Part 4
Lecture 2 Critical Review of AI/ML Publications

1

Who I am…

Pascal Tyrrell, PhD *Associate Professor*

Department of Medical Imaging, Faculty of Medicine

Institute of Medical Science, Faculty of Medicine

Department of Statistical Sciences, Faculty of Arts and Science

Why Critically Appraise AI/ML Studies?

- Researchers from other fields with diverse research backgrounds and publication cultures have entered the medical field
- The medical community has become accustomed to complying with agreed international standards of reporting
  - This appears to be much less prominent in other fields such as statistics, mathematics, or computational science

Enhancing the QUAlity and Transparency Of health Research (Equator Network)

- Reporting guidelines for main study types
- Look for AI extensions
  - https://www.equator-network.org/reporting-guidelines/tripod-statement/

Study Shows…

- In a recent systematic review, Faes *et al*. in 2020, conducted an evaluation of AI for disease diagnosis using medical imaging, focusing on deep learning models
  - The review identified more than 20,000 studies in the field
    - Less than 1% of these studies had sufficiently high quality design and reporting to be included in the meta analysis

Outline

- How can we evaluate AI/ML?
- Are models being evaluated in its intended stage in the care pathway?
- Do authors provide sufficient clarity on how the data was split?
- Are the image labels likely to reflect the true disease state?
- How is diagnostic accuracy reported?
- Is the dataset used in model development reflective of the setting in which the model will be applied?
- Is the output of the model interpretable and can it be interrogated?
- Is the performance reproducible and generalizable?

How Can We Evaluate AI/ML Systems?

Study Methodology Prespecificiation

Prespecification of study methodology should include…

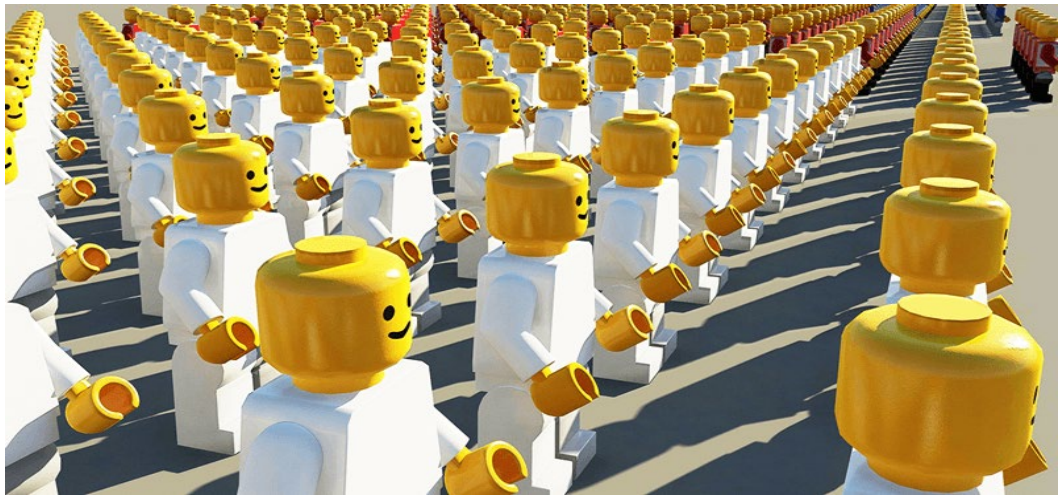
- A description of the unmet need
- The intended place of the model within a diagnostic pathway
- The inclusion/exclusion criteria
- The approach to validation
- Primary and secondary outcomes that will be evaluated
- Power calculation
- Statistical analysis plan

Look For a Clear Purpose

- The need to clarify a prior the purpose of the ML model
- A prior reporting of the study methodology helps tackle a number of biases, including publication bias
  - “Negative” studies (those failing to reject the null hypothesis) are less likely to be published, and where the evidence base may be skewed in favor of models showing high performance

Look For the Basics

- Although there is currently a lack of consensus on how to consider sample size in studies of ML models, it should still be prespecified according to the minimal clinical significant difference and the hypothesis of the study
- Sample size and a statistical analysis plan should be prespecified



A Few Things to Ask Yourself…

- Will this test be used for triage or diagnosis?
- If used in a triage situation, specific test requirements relevant to mass screening could apply?
- Will this model be used as an isolated test, used in combination with other diagnostic elements (e.g., multimodal imaging), or used as an addon or replacement test during the workup?
- If the ML model is a component of the diagnostic decision-tree, researchers should define how the information arising from the model fits within the overall diagnostic probability function.

