



# Time-to-event modeling with the MonolixSuite: a TTE model for the NCCTG lung cancer study

[Download data set only](#) | [Download all Monolix project files](#) | [Download simulx R script](#)

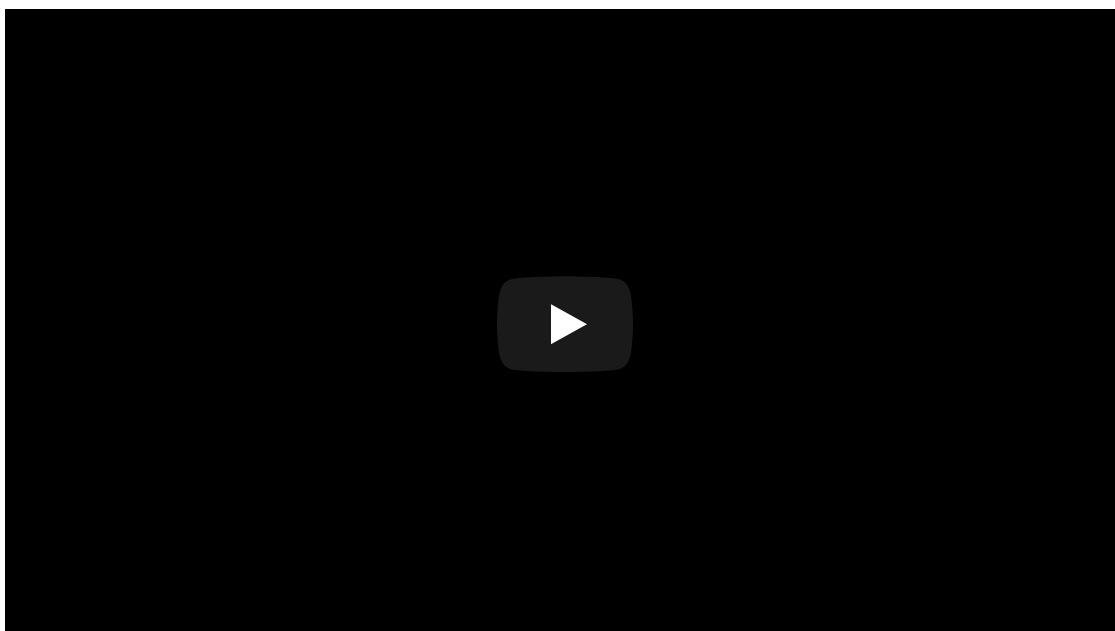


This case study is a detailed modeling and simulation workflow on a TTE data set. It is recommended to read before the [Introduction to TTE data and library of models in Monolix](#).

Another case study on time-to-event data is also available: [Case study on veteran lung cancer data set](#)

In this case study, we develop a model for the NCCTG survival data set and study the effect of covariates.

- [Introduction](#)
- [Data set visualization](#)
- [Modeling with Monolix](#)
- [Simulations with Simulx](#)
- [Conclusion](#)
- [Downloads](#)



---

## Introduction

The North Central Cancer Treatment Group (NCCTG) data set records **the survival of 228 patients with advanced lung cancer**, together with assessments of the patients performance status measured either by the physician and by the patients themselves. **The goal of the study was to determine whether patients self-assessment could provide prognostic information** complementary to the physician's assessment.

This data set has been originally presented and analyzed in:



Loprinzi et al. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 12(3), 601-607.

In this case study, we will **test several parametric models** to capture this data set and evaluate the prognostic performance of the recorded covariates.

## Data set visualization

The data set contains 228 patients, including 63 patients that are right censored (patients that left the study before their death). The original data set has been reformatted according to [the data set formatting guidelines](#) and includes both the starting time and the time of death or drop out.

To assess the importance of self performance assessment versus physician's assessment, the following covariates have been recorded:

