

# Internship and PhD Proposal

## Robust Determination of Similar Patient Cohorts from a Health Data Warehouse

**Keywords:** Multi-modal Medical Imaging, Text, Vision, Omics, Representation Learning

**Context** The construction of similar patient cohorts from a Health Data Warehouse (HDW) represents a major challenge for clinical research and personalized medicine. It enables the identification, for a given patient, of others sharing comparable characteristics, thereby supporting various applications: the creation of observational cohorts, the emulation of clinical trials, the selection of patients eligible for a therapeutic protocol, or even medical decision support.

The main difficulty is the definition of *similarity*, which can vary depending on the number of modalities (imaging, genetic, text, omics, etc.) and pathology (rare or common disease).

**Objectives** The goal of this internship, and possibly PhD thesis, will be:

- Benchmark current multi-modal foundation models on the APHP dataset consisting of textual (clinical reports) structured (laboratory results, diagnoses, prescriptions), and imaging data.
- Test existing distances/similarity measures (e.g., cosine, Mahalanobis) to compare patient profiles within this space
- Establish a new statistical framework to define optimal similarity thresholds ensuring robustness in all possible heterogeneous clinical situations (e.g., rare vs. common profiles) through adaptive mechanisms.
- The proposed method should be highly explainable, with clear reporting of the criteria that led to a patient's inclusion or exclusion from a similarity cohort, an essential requirement for medical use.

**When** As soon as possible for 6 months

**Team** The student will work at Télécom Paris and at the Hôpital Tenon (APHP). Internship supervisors will be Pietro Gori, Loïc Le Folgoc (Télécom Paris) and Christel Gerardin (APHP).

**Salary** Salary will be of around 600 euros/month

**Required background** Master student in applied mathematics, statistics, computer science, engineering with a good knowledge of Python and deep learning. Previous experience in Medical Imaging or Medical Data is not mandatory but highly valued.

**How to apply** Candidates are invited to send a CV to [pietro.gori@telecom-paris.fr](mailto:pietro.gori@telecom-paris.fr), [loic.lefolgoc@telecom-paris.fr](mailto:loic.lefolgoc@telecom-paris.fr) and [christel.gerardin@aphp.fr](mailto:christel.gerardin@aphp.fr) detailing their academic background with courses and grades.