What Not to Do When Developing and Publishing ML Models

Khaled El Emam

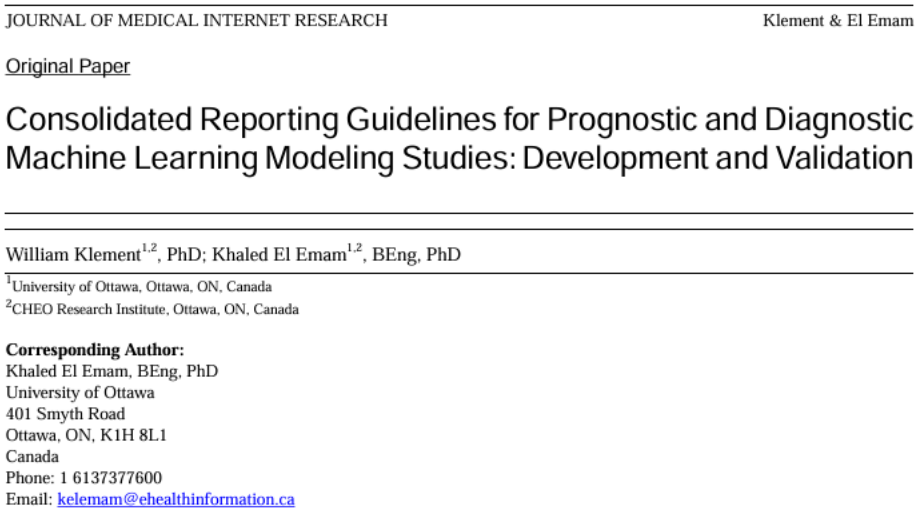
Electronic Health Information Laboratory, Children's Hospital of Eastern Ontario Research Institute

Preliminaries

- Assuming an audience with good knowledge of data analysis and training + evaluating machine learning models
- Based on editing JMIR AI, teaching ML to graduate students, and supervising research projects
- Highlight common mistakes and issues that come up that the analyst should watch out for
- This is not a comprehensive list of everything to watch out for – just common problems that were observed (i.e., the basics)
- Data collection and preparation takes significant resources – analysts and authors should then optimize the use of that data
- Focus on prognostic models / predictions with tabular data, but many issues are more broadly applicable to other types of studies

Reporting

Follow Reporting Guidelines



JOURNAL OF MEDICAL INTERNET RESEARCH Klement & El Emam

Original Paper

Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies: Development and Validation

William Klement[1,2], PhD; Khaled El Emam[1,2], BEng, PhD

[1]University of Ottawa, Ottawa, ON, Canada
[2]CHEO Research Institute, Ottawa, ON, Canada

Corresponding Author:
Khaled El Emam, BEng, PhD
University of Ottawa
401 Smyth Road
Ottawa, ON, K1H 8L1
Canada
Phone: 1 6137377600
Email: kelemam@ehealthinformation.ca

- Ensures that relevant issues are documented
- Not a quality checklist
- Makes it easier to review papers
- Can reduce the number of review iterations (faster time to decision)
- Improves reproducibility
- Broadly applicable checklist
- Include the checklist in an appendix to the paper

Klement W, El Emam K. Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies: Development and Validation. J Med Internet Res. 2023;25:e48763. doi: 10.2196/48763

Information About Data Sources

- It is surprising that sometimes authors do not report where their data comes from
- All of the institutions or repositories where data comes from must be reported, as well as, where relevant:
  - the dates of data collection or collection periods
  - how data was collected
  - data collection criteria (e.g., index dates or anchor dates)
- Information about where others can access or request the data should be included, or a statement that the data cannot be made available with justifications

Documenting Ethics Reviews

- Always ensure that the ethics review information is included, as well as the name of the ethics board and protocol number
- If it is determined that no ethics review is needed, provide the concrete justification for that decision (so still have a section on ethics review)
- While jurisdictional rules differ, examples of situations that require explanations for no ethics review:
  - data was public
  - data was de-identified / anonymized
  - participants are physicians not patients
- Cannot just state that the REB/IRB deemed the study not human subjects research – explanations and justifications are needed

