BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2023

Lecture 3: Introduction to AI/ML on clinical datasets

HARVARD
MEDICAL SCHOOL

Marinka Zitnik
marinka@hms.harvard.edu

Responses to L2 Quick Check

Examples of inferential gaps in clinical decision making

An example would be deciding whether to give a new treatment to a patient who is suffering from multiple comorbidities. Let's assume that the randomized controlled trial (RCT) that ascertained the efficacy of the new treatment had an exclusion criteria that would have excluded the patient in question [1]. Thus, there is no available knowledge that would help inform how the patient in question would respond to the new treatment. This means that clinicians would need to infer, and fill in the gap, to make a decision regarding the application of the new treatment for the patient with comorbid conditions. Comorbidities

Knowing when a particular lab value is relevant to the diagnosis or other clinical decisions. Often this results in sweeping lab panels becuase it is unclear which lab values might change treatment Lab panels

Patients with severe brain injury may remain comatose for weeks, requiring aggressive life support. Families often want to know if the patient will recover to a quality of life that they would find acceptable, as this has major impacts on whether the family chooses to continue life sustaining treatment. We currently have only general and imprecise data on what factors predict return to consciousness and independent living, so families must be counseled and make decisions with a great deal of uncertainty. The inferential gap is in having enough data and good enough models to accurately predict functional outcome 6-12 months after traumatic brain injury. Return to consciousness

If a drug has not been tested on pediatric populations, but a doctor wants to prescribe that drug to a child, the doctor will have to make a guess about the drug's effects without knowing information about how the drug works in pediatric populations. In this case, the doctor is experiencing an inferential gap. Pediatric patients

Physician seeing patients in the clinic not being fully equipped on how to interpret results from genetic testing. Genetic testing

Responses to L2 Quick Check

Examples of ML workflows on healthcare and clinical data

EHR-based research project: Anticipate disease (e.g. UC) flare-ups 1. Gather relevant features such as diagnosis code / primary note for UC, and identify other codes that relates to flare-up symptoms such as blood in stools, inflammation marker upon lab test and so on. 2. QC: if we use multiple data standard, make sure that the features are in same measuring unit, then scale/normalise, clean and so on before traning the data. 3. Missing values: depending on its context, missing values can be imputed/removed/predicted(i.e. use proxy such as mean value, etc), or sub classify as NA. 3-1. Choose which model is appropriate for anticipating flareups (i.e. prediction model) 4. From the chosen model, select the phenotypes (expressed features) that are typical for UC patients, that can be used to feed into the selected model in the previous step. 5. Outcome can be defined as disease flareup (e.g. heavy blood in stool for more than 5 consecutive days), and the study period can be 6 months/12 months as applicable to the data. 6&7. We could choose an appropriate predictive ML model, for which could be used to a test-train split datasets, where manual review for flareup identification can be done on the dataset for performance measure purposes. 8. We could refine the model by choosing/dropping new features, confounding factors and so on to improve model performance and to avoid possible "cheating of ML models". Then repeat.

Predicting disease flare-up

Responses to L2 Quick Check

Examples of ML workflows on healthcare and clinical data

For the task of predicting medication response we would first identify a cohort of patients on that medication, and ensure that the indication matches the indication of interest for our study, and remove any patients meeting exclusion criteria. We would check for availability of data indicating medication response - i.e. blood pressure measurements for an antihypertensive drug. For quality control we would check units and make sure they could all be converted to standard mmHg, and remove impossible/erroneous values (i.e. 0). We would check the number and timing of available BP measurements for each patient to select missingness, and look at the metadata around those measurements (i.e. outpatient or inpatient). Input features that may correlate with response would be extracted/cleaned - i.e. demographics/co-morbidities, and responders and non-responders would be labeled by selecting a BP and time point threshold. The cohort would be split into training, validation and testing sets sets, followed by pre-processing (i.e. scaling, numeric encoding). The ML model (e.g. LASSO regression, random forest) would be trained on the training set, and tuned with cross validation. Finally, the model would be tested on the hold-out test set.

Predicting drug treatment response

Responses to L2 Quick Check

Examples of ML workflows on healthcare and clinical data

Sub-phenotyping patients with a disease: an example of more risk-stratification, patients with a disease at low, medium, high risk of further organ dysfunction: relevant features are the disease label itself, demographic information, family history, medications, comorbidities, labs etc. QC to determine if unexpected min/max, percentage of missing data etc; would choose study period for exposures and outcomes here when doing the feature selection. Create the broader phenotype (patients with the disease; this definition/case classification step is nuanced. Define the outcomes (organ dysfunction grades for example, ie CKD 1 -5). Determine if the ML goal is prediction, classification etc. In this case, if prediction of risk for further organ dysfunction, may start with regression; perhaps also use random forests to classify the risk subphenotypes. Split the dataset into training, testing and validation; process the data by the features, develop the model and apply.

Disease sub-phenotyping

Guide triage decisions: First we collect all data from the EHR of patients that presented to an emergency room over the course of 2 years, which in a triage decision could be clinical notes, vital signs, images if done, medications given and laboratory data, then we check that all information is entered the same way, same units , or for example location to be accurate, or if a patient was moved to the ICU in less than 24 hours we could decide to have that as the initial disposition for training. (quality check), then we make sure the data is complete, we know have structured and unstructured data (the unstructured data may need to be pre processed), the outcome for the model/ label is where did the patient end up going, home vs observation unit, vs admitted to the hospital on a regular floor vs admitted to an intensive care unit. We could use a deep learning model to analyze these data and predict the outcome. That could be helpful to know if the patient will go home, maybe they could be seen as fast track vs regular path. Although there may be biases in clinicians, some clinicians may decide to keep more patients than others.

