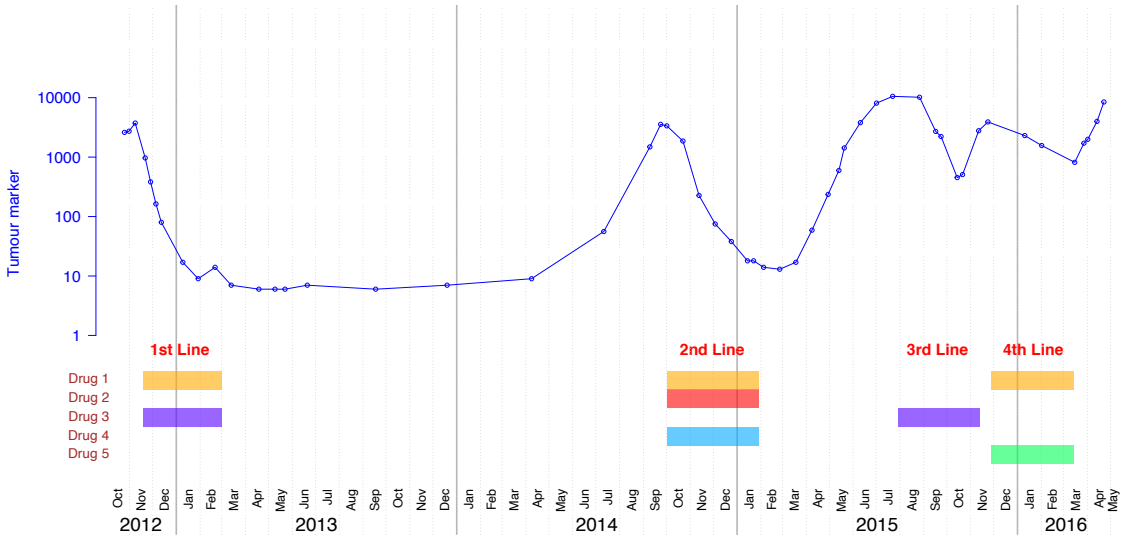
Presentation of the data to analyse

1 Background

line 1 of patient 1 is not the same for line 1 for patient 2

The following Figure shows the level of a tumour marker for a patient (y-axis, blue colour) as a function of time (x-axis, years and months) and treatment (coloured horizontal bars). This patient was diagnosed with a stage IV cancer in October 2012. A sample of the tumour tissue (biopsy) was immediately taken, and this patient had his first line of chemotherapy treatment, which consisted in the combination of drugs '1' and '3', between November 2012 and February 2013. The level of the tumour marker (on the log2 scale) strongly decreased with the treatment. The tumour was then carefully monitored. It was noted that the tumour was growing in July 2014 so that the patient had a second line of chemotherapy (composed of drugs '1', '2' and '4') from October 2014 to January 2015, leading to a strong decrease of the level of the tumour marker. The tumour progressively increased leading to a 3rd chemotherapy treatment (exclusively composed of drug '3') from August to October 2015. During this treatment line, the level of a tumour marker initially decreased before to increase again so that a 4th line of chemotherapy (composed of drugs '1' and '5') was considered. The patient died in May 2016.



2 Data

The dataset 'Chemotherapy.csv' contains the following information:

- **patient**: a unique patient id,
- **tumour**: level of the tumour marker increase (if positive) or shrinkage (if negative) over a given month of treatment (on the log2 scale), defined as the level of the tumour marker at the end of the month minus its level at the start of the month of interest,
- **line**: chemotherapy line. a patient may have several treatments over years, starting with 1st line, then 2nd line a few months later, aso.
- **month**: month of treatment for a given line of chemotherapy. Standard chemotherapy treatment last 3 months (per line) but may be shorter/longer depending on the patient health and on the drug combination,
- **sensitivity**: each patient had a biopsy of the tumour after diagnosis. The sensitivity of the tumour to each drug combination was assessed in-vitro. The score is high (around 1 and above) for tumour samples which are resistant to the drug combination of interest and low (around 0) for tumour samples which are sensitive to the combination of drug of interest.

3 Analysis

It is believed that the in-vitro assessment of the sensitivity of the tumour to a drug combination is a reliable predictor of the treatment effect of the same drug combination on the patient. However, the sensitivity score is believe to best work for initial chemotherapy lines (as the tumour changes afterwards and become resistant to the initial drug combination).