Beware: Selective Reporting

- Selective reporting of outcomes may occur, whereby the study is reported but only includes those outcomes that show the model in the best light
- This may be a particular pressure where a company holds a financial interest in a model and may profit from the exclusive reporting of positive outcomes
- Both challenges may be addressed by the prospective registration of studies

Avoiding Selective Reporting

- Prospective Registration of Studies
- Pre-Specific of Outcomes
- Transparent Reporting
- Follow Reporting Guidelines
- Independent Analyses
- Peer Review

Evaluation: From the Papers to the Clinic

- Understand the intended use of the ML model in the diagnostic process
- Documenting the intended study methodology enhances transparency
- When evaluating the clinical implementation of AI/ML systems, it is important to know whether it has been…
  - Tested in an experimental setting
  - Shown a meaningful impact in a population similar to the one for which it is being considered

Is the Model Being Evaluated in its Intended Stage in the Care Pathway?

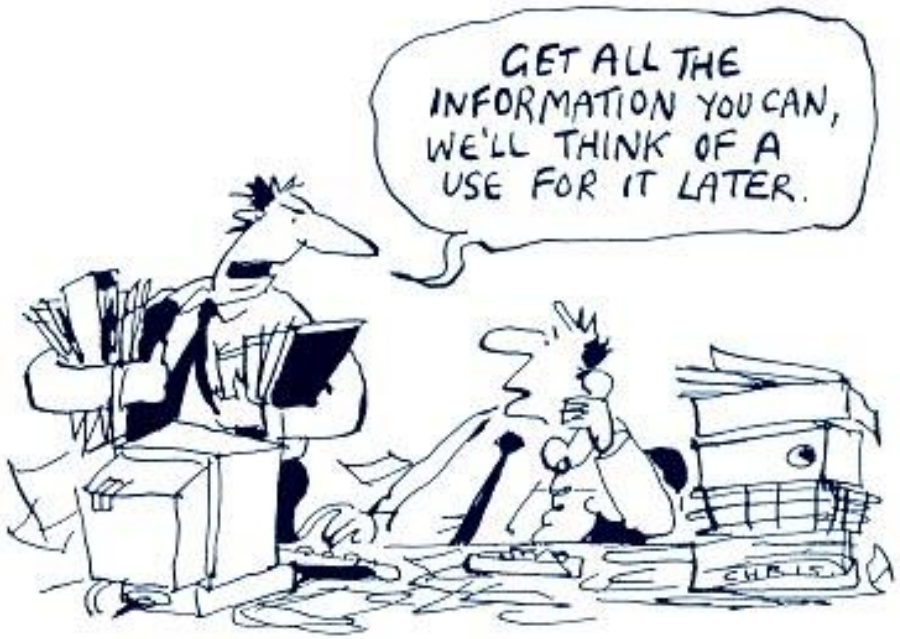
Diagnosis

- At each stage, whether it be the presenting history, clinical examination, or a series of investigations, constitutes an individual data point along a stepwise diagnostic process

- Diagnosis is a process of integrating information derived from various stages in the patient pathway

Every Step is Important…

- At each step, there is a transition from a pretest to posttest disease probability, and it is the combination of information derived at each step that makes up the final diagnosis decision
- It is important to understand where the dataset was generated from within a care pathway
- Any new test developed using a given dataset should not be considered in isolation from its clinical pathway



Consider this…

When considering the validation of models based on a pre-curated dataset, it is important to ask:

- For what purpose was this dataset originally curated?

- Does the disease probability within this cohort differ to the setting in which the model will be deployed?

Do the Authors Provide Sufficient Clarity on How the Data Were Split?