Not Reporting Missingness

- Already a part of the reporting guidelines, but wanted to highlight it separately as a separate issue
- Missingness values for all of the variables should be reported, ideally as part of Table 1 or other descriptive summary of dataset(s) used
- If any specific actions are taken to handle missingness, these must be reported as well
- Even if a complete case analysis is used, this should be explicitly noted as that has implications on the results and their interpretation, and the n's after removing observations due to using complete cases should be reported
- Some ML algorithm implementations do not handle missingness and therefore discussing this issue is important

Documenting the Hyperparameters

- This is in the reporting guidelines, but also wanted to highlight it here
- Some authors do not perform any hyperparameter tuning at all and just use the default hyperparameter
- It is generally recommended to tune the models – it will not always make a huge difference but may
- The method used for hyperparameter tuning should be stated (e.g., grid search, random search, Bayesian optimization)
- You want to get the best models for your data
- Always document the final hyperparameters to improve reproducibility (e.g., in an appendix / supplementary materials)

Methodology - General

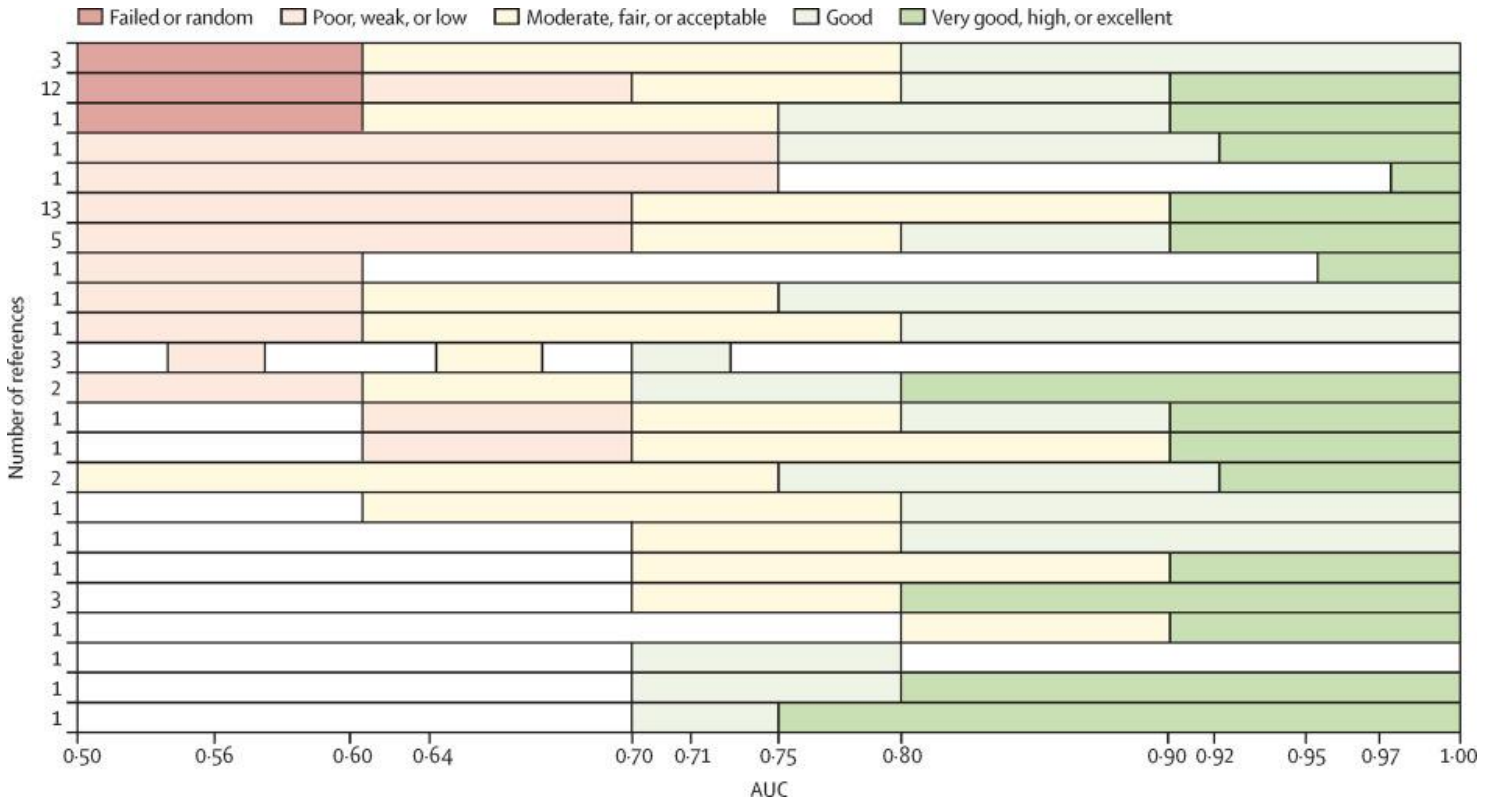
Dropping Observations

- A study starts with a large dataset, and then a proportion of that dataset (sometimes a substantial proportion) is dropped to get an analysis dataset
- There has to be a clear explanation for why cases have been dropped as that may introduce significant bias in the results
- Dropping extreme cases from the test set is also potentially problematic as the test set may no longer be representative of future unseen cases
- Selection criteria (inclusion / exclusion) for cases must be documented clearly

Having a Baseline to Compare Against

- A baseline can be very helpful for interpreting model performance results
- The baseline can be:
  - results from a simpler model (e.g., logistic regression)
  - results based on a review of the literature in the relevant area indicating performance obtained from previous studies
  - the performance of the existing decision-making process at the institution
- Great performance results may not be so great if they are not an improvement over the baseline and moderate performance results may be a clinically important improvement over the current baseline
- Discrepancies from the baseline should be explained

Poor Model Performance



Hond AAH de, Steyerberg EW, Calster B van. Interpreting area under the receiver operating characteristic curve. The Lancet Digital Health. 2022;4:e853–5. doi: 10.1016/S2589-7500(22)00188-1

Justifications for Selection of Predictors

- Predictors used in models must be a priori meaningful features for prediction
- Each one must be explained and justified, usually based on existing literature
- Where relevant, where in the workflow that information would be collected is also important to describe