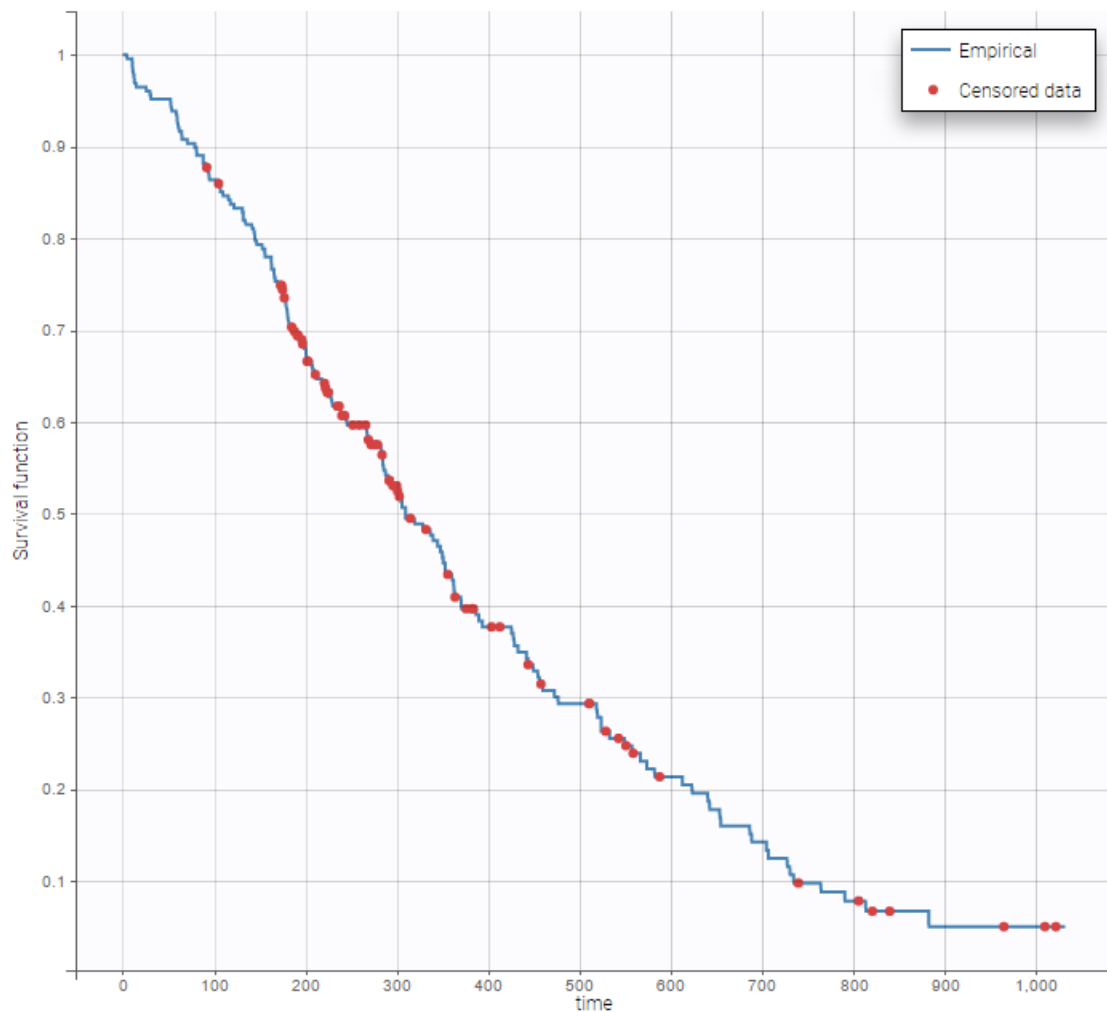
- **ecogPH**: ECOG (Eastern Cooperative Oncology Group) performance status assessed by the physician, on a scale from 0 (fully active) to 5 (dead). For information on the scale, click [here](#).
- **karnoPH**: Karnofsky performance status, assessed by the physician, on a scale from 0 (dead) to 100 (completely healthy). More details about the scale can be found [here](#).
- **karnoPAT**: Karnofsky performance status, assessed by the patient
- **sex**: sex of the patient (F for female, M for male)
- **age**: age of the patient (years)

ID	TIME	Y	age	sex	ecogPH	karnoPH	karnoPAT
1	0	0	74	M	1	90	100
1	306	1	74	M	1	90	100
2	0	0	68	M	0	90	90
2	455	1	68	M	0	90	90
3	0	0	56	M	0	90	90
3	1010	0	56	M	0	90	90