Splitting Datasets

- The terminology around datasets has been a common source of confusion in ML studies as authors have used many of the key terms interchangeably
- Common practice in developing ML diagnostic algorithms is to split a dataset for development into training, tuning, and internal validation test sets (split sample validation)
- Subsequent external validation test sets, for out-of-sample external validations, are also often sought to test for generalizability of the model

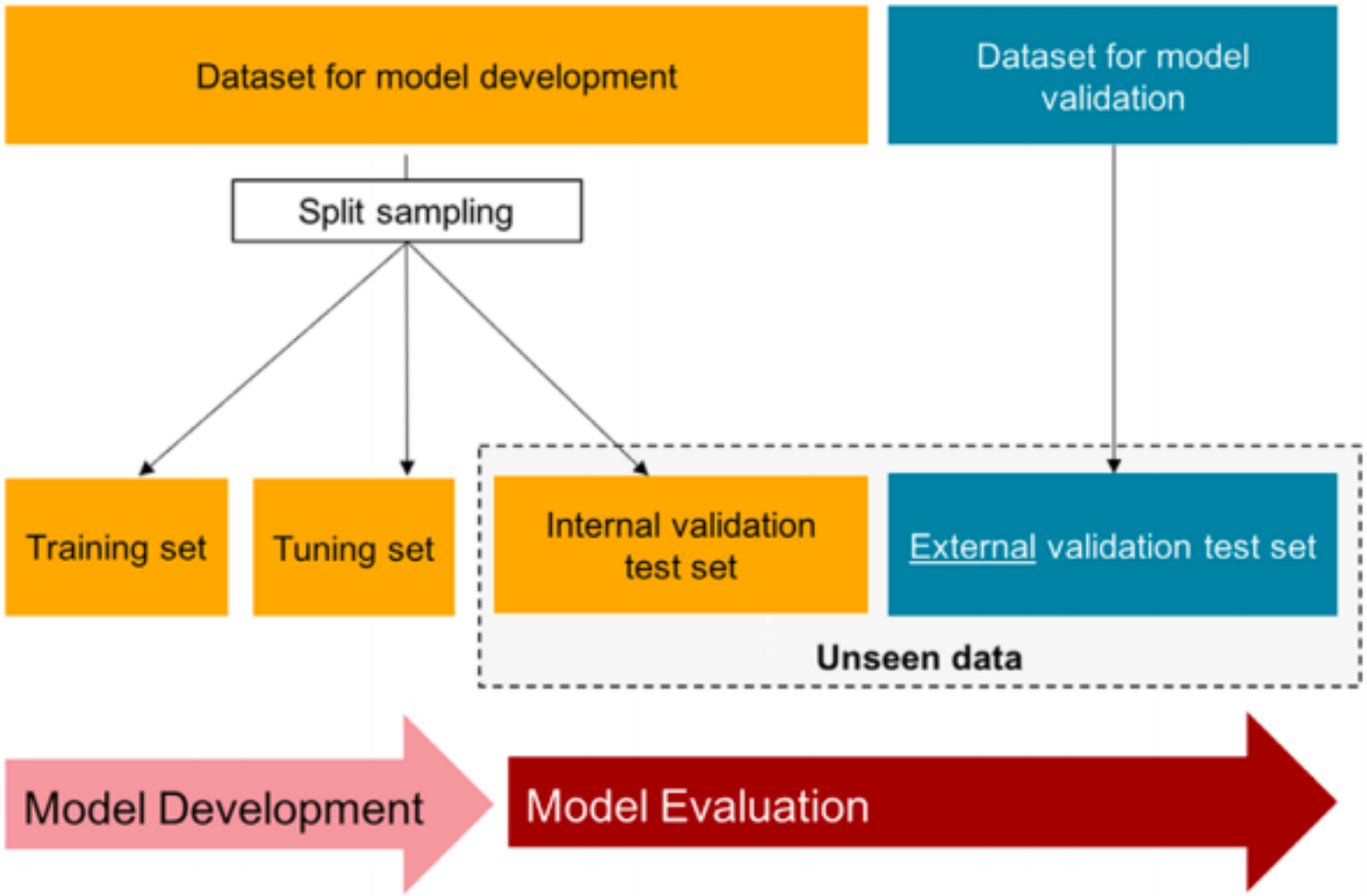


Figure 1. Overview of datasets involved in a machine learning diagnostic algorithm: model development and evaluation.

