Case Study #1

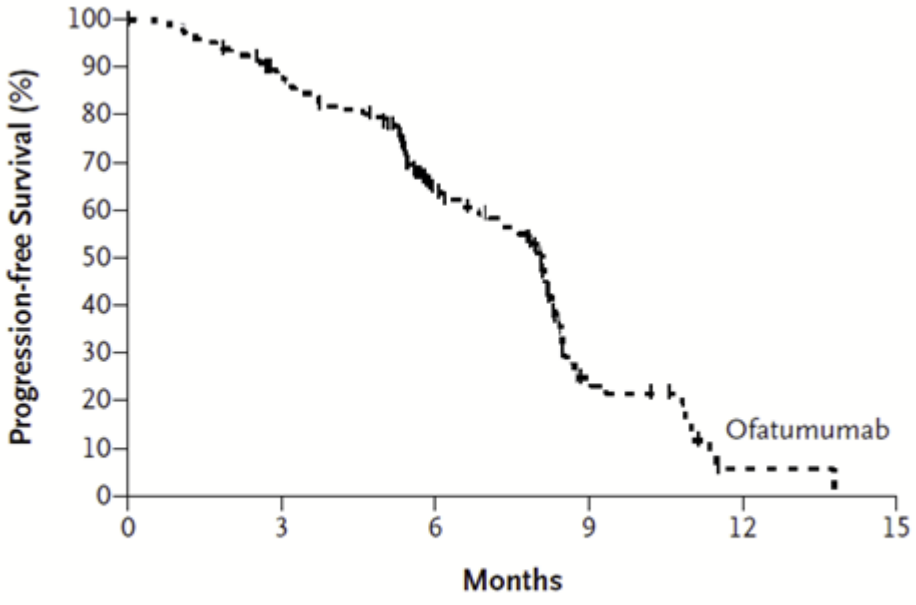
Chart/Plot Digitization

During data sourcing, the desired data may not always be available in the required format. Some datasets may only be accessible in image form, which must be converted to a tabular format with caution, as even minor errors could lead to significant issues.

We suggest using WebPlotDigitizer, an open-access tool, to digitize charts and plots into a tabular format. It's an efficient solution for data digitization and extraction processes. You can find useful information about the tool at:

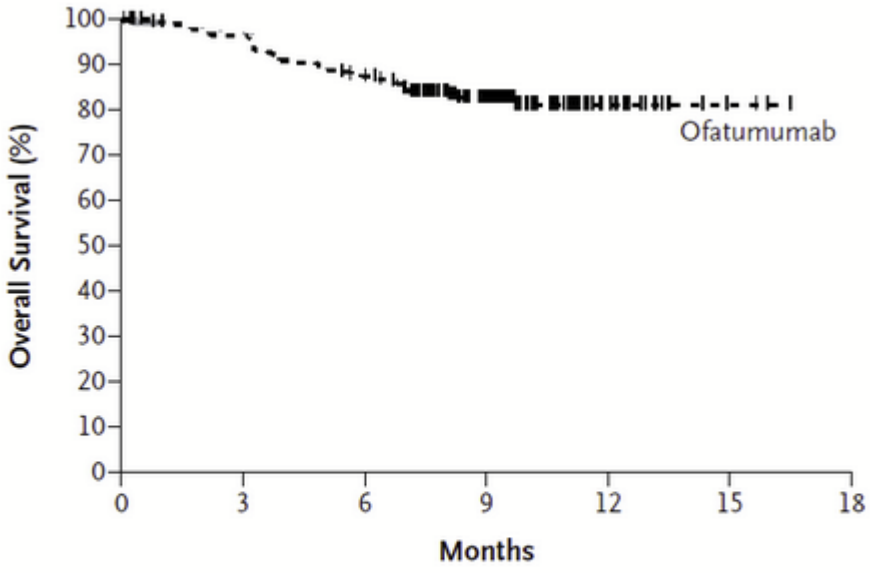
- User Manual
- Tutorials

Progression-Free Survival



No. at Risk						
Ofatumumab	196	161	83	15	1	0

Overall Survival



No. at Risk						
Ofatumumab	196	183	164	88	21	3

Challenge

In this case study, you are asked to digitize the above two charts using WebPlotDigitizer.

1. The output files should be in `.csv` format.
2. The files should include only two columns:
   a. `Time`: *Integer (1, 2, 3, ...)*
   b. `SurvivalProbability`: *Float with three decimal places (1.000, 0.998, 0.984, ...)*
3. Be sure that you extracted monthly data.
4. There may be some errors after the digitization process. You can manually fix those points when needed.