Predictors Available After Clinical Decision

- The clinical decision-making scenario needs to be described clearly (unless it is obvious)
- The predictor variables (features) must be:
  - only available before the decision is made and not after the decision is made
  - not a proxy for the outcome
    - e.g., a prescription is made during a visit, but the administration requires a subsequent visit and the outcome is another visit within the next 12 months
  - available before the outcome occurs
    - e.g., using total charges for a hospital stay as a predictor of LoS
- This is why all predictors need to be defined clearly

Discretization of Outcomes

- If the outcome is discretized (e.g., continuous to binary) then the dichotomization criteria need to be documented and justified
- Ideally there is a clinical or systems justification for the dichotomization that is related to the decision-making context
- Discretization can also be justified based on previous literature or guidelines
- Sometimes discretization is based on the median or mean; not ideal but should at least be documented (may be justified in the context of simulation studies, for example)

Choice of Cutoff

- This is the cutoff / threshold value that is used to convert a predicted (pseudo) probability from a binary classifier into a class
- Many authors / analysts just use a 0.5 cutoff, however, this is not always a good choice because the prevalence of the data may be very different
- The cutoff may be determined based on clinical criteria, cost criteria, resource availability criteria, patient burden criteria, or prevalence
- However that cutoff is determined, it must be documented and should be justified
- There are further issues with the choice of cutoff if data is rebalanced without subsequent calibration

Comparison of Models

- Reviewers generally ask for the modeling results across multiple models
- In addition to a simple baseline (e.g., a logistic regression model), multiple models should be evaluated
- This is not a strict requirement, but if not done then the authors need to provide some justifications for using a single ML model
- If multiple models are compared, then the results for all of the models must be included in the results as well

Very Low Response Rates

- For on-line surveys, we see studies with very low response rates
- Low response rates, in general, reduce the generalizability of the results
- The minimum expectation is a 40% response rate to be consistent with current trends, but ideally 60% or higher
- The authors must perform an analysis of non-response bias (e.g., comparing early and rate responders with late responders standing as a proxy for non-responders)
- Also, authors need to be able to calculate the response rates – the design of the study should enable that calculation to happen

Wu M-J, Zhao K, Fils-Aime F. Response rates of online surveys in published research: A meta-analysis. Computers in Human Behavior Reports. 2022;7:100206. doi: 10.1016/j.chbr.2022.100206