Are the Image Labels Likely to Reflect the True Disease State?

The Gold Standard

- To assess the accuracy of any model, we need to assess against the ground truth (more commonly known to clinicians as the gold standard)
- More often an issue in ML compared with other diagnostic studies because of the sizes of datasets involved and the demands this may place on any manual labeling process

An Experts Opinion?

- For most models, the best ground truth available is usually expert opinion
- However, the reliability of expert opinion should be critically appraised
  - It may vary considerably in robustness from, single expert to multiple expert majority vote, multiple expert consensus and multiple independent expert opinion with disagreements escalated to an adjudicator
  - Subspecialist for a certain number of years, board-certified specialist, or certified readers from a reading center
- By reporting interobserver agreement, readers can at least make a judgment on the likelihood that the ground truth label is correct

Consider this…

- Were the images labeled prospectively or retrospectively?
  - Note that in some situations, retrospective labeling may be beneficial as it benefits from additional information (such as further follow-up data confirming a diagnosis)
- A fundamental question is, how confident we are that these labels are indeed ground truth?

How Is Diagnostic Accuracy Reported?

Terminology For Everyone

- There is a clear need for both communities to understand each other's terminology: in medical applications, diagnostic accuracy is usually reported as sensitivity, specificity, and area under the curve; in ML applications, models are also commonly reported in terms of accuracy, F1 score, and dice coefficient

- The provision of the actual contingency tables ensures clarity, and to some extent bypasses this issue

		Ground truth / label *Gold standard / Reference test*		
		Condition Positive	Condition Negative	
ML model *Index test*	Predicted Positive	***True Positive*** ***TP***	***False Positive*** ***FP***	**Precision** *Positive predictive value* $\frac{TP}{(TP + FP)}$
	Predicted Negative	***False Negative*** ***FN***	***True Negative*** ***TN***	***Negative predictive value*** $\frac{FN}{(FN + TN)}$
		Recall *Sensitivity* $\frac{TP}{(TP + FN)}$	***Specificity*** $\frac{FP}{(FP + TN)}$	

Accuracy	**F1 Score**
$\frac{TP + TN}{(TP + FP + TN + FN)}$	$\frac{2TP}{(2TP + FP + FN)}$

Differences in nomenclature for machine learning (boldface type) and classical statistics (italic type) and where overlapping (boldface and italic) are highlighted

Is the Dataset Used in Model Development Reflective of the Setting in Which the Model Will Be Applied?

