Enrollment and event prediction R shiny app manual

1 Introduction

The R Shiny application developed in this study provides a reliable and flexible tool for predicting enrollment and events in clinical trials. The application can be utilized at different stages of a clinical trial, including the design stage, real-time before enrollment completion, and real-time after enrollment completion.

The app's versatility is due to its ability to accommodate various enrollment and event models. These models' assumptions are clearly outlined to ensure the app's predictions are accurate and reliable.

Users must provide relevant study data as input to the application, and it provides predictions as output.

2 Enrollment models

In this study, we adopt a Poisson enrollment process to model the number of subjects enrolled in a clinical trial over different time periods. This process assumes that the number of subjects enrolled during each period is statistically independent.

We use the function $a(t)$ to represent the enrollment rate on day t since the start of the trial. The number of subjects enrolled between day t_0 and day t_1 follows a Poisson distribution with mean

$$\mu(t_1) - \mu(t_0) = \int_{t_0}^{t_1} a(u)\, du$$

where $\mu(t)$ is the integral of $a(u)$ from 0 to t.

Different enrollment models assume different functional forms for $a(t)$ and $\mu(t)$. By selecting an appropriate enrollment model, we can estimate the enrollment rate and predict the number of subjects likely to be enrolled at different stages of the trial.

2.1 The Poisson enrollment model

The homogeneous Poisson enrollment model assumes a constant enrollment rate, i.e., $a(t) \equiv \mu$. The mean number of subjects enrolled by time t is given by $\mu(t) = \mu t$.

2.2 The time-decay enrollment model

The time-decay enrollment model assumes that $a(t) = \frac{\mu}{\delta}\left(1 - e^{-\delta t}\right)$, where μ is the base rate parameter and δ is the decay rate parameter. The enrollment rate begins at $a(0) = 0$ and increases to a steady state value of $a(\infty) = \mu/\delta$ as t approaches infinity. The mean number of subjects enrolled by time t is given by $\mu(t) = \frac{\mu}{\delta}\left(t - \frac{1}{\delta}\left(1 - e^{-\delta t}\right)\right)$.

2.3 The B-spline enrollment model

The B-spline enrollment model is proposed to address the limitations of the time-decay enrollment model, particularly in capturing complex enrollment patterns where the rate of enrollment initially

increases and then decreases. The B-spline function is employed to model the log enrollment rates to maintain the positivity of the enrollment rate. The B-spline model requires users to specify the number of inner knots and the number of days used for averaging enrollment rates before the last enrollment date (lag days) to make predictions. The application of log transformation and lag days are introduced to enhance the B-spline enrollment model proposed by Zhang and Long (2010).

The B-spline enrollment model can only be used after the trial has started and the enrollment is ongoing. It cannot be used at the design stage.

2.4 The piecewise Poisson enrollment model

The piecewise Poisson model is a widely used enrollment model that segments the time axis into multiple intervals, each characterized by a constant enrollment rate. Despite its lack of smoothness, the piecewise Poisson model is a flexible and powerful tool for specifying and analyzing enrollment trends in clinical trials.

2.5 Generation of enrollment times

Suppose that the study is in progress at time t_0, and $n(t_0)$ subjects have already been enrolled, with a target enrollment of n subjects. Therefore, the number of new subjects to enroll is $r = n - n(t_0)$. The Poisson enrollment process assumes statistical independence of the number of enrollments in separate time intervals. Let $n(t)$ represent the total number of enrolled subjects by time t, and $V_{(i)}$ denote the enrollment time for the ith new subjects. It is evident that

$$P\big(V_{(i)} > v_1 \big| V_{(i-1)} = v_0\big) = P(n(v_1) - n(v_0) = 0) = \exp\big(-\mu(v_1) + \mu(v_0)\big)$$

Using the inverse transform method, we can generate the enrollment times for the r new subjects sequentially as follows:

- Generate e_1 from a standard exponential distribution, and calculate $V_{(1)} = \mu^{-1}(\mu(t_0) + e_1)$.
- For $i = 2, \ldots, r$, generate e_i from a standard exponential distribution, and set $V_{(i)} = \mu^{-1}(\mu\big(V_{(i-1)} + e_i\big)$.

3 Event models

Let W denote the time between enrollment and event for a subject. We can characterize the random variable W using either the survival function, $S(t) = P(W > t)$, or the hazard rate function,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(T \le t + \Delta t | T > t)}{\Delta t}$$

The hazard rate function tells us the instantaneous rate of having the event at any given time, given that the subject has not had the event before that time.

