Phase II Overview

Phase II Quick Outline of Work

- **Present progress and respond to Phase I feedback**
  - Read and adjust project based on reviewer feedback
  - Create high level overview presentation of project goals and progress
  - Identify project adjustments and concerns for discussion
- **Obtain and explore core analysis dataset**
  - Download, extract, and examine the contents of your code dataset
  - Decide on and describe a data selection or exclusion strategy
  - Calculate summary statistics (counts, averages, ranges, percentages, etc) for important variables or subpopulations in your likely analysis
- **Investigate and formalize your analysis plan**
  - Update literature reviews with more investigation to appropriate analysis methods and benchmarks
  - Describe individual steps of proposed analysis workflow
  - Identify compute, software, and packages intended to complete analysis
  - Identify project risks and potential alternatives strategies
- **Submit and review Phase II reports**

Phase II Submissions and Deadlines [**year-long overview (https://canvas.illinois.edu/courses/57094/pages/assignments-overview)**]

Deliverable	Type	Due Date	Pts	Submit	Evaluators
Phase II: Dataset Characteristics and Project Proposal					
Phase II Progress Presentation/Discussion	18 min present & discuss		5	[Zoom]	CD
Phase II Report and Proposal	around 4 pg team report	Fri, Jun 20	15	[link]	SP, ME, CD
Phase II Peer Evaluations	3 report evals	Wed, July 2	9	[link]	CD

Progress Presentation Sessions

- **Phase Objectives**
  - PP-1. Present your project importance and progress quickly to a generalized audience
  - PP-2. Prepare responses to written reviews and respond to feedback and questions from a live audience
  - PP-3. Identify project risks and weaknesses and solicit assistance from other scientists/researchers
  - PP-4. Participate in and provide valuable contributions to data-related scientific discussions science
- **Session Format**. Each one-hour Zoom session will focus on three projects with twenty minutes allocated to each project. Course staff and all students and mentors from the three teams will be encouraged to attend, share their video, and engage in the discussion throughout. With a team's twenty minutes, the following is expected to occur:
  - **Team presentation [~ 12 minutes]**
  - [~5 min] A quick overview of the project importance, goals, and progress through previous phases
  - [~4 min] Responses based on the most valuable feedback from the previous phases, either answers to important questions asked or significant alterations to project plan based on reviewer comments
  - [~2 min] Plans to complete the objectives for the current phase and updates on progress
  - [~1 min] Possible discussion questions for the faculty audience or peers about project struggles or unknowns
  - **Group Discussion [~ 7 mins]**
  - Audience questions and feedback and discussion of team concerns
- The **5pts** of this session will be based on student participation in the session throughout the hour.

Data Preparation and Characteristics

- **Phase Objectives**
  - P2-1. Download, extract, and explore the primary dataset(s). Apply data transformations to ready the data for analysis.
  - P2-2. Characterize the scales of measurement (categorical, ordinal, numerical), distributions of data (normal, skewed), and extent of missing data for key dataset features.
  - P2-3. Define and apply inclusion and exclusion criteria for selection of key samples and variables for analysis of specific research question(s).
  - P2-4. Construct tables and/or visualizations that convey the summary statistics (counts, means, percentages, data missing) for the sample cohort interest and for important variables related to the proposed data analysis.
- **Overview**: In this phase, your team is expected to acquire and explore the dataset(s) you intend to use for your analysis project. As part of that exploration we expect you to identify key subsets of samples and variables/features in the data, and characterize and summarize their values. In terms of the Phase 2 report, you can separate the description of the steps for gathering and

preparing the data into an initial methods section and the description of the dataset counts, characteristics, and key features into an initial results section. It is strongly encouraged to visualize your data selection/exclusion criteria and counts as a flow diagram figure. There should be a main text table that contains the most important data characteristics