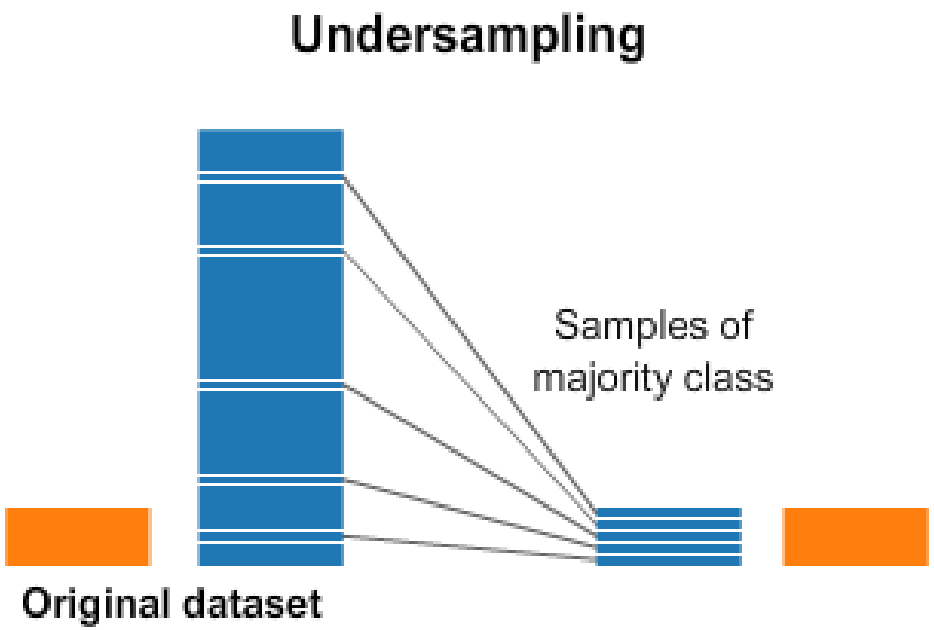
Spectrum Bias

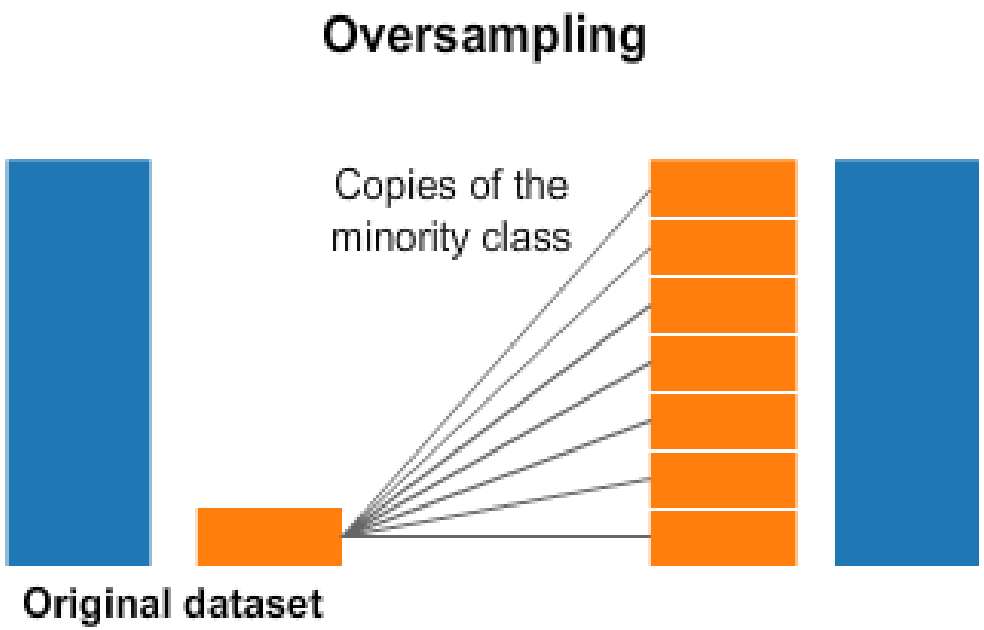
- Datasets (i.e., disease severity, stage, distribution of alternate diagnoses) do not adequately reflect the target patient population
  - Important to maintain transparency
- Is a common problem because many investigators may opt for datasets which represent extremes (i.e., normal vs severe disease)
  - Strive for Generalizability
- Underrepresentation of important diagnostic features or disease states during development may profoundly limit its performance once it is released into its intended clinical arena.

To Tackle This Problem…

- Consider whether the dataset represents the complete spectrum of diagnostic cues for the target population
- Algorithm developers often adopt various methods to balance the classes
  - Oversampling: Adding copies of the underrepresented class
  - Undersampling: Taking away instances of the overrepresented class
- Although this commonly used technique is helpful in algorithm training, investigators sometimes replicate the class distribution in the validation test set, which is most likely to ensure optimum model performance, even if it is an unrealistic disease prevalence

Oversampling vs Undersampling





Is the Output of the Model Interpretable and Can it Be Interrogated?

- In non-ML predictive modeling, input parameters of a model may have been chosen in a hypothesis-driven and rule-based manner

- On the contrary, common ML techniques for image-based diagnosis in deep learning, may potentially use thousands of input parameters fed into a complex model of weighted connections to create data-driven predictions without any supporting evidence