3.1 The exponential distribution

The exponential distribution is the most basic time-to-event distribution that assumes a constant hazard rate over time, which can be denoted as $h(t) \equiv \lambda$. The corresponding survival function is $S(t) = e^{-\lambda t}$. For instance, if we have an event rate of 5% in one year, this can be translated to an exponential event

distribution with a hazard rate of $\lambda = -\frac{\log(S(t))}{t} = -\frac{\log(1-0.05)}{365} = 0.00014$ per day. The median of the exponential distribution is $\frac{\log(2)}{\lambda}$.

3.2 The Weibull distribution

The Weibull distribution is a more versatile version of the exponential distribution. Unlike the exponential distribution, it does not assume a constant hazard rate, making it more widely applicable. This distribution is defined by two parameters, κ and λ, where κ determines the shape of the distribution curve and λ determines its scaling. These parameters are referred to as the shape and scale parameters, respectively.

The hazard function of the Weibull distribution can be expressed as

$$h(t) = \frac{\kappa}{\lambda}\left(\frac{t}{\lambda}\right)^{\kappa-1}$$

When $\kappa = 1$, the hazard rate remains constant over time, which is the same as the exponential case. However, when $\kappa > 1$, the hazard rate increases as time goes on, whereas it decreases when $\kappa < 1$.

The survivor function for the Weibull distribution is

$$S(t) = e^{-\left(\frac{t}{\lambda}\right)^{\kappa}}$$

The mean of the Weibull distribution is $\lambda\Gamma\left(1+\frac{1}{\kappa}\right)$ and the variance is $\lambda^2\left(\Gamma\left(1+\frac{2}{\kappa}\right) - \Gamma^2\left(1+\frac{1}{\kappa}\right)\right)$, where $\Gamma(\cdot)$ is the gamma function.

3.3 The log-logistic distribution

The log-logistic distribution is a probability distribution that models a variable whose logarithm follows a logistic distribution, i.e., $T \sim llogis(\kappa, \lambda)$ if $\log(T) \sim logis(\mu, \sigma)$, where $\mu = \log(\lambda), \sigma = 1/\kappa$. Unlike the Weibull distribution, which has a monotonically increasing or decreasing hazard rate, the hazard rate function of the log-logistic distribution initially increases from zero to a maximum and then decreases to zero as time approaches infinity. The log-logistic distribution generally has heavier tails than the Weibull distribution. This means that there is a relatively higher probability of observing extreme values for a log-logistic random variable than for a Weibull random variable. The survival function of the log-logistic distribution is

$$S(t) = \frac{1}{1+\exp\left(\frac{\log(t)-\mu}{\sigma}\right)} = \frac{1}{1+\left(\frac{t}{\lambda}\right)^{\kappa}}$$

where $\kappa = \frac{1}{\sigma}$ is the shape parameter of the log-logistic distribution, and $\lambda = \exp(\mu)$ is the scale parameter of the log-logistic distribution.

The mean of the log-logistic distribution exists if $\kappa > 1$ and the variance of the log-logistic distribution exists if $\kappa > 2$.

3.4 The log-normal distribution

The log-normal distribution is a probability distribution that models a variable whose logarithm follows a normal distribution, i.e., $T \sim lnorm(\mu, \sigma^2)$ if $\log(T) \sim N(\mu, \sigma^2)$. Unlike the Weibull distribution which has a monotonically increasing or decreasing hazard rate, the hazard rate function of the log-normal distribution initially increases from zero to a maximum and then decreases to zero as time approaches infinity. The log-normal distribution is similar to the log-logistic distribution, with the latter having heavier tails. The survival function of the log-normal distribution is

$$S(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right)$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution.

The mean of the log-normal distribution is $\exp\left(\mu + \frac{1}{2}\sigma^2\right)$ and the variance of the log-normal distribution is $(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$.

3.5 The piecewise exponential distribution

The piecewise exponential distribution divides the time axis into multiple intervals, each characterized by a constant hazard rate. This allows the hazard rate to change over time and hence is more flexible than the exponential distribution.

3.6 The model-averaging event distribution

To perform model-averaging, we model the time-to-event using two distributions: Weibull and log-normal. The weights for each distribution are determined based on the Bayesian Information Criterion (BIC) score. This approach seeks to balance and improve the robustness of the model by combining the strengths of both parametric models. The survival function of the resulting averaged model takes the following form

$$S(t) = w_{WB}S_{WB}(t) + w_{LN}S_{LN}(t)$$

where w_{WB} and w_{LN} are the weights for the Weibull and log-normal distributions, respectively,

$$w_{WB} = \frac{\exp\left(-\frac{1}{2}BIC_{WB}\right)}{\exp\left(-\frac{1}{2}BIC_{WB}\right) + \exp\left(-\frac{1}{2}BIC_{LN}\right)}$$

$w_{LN} = 1 - w_{WB}$, and BIC_{WB} and BIC_{LN} are the BIC scores for the respective models.