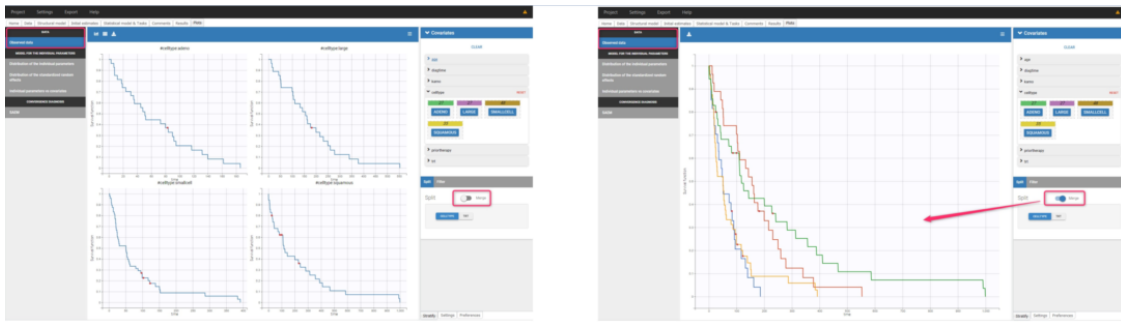
We first use Monolix to visualize the data set. After having opened Monolix,

we **start a new project and load the data** (the covariates must be assigned to COV for continuous covariates or CAT for categorical covariates). Before using the data visualization feature of Monolix, we must indicate that the data is TTE data. This is done via the choice of a structural model, that defines an event as output. For the moment, we can just select any model from [the TTE model library](#) via the model selection window that opens when clicking on “model file”. We next click on the **data visualization button**, next to Data button.

The **Kaplan-Meier (KM) estimate of the survival curve** as well as the mean number of events curve appear in the figure window. In the “Settings”, we can choose to remove the mean number of event curve and **display the censored data**.

