**EBMT Statistical Committee Meeting: Missing data group**

**Friday 10th of May from 10:15 to 15:45, Leiden**

**LUMC University (room location to be define)**

**Participants:**

From Paris: Ariane Boumendil (AB), Christophe Peczynski (CP), Jacques-Emmanuel Galimard (JEG) and Matthieu Resche-Rigon (MRR).

From Leiden/Roma: Simona Iacobelli (SI), Giulia Sbianchi (GS), Vittoria Malpassuti (VM), Dirk-Jan Eikema (DJE), Junfeng Wang (JW, Liesbeth de Wreede (LdW), Hein Putter (HP), Ed Bonneville (EB) and Bart Mertens (BM).

**Main focus of the study group**

Dealing with missing covariates in the context of multiple endpoints (OS, PFS, RI, NRM and GVHD)

In the EBMT context, several outcomes are analysed, namely: OS, PFS, RI and NRM, and GVHD (acute, chronic and/or extensive). To handle adequately missing data on covariates, all endpoints have to be taken into account. In the case of survival endpoints, using likelihood based method or multiple imputations; it is not an easy issue.

**Agenda**

1. Examples of EBMT studies with missing data and method used (ALL)

A/ Cytogenetics in AML patients

Cytogenetics is often mandatory in the analysis of AML patients and has often a large rate of missing.

Example: Comparison of outcomes of BM without ATG and PBSC with ATG in AML patients receiving a first allogeneic SCT (JEG)

In this study, cytogenetics is missing for approximatively 30% of patients; the complete case analysis thus requires to exclude 30% of the patients.

We present the comparison of the CC analysis with two other methods to handle missings,

1 - using a NA-level

This method makes it difficult to interpret the coefficient associated to the imputed covariate. Does it bias the results if used only for covariates adjustment in a multivariate analysis?

2 – Multiple imputation

For a single endpoint, the Nelson-Aalen estimator is used in the imputation model. In the competing risk setting, there is no consensus in the literature. However, some authors proposed to include two Nelson Aalen estimations for each cause-specific model.

B/ Number of treatment lines before transplant for auto SCT in lymphoma patients.

C/ Other ??

1. Practical session in R on MI (DJE)
2. Litterature review (ALL)

A brief review of main suggestions on missing values and methods (*MI*  and alternatives) based on existing literature should be conducted.

Focus on methods to handle missing in the context of multiple endpoints/competing risks.

To be done by a student/PhD ?

1. Simulation plan (ALL)

Consequences of missing values in covariates at baseline, with reference to the analysis of multiple endpoints. To be done by a student/PhD ?

To perform a simulation study, we need to determine the way to generate the data.

There are two possibilities:

* Use a real complete dataset and generate missing data
* Generate the entire dataset. It could prove difficult to generate several outcomes in the same procedure.

How to generate the missing data?

If we use a model, which model? Logistic? Probit?

How many variables with missing data?

What type of variable? Continuous, binary or categorical (ordered or not)?

Which method to compare?

* CC analysis
* NA-levels
* Multiple imputation
  + One Nelson-Aalen by outcome in the imputation model?
  + Which imputation model?
* Other …