Title of the thesis	Semi-supervised learning of high dimension composite information for precision medicine in DIAbetes and DYSlipidemia
Acronym	DIADYS
Reference number	021

Hosting institution	**Employer**
Université de Lille Website: https://www.univ-lille.fr/home/	Inria Website: https://www.inria.fr
Hosting research unit 1	**Hosting research unit 2**
Name: Laboratoire Paul Painlevé Acronym: LPP Identification number: UMR 8524 Address: Université de Lille Cité Scientifique Bâtiments M2 et M3 59625 Villeneuve d'Ascq Website: https://math.univ-lille1.fr/	Name: UMR 1283 Inserm/8199 CNRS - European Genomics Institute for Diabetes Acronym: EGID Identification number: UMR 1283/8199 Address: Université de Lille Faculté de Médecine, Pôle Recherche 1 place de Verdun 59045 Lille Cedex - France Website: http://www.egid.fr/1/home/
Principal supervisor	**Co-supervisor**
Name: Christophe Surname: BIERNACKI Email: Christophe.biernacki@inria.fr Phone: 03.59.57.78.58 / 06.62.56.61.60	Name: Martine Surname: VAXILLAIRE Email: martine.vaxilaire@cnrs.fr Phone: 03.74.00.81.03 / 06.77.74.18.88

Thesis information

Keywords	Model-based hybrid co-clustering, Unsupervised analysis, Type 2 Diabetes, Cardio-metabolic disease risk stratification, Precision medicine
Abstract	In order to cope with the acquisition of massive genomic data and the limitations of current tools to predict individual risk of common multifactorial diseases, there is a pressing need for innovative artificial intelligence methods integrating personal genome data in precision medicine. This is challenging in the case of cardiovascular diseases and their silent surrogates: diabetes mellitus and dyslipidemia. The **DIADYS project** aims to implement a multiple-layered approach supported by statistical-learning algorithms driving disease entities and risk prediction for cardio-metabolic disorders from high dimension multi-modal composite human datasets. Through a transdisciplinary research project, three main objectives will be followed up: **1/** to build up curated multi-modal composite human datasets (i.e. discrete variables with several levels of uncertainty) of quantitative traits (e.g. biological and anthropometric measures) and binary clinical outcomes, and of large-scale genomic data in independent population-based longitudinal studies of European ancestry and in patient cohorts with clinically defined diabetes and/or dyslipidemia (**Task-1**), **2/** to develop hybrid co-clustering models and related algorithms enabling to improve parameters of risk prediction and of patient stratification, and feature weighting at a population level (**Task-2**), and **3/** to assess the performance and robustness of the novel models in internal and external validation datasets, compare their accurateness against existing disease clusters and PRS, and to design novel integrated tools for decision making in precision medicine (**Task-3**).

	Here, the originality is to address the high dimension and mixed type features supervised problem related to the initial medical questioning in a specific model-based hybrid co-clustering context. Indeed, most of the related works are usually defined in a more empirical environment from a methodological point of view. Ambitiousness is thus to rely on the methodological fundament of the proposed approach to gain high visibility in the medical community. In addition, for statisticians/mathematicians it is an opportunity to address some general and recurrent methodological/theoretical problems (variable weight, sub-clustering) which can then be applicable beyond the initial medical motivation. Thanks to an innovative interdisciplinary research, the ambition of the **DIADYS** project is to develop usable workflows and novel integrated tools relied on mixed datasets for decision making in precision medicine, applicable beyond the medical field; provided that feasibility can be relevant for the medical community (especially for the field of cardio-metabolic diseases).
Expected profile of the candidate	Master degree in Data Science with some knowledge (**basic skills**) or at least interest in Genomics, **Human Genetics or Statistical Genetics**. Master degree in Biostatistics with a good knowledge in Theoretical Statistics. In both cases, a good skill for scientific programming is required (at least R or Python)
Application procedure	The application procedure is detailed on the European programme PEARL website www.pearl-phd-lille.eu. The funding is managed by the I-SITE ULNE foundation which is a partnership foundation between the University of Lille, Engineering schools, research organisms, the Institut Pasteur de Lille and the University hospital. The application file will have to be submitted before April 15, 2020 (10h Paris Time) and emailed to the following address : international@isite-ulne.fr.
Net salary and Lump Sum	A net salary of about €1,600 + €530 per month to cover mobility, travel and family costs.