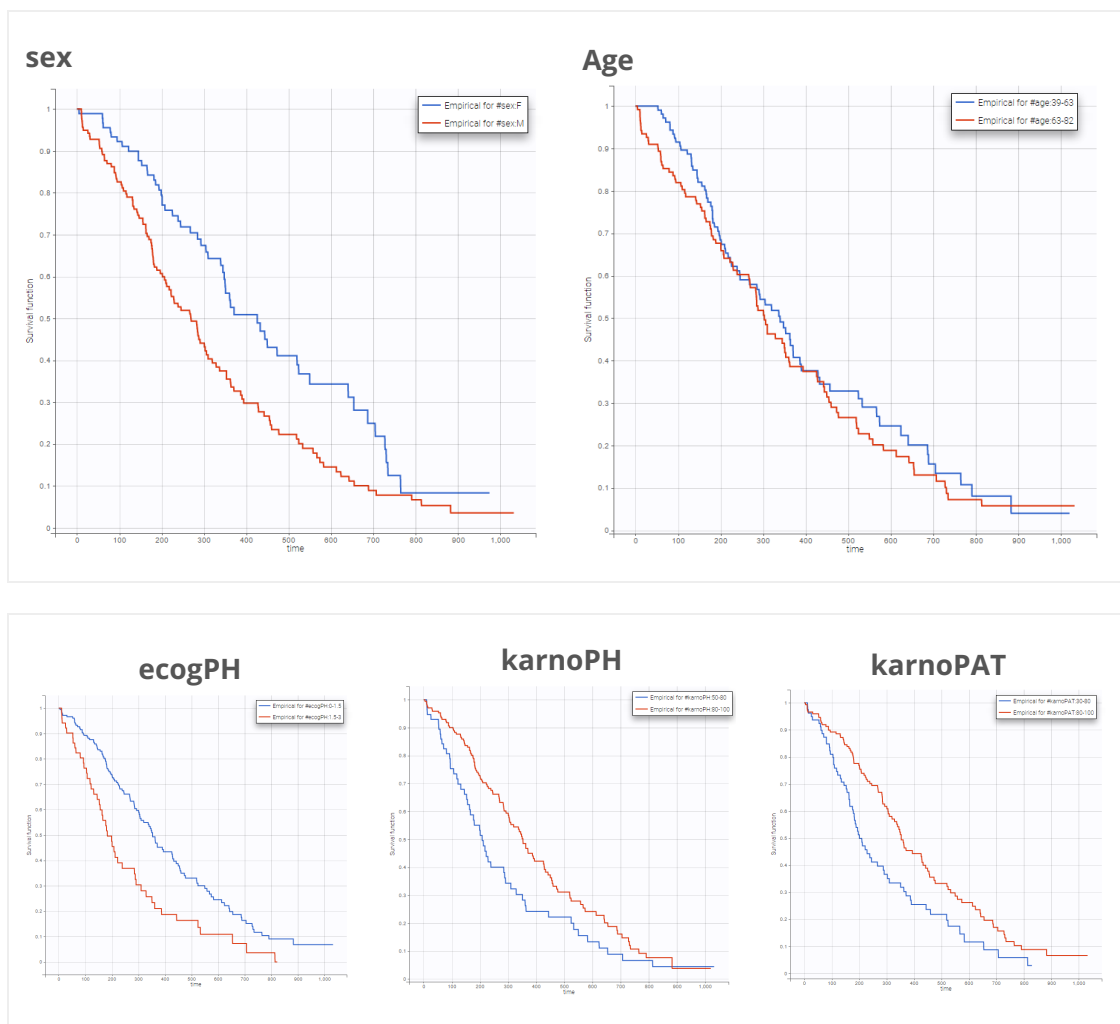


**Remark:** In Datxplore KM curves can not be stratified by color. However, in the Monolix plot “Observed data” (in the Plots tab), you can split a KM curve and then a “merge” button appears. It will merge the plots on a one figure with different colors (fixed) for each curve.



The plot for “Observed data” is available without running Monolix tasks, but it requires choosing a model. You can take the simplest model from the TTE library, if you are interested only in the data visualization. Two following figures below are obtained in Monolix with the “merge” option.

In the “Stratify” part, we can **split the KM curve according to categorical covariates, or categorized continuous covariates** (groups can be changed), in order to visually check the impact of the covariates. At first sight, it seems that sex, ECOG, karnoPH and karnoPAT influence the survival, while age does not.



Note that ECOG, karnoPH and karnoPAT are all performance scores and that they are probably strongly correlated.

## Modeling with Monolix

We now would like to **develop a model for this data set**, and next analyze which

covariates have the highest prognostic performance. Within the MonolixSuite, only parametric models are possible (no non-parametric or semi-parametric approaches).

## Structural model

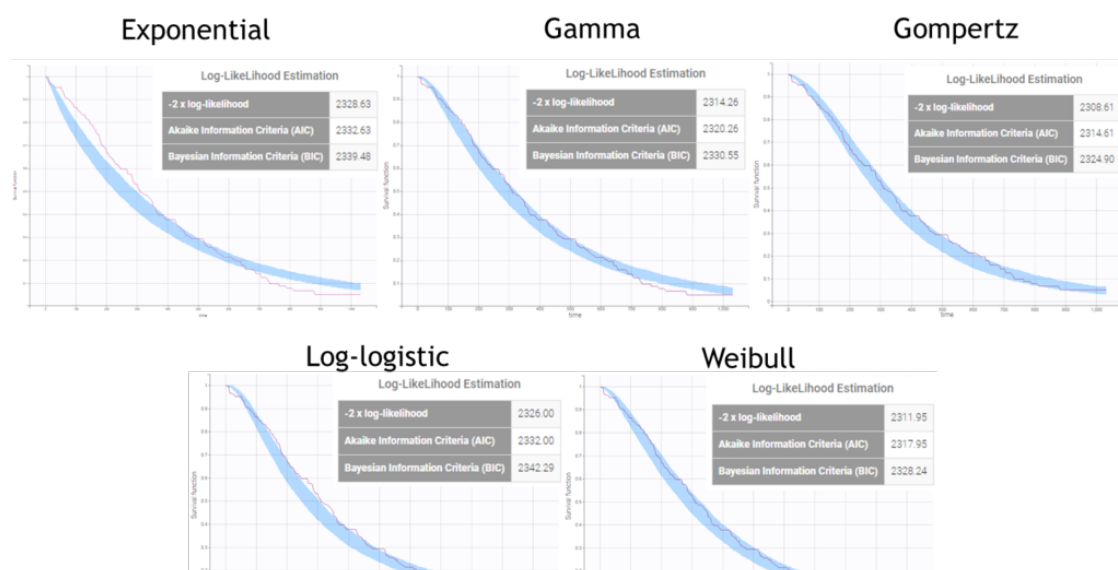
We start with the development of the structural model. The typical shape of common parametric survival models has been presented in the Part 1 of the TTE case study, together with the [library of common TTE models](#). **From the KM curve visualization, the Weibull, log-logistic, gamma or Gompertz models could be appropriate.** We will try each in turn, choosing the files with extensions “\_singleEvent.txt” from the library.

