# Machine Learning for Healthcare - Final Project

0368-4273

**Submission date: 18.09.25**

## Overview

In this project, you will use observational health data to answer a research question on a course-related topic. We offer you one of the following options:

- **Option 1 - ICU prediction models:** Develop prediction models for three ICU outcomes according to predefined guidelines (see section 1).
- **Option 2 - Suggested project:** Any other course-related research question - requires approval in advance! (see section 2).

In **both cases**, your final submission must include **(1) a paper** and **(2) your code**, and comply with the submission requirements in section 3.

## 1 Guidelines for Option 1: ICU Prediction Models

In this project, you will use a **subsample of MIMIC-III** to predict **three clinical targets** defined in section 1.2. You should implement an **end-to-end pipeline** that includes data extraction, data preprocessing, target annotations, model(s) training, and evaluation, based on methods introduced in the course or from the literature.

### 1.1 Data

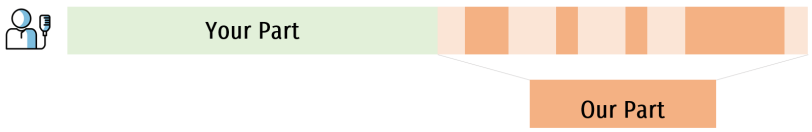
You will develop your pipeline solely based on a predefined patient subsample extracted from MIMIC-III (**your part**). After submission, we will evaluate your trained submitted model(s) on a disjoint **unseen** patient subsample (**our part**) from MIMIC-III with similar characteristics. Specifically:

- **Your part:** *initial_cohort.csv* contains the list subject IDs representing your *initial cohort*.

  **This is the only cohort you can use for the project** – Use it properly to perform inclusion & exclusion criteria, data preprocessing, data partition, model development, and evaluation.

  **Note:** We did not perform any filters on this subset!

- **Our part:** Your trained model(s) will be evaluated on unseen patient data. This entails some technical requirements, which are listed in section 1.5.



## 1.2 Targets Definition

Your model(s) should predict each of the following ICU targets:

- **Mortality** during hospitalization or up to 30 days after discharge.
- **Prolonged stay:** length of stay $> 7$ days.
- **Hospital readmission** in 30 days after discharge (not to be confused with ICU readmission within the same hospital admission).

**Notes**

- Each patient can experience more than one clinical target.
- The tasks differ in prediction complexity, as you will likely observe during your work.
- In practice, 30-day readmission is typically predicted at discharge. Here, you are asked to predict it earlier— using only data from early hospitalization (see Section 1.3). While this setup is less common, in this project, we will explore early risk stratification for ICU-related outcomes.

## 1.3 Prediction Timeline

For each patient, we will perform a **single prediction** for each clinical target during their **first hospital admission**, using data collected at the **first** 48 **hours** with a **6-hour gap**.

**Input representation:** You are free to choose how to represent and model the 48-hours data input — e.g., as an aggregated feature vector or as a time series (sequence-to-target).

Implement your pipeline according to the following timeline (See Fig. 1):

- Focus only on **first** hospital admissions in cases of multiple admissions available.
- Consider only patients with **at least 54 hours** of hospitalization data.
- To preserve a **6-hour prediction gap**, use only data collected in the **first** 48 **hours** for prediction. Specifically, the **prediction time** for each patient is 48 hours since admission ($t = 48$), using only data collected up to that time point.

