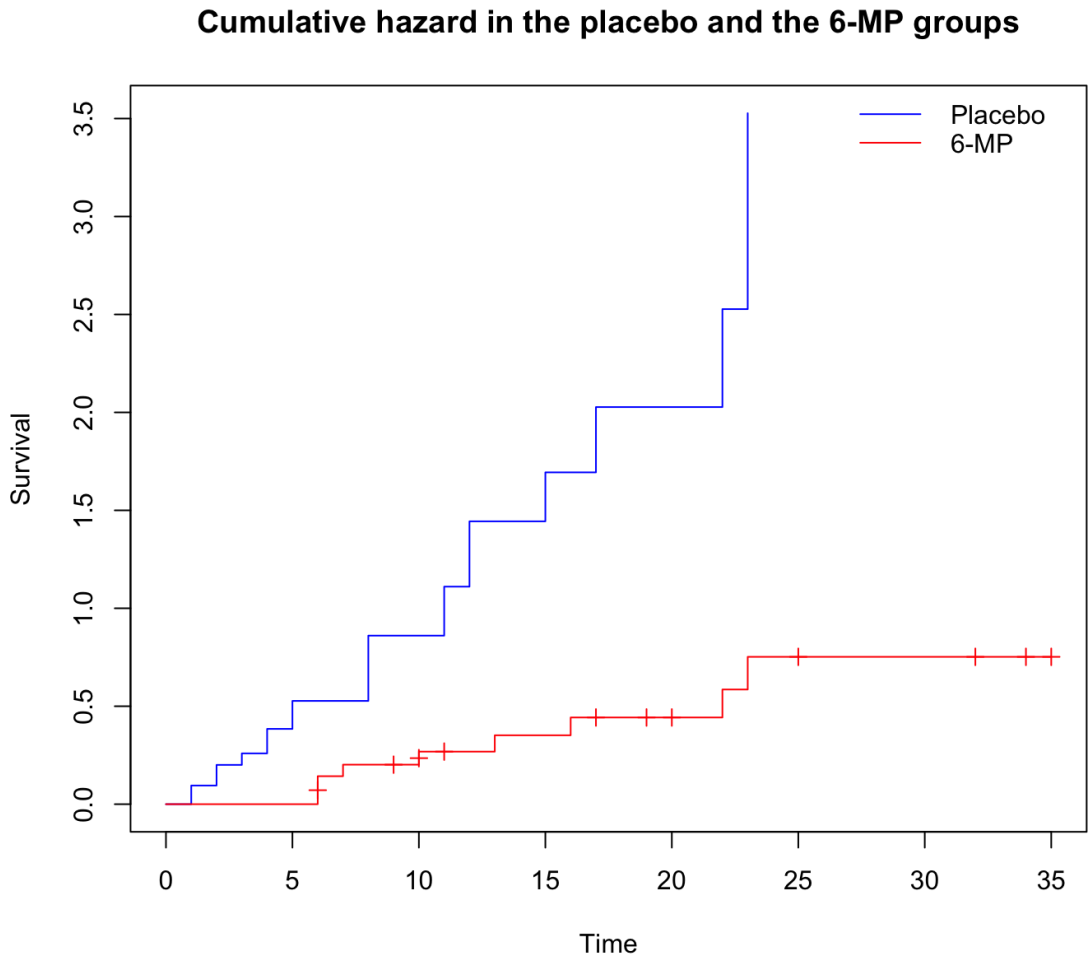

Figure 3: Nelson-Aalen plot of the placebo and 6-MP groups. Censoring in the treatment group is denoted with ticks.

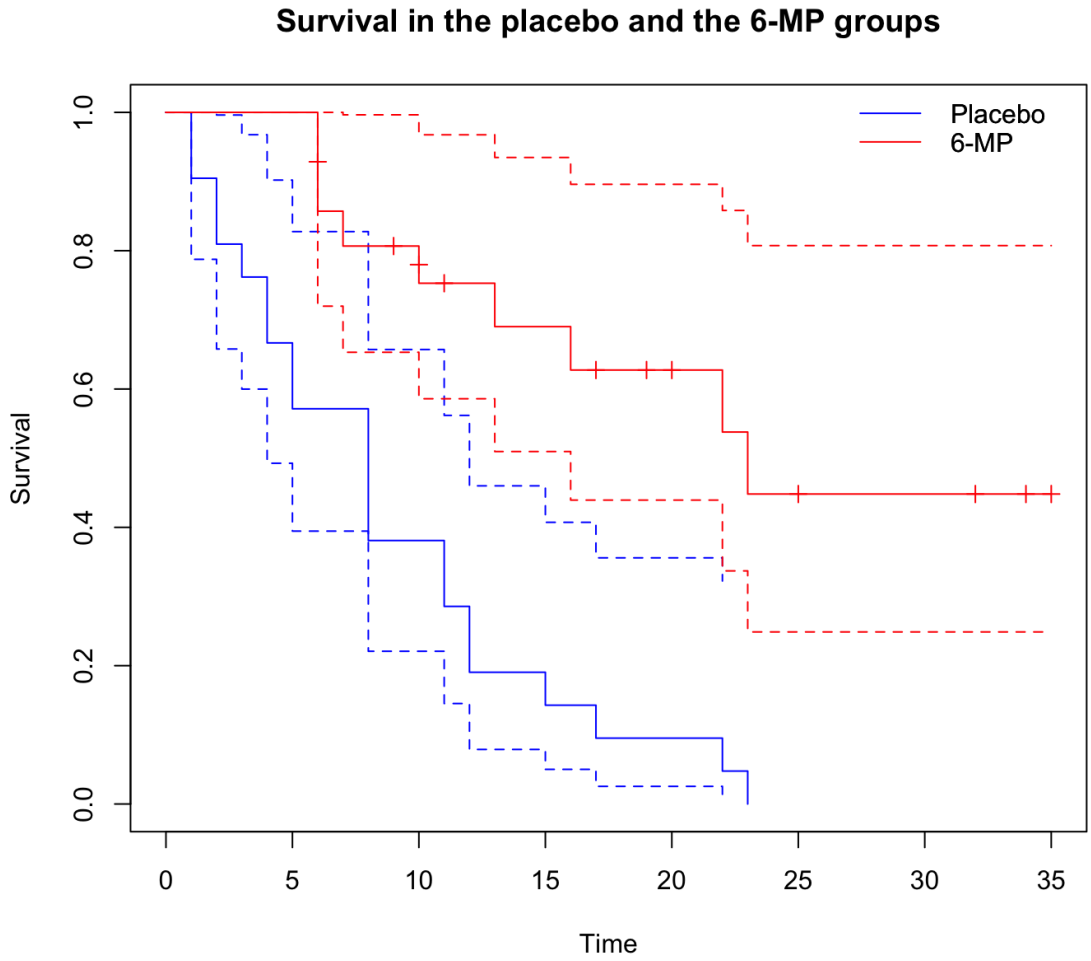

Figure 4: Kaplan-Meier plot of the placebo and 6-MP groups. Censoring in the treatment group is denoted with ticks. The 95% CI-s are denoted as dashed lines.