SUPPLEMENTARY MATERIAL

APPENDIX 1: Mathematical background of Kaplan-Meier estimator and competing risk analysis. Classical survival techniques. In classical survival analysis, the survival time (T) denotes the time from a well-defined time origin to the occurrence of an event of interest. The observation window during which data are collected causes individuals to have part of their disease history unobserved. If the endpoint of interest has not yet occurred at the end of the observation window, we said that the event time was censored. Each individual (i) is assumed to have an event time (t_i) and a censoring time (c_i). We observe the minimum between these two times. The basic assumption of the standard models for censored data is that the censoring distribution and the event time distribution are independent. This implies that at each point in time, the individuals who are censored can be represented by those who remain under observation.

The number at risk (subjects that are in follow-up and have not experienced their event at time t) and the number of observed events at time t_j are denoted by n_j and d_j, respectively. A crucial quantity is the hazard λ, defined as the conditional probability of failing at t_j, given still alive just before time t_j.

The independence assumption between the censoring mechanism and the event time distribution implies that the hazard of the individuals that are censored is equal to the hazard of the individuals that remain in follow-up. This implies that subjects in the risk set are representative for all subjects alive and therefore the hazard $\lambda(t_j)$ can be estimated proportion of individuals that fail at time t_j; the estimated hazard λ is given by:

$$\hat{\lambda}(t_j) = \frac{d_j}{n_j}$$

The Kaplan-Meier methodology estimates the probability of surviving $S(t)$ up to time t_j. The probability of surviving up to t_j is the product of the probability of surviving up to the previous time, and the conditional probability of surviving up to t_j given you are still alive beyond t_{j-1}.

$$\hat{S}(t)_{KM} = \prod_{j:t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$

where ∏ means the product of all terms (thus the Kaplan-Meier estimator is also known as the product limit estimator).

Competing risks. Estimating the probability of the event of interest in the presence of competing risks. The situation for each individual can be summarised in this way: 1) the patient fails from the event of interest; 2) the patient fails from a competing event at time t_j; or 3) the patient has not failed from either causes but has follow-up only to time t_j.

The fundamental concept in competing risks models is the cause-specific hazard function, the hazard of failing from a given cause in the presence of the competing events. This quantity is estimated as the proportion of subjects at risk that fail from cause k:

$$\hat{\lambda}_k(t_j) = \frac{d_{kj}}{n_j}$$

where d_{kj} denotes the number of patients failing from cause k at time t_j. Note that the Kaplan-Meier estimator can be written as:

$$\hat{S}(t) = \prod_{j:t_j \le t}\left(1 - \sum_{k=1}^{K} \hat{\lambda}_k(t_j)\right) = \prod_{j:t_j \le t}\left(1 - \sum_{k=1}^{K} \frac{d_{kj}}{n_j}\right)$$

where the sum (indicated as ∑) is over all K possible competing events. The survival function in this context is the probability of not having failed from any cause at time t. The cumulative incidence function is defined as the probability of failing from a specific cause k before time t and it is denoted by $I_k(t)$. In the medical literature the cumulative incidence is also known as cause specific failure probability, crude incidence or cause specific risk. The cumulative incidence $I_k(t)$ of cause k at time t is estimated as:

Equation 1

$$\hat{S}(t) = \prod_{j:t_j \le t}\left(1 - \sum_{k=1}^{K} \hat{\lambda}_k(t_j)\right) = \prod_{j:t_j \le t}\left(1 - \sum_{k=1}^{K} \frac{d_{kj}}{n_j}\right)$$

$S(t_{j-1})$: estimated probability **free of any event** at time t_j (or the probability of not having failed from any cause at time i).

When calculating the cumulative incidence by using the Kaplan-Meier methodology, events from causes other than k are treated as censored, therefore the naive Kaplan-Meier estimates the cumulative incidence as:

Equation 2

$$\hat{I}_k(t) = 1 - \hat{S}(t) = \sum_{j:t_j \le t} \hat{\lambda}(t_j)\hat{S}(t_{j-1}) = \sum_{j:t_j \le t} \frac{d_j}{n_j}\hat{S}(t_{j-1})$$

Note that in Equation 1 the estimated cause-specific hazard $\lambda_k(t_j)$ is used in the estimation of the cumulative incidence while in Equation 2 the estimated hazard $\lambda(t_j)$ is employed.

APPENDIX 2: Which data is necessary to perform a competing risk analysis?. In order to obtain an unbiased estimation of the probability of revision surgery, an analysis using a competing risks model is necessary whenever competing risks are present. But which data should be gathered for such a competing risk model? With standard survival data, there is only one type of event, and the number of events is either 0 or 1. An individual's survival data is expressed by three variables: 1) the time the individual becomes at risk (entry time); 2) the time the individual experiences the event or is censored (event time); and 3) a variable denoting whether the event has occurred or was censored (variable status). According to this representation in our example the entry time is the date of initial surgery, the event time is time to revision surgery and status will be 1 if the event has occurred, or 0 if this is not the case. In a competing risk setting, there are multiple events. In our example, we have revision surgery and death. Again, three variables express an individual's survival data: the time the individual becomes at risk (entry time), the time the individual experiences either event or is censored (event time) and a variable denoting whether any event is observed or censored (status). In our setting the variable status will take value 0, 1, or 2, which represent censored observation, revision surgery and death, respectively. The event time is time to revision, time to death or time to last known contact.