Supplementary file 1

Mixture split-population (MSP) models

Let W be the indicator denoting a subject is non-susceptible (W = 0) or susceptible (W = 1) to the desired event and T is a nonnegative random variable indicating the survival time. The MSP model is given by

$$S(t|\mathbf{x},\mathbf{y})=1-\pi(\mathbf{x})+\pi(\mathbf{x})\times S_{W=1}(t|\,\mathbf{y}) \qquad \text{(I)}$$

where $S(t|\boldsymbol{x},\boldsymbol{y})$ is the unconditional survival time function for the entire population, $\pi(\boldsymbol{x})$ is the probability of being long-term survivors given a covariate vector $\mathbf{y}$, and $S_{W=1}(t|\,\mathbf{y})$ is the survival time function for short-term survivors given a covariate vector $\mathbf{y}$.[28-30] Note that $S(t|\mathbf{x},\mathbf{y})\cong 1-\pi(\mathbf{x})$, i.e. the long-term OS rate, as *t* approaches infinity. As seen in the formula (I), the MSP models consist of two components: incidence component ($\pi(\boldsymbol{x})$) and latency component ($S_{W=1}(t|\,\mathbf{y})$). An advantage of the MSP models is that the proportion of long-term survivors and the survival distribution of the short-term survivors are modeled separately.[29] Different link functions can be utilized to the incidence component such as logit, complementary log-log, and probit link functions. In this study, the logit link function is

$$\operatorname{logit}(\pi(\mathbf{x}))=\beta_0+\sum_{j=1}^{p}\beta_j x_j \Leftrightarrow \pi(\mathbf{x})=\frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}} \qquad \text{(II)}$$

where $\beta_j s$ are unknown parameters, is applied to model the effects of $\boldsymbol{x}$.[29,30,34]

In the parametric acceleration failure time MSP (AFTMSP) models, the survival function for short-term survivors, $S_{W=1}(t|\,\mathbf{y})$, takes the form of parametric distribution like exponential, Weibull, lognormal, loglogistic, EGG etc.[35] By using the AFTMSP with logit link function, the impact of the different covariates on the survival of susceptible and non-susceptible individuals can be interpreted via the ETR and odds ratio (OR) statistics, respectively. The long-term OS rate can be estimated based on the results from the incidence component. Using formula (II), we can obtain the following formula for the long-term OS rate:

$$\text{Long-term OS rate}=1-\pi(\mathbf{x})=1-\frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}} \qquad \text{(III)}$$

The model parameters are estimated by maximizing the observed log-likelihood function directly and the standard error estimates are obtained from minus the second derivatives of the log-likelihood function.[28,34,35]