

## Intelligent Data Analysis of Clinical Trials with Stata

Antonio Gouveia de Oliveira, M.D., Ph.D

Departamento de Bioestatísticas e Informática, Faculdade de Ciências Médicas,  
Universidade Nova de Lisboa, Lisboa, Portugal

Clinical trial statistical analysis and reporting is a formidable task. A final study report requires the creation of hundreds of tables and data listings, and the calculation of over one thousand statistical significance levels, difference estimates and confidence limits. Typically, several database programmers, statistical programmers and biostatisticians are needed to perform this task over a period of time that is measured in months. We describe the design approaches and the evaluation of an intelligent data analysis system that automates the creation of clinical trial statistical reports, DART, which is one component of an integrative Clinical Trials Information System. This application was developed in Stata programming language and has about 9,000 lines of code. This unsupervised knowledge-based system is able to select, according to the characteristics of the study design, the study statistical analysis plan and the type of baseline and efficacy variables used, which are all encoded and stored in the database, the statistical methods adequate for each analysis and the results that need to be reported. The entire process of data analysis and reporting can be performed automatically, or the user may specify some parameters of the analysis (e.g., scale transformations, adjustment for confounding). The application can handle commonly used statistical methods applied to clinical trials analyses for nominal, multi-valued, ordinal, interval and event/count data in one-, two- and multiple-arm trials, crossover studies and factorial designs, with or without stratification. It handles imputation of missing data, scale transformations and regrouping of study centers. It can automatically select baseline variables for inclusion as covariates, and to conduct post-stratification analyses and subgroup analyses. So far, DART has been successfully used for the automated statistical reporting of 35 pharmaceutical clinical trials. In a validation study, the statistical methods used in a random sample of 51 clinical trials published in *The New England Journal of Medicine* and in *The Lancet*, reporting 97 different analysis, the analytical methods were identical or equivalent to those selected by DART in 84.5%, different in 6.2% and not supported by DART in 9.3%.