For models with 2 parameters, **it is common to assume that the shape parameter is the same for all individuals, while the scale parameter can vary from individual to individual.** This inter-individual variability can be captured via the effect of covariates and/or via random effects (often called frailty models). Following this strategy, the random effects are disabled for the shape parameters  $p$ ,  $s$ ,  $k$  and  $\alpha$  (by clicking on the corresponding diagonal element of the variance-covariance matrix) and kept for the scale parameter  $T_e$ . We also keep the default log-normal distributions, to ensure positive values of the parameters.

In each model, the scale parameter is the characteristic time  $T_e$ . **Using the KM data visualization, we choose 300 as initial value**, i.e the value for which the survival is around 50%. For the shape parameter, we use as initial value 1, according to [the typical value chart given on the page of the TTE model library](#).

**We next launch the estimation tasks for each structural model in turn:** estimation of the population parameters, estimation of the standard errors via the Fisher Information Matrix, estimation of the individual parameter by mean, estimation of the log-likelihood and generation of the graphics. **The performance of each structural model can then be assessed and compared using the TTE graphic and the log-likelihood values.**

After each run, in the Time to Event data graphic, one can calculate the **90% prediction interval for the KM curve given the model** (in Settings > Prediction interval) and overlay it on top of the empirical KM curve. It can be interpreted in the same way as a VPC. A summary is presented below:





From a visual point of view, **the exponential, log-logistic and gamma models can be excluded as they do not capture the shape of the KM curve**. The Weibull and Gompertz models are satisfactory, with a slight preference for the Gompertz model, as indicated by the BIC values. **We thus choose the Gompertz model as structural model.**

## Covariate model

**We next investigate if considering covariates on the  $T_e$  parameter can help explaining its inter-individual variability.** For didactic purposes, we will perform the covariate search by hand and stepwise.

**In Monolix, using a backward covariate search approach is especially powerful.** Indeed, after having calculated the s.e, a Wald test is performed to test the significance of each covariate. Thus it is sufficient to estimate the parameters of the model including all covariates relationships to get a p-value for each relationship, without having to estimate the submodels with one covariate relationship less. **Following this strategy, we will estimate the model with all available covariates on  $T_e$  and stepwise remove the less significant relationship, until all remaining covariates are significant.** The AIC and BIC will also be monitored in parallel.

We thus add all covariates on the  $T_e$  parameter in the "Covariate model" section, which corresponds to the following model for  $T_e$ :

$$T_{e,i} = T_{e,\text{pop}} e^{\beta_{\text{sex}}[\text{if sex=M}] + \beta_{\text{age}} \times \text{age} + \beta_{\text{ecogPH}} \text{ecogPH} + \beta_{\text{karnoPH}} \text{karnoPH} + \beta_{\text{karnoPAT}} \text{karnoPAT}} e^{\eta_i}$$

The table below summarizes the stepwise covariate removal, based on the p-values, and AIC/BIC:

Covariates on $T_e$					AIC	BIC	highest p-value
age	sex	ecogPH	karnoPH	karnoPAT			
x	x	x	x	x	2288	2315	0.41 for karnoPH => to remove
x	x	x	-	x	2285	2309	0.68 for age => to remove
-	x	x	-	x	2283	2304	0.043 for karnoPAT => at limit
-	x	x	-	-	2285	2303	all significant
-	x	-	-	x	2290	2308	all significant

The models with (sex, ecogPH, karnoPAT) and (sex, ecogPH) have similar AIC and BIC values. **Yet the model with (sex, ecogPH, karnoPAT) has a high condition number (around 300), we thus prefer the (sex, ecogPH) model.**

## Final model

Our final model includes a Gompertz structural model, and the covariates sex, and ecogPH on the scale parameter  $T_e$ . The model improvement when karnoPAT is included in addition to ecogPH is very small indicating that **a self-assessment of the performance status by the patient permits only a slightly better prognosis, compared to using the physicians ECOG performance status evaluation only**. In the original study including more patients and different types of cancer, the value of patient self-assessment was higher.

