Clustering patient trajectories in electronic health records

Project Supervisor: Christel Sirocchi

Electronic Health Records (EHR) provide a rich source of medical data including clinical measurements, diagnoses, procedures, and drug administrations reported over time. While EHR data has been extensively leveraged for predictive modelling of clinical outcomes, most efforts have focused on supervised learning, such as classification tasks for predicting in-hospital mortality or regression tasks for estimating time to the next hospital admission. In contrast, unsupervised approaches, which cluster patients based on similarities in clinical trajectories, have received significantly less attention. These methods can reveal distinct patient subgroups, offering insights that support personalised treatment strategies and advance precision medicine.

In this project, unsupervised learning techniques are applied to time series data of oncological patients from the MIMIC-IV dataset to identify clinically relevant patient subgroups and characterise their varying responses to chemotherapy treatment. The study explores different representation learning methods and clustering algorithms, trained either as a two-step process or end-to-end, followed by interpretability analysis to extract discriminating features and patterns.

Specifically, the project will proceed in the following steps:

1. Install and explore the MIMIC-IV database, writing SQL queries to extract relevant patient cohorts.
2. Leverage state-of-the-art representation learning techniques (e.g., based on variational autoencoders) to model temporal patterns in EHR data, also adapting and repurposing methods originally developed for supervised settings.
3. Apply clustering algorithms and evaluate their performance and robustness using appropriate metrics.
4. Conduct feature importance analysis and leverage other post-hoc explainability tools to identify key features and temporal patterns that differentiate clusters.
5. Derive prototype patient trajectories, summarising the average course of treatment for each cluster.

Requirements

To complete this project, the following prerequisites are recommended:

- Knowledge of Python and machine learning concepts.
- Familiarity with UNIX environments and command line usage.
- Willingness to independently dive into the machine learning literature and test new approaches.

Depending on the results, continuation of the project in the context of a MSc thesis is possible.

References

Ghahremani, Y. and Metsis, V. (2025). Time series embedding methods for classification tasks: A review. *arXiv preprint arXiv:2501.13392*.

Merkelbach, K., Schaper, S., Diedrich, C., Fritsch, S. J., and Schuppert, A. (2023). Novel architecture for gated recurrent unit autoencoder trained on time series from electronic health records enables detection of icu patient subgroups. *Scientific reports*, 13(1):4053.