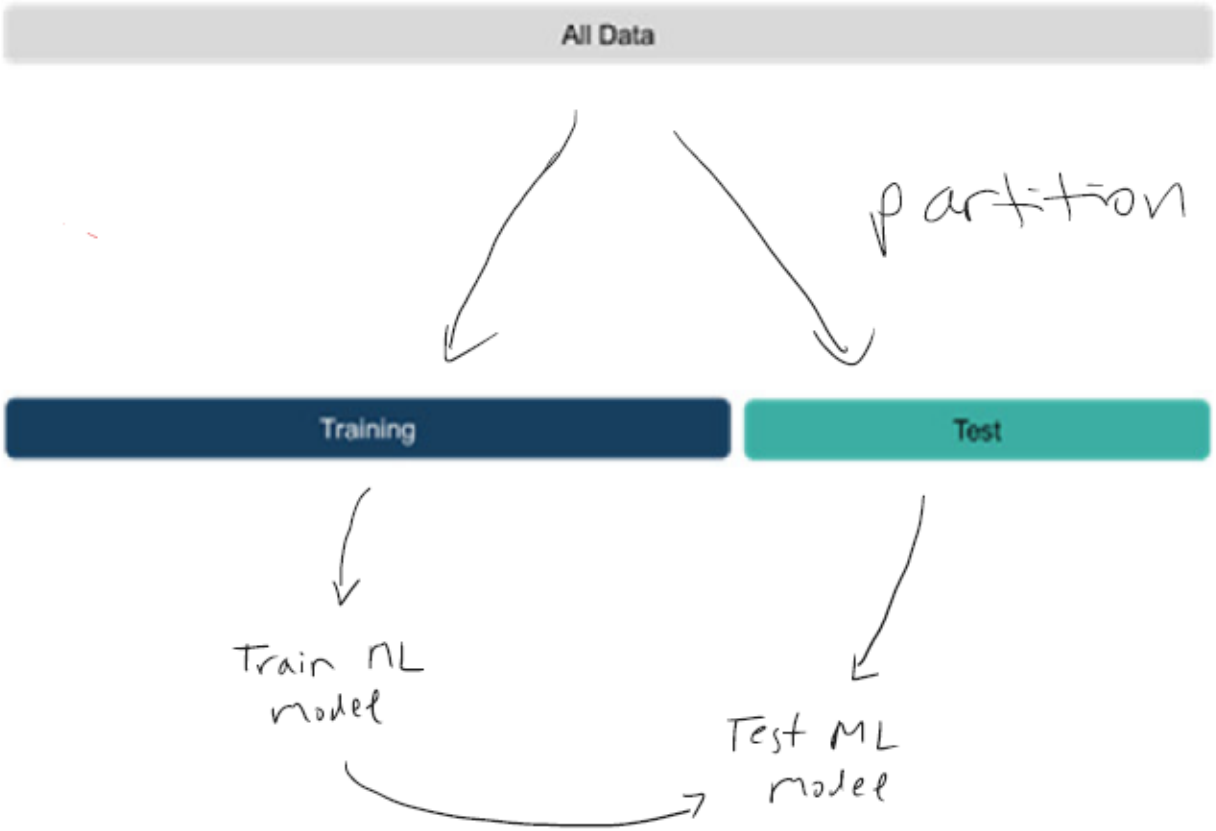
Methodology – Data Leakage

Data Leakage

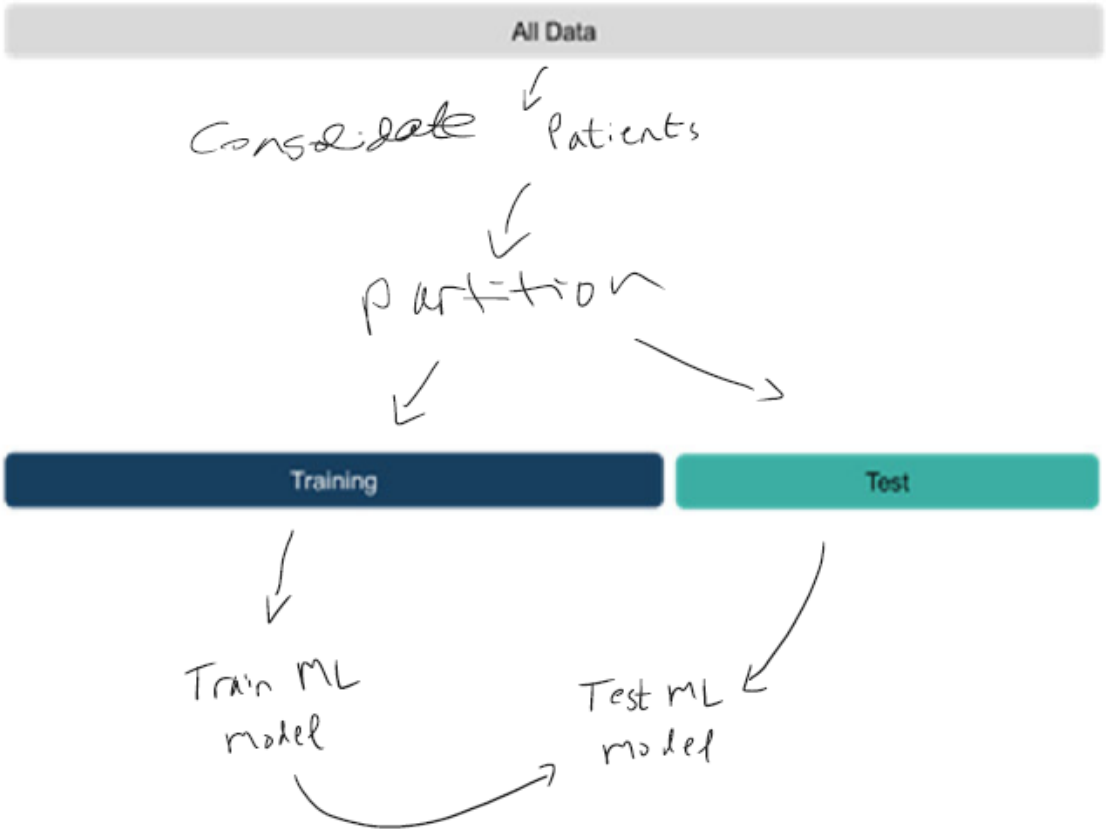
Data leakage is a spurious relationship between the independent variables and the target variable that arises as an artifact of the data collection, sampling, or pre-processing strategy. Because the spurious relationship will not be present in the distribution about which scientific claims are made, leakage usually leads to inflated estimates of model performance.

Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns,. 2023;0. doi: 10.1016/j.patter.2023.100804