Case Study #2

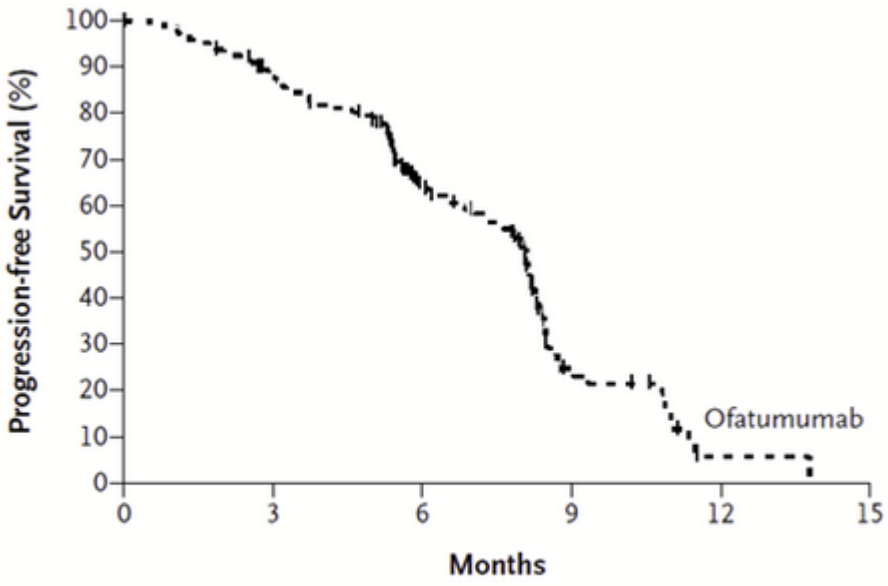
Survival Analysis

Survival analysis is a branch of statistics for analyzing the expected duration of time until one event occurs, such as death in biological organisms and failure in mechanical systems. In health economics literature, it is mostly used for reporting clinical trial results and used to generate death and disease progression probabilities. In this case, you are asked to perform some basic tasks to generate time-dependent death and disease progression probabilities using different tools.

For more information about the survival analysis and its use areas you can checkout these sources:

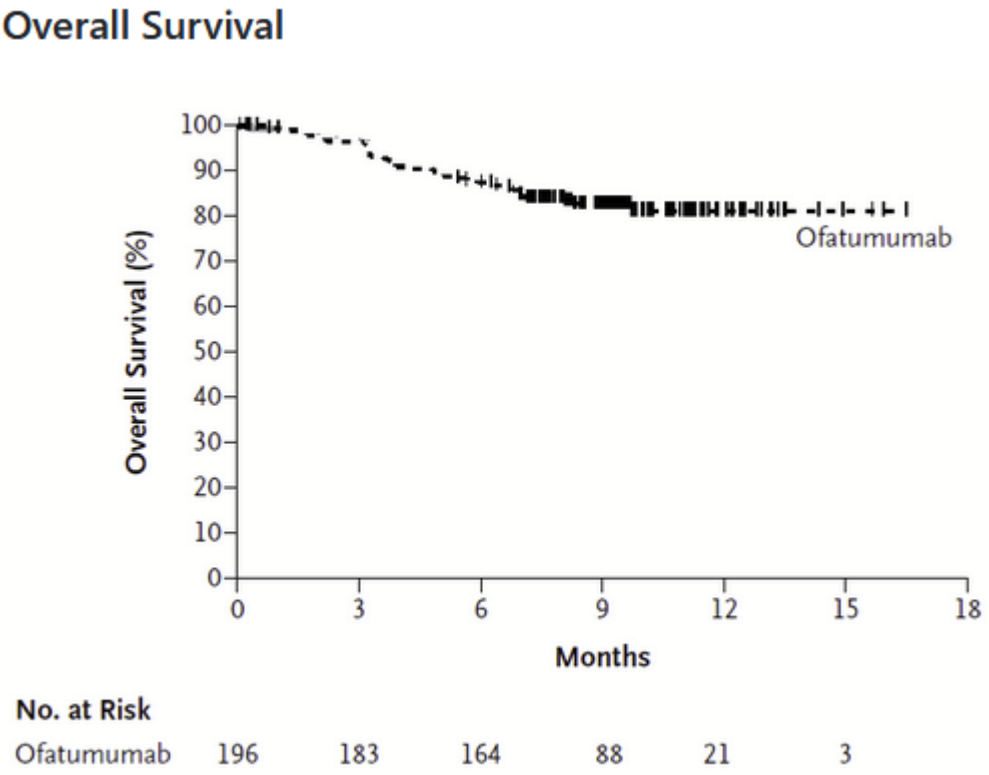
- Survival Analysis Part I: Basic concepts and first analyses, Clark et al, 2003
- Survival analysis - Wiki

Progression-Free Survival



No. at Risk						
Ofatumumab	196	161	83	15	1	0

Overall Survival



Challenge

In this case study, you are asked to conduct a survival analysis to generate input data that will be used to build a microsimulation model for a specific type of cancer. Building the microsimulation model is not part of this case study.

Our initial step is to gather time-dependent survival probabilities for patients, utilizing published clinical data as our primary source. The figure below, taken from a trial paper, will serve as the main data source for probability calculations.

Digitized survival data is the main prerequisite of this case. Please use the digitized data In Case Study #1 and proceed with the next steps then.

Question-1

Using the digitized data, please generate individual patient level data using the technique outlined by Guyot et al.

Question-2

Using the individual data generated in Question 1, please find the best fitting distributions for each curve using survival analysis techniques.

Question-3

Using the best fitting distribution Question 2, please generate time-dependent death progression probabilities for a period of 250 months. The output files should be in csv format and should include the following columns:

- **Time**
- **Probability**

Note that, progression-free survival curve has both 'death' and 'progression' events while overall survival has only 'death'. To calculate progression probability, you should first calculate the death probability then subtract it from the event occurrence probability calculated from the progression-free survival curve.

Notes:

- *Please use R or Python to process data and do not do any manual work on csv/xls/xlsx files.*
- *Please use a naming convention for variables/functions in your code. Please refer to: https://medium.com/wix-engineering/naming-convention-8-basic-rules-for-any-piece-of-code-c4c5f65b0c09*
- *Please use a naming convention for column names (at least for the final datasets).*
- *Make sure that your code generates reproducible results. If you have any randomness in your code, please use a seed. Please refer to: https://en.wikipedia.org/wiki/Random_seed*
- *You can utilize lifelines package to generate time dependent probabilities requested in Question-3.*