- **Know Your Data,** related **Data Analysis Skills (https://canvas.illinois.edu/courses/57094/pages/expected-data-analysis-skills)** :
  - **Data Provenance:** Where does the data come from? How and when was it obtained? (Velocity) What assumptions were made in its acquisition? What aspects of reality were captured well/poorly? How was the data transformed for sharing and preparation for data analysis? What information was lost in that transformation?
  - **Data Characteristics:** What scales apply to the different data features in the dataset? Are they categorical? ordinal? numerical? What are the underlying distributions of those features (normal, skewed)? What is the variation in the data features and values? (Variety) Are there extreme outliers? How do the different features relate to each other?
  - **Data Quality:** How can you be assured of the data quality? (Validity) Are there missing features that can be created and would be useful? Are there missing or noisy data that can be imputed/corrected? Are there outside data sources that can validate or expand the set of reliable features?
- **Evaluation** of the report for its descriptions of data preparation and characteristics will be based on if it:
  - Clearly identifies at least one dataset suitable for proposed clinical question and provides sufficient background info
  - Provides a clear description of how the data was acquired, pre-processed, and transformed to an analysis-ready format
  - Describes the inclusion / exclusion criteria for selecting data samples (with figure)
  - Identifies specific features / variables relevant to the proposed clinical question
  - Provides simple summary statistics table(s) for key sample populations (cohorts) and important variables

Methods Review and Analysis Plan Proposal

- **Phase Objectives**
  - P2-5. Synthesize from related medical literature information about the common analysis approaches, their advantages and disadvantages, and any previous benchmark values in similar prior studies
  - P2-6. Propose a step-by-step data analysis plan that uses the appropriate methods and evaluation criteria and explains how they will address the selected clinical question(s).
  - P2-7. Identify the existing and needed resources to conduct an analysis project and estimate the time and costs associated with each phase of the project
  - P2-8. Enumerate known issues or risks with the project proposal with accompanying ideas about how to handle them. Also clearly enumerate any major project directions that you

consider out of scope in the analysis proposal.

- **Overview**: In this phase, your team is expected to develop and add a data analysis project proposal into your report. This will mean narrowing down on a specific clinical question and related datasets and carefully defining the objectives and scope of your analysis project. Once you have your project objectives, you should augment your literature review with details about the analysis methods and benchmarks that are related and current for those objectives. This will help you develop a plan for processing, transforming, and augmenting the data, before passing it to a multi-step analysis with specified methods and evaluation criteria. You should be able to describe the step-by-step details of your analysis plans in the report, as well as provide a diagram that visualizes the overall plan for readers. Finally, you will need to estimate what it will take to complete the analysis plan you envision as well as comment on potential hurdles you will encounter.
- **Know Your Analysis Methods**, related **Data Analysis Skills (https://canvas.illinois.edu/courses/57094/pages/expected-data-analysis-skills)** :
  - **Method Details:** What statistical or machine learning data analysis methods are applied to the data? What are some advantages and shortcomings of those analysis types? What makes them best suited for the data and the research questions? What assumptions do these analysis models make? especially about the number of training samples and features? (Volume) How many parameters do these analysis models have? What are other alternative analysis techniques and the rationale for not conducting them?
  - **Training Procedures:** How are data samples stored, extracted, transformed, and used in data analysis and model training? What technologies and software are used? How are the training data samples selected? Are there biases in the training data and can they be accounted for?
  - **Evaluation Metrics:** What metrics are being used to evaluate model performance? What are alternative evaluation metrics and the advantages and shortcomings of the selected ones? Are there biases in the evaluation? How are hyperparameters to the model optimized? How does performance change as these parameters vary? How reproducible is the model performance on a different training set? How does model performance change with more training data and time?
- **Evaluation** of the report for its methods review and analysis plan will be based on if it:
  - Updates the literature review to contain more background on selected analysis methods and related benchmarks
  - Proposes and justifies how overall analysis or modeling method and evaluation criteria address clinical question
  - Describes step-by-step details to conduct the necessary analysis (with figure)
  - Estimates required computational environment, software applications and packages, and other project needs
  - Discusses potential risks of analysis plan and alternative strategies to address them

