**Project Guidelines:**

1. Projects will include a proposal, in-class final presentation, and final report.
2. The project should be done individually or in a team can have at most two members. **You must inform me via email by** **March 9th** **about 1-page proposal summary, and your teammates or if you prefer to work individually on the project.** Your project should be about a problem related to healthcare analytics.
3. **Important note:** Submitting any project or results that were done by others (e.g., copying from Kaggle or any other website, etc.) will be treated as a violation of academic integrity. If occurred, the disciplinary consequences are applied to all parties involved.
4. Each student should select a project falling in at least one of the following categories:

(i) Solving a specific clinical question and a direct application of a machine learning model on healthcare data. For example, MIMIC-III provides a wealth of data to tackle a variety of clinical tasks in the ICU. Here are a few instances of potential clinical questions:

* Identify organ failure
  + In the ICU, we are concerned about organ failure in the heart, kidney, liver, and lung. What predictors could be useful for organ failure?
  + We would want an early enough prediction to be productive and sufficient data in the lab tests (e.g., liver enzymes, creatinine and BUN, blood pressure) in earlier data.
* Effects of interventions
  + Which interventions should we do and at what time?
  + Potential interventions include vasopressors, mechanical ventilation, dialysis, and cardiac assist devices.

Below are some examples of research papers that used the MIMIC-III dataset:

* Che et al, 2018. "[Recurrent Neural Networks for Multivariate Time Series with Missing Values](https://www.nature.com/articles/s41598-018-24271-9)." *Nature Scientific Reports.*
* Dernoncourt et al, 2016. "[De-Identification of Patient Notes with Recurrent Neural Networks](https://academic.oup.com/jamia/article/24/3/596/2769353)." *JAMIA*.
* Ghassemi et al, 2014. "[Unfolding Physiological State: Mortality Modelling in Intensive Care Units](https://www.ncbi.nlm.nih.gov/pubmed/25289175)." *KDD*.
* Boag et al, 2018. "[Racial Disparities and Mistrust in End of Life Care](https://arxiv.org/abs/1808.03827)." *MLHC*.
* Johnson et al, 2017. "This resource may not render correctly in a screen reader.[Reproducibility in Critical Care: A Mortality Prediction Case Study](http://proceedings.mlr.press/v68/johnson17a.html)." *MLHC*.
* Schulam et al. 2016. "This resource may not render correctly in a screen reader.[Integrative Analysis using Coupled Latent Variable Models for Individualizing Prognoses (PDF)](http://jmlr.org/papers/volume17/15-436/15-436.pdf)." *JMLR*.
* Young et al. 2018. "[Uncovering the Heterogeneity and Temporal Complexity of Neurodegenerative Diseases with Subtype and Stage Inference](https://www.nature.com/articles/s41467-018-05892-0)." *Nature Communications*.

NOTE: Samples taken from MIT OpenCourseWare (https://ocw.mit.edu).

(ii) Improving and extending the results of a given study for a more realistic solution. Here, you may use the same healthcare data available in the study and compare your results to the existing one.

(iv) Developing a novel machine learning model or different machine learning models on healthcare data. For example, you can also explore how to extend known machine learning methodology given the challenges of clinical data:

* Clinical natural language processing (NLP)
  + How can we better extract entities from the clinical notes such as diseases, symptoms, and treatments?
  + Can we construct a timeline from the clinical notes? How would these extracted entities relate with the coded events in the patient chart?
* Time series
  + How can we better predict a patient’s progression? Challenges could include missing data, unknown alignment of patients, and heterogeneity of conditions.
  + How can we interpret a patient’s progression? Clinicians may be interested in how a patient progresses through known concepts (e.g., ICD-9 codes) and also what the specific stages of progression might be.