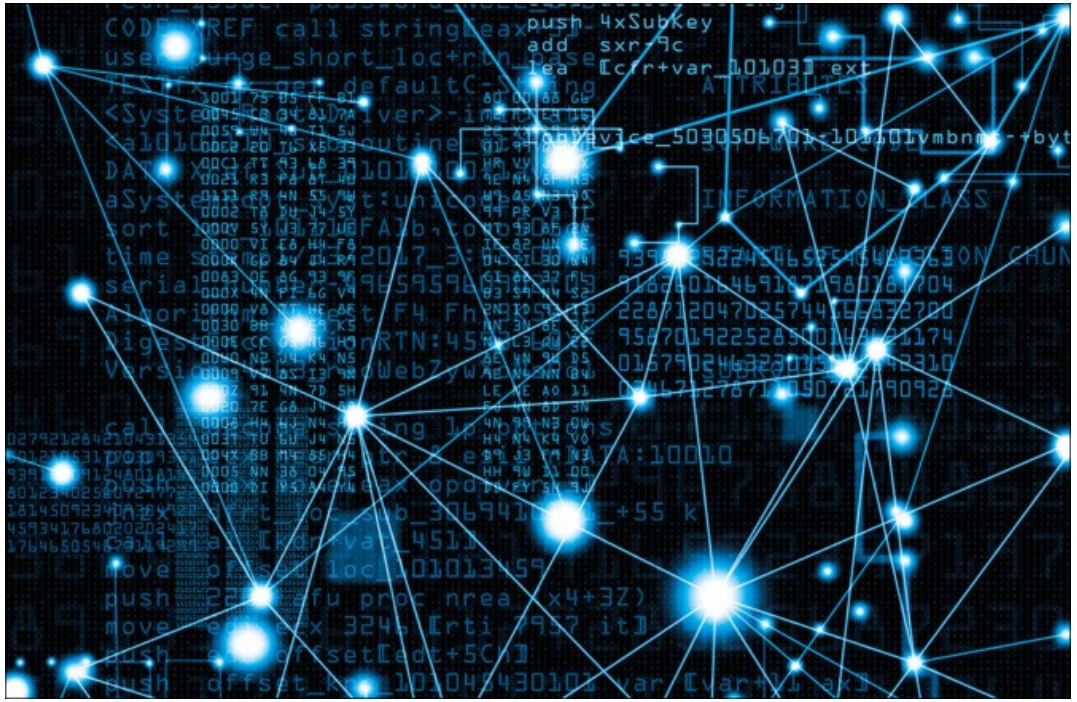
- This way of modelling stays fairly abstract to the human mind ("black box" decision making) and makes it harder to detect bias, overfitting, and confounding

Is the Performance Reproducible and Generalizable?

Missing Piece…

- Important that the predictive accuracy has been shown to be robust beyond the cohort they have been developed in

- It is a known phenomenon that classification performance of predictive models, including ML models, can be overestimated in internal validation alone…



...External Validation

- Assessing the performance with separate data not used for model development
- Evaluation in a dataset that is independent, but differs in either the population or the setting evaluation in the same or new populations over time to test for degradation of the model performance as the population evolves
- These factors may profoundly affect the performance of a model and highlight the need for reproducibility and generalizability to be evaluated

External Validation

- External validation should be considered as a continuum rather than a single event
- External validation in ML-based diagnostic models is arguably even more important because of the “black box nature” of these systems and the inability to interrogate the models’ decisions

External Validation

To test the generalizability of the algorithm's performance, authors should therefore seek to externally validate their results in an out-of-sample external validation to avoid overly optimistic estimates

This should be done in a temporally, or preferably geographically, separate study population, and ideally by an independent research group

Outline

- How can we evaluate AI/ML?
- Are models being evaluated in its intended stage in the care pathway?
- Do authors provide sufficient clarity on how the data was split?
- Are the image labels likely to reflect the true disease state?
- How is diagnostic accuracy reported?
- Is the dataset used in model development reflective of the setting in which the model will be applied?
- Is the output of the model interpretable and can it be interrogated?
- Is the performance reproducible and generalizable?

New Standards

- New standards specific to reporting studies of ML interventions in health care are in development
  - TRIPOD-ML (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis)
  - SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence)
  - CONSORT-AI (consolidated standards of reporting trials)
- It is hoped that they will lead to improvements in the design and reporting of such studies

End of Lecture 2

Next up Part 5 Lecture 1: Data ownership, data sharing, and ethics