Final Project - Learning and Evaluating Clinical Decision Rules, Stat 215A, Fall 2024

These are the **final project specific** instructions. Please also see the general lab instructions in **lab-instructions.pdf**.

Contents

1 Submission

Push a folder called `final-project` to your `stat-215-a GitHub` repository by **23:59 on Friday December 13**. Unlike lab4, **only one group member needs to submit the project**. I will run a script that will pull from all of your GitHub repositories promptly at midnight so take care not to be late as late labs will not be accepted.

The page limit for this lab is 15 pages. The bibliography and academic honesty sections don't count towards this limit.

Follow the general lab instructions in stat-215-a-gsi/disc/week1/lab-instructions.pdf for more details. Please do not try to make your project fit the requirements at the last minute!

I have not provided a template as a guideline for the writeup. In your `final-project` folder, please provide everything (except the data) that is needed for someone else to be able to compile your report and receive the exact same pdf.

Special coding instructions

For any neural networks you discuss in your report, please save the trained model (as a model checkpoint) in your `results` directory. It should be a `.pt` file. If you use YAML files to configure training runs (hint: you should), please include them somewhere in your `code` directory.

2 Academic honesty and teamwork

Academic honesty statement

You can use your statement from lab1 or lab2 or lab3 or lab4

Collaboration policy

Within-group: In a file named `collaboration.txt` in your `report` directory, you must include a statement detailing the contributions of each student, which indicates who worked on what. After the labs are submitted, I will also ask you to privately review your group members' contributions.

With other people: You are welcome to discuss ideas with me. Please avoid discussing specific ideas with other groups, as I don't want to receive 5 of the same lab. If you do consult with students in other groups, you must acknowledge these students in your lab report.

LLM usage policy

You are allowed to use LLMs (ChatGPT, GitHub Copilot, etc.) to help with the **coding** for this lab. If you use an LLM to help with coding, you must do the following:

- At the end of the report in the academic honesty section, state which tool(s) you employed and for what parts of the coding.
- Describe the nature of your interactions with the LLM. e.g. did you ask ChatGPT to code things in English? Did you use Copilot by writing a comment then tab-completing individual lines? Did you get Copilot to write whole functions? Did you mostly use LLMs for debugging? For writing comments? etc.
- If you used a prompt-based LLM like ChatGPT, add a link to your ChatGPT session, e.g. `https://chatgpt.com/share/f7a9094a-9eda-47a0-9103-4aa6345284b0`.

You are also allowed to use an LLM for polishing the writing of your report at the sentence or paragraph level. As an example, you **may**, for example, plug in a sentence or paragraph to ChatGPT and ask it to be checked for grammar and general sentence flow. You **must not**, for example, ask ChatGPT to expand some bullet points into a paragraph, or have it completely rewrite a paragraph, or plug in your whole report, or make something "sound better," etc. The report should still be in your "voice" and not in the (very obvious) ChatGPT voice. And you **must not** submit any writing directly by an LLM without reviewing/editing it yourself. Please ask me if you are unsure if something is allowed.

Any LLM usage for writing must be reported in the LLM usage section of your report—e.g. "I used ChatGPT to revise the grammar of section 2.1." It must be clear to me how much you used LLM tools.

3 What this lab is

The Pediatric Emergency Care Applied Research Network (PECARN) performs research into acute injuries and illnesses among children in a wide range of demographics and institutions. Practitioners in clinical medicine must often prescribe treatment based on imperfect information and often weight tradeoffs when the treatment has the possibility of adverse side effects. For example, the benefit of a CT scan must be weighed against the risk that ionizing radiation from the scan will cause further health problems in the patient. In order to aid clinical practitioners, it is of interest to develop algorithmic risk scores in order to screen patients based on their clinical characteristics and identify patients of high risk. In this project, you

will be able to choose among three PECARN datasets and will develop an interpretable clinical decision rule for your dataset and evaluate the rule in a range of settings. Each of the three datasets explores a different injury common among pediatric patients, one of which you have explored in Lab 1: traumatic brain injury (TBI).

Datasets

Your group must select one of the three datasets detailed below in order to carry out the project. Note that you will need to clean the data in each case, but if the TBI data is selected, and you agree with the cleaning steps done in your Lab 1, you can use that method of cleaning. As an incentive for the groups that decide to use a new dataset, if your group uses one of the new datasets, you will receive 6% of the total points for the lab as extra credit. The data available comes from three different injuries:

1. (CSI): Predicting Cervical Spine Injury. Data can be downloaded here: `https://pecarn.org/datasets/` "Predicting Cervical Spine Injury (CSI) in Children: A Multi-Centered Case-Control Analysis". Original papers:
   (a) Factors associated with cervical spine injury in children after blunt trauma. `https://pubmed.ncbi.nlm.nih.gov/21035905/`
   (b) Utility of plain radiographs in detecting traumatic injuries of the cervical spine in children. `https://pubmed.ncbi.nlm.nih.gov/22531194/`
2. (IAI): Identifying children at very low risk of clinically important blunt abdominal injuries . Data can be downloaded here: `https://pecarn.org/datasets/` "Identifying children at very low risk of clinically important blunt abdominal injuries". Original paper:
   (a) Identifying children at very low risk of clinically important blunt abdominal injuries. `https://pubmed.ncbi.nlm.nih.gov/23375510/`
3. (TBI): Identification of children at very low risk of clinically-important brain injuries after head trauma. Data can be downloaded here: https://pecarn.org/datasets/ "Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study". Original paper:
   (a) Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. `https://pubmed.ncbi.nlm.nih.gov/19758692/`

Prior to starting your project, your group should decide which of the above datasets you will use. If you have trouble downloading one of the datasets, please let me know.