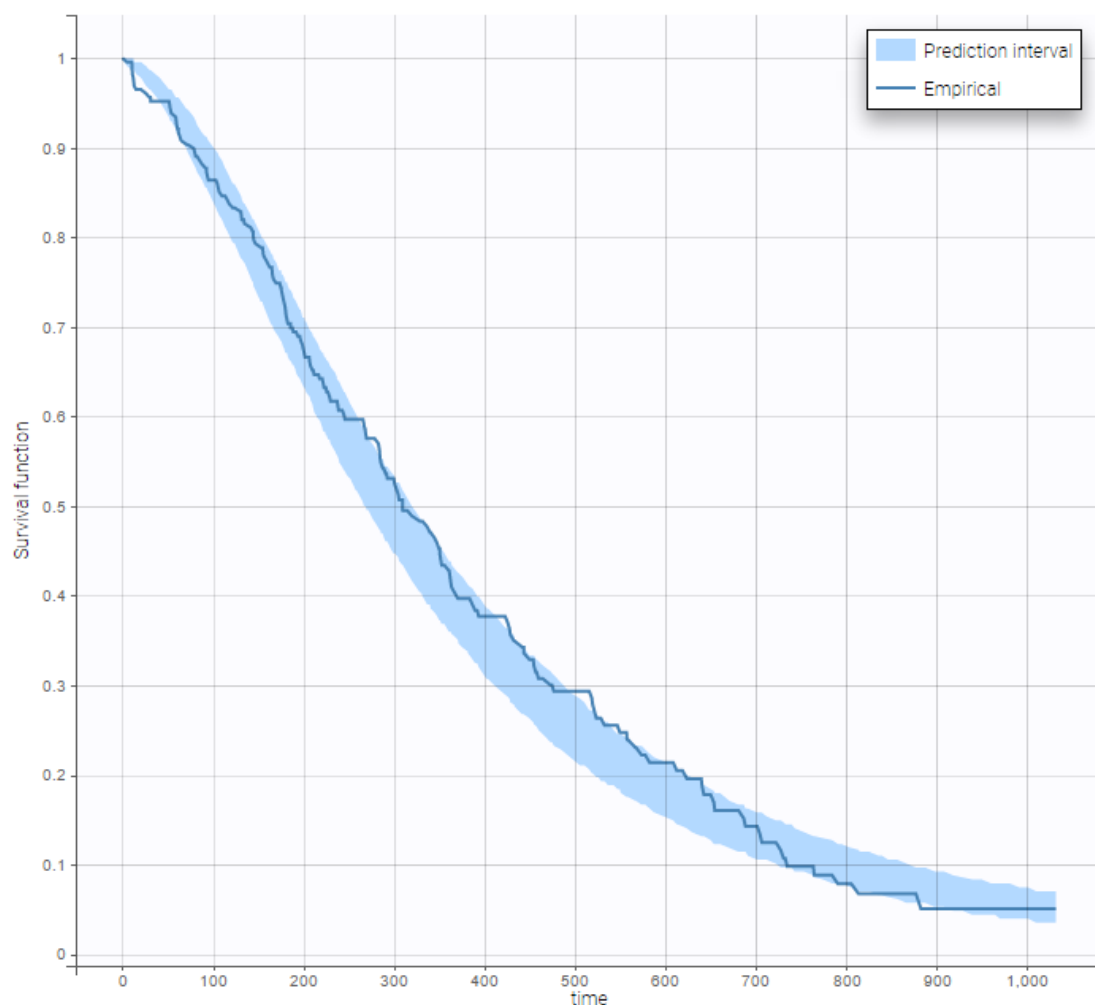
The estimated parameters are below. The r.s.e are reasonable.

## STOCH. APPROX.

	VALUES	S.E.	R.S.E.(%)	P-VALUE
<b>Fixed Effects</b>				
Te_pop	611	79.2	13	
beta_Te_ecogPH	-0.369	0.08	21.7	4.03e-06
beta_Te_sex_M	-0.391	0.117	30	0.000843
k_pop	0.114	0.0531	46.5	
<b>Standard Deviation of the Random Effects</b>				
omega_Te	0.524	0.0629	12	



The 90% prediction interval for the KM curve shows that the data is properly captured. The other graphics do not hint at any model mis-specification.