BIC is a statistical measure used for model selection among a set of candidate models. It is a criterion for model selection that balances model fit against model complexity. Among competing models, the one that achieves the lowest BIC value is typically preferred as it indicates a better balance between model complexity and goodness of fit.

We utilize a weighted BIC to evaluate the performance of the averaged model. Specifically, we calculate the weighted BIC as $w_{WB}BIC_{WB} + w_{LN}BIC_{LN}$.

3.7 The spline event model

In the spline event model developed by Royston and Parmar (2002), a transformed survival function, $g\big(S(t)\big)$, is modelled as a natural cubic spline function of log time $x = \log(t)$,

$$g\big(S(t)\big) = s(x, \gamma)$$

In the proportional hazards model (`scale = "hazard"`), $g\big(S(t)\big) = \log\big(-\log\big(S(t)\big)\big)$.

In the proportional odds model (`scale = "odds"`), $g\big(S(t)\big) = \log\left(\frac{1}{S(t)} - 1\right)$.

In the probit model (`scale = "normal"`), $g\big(S(t)\big) = -\Phi^{-1}\big(S(t)\big)$.

The natural cubic spline is constrained to be linear beyond boundary knots, k_{min} and k_{max}, and is defined as

$$s(x, \gamma) = \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + \cdots + \gamma_{m+1} v_m(x)$$

where $v_j(x)$ represents the jth basis function:

$$v_j(x) = \big(x - k_j\big)_+^3 - \lambda_j (x - k_{min})_+^3 - \big(1 - \lambda_j\big)(x - k_{max})_+^3$$

Here, k_j is the jth inner knot, $\lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}}$, for $j = 1, \ldots, m$. The knots are chosen as equally spaced quantiles of the log uncensored survival times. The boundary knots are chosen as the minimum and maximum log uncensored survival times. In addition, x_+ denotes the positive part of x.

With no knots ($m = 0$), the spline reduces to a linear function, and these models are equivalent to Weibull, log-logistic and lognormal models, respectively. As noted in Royston and Parmar (2002), experience suggests that a worthwhile improvement in fit over a straight-line model is often obtained by using a spline model with a single internal knot, but often little is gained by adding further knots.

3.8 Generation of event times

Assuming a data cutoff time of t_0 for the study, we can generate the underlying event time, W_i, for an ongoing subject i. We know that the enrollment time $U_i \le t_0$, and that $W_i > t_0 - U_i$. We use the inverse transform method to generate W_i by setting the conditional probability

$$P(W_i > t | W_i > t_0 - U_i, U_i) = \frac{S(t)}{S(t_0 - U_i)}$$

equal to a uniform random variable p_i, so that

$$W_i = S^{-1}(S(t_0 - U_i)p_i)$$

For instance, for the Weibull distribution with a shape parameter κ and a scale parameter λ, the following equation can be used to generate W_i:

$$W_i = \lambda\left(\left(\frac{t_0 - U_i}{\lambda}\right)^{\kappa} + e_i\right)^{1/\kappa}$$

Here $e_i = -\log(p_i)$ is a random variable generated from a standard exponential distribution.

When dealing with the log-normal distribution, it is more efficient to utilize specialized algorithms designed to generate random variables from truncated normal distributions.

To generate the event time from the model averaging event model, we begin by generating the component indicator Y_i from the following Bernoulli distribution,

$$Y_i \sim b\big(1, P(Y_i = 1 | W_i > t_0 - U_i, U_i)\big)$$

where

$$P(Y_i = 1 | W_i > t_0 - U_i, U_i) = \frac{W_{WB} S_{WB}(t_0 - U_i)}{W_{WB} S_{WB}(t_0 - U_i) + W_{LN} S_{LN}(t_0 - U_i)}$$

If $Y_i = 1$, then we generate W_i from the truncated Weibull distribution. If $Y_i = 0$, then we generate W_i from the truncated normal distribution.