Guiding triage decisions

Outline for today's class

1. Highlights of ML on EHR data:
   - Polypharmacy and adverse drug events
   - Modeling disease progression
   - Mortality and critical event prediction

2. Federated learning in healthcare
   - What is federated learning?
   - Why is it important for medical data?

3. What's next for clinical informatics research

Prediction of critical patient events

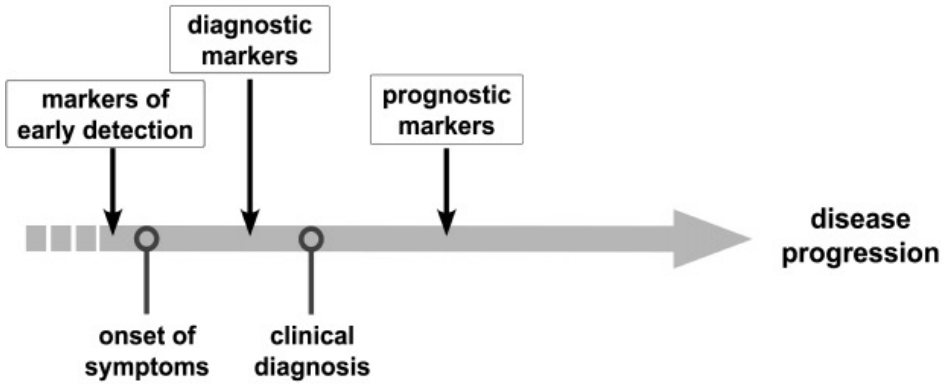
Objectives:

- Analyze EHRs of patients who tested positive for COVID-19 and were admitted to hospitals in the Mount Sinai Health System in New York City
- Develop ML models for making predictions about the hospital course of patients over time horizons based on patient characteristics at admission
- Assess performance at multiple hospitals & time points
- Identify key patient characteristics that govern the course of disease across a large patient cohort

Vaid et al., Journal of Medical Internet Research, 2020

Prognostication using ML

- Prognostication with ML can aid physicians in predicting disease trajectory, allocating essential resources effectively, and improving outcomes
- However, efforts have been limited by small sample sizes, lack of generalization to diverse populations
- Studies lack:
  - Temporal benchmarks
  - Interhospital or prospective validation
  - Systematic evaluation of multiple models
  - Consideration of covariate correlations



Vaid et al., Journal of Medical Internet Research, 2020

Critical patient events: Approach

- Boosted decision tree-based ML model trained on EHRs from patients confirmed to have COVID-19 at a single center at Mount Sinai in New York City to predict critical events and mortality
- Analyses:
  - Generalizability: External validation at four other hospital centers
  - Prospective evaluation: New set of patients from all five hospitals
  - Feature importance: Saliency analysis using SHAP (SHapley Additive exPlanation) values to identify the most important features for outcome prediction

Vaid et al., Journal of Medical Internet Research, 2020

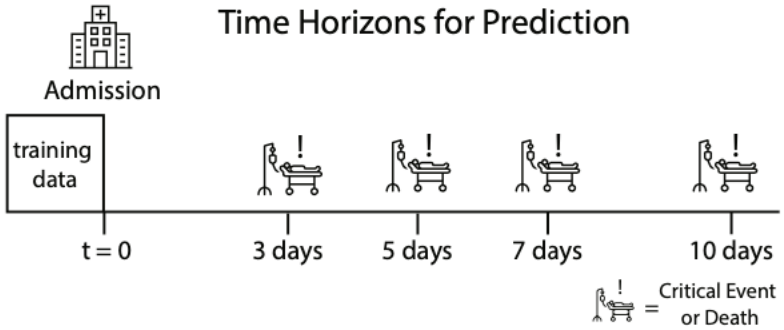
Study population

- Retrospectively included:
  - Patients who were over 18 years of age
  - Had laboratory-confirmed COVID-19 infection
  - Admitted to Mount Sinai hospitals between March 15 and May 22, 2020
- Confirmed COVID-19 infection:
  - Positive RT-PCR assay of a nasopharyngeal swab
  - Restrict data to only primary COVID-19-related encounters
  - Patients who had been discharged, had died, or were still admitted and had stayed in the hospital for at least the amount of time corresponding to the outcome

Vaid et al., Journal of Medical Internet Research, 2020

Study design

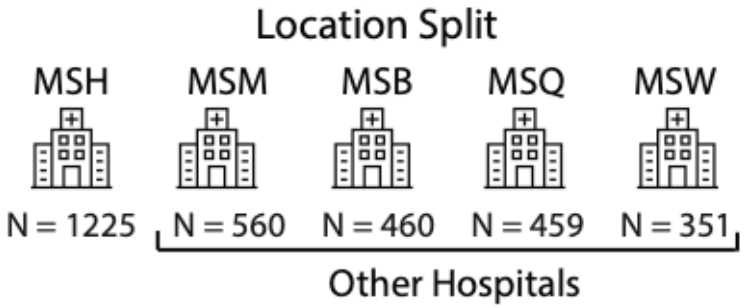
- **Internal validation:** Predictive models were built using data from patients who were admitted from March 15 to May 1, 2020
  - Models were trained and evaluated through stratified k-fold cross-validation to mitigate the variability of a single train-test split
  - Final model was trained for each outcome and time window using all patients
  - The model was assessed through a series of validation experiments
  - **Cutoff time for prospective evaluation:** May 1