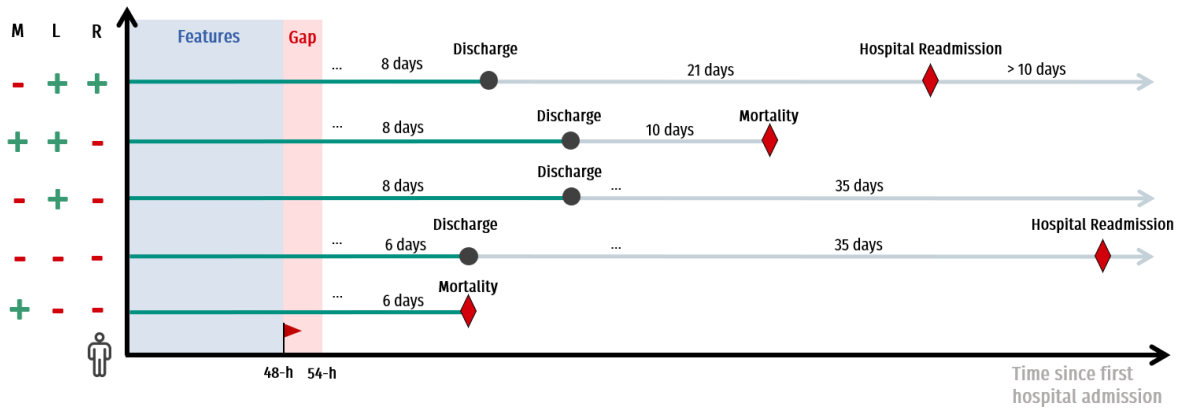



Figure 1: Example of labeled patient timelines. The *y-axis* indicates the label for each of the targets: Mortality (M), Prolonged Length of Stay (L), and Hospital Readmission (R); '+' and '–' denote positive and negative labels, respectively. The *x-axis* shows time since first hospital admission. The red flag marks the prediction time at $t = 48$ h. Green line: during hospitalization, gray line: after discharge.

## 1.4 Pipeline

Your pipeline must contain some **mandatory pre-defined pipeline requirements**, described below. Besides, you are recommended to design and implement your ideas beyond the requirements, using methods introduced in the course and related literature (with credit). You can find some suggestions in the *course summary* slides on Moodle.

**Pipeline Requirements:**

1. **Data Extraction:**
   You must extract and preprocess at least several features of each of the following types:
   (a) Demographic features (e.g., age, gender).
   (b) Vital signs.
   (c) Laboratory test results
   **Note:** Use only *Labevents* table, although some lab tests exist also in the *Chartevents* table.
   (d) At least two additional data modalities:
   - Explore MIMIC and pick at least **two additional data modalities**, e.g., microbiology, habits, data extracted from free-text notes (such as background diseases), non-vital and non-lab chart events, inputs, outputs, prescriptions, etc.
   - Use MIMIC documentation and a broad literature review to:
     - Extract a few features per modality (e.g., a few medications, a few microbiology tests, etc.).
     - Preprocess the new features and include them in your analysis.

   Explain your choices and analyses in the paper.

   **Notes:**
   - If you wish, you can later perform automatic feature selection to reduce feature dimensions (e.g., using `scikit-learm.feature_selection`).
   - If you need help with defining new SQL queries you can contact us.
   - You can look over the literature on available tools for extracting structured data from free-text medical notes.
2. **Target Definition:** Anotate the targets according to section 1.2.
3. **Data Partition:** Your choices.
4. **Data Preprocessing:** Your choices.
5. **Modeling:** Your choices.
   - Your final model must return **calibrated probabilities**.
6. **Evaluation:** Your model must be evaluated using **at least** each of the following aspects:
   - **Classification performance:** e.g., plot ROC and PR curves.
   - **Calibration performance:** e.g., plot calibration curve.
   - **Feature importance:** e.g., analyze the important features using an interpretability method (e.g., SHAP).

**Note:** You are free to apply **inclusion and exclusion criteria** over patients or variables, but document and justify them clearly in the paper.

### 1.5 Evaluation on Unseen Data & Related Code Requirements

Your trained model(s) will be evaluated on an unseen test set, disjoint from your training data but with similar characteristics (see Section 1.1).

To support this, your repository should include a folder named *project* containing:

- **Your code:** All the files and resources required for running your pipeline, e.g., PY files, trained model(s), training parameters for preprocessing, etc.
- README file and `environment.yml` as stated in Section 3.2.
- **unseen_data_evaluation.py**: To be able to run your pipeline on our unseen subset, we provided a file named *unseen_data_evaluation.py* that defines the function *run_pipeline_on_unseen_data*. This function should take subject IDs as input (a *test* set), extract their data, apply your pre-trained model(s), and generate prediction probabilities of each patient to have each of the outcomes at $t = 48$-h. **Note:** It should also apply the required data transformation and processing.

  See the function's documentation and **implement** *unseen_data_evaluation.py* accordingly.

Tester notebook: We've provided a notebook running your final pipeline. **Use it to make sure your code is ready for testing!** To run it:

- Create a copy of the notebook and insert your BigQuery project id in the relevant place.
- Upload your zipped *project* folder.
- Upload the *test_example.csv* and run the code.