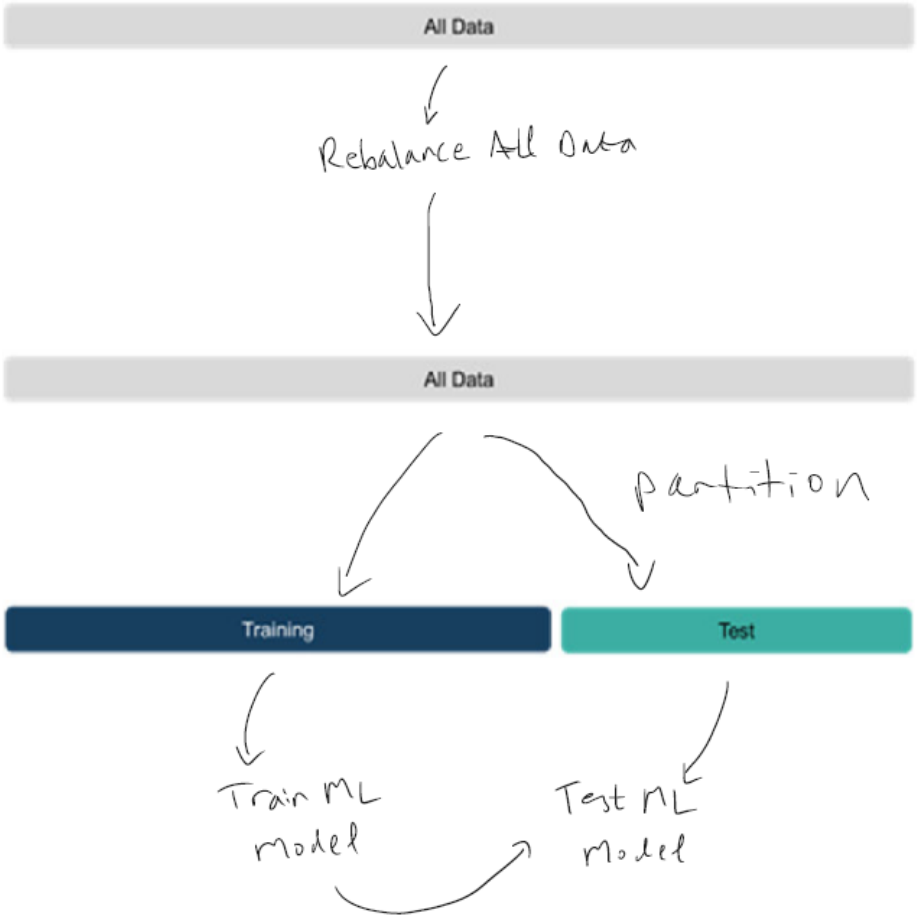
Multiple Records per Patient



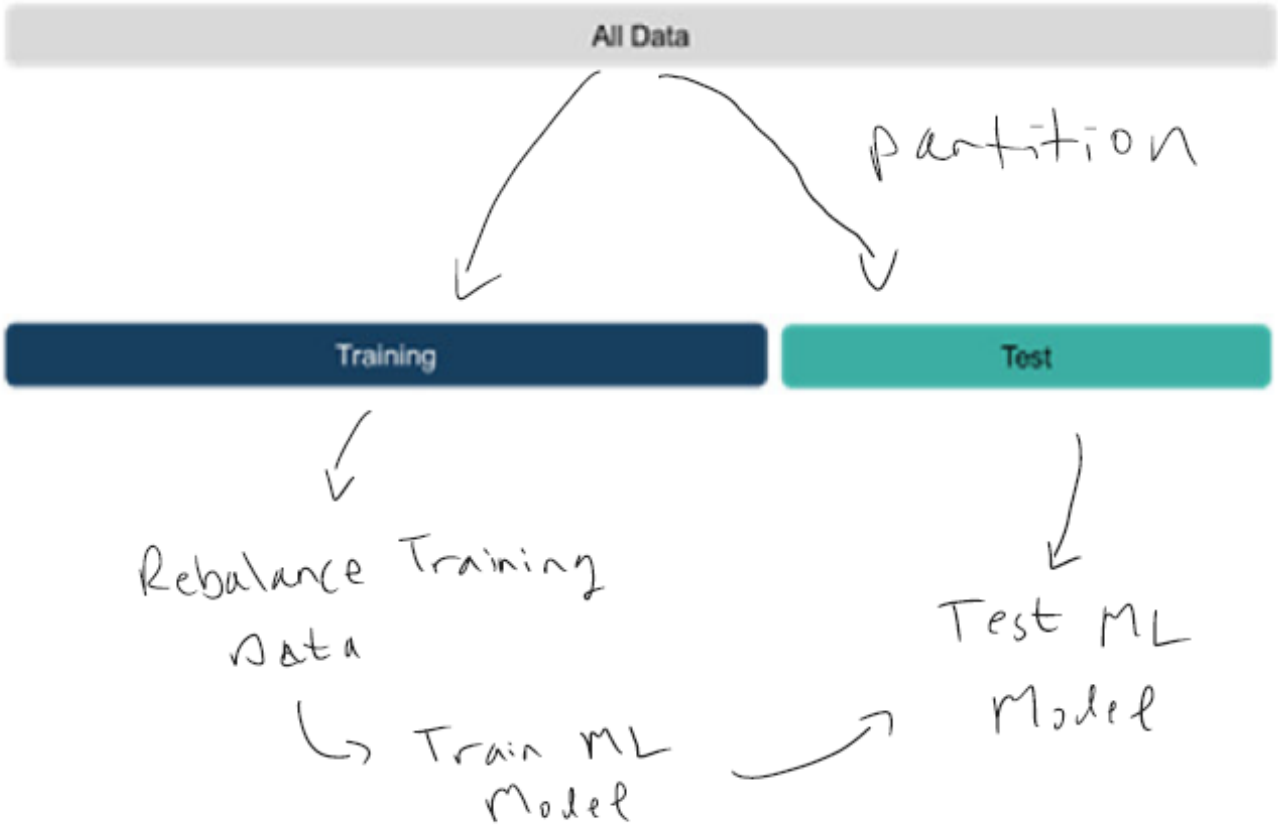