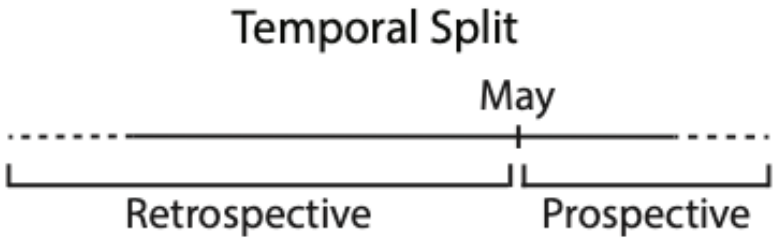


Vaid et al., Journal of Medical Internet Research, 2020

Study design

- **Generalizability of the model in a new setting:** Externally validated patients from other hospitals from March 15 to May 1, 2020, which was the same time frame used to train the model
- **Temporal generalizability:** Prospective validations of the model independently on both MSH and OH patients admitted from May 1 to May 22, 2020





Vaid et al., Journal of Medical Internet Research, 2020

Study data

- Demographics: Age, sex, race, and ethnicity
- Diagnosis codes based on ICD-9/10-CM codes and procedures to identify pre-existing conditions
- Laboratory orders within the timeframe of interest
  - Laboratory data below the 0.5th percentile and above the 99.5th percentile removed to avoid inclusion of obvious outliers that could represent incorrect documentation or measurement errors

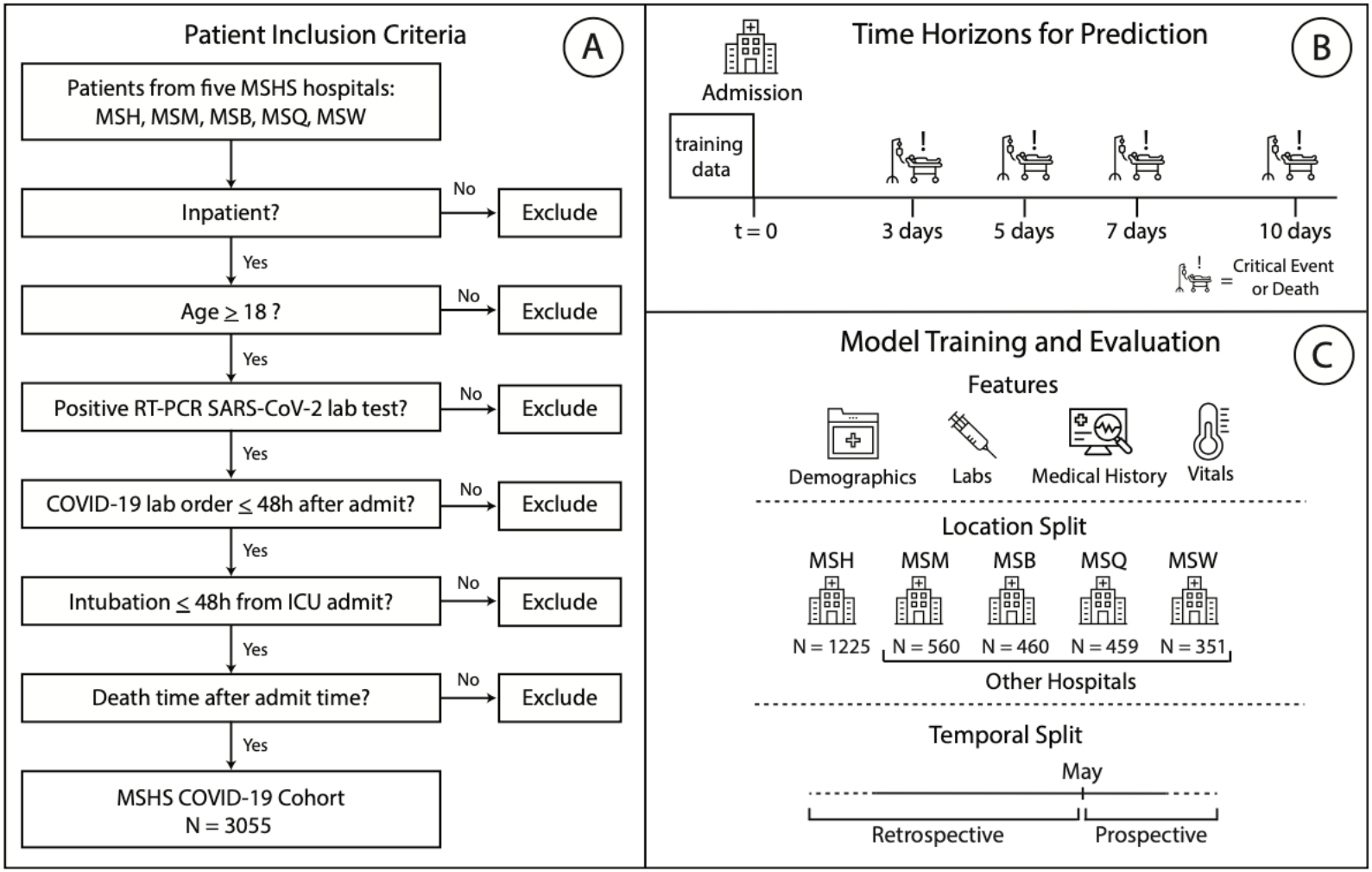
Vaid et al., Journal of Medical Internet Research, 2020