1. 1-page summary should include the problem description and motivation, data, and methods that you may use, and the evaluation approach.
2. The due date to submit the Final Report is May 6th. The final report should be limited to 15 pages (excluding references and appendix). Your final report should discuss the model formulation, assumptions, solution, and results, highlighting the major contribution made by you through the project work, and issues encountered, deviation from the preliminary objectives, and major conclusions. Please be note that you are welcome to discuss with me the progress of your project from time to time.
3. You are expected to turn in your write-up in a PDF format and your source code. You should include a readme file with instruction on how to reproduce your results as well as all the data pre-processing and analysis code. Please do not include any proprietary data in your submission.
4. If you are working in a team of two, you are required to include a section that clearly outlines the contributions of each team member.
5. I encourage you to include the following sections in your writeup:
   * *Introduction*: This section should include a brief explanation of your problem and its clinical importance. You should briefly explain your basic approach and your main conclusions. A figure is often helpful to motivate the work.
   * *Related work*: This section should highlight previous work related to your problem and should put your work in a broader context. It may also include a comparison of why previous approaches could not be used to solve your particular problem.
   * *Methods*: Here you should formally define your problem and describe the method you implemented in detail. Include any simplifying assumptions that you make about your data or the general problem. You should enumerate any modelling choices that you had to make and justify your choices. A main figure illustrating the overall methodology often adds a lot.
   * *Data and experiment setup*: Include details about your data, what variables you have access to, your cohort selection criteria, and your preprocessing choices. You might find it useful to include a table with population characteristics, or an example of the data available for a specific individual, both before (i.e., the original data) and after any pre-processing (i.e., feature construction), to make the discussion concrete. Describe your benchmarks.
   * *Results*: Report the quantitative results of your analyses. You may choose to present graphs or tables, the important thing is that your tables and plots should summarize the relevant results that you got out of the analysis. Comment on these results: are they statistically significant? Are there interesting trends? Do you do significantly better than your benchmarks? Is there a significant treatment effect? You may also present qualitative results such as an in-depth analysis of what the approach would do for a few randomly chosen patients.
   * *Discussion*: Highlight how your results relate to your original question formulation. Do they support your hypothesis? Do they reveal interesting insights about existing medical practices, global health outcomes, the nature of diseases, etc.? Discuss limitations with your analyses and how they might motivate future research directions.
6. The following points will be taken into consideration while awarding the project grade: 1) Complexity of the project, 2) Adherence to the project guidelines, 3) Presentation of final report, 4) Results and major contributions.
7. The Final presentation day will be on **May 6th** **at 1:30 - 3:30 pm.**

## List of Medical Datasets

* [Multiple Myeloma Research Foundation Compass](https://research.themmrf.org/)
* [The Parkinson’s Progression Markers Initiative](http://www.ppmi-info.org/)
  + Specific project suggestion: Learning a representation of MRI images that will help with downstream tasks, e.g., classifying healthy, prodromal, early-stage Parkinson’s, vs late-stage Parkinson’s, or time to outcome. Possible methodologies to consider include [3-D CNN](https://arxiv.org/abs/1806.05233), [voxel-based logistic regression](https://www.sciencedirect.com/science/article/pii/S1053811917301787?via%3Dihub), or principal components. Can you interpret what the representation is capturing?
* [eICU](https://eicu-crd.mit.edu/about/eicu/)
  + Same access requirements as MIMIC-III. eICU has clinical ICU data from 200+ hospitals.
* Chest X-rays
  + [The MIMIC-CXR Database](https://archive.physionet.org/physiobank/database/mimiccxr/)
  + [CheXpert](https://stanfordmlgroup.github.io/competitions/chexpert/)
  + [Open-i service of the National Library of Medicine](https://openi.nlm.nih.gov/faq)
* [The Osteoarthritis Initiative](https://data-archive.nimh.nih.gov/oai/)
* A more extensive [list of medical datasets](https://github.com/beamandrew/medical-data).

NOTE: The list is taken from MIT OpenCourseWare (https://ocw.mit.edu).