Data-driven and knowledge-guided representations of diagnostic codes

Project Supervisor: Dr. Christel Sirocchi

In electronic health records, diagnostic codes, typically recorded as ICD codes within the ICD hierarchy, summarise conditions diagnosed across a patient's medical history. Because prior medical history is often highly predictive of future outcomes, ICD codes offer great potential for use in predictive modelling. However, effective representation of ICD codes is challenging due to their high dimensionality, sparsity, and noise. Several strategies have been proposed, each with limitations:

- **Baseline encodings:** One-hot or multi-hot vectors summarise patient admissions as binary or count vectors of codes, but ignore relationships between codes.
- **Data-driven embeddings:** Dense vector representations, analogous to word embeddings, are learned from co-occurrence patterns but often do not explicitly leverage the ICD hierarchical structure.
- **Knowledge-guided embeddings:** Incorporate hierarchical relationships from the ICD ontology or external medical knowledge graphs, but often fail to capture cohort-specific co-occurrence patterns.

This project aims to investigate how different strategies for representing ICD codes affect the predictive performance of machine learning models in clinical tasks. The main application will be the prediction of chemotherapy side effects using diagnostic codes from the MIMIC-IV dataset, considered both in isolation and in combination with other clinical data modalities. Beyond testing existing approaches, there is an opportunity for designing novel hybrid representations that balance data-driven patterns with structured medical knowledge.

Methodological Steps. The project will proceed through the following steps:

1. Review the literature on ICD code representation learning and develop a taxonomy categorising methods according to their reliance on data, knowledge, or neither.
2. Evaluate the predictive performance of models trained on these representations for chemotherapy side effect prediction, highlighting the strengths and limitations of each strategy.
3. Investigate the integration with other clinical data modalities (e.g., longitudinal laboratory measurements), either at the representation level (e.g., concatenating embeddings) or at the model level (e.g., ensemble approaches), and evaluate resulting performance gains.
4. (Optional/Advanced) Develop a novel hybrid representation of ICD codes that combines the most effective elements of data-driven and knowledge-guided strategies.

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 ML literature and test new approaches.

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

References

Johnson, R., Gottlieb, U., Shaham, G., Eisen, L., Waxman, J., Devons-Sberro, S., Ginder, C. R., Hong, P., Sayeed, R., Su, X., et al. (2024). Clinvec: Unified embeddings of clinical codes enable knowledge-grounded ai in medicine. *medRxiv*, pages 2024–12.

Lee, Y. C., Jung, S.-H., Kumar, A., Shim, I., Song, M., Kim, M. S., Kim, K., Myung, W., Park, W.-Y., and Won, H.-H. (2023). Icd2vec: Mathematical representation of diseases. *Journal of Biomedical Informatics*, 141:104361.