You have been asked by the group leader of a research group at CRUK to check

- if the in-vitro sensitivity score can be used to predict the treatment effect on the patient at the hospital,
- if the in-vitro sensitivity score decreases with time and is stronger for line 1 than for lines 2 and 3+ (you can consider the variable 'line' as a 3-level factor with levels '1st line', '2nd line' and '3rd line and more').

In a report (maximum of two one-sided A4 pages to which you will add an appendix including R outputs and important figures), describe the different steps of your data analysis using the methodology taught in the course and the practicals, and provide an answer to the research question using the dataset 'Chemotherapy.csv'.

Further comments:

- if you decide to perform the analysis with other students, please mention the name of all students of your group on the report,
- when referring to R outputs and figures, make sure to use a clear reference system, like 'left plot of Figure A3' (i.e., Figure 3 of the Appendix), for example,
- send your analysis report by email to the teaching assistant (Younes.Boulaguiem@unige.ch) and Lecturer (Dominique-Laurent.Couturier@unige.ch) with each member of your group cced before **2pm (Swiss time) on the 14th of January 2021** (reports received afterwards will *not* be considered). The teaching assistant will confirm reception of your report by means of a reply to all.
- we strongly advise you to follow the analysis steps described in the following sections.

4 Suggested procedure for data analysis

The procedure of for data analysis highly depends on the aims and questions which are at the origin of the data collect and on the theoretical background available for the relevant research field.

The analysis is not the same if the original question is about a significant difference between the levels of a fixed factor or if the goal is to determine the variables which influence significantly the response. The analysis is also different if it is known that a covariate enters in a quadratic form in the model or if the model is completely unknown.

Thus, the procedure that we suggest here isn't a rule of thumb that one may apply in every situations but a general approach which must be adapted to each analysis.

4.1 First look of the data

- Define:
  - what are the independent experiment units (subjects, households, lots, ...)
  - the role of the different variables in your analysis:
    - which variable is the response ?
    - which variables are covariates ?
    - which variables are factors ?
    - Are the covariates and factors fixed or random effects ? (justify)
    - Are the factors crossed, nested, ... ?

- Check if the class of your variables correspond to your needs:
  - class **numeric** for the response
  - class **numeric** or **integer** covariates like age, ...
  - class **factor** or **ordered** for factors like subject, gender, ...
- Write a first version of your model that include potential interactions

4.2 Exploratory data analysis

- check if the model's assumptions hold:
  - linearity of the relation between the covariates and the response
  - equal variance of the responses around the different factor's levels
  - presence or absence of interactions between the different "explanatory variables"
  - correlated or uncorrelated random effects ?
- get an impression of the variables potentially explanatory

4.3 Model building and hypotheses

- Rewrite your model to take into account the results of the exploratory data analysis. If you have a doubt about the presence of a parameter (interaction between fixed factors, correlation between random effects), include it in the full model. It's pertinence in the model will be tested.
- Define precisely the hypotheses you want to test depending on the questions you wish to answer (Are you interested on the main effect of a factor or on its marginal effect ?).

4.4 Model estimation

- Estimate the postulated model.

4.5 Diagnostic analyses of the full model

Before analyzing the results of your estimations, you have to determine the model is valid

- check the residuals : are they iid $N(0, \sigma^2_\epsilon)$?
- check the random effects : are they iid $N(0, \sigma^2)$?

If not, try

- a scale's transformation of some variable (log of the response in the case of heteroscedasticity for example)
- a quadratic form if non linearity of the influence of a covariate is suspected
- to relax the assumption of equal variance of the observations around the levels of a factor
- ...

4.6 Model selection

Once the model is estimated and its validity checked, you may wish to determine with likelihood ratio tests if some of its parameters are significant or not with the aim of selecting a more parsimonious model. Suggestions:

- if the variance of random effect is small compared to the residual variance, this random effect may not be useful.
- if the means of the responses for the levels of a factor seem the same, a model without that fixed factor may be more adequate.
- ...

The selected model should be re-checked as in 2.5.

4.7 Hypotheses

Once you have selected the appropriate model, test the hypotheses you defined earlier.