Master's Internship Offer (or equivalent)

4 November 2026

Background:

The generation of synthetic data from real-world data has grown considerably, particularly with the development of machine learning methods and generative models. These methods make it possible to produce artificial data that reproduce, at least in part, the statistical properties of the data used to train them. In medical research, they offer promising opportunities, particularly for facilitating data sharing, developing new methodological approaches, and exploring innovative strategies for the design and analysis of clinical studies.

One particularly promising application is the use of virtual patients in clinical trials. Synthetic data could be used to simulate entirely virtual clinical trials, but also to supplement trials based on real-world data by artificially increasing their sample size. This second approach, known as an augmented clinical trial, could help improve the precision of estimates or explore the consequences of different recruitment strategies.

However, the use of synthetic data raises fundamental statistical questions. In particular, the generated data are not independent of the real-world data used to train the generator. They may also differ from the target population distribution, potentially introducing bias into the estimation of parameters of interest. In the context of clinical trials, a central question is therefore to determine the extent to which the inclusion of virtual patients can improve precision without compromising the validity of treatment effect estimates.

Work objectives:

Preliminary work is currently underway to study the statistical properties of clinical trials simulated entirely from synthetic data. The proposed internship aims to extend this work by focusing more specifically on augmented trials, combining real and synthetic patients. The objective will be to study the statistical properties of treatment effect estimators obtained from datasets combining real observations with synthetic observations generated by a model trained on the real data. The work will address several areas:

- Review and understanding of the existing work.
- Development of an augmented trial strategy: setting up a simulation framework to generate synthetic data using a generator trained on a sample of real-world data, and then constructing datasets combining different proportions of real and synthetic patients.
- Study of the statistical properties of estimators: evaluation of the performance of different treatment effect estimation strategies, focusing in particular on their bias, variance, mean squared error, and confidence interval coverage.

- Analysis of the influence of generator characteristics: study of the impact of generator quality, training sample size, and the proportion of synthetic data introduced into the trial on the properties of the estimators.
- Comparison of strategies to identify the conditions under which the use of synthetic data may offer statistical benefits.

The work will be based primarily on a simulation approach, involving the development of reproducible numerical experiments. Depending on the progress of the internship, an application to data from a clinical study may also be considered.

Keywords: Synthetic data; Augmented clinical trials; Treatment effect; Generative models; Statistical simulation.

Candidate profile:

Master's student (second year) or final-year engineering student in applied mathematics, with a focus on statistics, biostatistics, data science, or machine learning. The desired skills are:

- Statistics: good knowledge of inferential statistics, estimation, and simulation methods.
- Machine learning: knowledge of statistical learning methods, with an interest in generative models and synthetic data generation.
- Programming: good command of the R programming language and the ability to develop reproducible simulation programs.
- Personal qualities: scientific rigour, independence, curiosity, and an interest in methodological issues at the interface of statistics, artificial intelligence, and healthcare.

Internship details:

Host institution: Institut de Mathématiques de Toulouse

Supervisors: Nicolas SAVY - nicolas.savy@math.univ-toulouse.fr

Philippe SAINT-PIERRE - philippe.saint-pierre@math.univ-toulouse.fr

Duration: 4 to 6 months

The internship may lead to a PhD opportunity.