Definition of outcomes

- Primary outcomes
  - Death versus survival or discharge
  - Critical illness versus survival or discharge through time horizons of 3, 5, 7, and 10 days.
    - Critical illness was defined as discharge to hospice, intubation ≤48 hours prior to intensive care unit (ICU) admission, ICU admission, or death
    - Composite outcome (i.e., mortality as opposed to discharge or survival) was chosen to bypass issues of competing risks

Vaid et al., Journal of Medical Internet Research, 2020

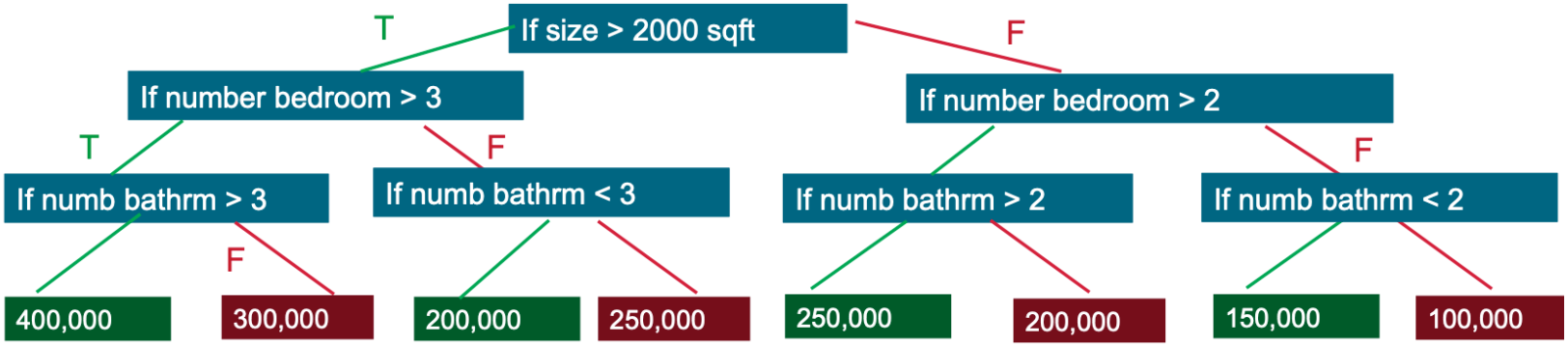
Study design and workflow: Recap



Vaid et al., Journal of Medical Internet Research, 2020

Boosted decision trees (XGBoost)

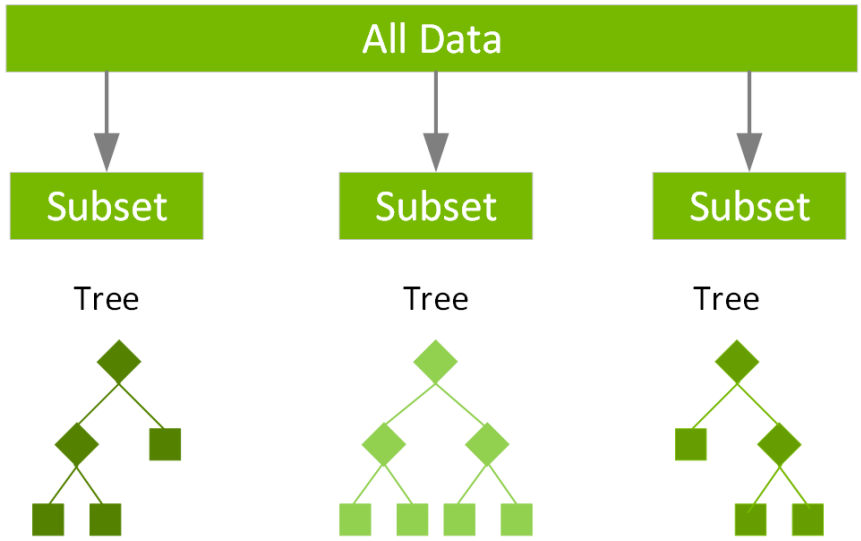
- Decision trees create a model that predicts the label by evaluating a tree of if-then-else true/false feature questions, and estimating the minimum number of questions needed to assess the probability of making a correct decision



https://xgboost.ai - XGBoost is a scalable and highly accurate implementation of gradient boosting

Boosted decision trees (XGBoost)

- XGBoost is a decision tree ensemble learning algorithm similar to random forest
  - Ensemble models combine multiple ML models to obtain a better model
  - Random forest and XGBoost build a model using multiple decision trees
    - The difference is in how the trees are built and combined



https://xgboost.ai - XGBoost is a scalable and highly accurate implementation of gradient boosting

Boosted decision trees (XGBoost)

Ensemble models:

- Bagging:
  - Build decision trees in parallel from random bootstrap samples of the dataset
  - Final prediction is an average of all decision tree predictions
- Boosting:
  - Build many weak models
  - Combining weak models to generate a collectively strong model