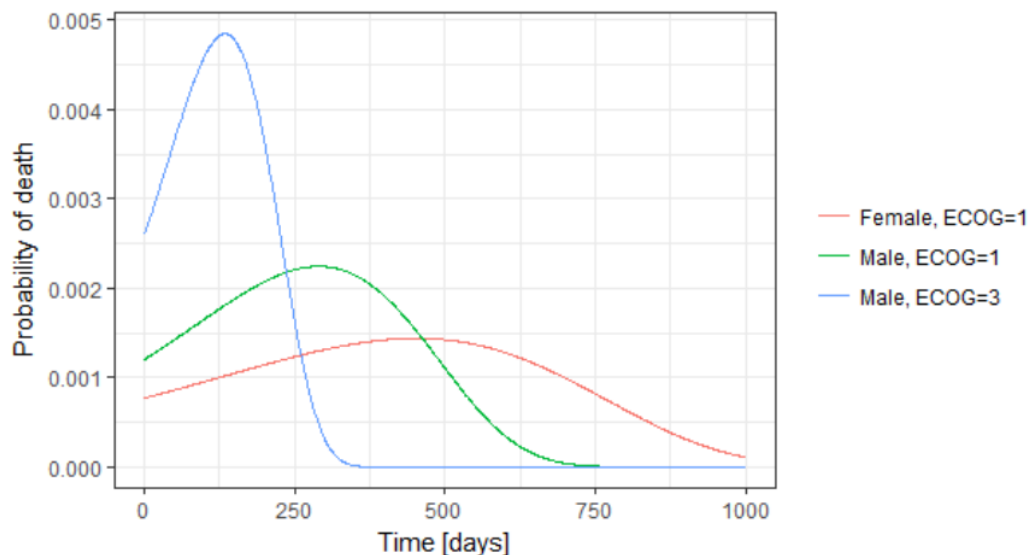


For a given sex and a given ecogPH score, we can easily calculate analytically the typical hazard and the associated Gompertz model survival. **From the survival function, we calculate the probability to survive at least 6 months (180 days), 1 year (365 days), or 2 years (730 days) depending on the sex and ECOG score:**

		Probability to survive at least		
		6 months	1 year	2 years
Male	ECOG 0	82%	57%	9%
	ECOG 1	71%	31%	≈ 0%
	ECOG 2	53%	5%	≈ 0%

	ECOG 3	26	$\approx 0\%$	$\approx 0\%$
Female	ECOG 0	89%	75%	41%
	ECOG 1	83%	60%	13%
	ECOG 2	73%	35%	0.2%
	ECOG 3	56%	8%	$\approx 0\%$

The probability density function of the death event can also be easily calculated analytically as the product of the hazard and the survival functions, for a given sex and ECOG score. Below we show **the probability of death with respect to time for three different cases**. The plot can also be interpreted as the distribution of death times in each of the three sub-populations.



## Simulations using Simulx

Simulx is part of the [mlxR package](#). To run the scripts, mlxR version  $\geq 3.3.1$  is required.

**We would like to simulate three new patients cohorts (with different covariates compared to the original data set).** The three cohorts have the following characteristics:

- cohort 1: 50% male / 50% female, good ECOG scores (0 or 1)
- cohort 2: 50% male / 50% female, poorer ECOG scores (2 or 3)
- cohort 3: 90% male / 10% female, good ECOG score (0 and 1)

To simulate these three cohorts, we create three groups, each defined via a data frame of the individuals covariates, and pass them as simulx input argument. **The simulation is done via the simulx function** which returns a R object containing the result of the simulation (time of death for each individual). We can then pass this object to `kmplotmlx` to obtain the Kaplan-meier survival curves.

For cohort 1, the R code reads:

```
# defining the path to the mlxtran project file
project.file <- "../monolix_project/43_gompertz_all_nokarnoPH_noage"
```



```
#===== defining group 1
# covariate data frame for group 1, with column id and one column
cov1 <- data.frame(id=1:228,
  sex = c(rep("F",114),rep("M",114)),
  ecogPH = rbinom(n=228, size=1, prob=0.5),
  age = NaN,          #unused in model but must be present
  karnoPH = NaN,      #unused in model but must be present
  karnoPAT = NaN) #unused in model but must be present

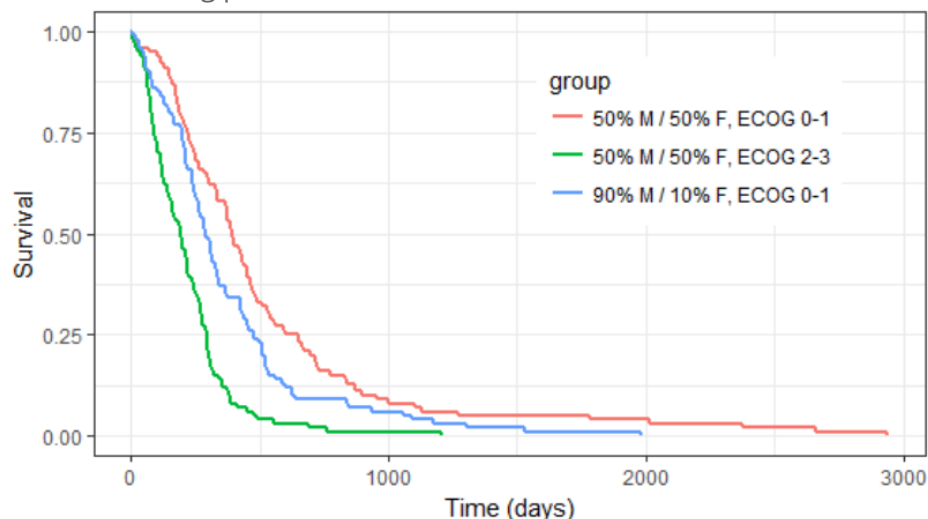
#===== calling simulx for group 1
res1 <- simulx(project = project.file,
  parameter = cov1)
```