Split Patients Across Train-Test



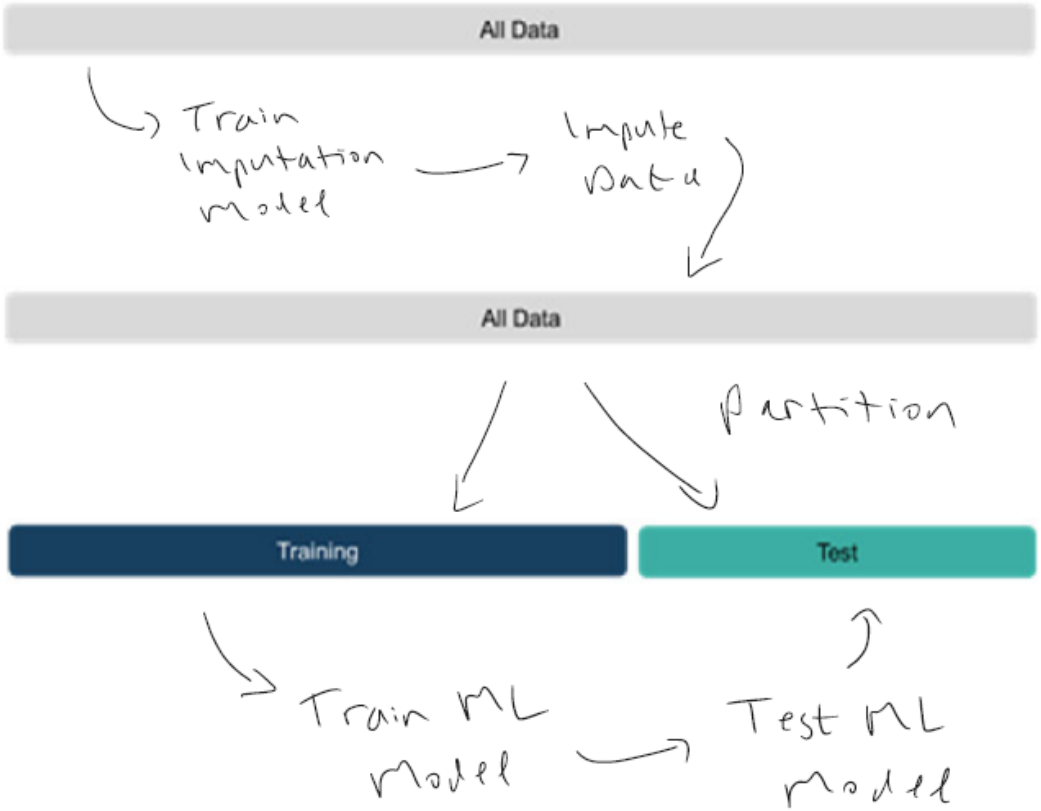
Rebalance (oversampling) Before Partitioning