**Note:** Additional submission guidelines are listed in Section 3.

## 2 Option 2: Suggested Projects

This option requires a prior proposal and approval. Your project should connect to core topics covered in the course or from the MLHC literature — e.g., clinical prediction models, time-series modeling, disease subtyping, medical imaging, data challenges, interpretability, and others. You can find some optional datasets in the *course summary* slides on Moodle.

Your project must follow the **final submission guidelines** in Section 3. In addition, your paper must contain **literature review** - clearly describe the research question, motivation, and related work in the area.

## 3 Final Submission Requirements (All Projects)

The final submission of all projects must include **(1) a paper** and **(2) your code** (see below).

### 3.1 Paper

The paper should describe your project, including the main analyses and results:

- Use ML workshop format (e.g., ICML workshop). You can find LaTeX templates on the web.
- Page limit: Up to 4 pages in the main text (excluding references). An abstract is optional but not required. You can use an appendix (up to an additional 6 pages).
- Submission in a single PDF file.
- Writing is in English.

It is recommended to structure your paper with the following sections (flexible by project type):

1. **Introduction:** Motivation and background to your research question, related work.
2. **Data / Cohort Description:** A description of the dataset(s), including cohort characteristics (sample size, comparison between subgroups, etc.) and your extracted features.
3. **Methods:**
   (a) **Inclusion and exclusion criteria**
   (b) **Data exploration and preprocessing:** Explain your data exploration process, the extracted data modalities, and the subsequent data preprocessing steps (e.g., feature analyses, statistical testing, feature engineering, data transformations, etc.). Explain your choices.
   (c) **Models:** Describe your models. When applying existing models, cite them and briefly explain your modeling choices. If you develop something new – describe it in detail.
   (d) **Evaluation:** Explain how you evaluated your model (e.g., cross-validation, bootstrapping)
4. **Results:** Report and describe your results with relevant figures and/or tables.
   - The figures should be informative with clear legends and captions.
   - You are recommended to explore different aspects of results, e.g., model selection, evaluation over patient subgroups, etc.
   - Mandatory reports:
     - *Option 1:* You must describe your mandatory evaluation performance results, detailed in section 1.4), including model performance, calibration, and feature importance.
     - *Option 2:* Provide a comprehensive summary of your experiments and results. This includes a clear description of your experimental setup, performance metrics, main results, and any follow-up analyses.
5. **Discussion:** Summarize your results, clinical insights (if any), limitations, and main conclusions.
6. **References**

For projects of a more methodological nature, you may follow an alternative outline such as: (1) Introduction, (2) Related Work, (3) Problem Setting, (4) Proposed Method, (5) Experiments, (6) Discussion. (7) References.

## 3.2 Code:

You should provide the code used for your analysis, as follows:

- Attach a link to your **GitHub** repository in a footnote on the first page of the paper.
- The code should be written in **Python**. You might use open-source libraries such as Numpy, Pandas, Scikit-learn, TensorFlow, Pytorch, SciPy, Matplotlib, etc.
- For *Option 1*: It must contain `unseen_data_evalution.py` (Section 1.5).
- Add a README file providing high-level documentation of your project.
- Include an `environment.yml` file specifying the Python environment and required packages for your project. If you did not use a custom environment, you may leave it as follows:

```
name: mlhc_project
dependencies: []
```

# 4 Project Evaluation

Your project will be mainly evaluated based on:

- Compliance with the instructions.
- Novelty and creativity beyond the mandatory requirements.
- Incorporation of aspects discussed in the course or from the MLHC literature.
- Final evaluation and performance results.
- Paper quality: Writing quality, organization, clarity, etc.
- Code quality, readability, and documentation.

# 5 Notes & Tips

- **Plan you work**, e.g., before running fast, you can start simple (develop a baseline model)
- **Understand and explore your data:** Your patient cohort and features (e.g., distributions, outliers) – Preprocess them carefully!
- Try to account for **statistical, analytical, and ethical challenges** discussed in the course, e.g., potential biases, noise, missing data, ethics & discrimination, interpretability, robustness, generalizability, etc.
- **Literature review:** Look for similar research papers - *Can we do better?*
- Account for **data leakage**, e.g., perform data split carefully, perform pre-preprocessing separately in train/test folds, select your features carefully, etc.
- Provide **uncertainty measures** in your reports (confidence intervals, bootstrap, SD, etc.).

**You are more than welcome to consult with us throughout the project!**

# Good Luck!