Gradient boosting:

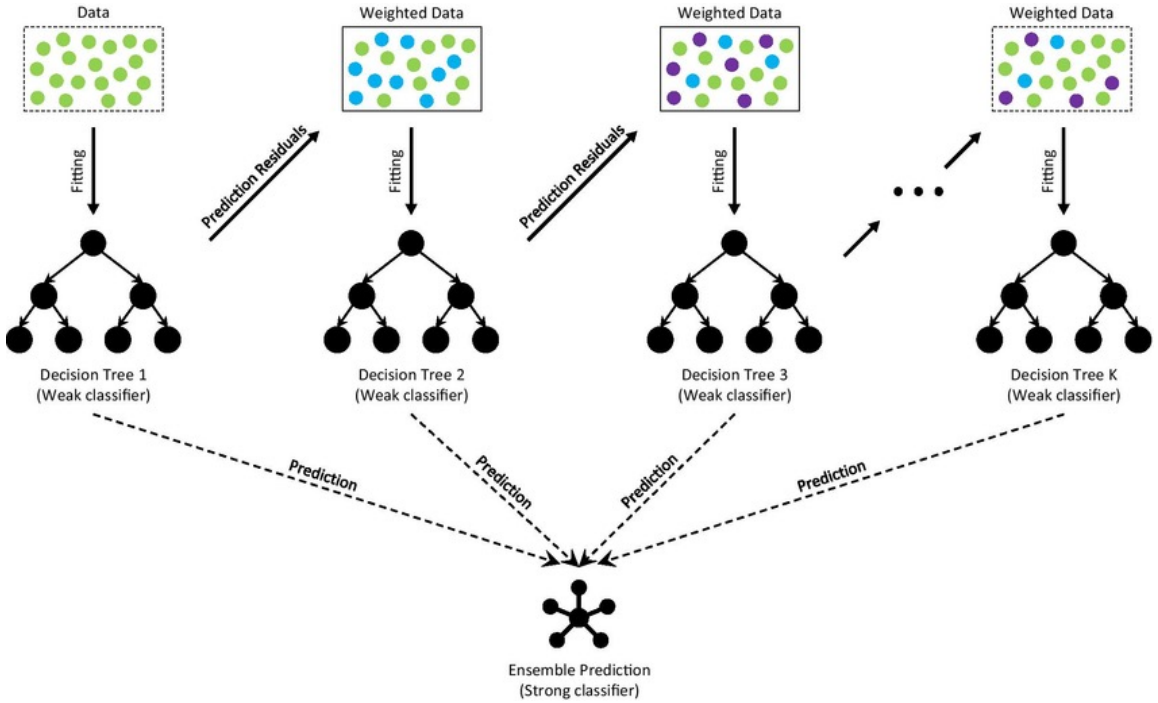
- Extension of boosting where the process of additively generating weak models is a gradient descent algorithm over an objective function
- Set targeted outcomes for the next model in an effort to minimize errors
- Targeted outcomes for each case are based on the gradient of the error (hence the name gradient boosting) with respect to the prediction
  - Iteratively train an ensemble of shallow decision trees
  - Each iteration uses error residuals of the previous model to fit the next model
  - Final prediction is a weighted sum of all tree predictions

https://xgboost.ai - XGBoost is a scalable and highly accurate implementation of gradient boosting

Boosted decision trees (XGBoost)

Outcomes for each tree in the ensemble are based on the gradient of the error w.r.t. prediction

1. Iteratively train an ensemble of shallow decision trees
2. Each iteration uses error residuals of the previous model to fit the next model
3. Final prediction is a weighted sum of all tree predictions



https://xgboost.ai - XGBoost is a scalable and highly accurate implementation of gradient boosting

Why do tree-based models still outperform deep learning on tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