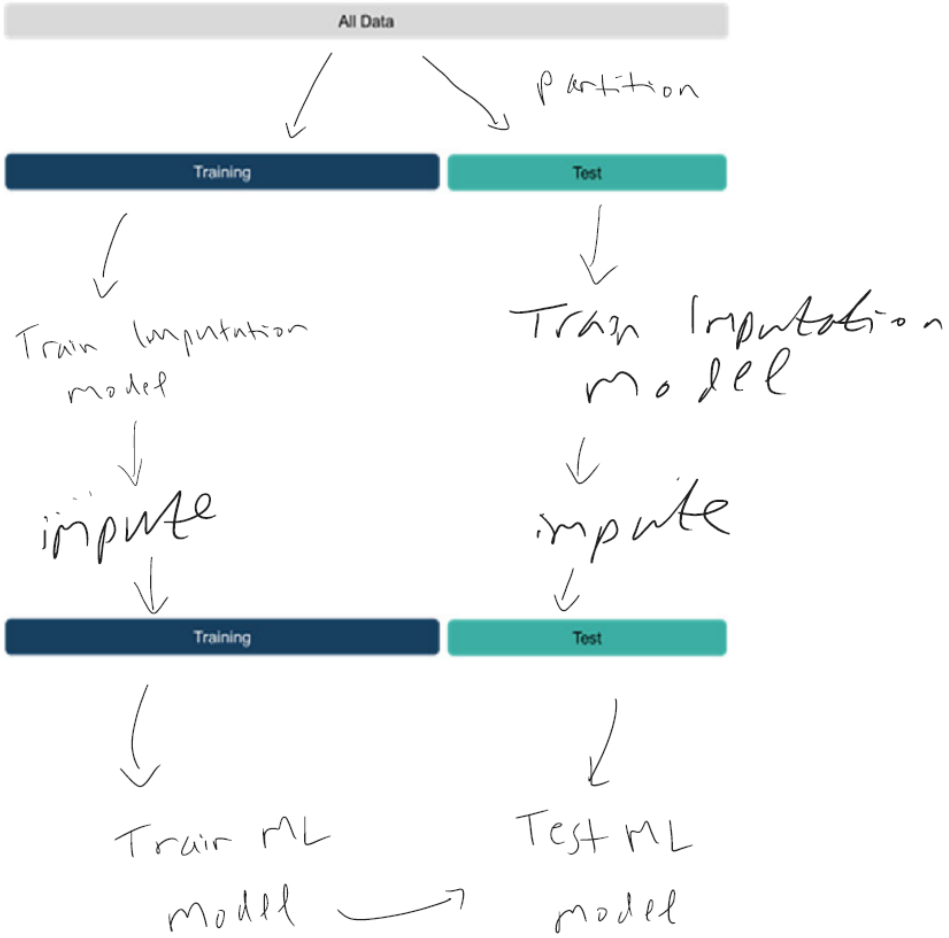
Rebalancing After Partitioning



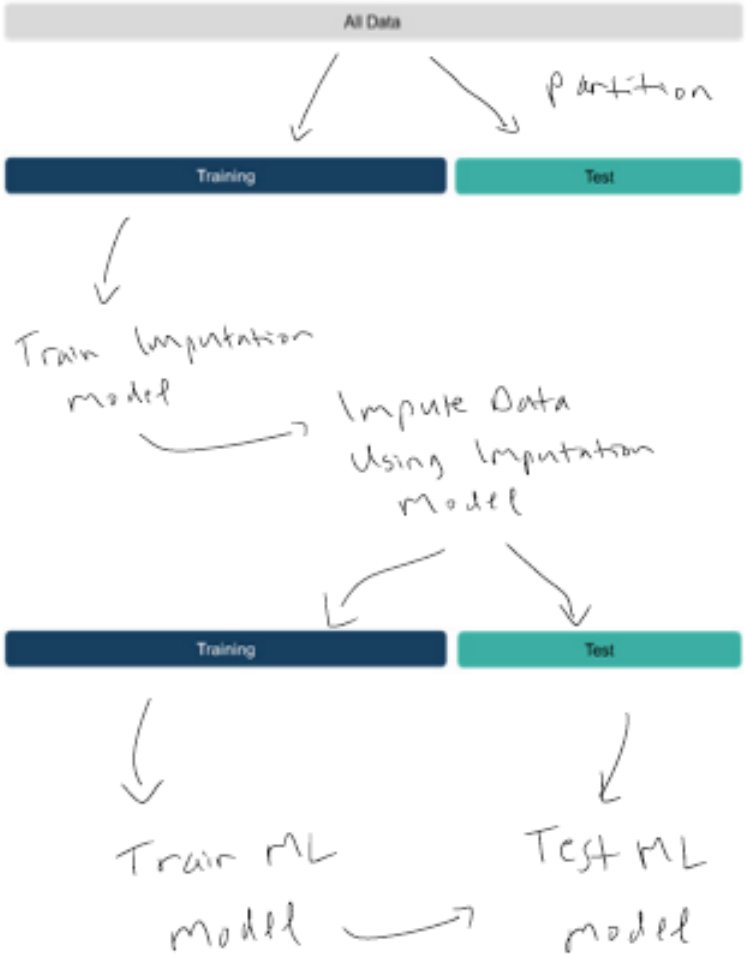
Impute Before Partitioning



Impute After Partitioning



Impute After Partitioning




QUESTIONS