PEPR digital health
Axis Methods and Models for Multimodal Data Integration (M4DI)

Defining patients' phenotype from a mixed approach based on data and expert knowledge

Keywords

Clustering, medical knowledge representation, rare diseases

IRP summary

Background: One of the challenges in using data from health databases is the characterization of phenotypes and the grouping of patients sharing the same phenotype using variables of different modalities and scales. For example, patients suffering from diabetes can be identified from the medication they take, examinations they have undergone, or a diagnosis of diabetes during hospitalization. This grouping can be done through an expert approach, using the expert metadata associated with the variables. In addition, the classification of the metadata, for instance in ontologies, can be exploited. This grouping can also be done in a data-driven way, using correlations within the data. Indeed, variables concerning the same phenotype are correlated with each other: for example, when a headache occurs, the management will involve several simultaneous actions. This is in line with the work that we already performed to identify patient subgroups from longitudinal data. Importantly, the identification of a group of patients who have the same phenotype using both data-driven variable correlations and expert metadata is still missing in the field. Such an approach is a task involving unsupervised learning methods with the challenge of calculating distances between patients incorporating all this information.

Objectives: The aim of this thesis project is to develop a generic method for identifying subgroups of patients with the same phenotype from health databases, using jointly variable correlations and expert data, and to implement it within a computer package.

Methods: We will implement and compare different approaches to integrate the observed correlations between variables and metadata ontologies. Among these approaches, we will evaluate different weighted clustering methods and latent variable modeling (Expectation-Maximization algorithm and its variants). The approaches will be compared by different performance criteria (silhouette score, Akaike criterion...). We will

endeavor to use and adapt methods that are as interpretable as possible, i.e., that allow us to characterize the weight given to each element in the definition of the phenotype.
Use-case: To illustrate the method, we will use data from the Dromos project (https://health-datahub.fr/partenariats/dromos). The DROMOS project is a project that uses the National Data Bank for Rare Diseases by linking it to National Health Insurance Data (SNDS) for the most frequent rare diseases, i.e., those with a minimal sample size of 100 patients.

Expected results: The expected result will be methods and an associated package composed of different approaches. These approaches will allow to obtain, for a given phenotype, a classification of individuals according to multi-scale data from the SNDS.

The PhD student will be part of the doctoral network of the project M4DI, one of the axes of the PEPR digital health. As such, they will have a host lab together with funding for a research stay of about 4 months in a secondment lab. The PhD student will further have the opportunity to participate in data challenges and collaborate with other axes of the PEPR digital health.

Host lab

HeKA team, INSERM/INRIA/Université Paris Cité, Centre de recherche des Cordeliers
Supervisors in the host institution: Anne-Sophie Jannot, Nicolas Garçelon
HeKA team has a strong experience in health database data modeling and machine learning approaches including clustering

Secondment lab

K Team, LORIA
Supervisors/collaborators in the secondment institution: Nicolas Jay, Aurélie Bannay
The K team has a strong experience in knowledge representation of SNDS data.

Expected profile

Master in bioinformatics/biostatistics
Fluency in English

References

Lambert J, Leutenegger AL, Jannot AS, Baudot A. Tracking clusters of patients over time enables extracting information from medico-administrative databases. J Biomed Inform. 2023;139:104309. doi:10.1016/j.jbi.2023.104309