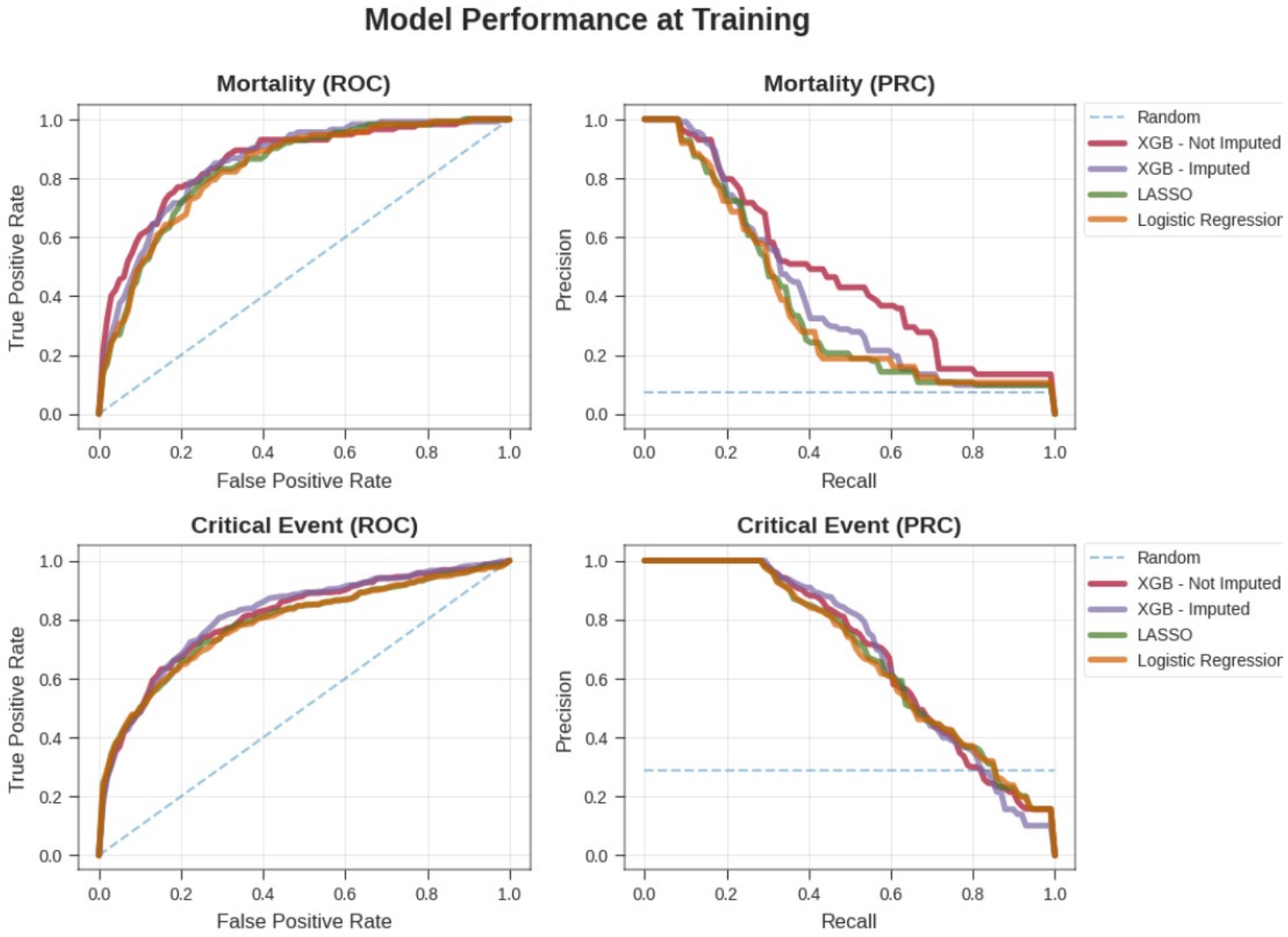
Edouard Oyallon
ISIR, CNRS, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay

Abstract

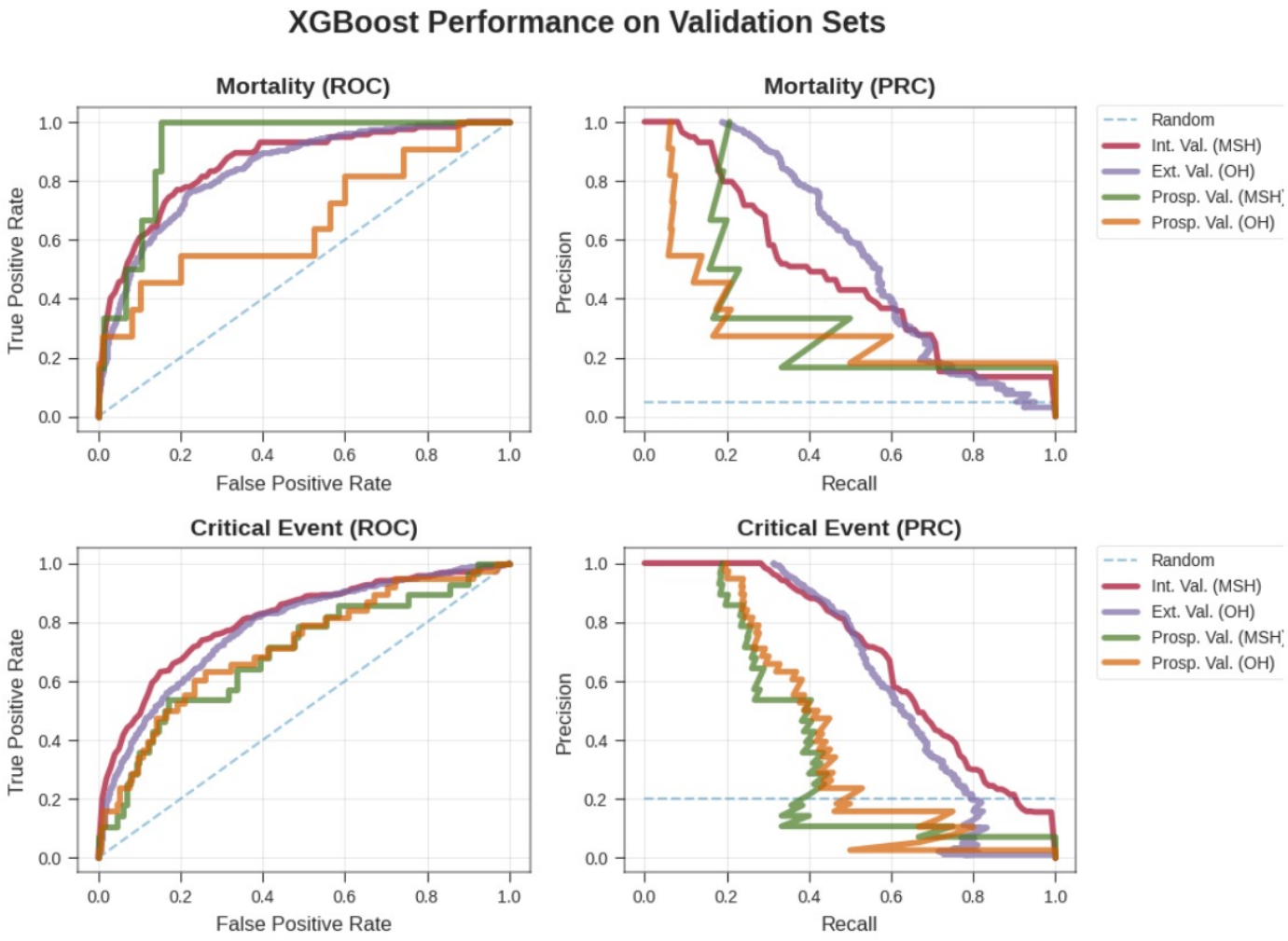
While deep learning has enabled tremendous progress on text and image datasets, its superiority on tabular data is not clear. We contribute extensive benchmarks of standard and novel deep learning methods as well as tree-based models such as XGBoost and Random Forests, across a large number of datasets and hyperparameter combinations. We define a standard set of 45 datasets from varied domains with clear characteristics of tabular data and a benchmarking methodology accounting for both fitting models and finding good hyperparameters. Results show that tree-based models remain state-of-the-art on medium-sized data ($\sim$10K samples) even without accounting for their superior speed. To understand this gap, we conduct an empirical investigation into the differing inductive biases of tree-based models and Neural Networks (NNs). This leads to a series of challenges which should guide researchers aiming to build tabular-specific NNs: **1.** be robust to uninformative features, **2.** preserve the orientation of the data, and **3.** be able to easily learn irregular functions. To stimulate research on tabular architectures, we contribute a standard benchmark and raw data for baselines: every point of a 20 000 compute hours hyperparameter search for each learner.

