A.D. 1308
unipg
DIPARTIMENTO
DI MEDICINA E CHIRURGIA

# PhD Program in Clinical and Molecular Medicine

## "Data Analysis and Biostatistics"

**Teacher: Prof. Giovanni Bellomo** (giovanni.bellomo@unipg.it)
**Date: October 2024**

### Description

The Data analysis and biostatistics course covers key statistical topics such as probability, tests, correlation, ANOVA, regression, power analysis, and sample size calculation. It also touches on practical machine learning applications for research with examples on R. The course balances theory with hands-on applications to refine analytical and computational skills crucial for doctoral candidates. The final group project lets participants apply their knowledge to create a data analysis pipeline for their own PhD projects.

### Course program

Lesson 1 (2h) - Probability and statistical tests

- Why do you need biostatistics?
- Probability density functions (Normal, Binomial, Student's) and statistical estimates (mean, median, variance, covariance etc.)
- Kolmogorov-Smirnov and Shapiro-Wilk normality tests.
- Parametric statistical tests: t-test
- Non-parametric statistical tests based on rank: U-test
- Examples on *R*

Lesson 2 (2h) – Multiple groups comparisons, covariate variables

- Pearson's and Spearman's correlation coefficients.
- One-way analysis of variance (ANOVA)
- Two-way ANOVA
- Analysis of covariance (ANCOVA)
- Linear regression
- Logistic regression
- Examples on *R*

Lesson 3 (2 h) – Power analysis & sample size calculation

- Type I & II statistical errors
- Power and effect size
- Sample size calculation (t-test, correlation, ANOVA)
- Introduction to survival analysis (Kaplan–Meier curves, Log-rank test, Cox regression)
- Examples on *R*

Lesson 4 (2h) – Introduction to machine learning 1

- Unsupervised vs supervised classifiers
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Clustering: EM-clustering, K-means, Hierarchical clustering
- Examples on *R*

Lesson 5 (2h) – Introduction to machine learning 2

- Confusion matrices
- Logistic regression and ROC analysis
- Support vector machines (SVM)
- Regularization techniques (Ridge, LASSO, Elastic-Net)
- Examples on *R*

Lesson 6 (2h) - Group exercise

Structure your own data analysis pipeline