Updated and Revised Phase II Report

- **Phase Objectives**
  - P2-9. Produce a concise, clear, and well documented written proposal for a data analysis that will be relevant to an important clinical problem and demonstrate that the primary dataset(s) are available and sufficient.
- **Phase II Report**: At the end of Phase II, all teams are expected to revise and update their previous submissions into a more complete report. By the end of Phase II, the report should contain at least the following sections:
  - **Project Title & Authors** (*Revised*)
  - **Proposal Abstract** (***New***) - short paragraph conveying the high level background, goal, and plan for how your data analysis will address an important clinical or medical issue
  - **Introduction** -
    - Literature review of **clinical relevance of** issue (*Revised*)
    - **Background** about primary dataset(s) (*Revised*)
    - Literature review of **analysis methods and benchmarks** (***New***) - Evidence from medical literature about the types of analysis methods, their advantages and disadvantages, and any previous benchmark values that have been applied and set in similar prior studies
  - **Methods**
    - Methods for **Basic Dataset Characterization** (***New***)
      - Methods for downloading, extracting, and pre-processing primary dataset
      - Description and **figure** of data selection inclusion or exclusion steps
  - **Results**
    - Results for **Basic Dataset Characterization** (***New***) -
      - Initial description of important summary values observed in selected data
      - **Table** with summary statistics for key sample populations (cohorts) and important variables related to the proposed data analysis
  - **Proposed Analysis Plan** (***New***)
    - Step-by-step description of multi-phase approach to answer clinical question with data analysis of selected data set
      - Indicating choice of analysis methods and evaluation criteria for each analysis step
    - **Figure** diagram to summarize that high-level analysis approach
    - Description of analysis **requirements**: computing environments, software applications and packages, other project needs or expenses
    - **Discussion on risks** and weaknesses of analysis plan (what will be the biggest challenges, where will the most time be spent, what are the greatest uncertainties) and possible approaches or alternatives that minimize those risks
  - **References** - cited throughout and listed at the end (*Revised*)
- The length of this report is expected to be equivalent to around 4 pages of 12 pt single-spaced text, not counting any figures, tables, or bibliography sections. As always, it is expected to

conform to a medical journal or technical report style and be submitted as a pdf following the team number and shortened title naming convention, e.g. “Team 01 - Analysis of Breast Cancer Readmission.pdf”. Additional material can be included as a single (or zipped directory) supplementary file.

- **Report Evaluation**: Submissions will be evaluated by faculty and peers on the following:
  - Previous content (clinical review, dataset background) appropriately updated
  - Clear and concise proposal abstract
  - Organization and clarity of the flow of thought, especially with integration of previous and revised materials
  - Tables and figures are numbered with titles and descriptions and cited in the text
  - Correct usage of grammar, punctuation, and spelling
  - Adhering to professional journal/report formatting and style

Critical Evaluation

- **Phase Objectives**
  - E-1. Read, understand, and think critically about data analysis reports
  - E-2. Corroborate and assess the soundness of proposed and reported research in domains outside your expertise by finding and comparing to external literature sources
  - E-3: Provide meaningful and professional peer review feedback that resembles a medical journal review process
- **Phase Peer Evaluations**: We will assign every student to review three submitted reports each phase and provide valuable feedback to their peers. The purpose of this exercise is to give reviewers exposure to the efforts and outputs of other teams and exercise the ability to read and think critically about analyses in other domains presented to them and practice communicating their questions or suggestions. For the teams reviewed, this provides additional outside perspectives on the presentation and direction of their project that they have the chance to consider and respond to. We expect peer reviews to contain **Meaningful Feedback**, defined as
  - advice for fixing content errors (not grammatical errors) in the presentation, organizing the information in different ways to make it easier for the audience to follow, or suggestions for alternative methodology, research questions, or interpretation of findings which may constitute a future improvement to the work.
- Some resources for how to perform and write a meaningful review can be found at the paper, “**How to Review a Clinical Research Paper (https://www.ahajournals.org/doi/full/10.1161/STROKEAHA.118.021286)** ” or the **JEE reviewer guidelines (https://cimed-dsp.github.io/files/JEE_reviewer_guidelines.pdf)** .