We then combine the output objects into one and plot with:



```
#===== plotting the KM survival curve
group.labels <- c("50% M/50% F, ECOG 0-1", "50% M/ 50% F, ECOG 2-3",
  "90% M / 10% F, ECOG 0-1")
kmplotmlx(EvRes,labels=group.labels,facet=FALSE) + xlab("Time (day)")
theme(legend.justification=c(1,1), legend.position=c(0.9,0.9))
```

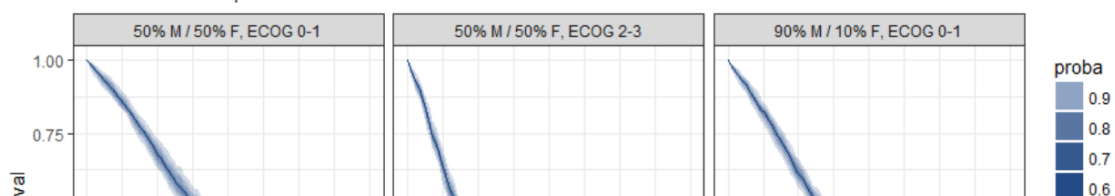
We obtain the following prediction:

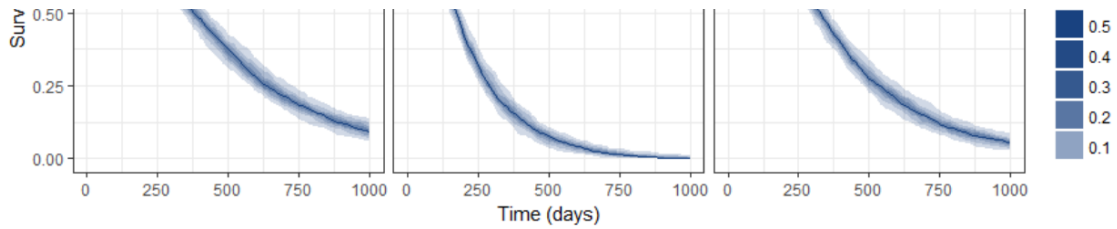


Because the time of death is a random variable, several simulations of the same population will lead to different survival curves. **We can assess the uncertainty of the survival curve by doing replicates.** This can be done very easily by adding an argument `nrep`:

```
res1 <- simulx(project = project.file,
  parameter = cov1,
  nrep = 100)
```

We then obtain a prediction interval for the three survival curves:







## Conclusion

The MonolixSuite is a powerful tool for modeling and simulation of time-to-event data via a parametric approach. **Covariates, frailty models and censoring can easily be incorporated.** Build-in statistical tests and diagnostic plots (in particular the “visual predictive check for TTE data”) render the model development process straightforward.

With the lung cancer data set of this case study, the risk of death increases over time (Gompertz distribution of death times). **Sex and the ECOG performance score are significant covariates and thus prognostic factors.** Performance scores assessed by patients rather than physicians can also serve as prognostic factors but their utility in addition to the physicians measured score is small.

Thank to the parametric formulation of the model, **survival probabilities depending on the sex and ECOG score can easily be computed.** In addition, **simulations of cohorts with combination of covariates can be performed using Simulx** and the uncertainty of the resulting survival curve can be visualized.

Overview   Data, model and mapping   Initialization   Tasks and results  
Model building   Plots   R-functions   Case studies   FAQ    



A web site of the network [Lixoft.com](https://lixoft.com) | Follow us on [LinkedIn](#)