The following are the instructions for the lab. As usual, please carefully document your analysis pipeline, justify the choices you make, and place your work within the domain context.

4 Tasks

How can we best vet and/or improve the clinical decision rule for your given problem? Most importantly, the clinical decision rule should be highly predictive and minimize the amount of missed diagnoses (i.e. have a very high sensitivity). It should also be easy-to-use, using variables that clinicians can readily have access to when making their decisions. Finally, the interpretability of the rule helps to check whether its predictions will make sense for new patients and makes it easier to apply in new settings. Specifically, for this project, your group will have to deliver the following:

- **Introduction**: Introduce the domain problem for the dataset you have selected. Motivate why this is a relevant problem to solve and give some background of previous work in the area.
- **Data**: Download the raw data. Detail your data cleaning process and document all judgement

calls made. Data cleaning will be graded about as harshly as in lab1. Describe the features in the processed data and the outcome features. Detail how you will split the data for the development and testing of your new clinical decision rule. Select a hold out test set that will reflect how the model will be used in practice.

- **Modeling**:
  - Implement the baseline clinical decision rule from the original paper for the dataset you selected.
  - Choose or create an appropriate metric (or set/family of metrics) to evaluate predictive models which will be used as decision rules. Carefully consider the domain context. (Note: this is one of the most important parts of the lab as it will largely determine your choice of model and how you will evaluate it)
  - Develop and implement a new model which is not a neural network.
  - Develop and implement a neural network model.
  - For both models, explain in detail how you developed your model, and describe any intermediate models you did not end up using.
  - Use your judgment to pick a final model to continue the lab with.
- **Interpretation**: Examine the interpretability of your model. This should help clinical practitioners use your model in practice. Is your model a simple interpretable form where you can easily tell how it's making its predictions? If not, how do you recommend interpreting how it obtains the predictions it does?
- **Stability under Model Perturbation**: Introduce a perturbation to your final model, and summarize the effects of this perturbation on the predictions of your model.
- **Stability under Data Perturbation**: Study your final model under three perturbations of the data:
  - What happens if the covariate distribution of the test set changes from what the model was trained on? Simulate this and present the implications for your model.
  - What happens if your model is used on only a subgroup of the patients the model was trained on? Simulate this and present the implications for your model.
  - Create an additional stability check and show how it affects your model.
- **Evaluation**: On your held out test set, evaluate and present the final performance of your model. Is the performance similar to that in the training/validation sets? If not, is there any pattern in the mistakes made?
- **The real world**: Consider the real-world use of your model. What are the implications of the evaluation? Describe how your model might be implemented into a doctor's workflow (this should be informed by the evaluation). Would a doctor trust its output? How would they use it? Would they try to carefully interpret and check each prediction, or use it without checking much? How would a patient feel about the model? If a patient could choose between being evaluated by a doctor alone, by this model alone, or by this model informing a doctor, which would they choose and why?

As you work on this project, keep the following ideas about the data from Lab 1 in mind:

Data Collection

What are the most relevant data to collect to answer the domain problem?

Ideas from experimental design (a subfield of statistics) and active learning (a subfield of machine learning) are useful here. The above question is good to ask even if the data has already been collected because understanding the ideal data collection process might reveal shortcomings of the actual data collection process and shed light on analysis steps to follow.

The questions below are useful to ask: How were the data collected? At what locations? Over what time

period? Who collected them? What instruments were used? Have the operators and instruments changed over the period? Try to imagine yourself at the data collection site physically.

Meaning

What does each variable mean in the data? What does it measure? Does it measure what it is supposed to measure? How could things go wrong? What statistical assumptions is one making by assuming things didnât go wrong? (Knowing the data collection process helps here.) Meaning of each variable—imagine being there at the ER and giving a Glasgow coma score, for example, and also a couple of variables. What could cause different values to be written down?

How were the data cleaned? By whom?

Relevance

Can the data collected answer the substantive question(s) in whole or in part? If not, what other data should one collect? The points made in the above subsection are pertinent here.

Comparability

Are the data units comparable or normalized so that they can be treated as if they were exchangeable? Or are apples and oranges being combined? Are the data units independent? Are any two columns of data duplicates of the same variable?

5 Note on Grading

As the final project is a group project, being a good collaborator on the project will be taken into account for each individual's grade on the final project (just as for lab4). After the project is submitted, we will send a Google form for each group member to evaluate the collaboration received from other members of their group. The final score for each student will be a combination of the student's collaboration score from their group mates, and the overall project score.