Results: Primary hospital



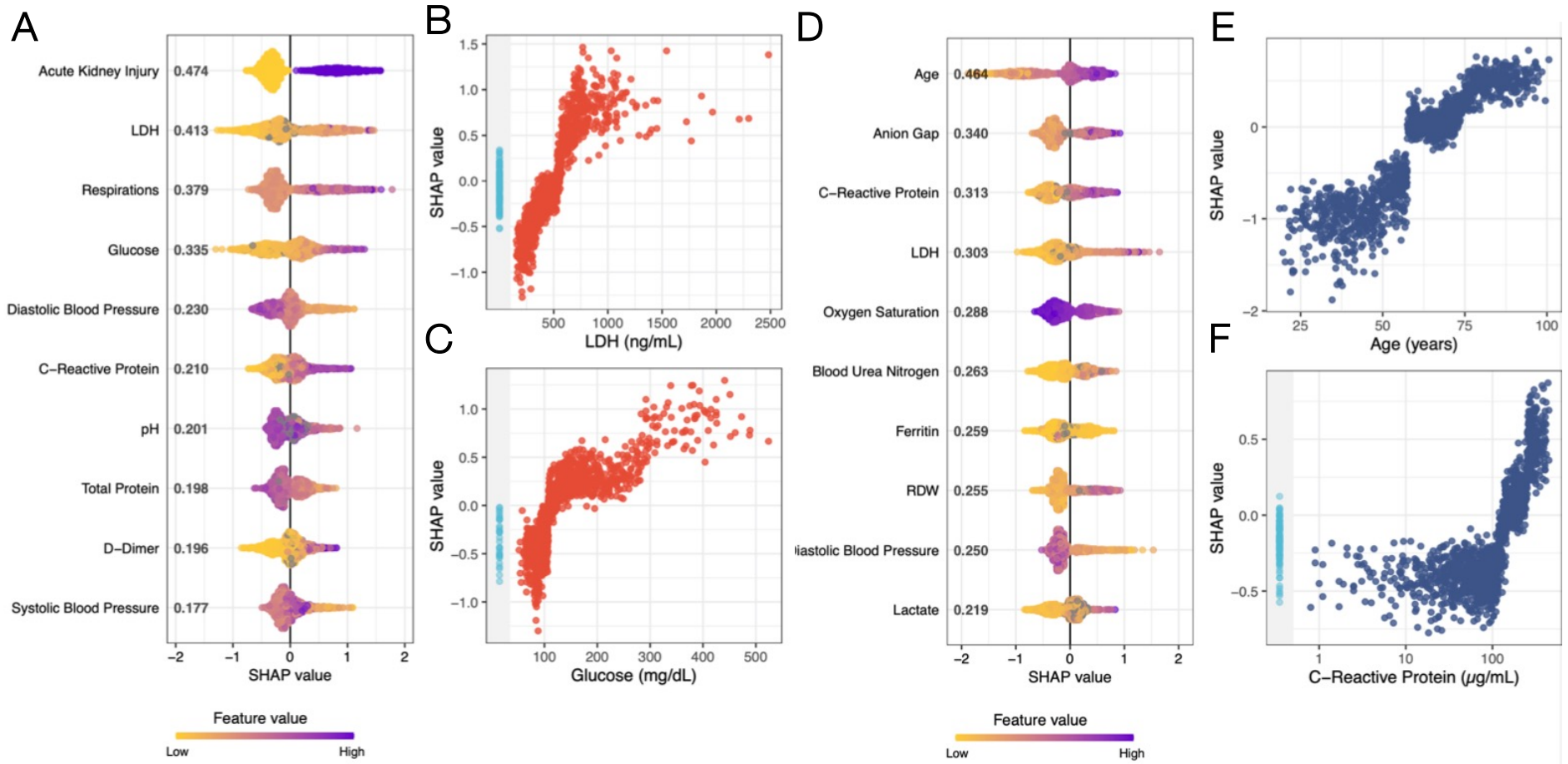
Vaid et al., Journal of Medical Internet Research, 2020

