THRESH - Knowledge-Guided and Data-Driven Threshold Discovery for Predictive Modelling in Longitudinal Laboratory Data

Project Supervisor: Dr. Christel Sirocchi

Clinical laboratory measurements are routinely used to support diagnostic and prognostic decision-making. These measurements are typically interpreted relative to reference intervals, with clinicians often reasoning in terms of whether a value lies within or outside an expected range rather than relying on its exact numeric magnitude. This suggests that interval membership may carry substantial predictive value. On the other hand, these reference intervals are generally static and defined at the population level, and therefore may not represent thresholds that are maximally predictive for specific outcomes or subgroups.

In parallel, the growing availability of electronic health records has advanced the development of ML models for predicting clinical outcomes from laboratory data. These models commonly use raw laboratory values as continuous features. Although this representation is expressive, it assumes that the exact numeric scale is always informative, an assumption that may not hold for sparse or irregularly sampled longitudinal laboratory measurements. Interval-based representations may provide a more robust and clinically aligned alternative, particularly when sample sizes are limited. Yet, relying exclusively on established population-based intervals may ignore outcome-specific structure present in the data.

This project investigates whether threshold-based representations can improve predictive performance and interpretability by systematically comparing three strategies: continuous representations, knowledge-driven intervals derived from clinical reference ranges, and data-driven thresholds identified directly from outcome supervised analyses. It evaluates cohort and subgroup-specific benefits and analyses whether binary or discretised representations are sufficient for accurate prediction.

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

1. Curation and preparation of laboratory datasets across multiple cohorts and prediction tasks.
2. Training of standard machine learning models using raw laboratory values as continuous features.
3. Transformation of laboratory variables into categorical or binary features based on established clinical reference intervals (e.g. below, within, above range). Evaluation of predictive performance relative to the continuous baseline.
4. Mining of alternative predictive thresholds using supervised procedures such as tree based split extraction, optimal binning, and cross validated search. Construction of discretised representations based on discovered cut points and assessment of threshold stability across resampling.
5. Systematic comparison of continuous, knowledge driven, and data driven thresholded representations across more than 4,000 datasets. Analysis of performance variation with respect to cohort characteristics, target definition, sample size, and data sparsity.
6. Fairness analysis to assess whether certain patient groups benefit disproportionately from specific threshold representations. Evaluation of robustness of learned thresholds across cohorts and their alignment or deviation from established clinical intervals.

Requirements. To complete this project, the following prerequisites are recommended:

- Knowledge of Python and ML concepts. Familiarity with UNIX environments.
- 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.