4 Dropout models

In survival analysis, dropout can act as a competing risk that may prevent the observation of the event of interest. The R shiny app models the time to dropout using various probability models, including exponential, Weibull, log-logistic, log-normal, piecewise exponential, model averaging, and spline. To generate the time-to-dropout data, we use the same algorithm that is applied to generate time-to-event in Section 3.6.

5 Number of events

Bagiella and Heitjan (2001) proposed a method to calculate the cumulative number of events by time t in a clinical trial using the following equation:

$$D(t) = D(t_0) + Q(t_0, t) + R(t_0, t)$$

where $D(t_0)$ represents the number of events that have already occurred by time t_0, $Q(t_0, t)$ represents the predicted number of events between t_0 and t from ongoing subjects, and $R(t_0, t)$ represents the predicted number of events between t_0 and t from new subjects. Here, the number of events reflects the observed events after accounting for dropouts and administrative censoring.

6 Input and output

To predict enrollment and events accurately, the required input and the resulting output vary depending on the stage of the study and prediction target.

6.1 Design stage enrollment prediction

The following input must be provided:

- The target enrollment (number of subjects)
- The level of prediction interval (95%, 90%, or 80%)
- The number of years after study start (prediction horizon)
- Whether to predict by treatment
- The number of treatment groups
- Treatment allocation in a randomization block

- Treatment description
- The number of simulations to be conducted
- The random seed used to initiate the simulations
- The enrollment model for the study (e.g., Poisson, time-decay, or piecewise Poisson) and the corresponding model parameters. These parameters can be based on previous studies, literature reviews, and estimations from sites

The following output will be produced:

- Predicted time from trial start until reaching the target number of subjects
- Plots of predicted cumulative number of subjects enrolled over time

6.2 Design stage enrollment and event prediction

The following input must be provided:

- The target enrollment (number of subjects)
- The target events
- The level of prediction interval (95%, 90%, or 80%)
- The number of years after study start (prediction horizon)
- What to show on prediction plot: enrollment, event, dropout, and/or ongoing
- Whether to predict by treatment
- The number of treatment groups
- Treatment allocation in a randomization block
- Treatment description
- The number of simulations to be conducted
- The random seed used to initiate the simulations
- The enrollment model for the study (e.g., Poisson, time-decay, or piecewise Poisson) and the corresponding model parameters. These parameters can be based on previous studies, literature reviews, and estimations from sites
- The event model for the study (e.g., exponential, Weibull, log-logistic, log-normal, or piecewise exponential) and the corresponding parameter values by treatment. These parameter values can also be based on previous studies and literature reviews
- The dropout model for the study (e.g., none, exponential, Weibull, log-logistic, log-normal, or piecewise exponential) and the corresponding parameter values by treatment. These parameter values can also be based on previous studies and literature reviews

The following output will be produced:

- Predicted time from trial start until reaching the target number of subjects
- Predicted time from trial start until reaching the target number of events
- Plots of predicted cumulative number of subjects enrolled and cumulative number of events over time

6.3 Enrollment phase enrollment prediction

The following input must be provided:

- The target enrollment (number of subjects)

- The subject level data set, which must include the following variables:
    - trialsdt: the trial start date
    - cutoffdt: the data cutoff date for analysis
    - usubjid: unique subject identifier
    - randdt: the randomization date (or the enrollment date for a non-randomized study) for the subject

  For prediction by treatment, the subject level data set should also include
    - treatment: treatment arm coded as 1, 2, and so on for the subject
    - treatment_description: treatment label corresponding to the numeric treatment code
- The level of prediction interval (95%, 90%, or 80%)
- The number of years after data cutoff (prediction horizon)
- Whether to predict by treatment
- The number of treatment groups required for prediction by treatment
- Treatment allocation in a randomization block required for prediction by treatment
- The number of simulations to be conducted
- The random seed used to initiate the simulations
- The enrollment model for the study (e.g., Poisson, time-decay, B-spline, or piecewise Poisson)

The following output will be produced:

- Summary of observed data in terms of trial start date, trial cutoff date, days since trial start, and the current number of subjects
- Plot of the observed cumulative number of subjects enrolled
- Plot of the daily enrollment rate with loess smoothing
- Plot depicting the enrollment model fit
- Predicted time from cutoff until reaching the target number of subjects
- Plots of observed and predicted cumulative number of subjects enrolled over time

6.4 Enrollment phase enrollment and event prediction

The following input must be provided:

- The target enrollment (number of subjects)
- The target events
- The subject level data set, which must include the following variables:
    - trialsdt: the trial start date
    - cutoffdt: the data cutoff date for analysis
    - usubjid: unique subject identifier
    - randdt: the randomization date (or the enrollment date for a non-randomized study) for the subject
    - time: days from enrollment to the event of interest or data cutoff, whichever comes first, for the subject
    - event: the event indicator for the subject, which takes the value 1 if the subject had the event of interest before the data cutoff date, and 0 otherwise

    - dropout: the indicator of competing risks to the event of interest for the subject, which takes the value 1 if the subject dropped out before having the event of interest, and 0 otherwise

  For prediction by treatment, the subject level data set should also include
    - treatment: treatment arm coded as 1, 2, and so on for the subject
    - treatment_description: treatment label corresponding to the numeric treatment code
- The level of prediction interval (95%, 90%, or 80%)
- The number of years after data cutoff (prediction horizon)
- What to show on prediction plot: enrollment, event, dropout, and/or ongoing
- Whether to predict by treatment
- The number of treatment groups required for prediction by treatment
- Treatment allocation in a randomization block required for prediction by treatment
- The number of simulations to be conducted
- The random seed used to initiate the simulations
- The enrollment model for the study (e.g., Poisson, time-decay, B-spline, or piecewise Poisson)
- The event model for the study (e.g., exponential, Weibull, log-logistic, log-normal, piecewise exponential, model averaging, or spline)
- The dropout model for the study (e.g., none, exponential, Weibull, log-logistic, log-normal, piecewise exponential, model averaging, or spline)

The following output will be produced:

- Summary of observed data in terms of the trial start date, cutoff date, days since trial start, the current number of subjects, events, dropouts, and ongoing subjects
- Plot of the observed cumulative number of subjects enrolled and cumulative number of events
- Plot of the daily enrollment rate with loess smoothing
- Kaplan-Meier plot for time to event
- Kaplan-Meier plot for time to dropout
- Plot depicting the enrollment model fit
- Plot depicting the event model fit
- Plot depicting the dropout model fit
- Predicted time from cutoff until reaching the target number of subjects
- Predicted time from cutoff until reaching the target number of events
- Plots of observed and predicted cumulative number of subjects enrolled and cumulative number of events over time

6.5 Follow-up phase event prediction

The following input must be provided:

- The target events
- The subject level data set, which must include the following variables:
    - trialsdt: the trial start date
    - cutoffdt: the data cutoff date for analysis
    - usubjid: unique subject identifier

    - randdt: the randomization date (or the enrollment date for a non-randomized study) for the subject
    - time: days from enrollment to the event of interest or data cutoff, whichever comes first, for the subject
    - event: the event indicator for the subject, which takes the value 1 if the subject had the event of interest before the data cutoff date, and 0 otherwise
    - dropout: the indicator of competing risks to the event of interest for the subject, which takes the value 1 if the subject dropped out before having the event of interest, and 0 otherwise

  For prediction by treatment, the subject level data set should also include
    - treatment: treatment arm coded as 1, 2, and so on for the subject
    - treatment_description: treatment label corresponding to the numeric treatment code
- The level of prediction interval (95%, 90%, or 80%)
- The number of years after data cutoff (prediction horizon)
- What to show on prediction plot: enrollment, event, dropout, and/or ongoing
- Whether to predict by treatment
- The number of simulations to be conducted
- The random seed used to initiate the simulations
- The event model for the study (e.g., exponential, Weibull, log-logistic, log-normal, piecewise exponential, model averaging, or spline)
- The dropout model for the study (e.g., none, exponential, Weibull, log-logistic, log-normal, piecewise exponential, model averaging, or spline)

The following output will be produced:

- Summary of observed data in terms of the trial start date, data cutoff date, days since trial start, the current number of subjects, events, dropouts, and ongoing subjects
- Plot of the observed cumulative number of subjects enrolled and cumulative number of events
- Kaplan-Meier plot for time to event
- Kaplan-Meier plot for time to dropout
- Plot depicting the event model fit
- Plot depicting the dropout model fit
- Predicted time from cutoff until reaching the target number of events
- Plots of observed and predicted cumulative number of subjects enrolled and cumulative number of events over time

Both summary data and subject data are available for download. Except for the input data set, the user inputs can be saved and reused later.

7 References

Emilia Bagiella and Daniel F. Heitjan. Predicting analysis times in randomized clinical trials. *Stat in Med.* 2001; 20:2055-2063.

Xiaoxi Zhang and Qi Long. Stochastic modeling and prediction for accrual in clinical trials. *Stat in Med.* 2010; 29:649-658.

Patrick Royston and Mahesh K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat in Med.* 2002; 21:2175-2197.