Results: External & temporal



Vaid et al., Journal of Medical Internet Research, 2020

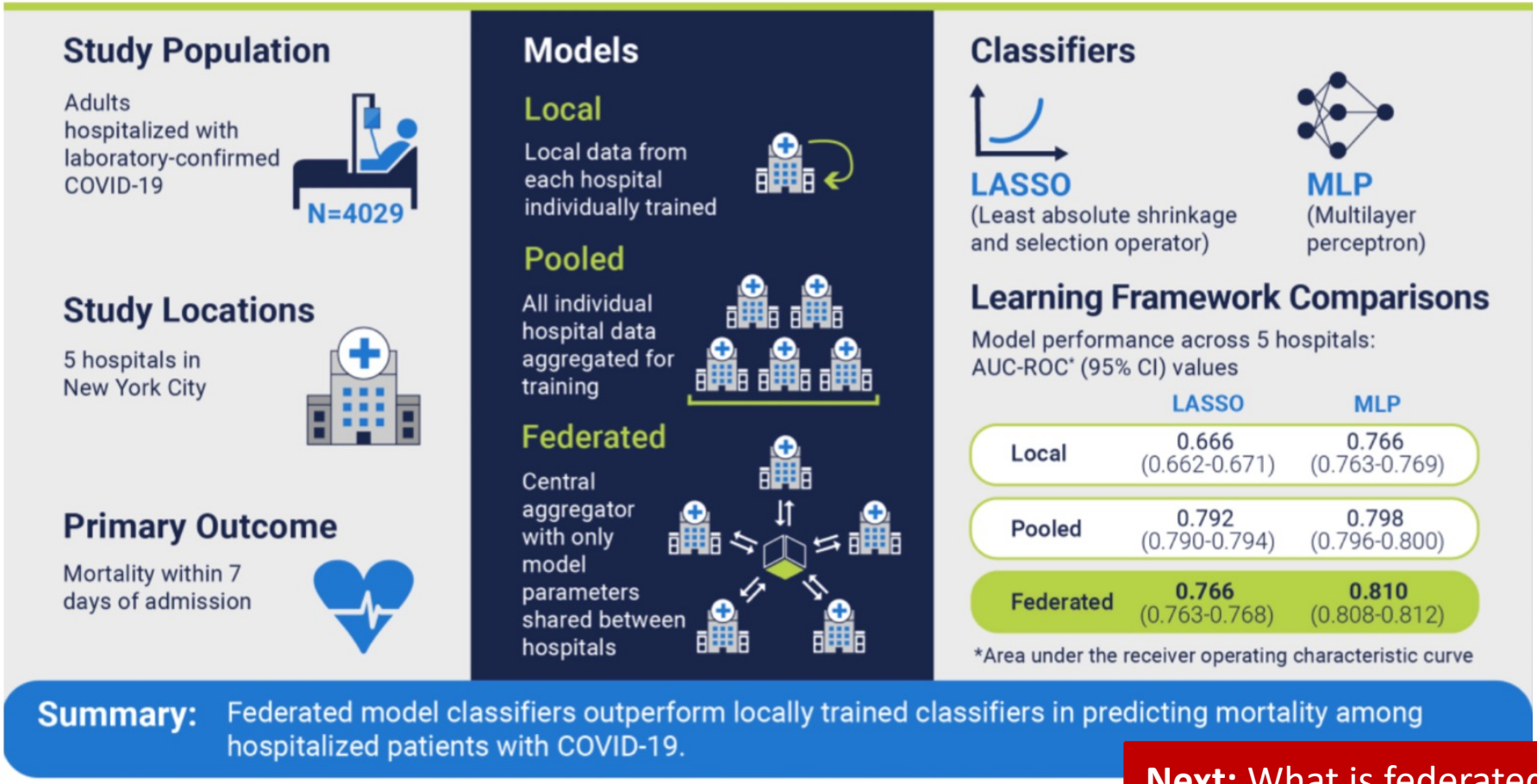
What did the model learn?



SHAP summary plots for critical event (A) and mortality (D) at 7 days showing the SHAP values for the 10 most important features for the respective XGBoost models. Features in the summary plots (y-axis) are organized by their mean absolute SHAP values (x-axis), which represent the importance of the features in driving the prediction of the classifiers for patients. (B) and (C) Dependency plots demonstrating how different values can affect the SHAP score and ultimately impact classifier decisions for LDH and glucose, respectively, for critical event prediction. (E) and (F) Dependency plots for age and C-reactive protein levels. LDH: lactate dehydrogenase; RDW: red cell distribution width; SHAP: SHapley Additive exPlanation.

Vaid et al., Journal of Medical Internet Research, 2020

Improving predictive capabilities using federated learning



Next: What is federated learning?

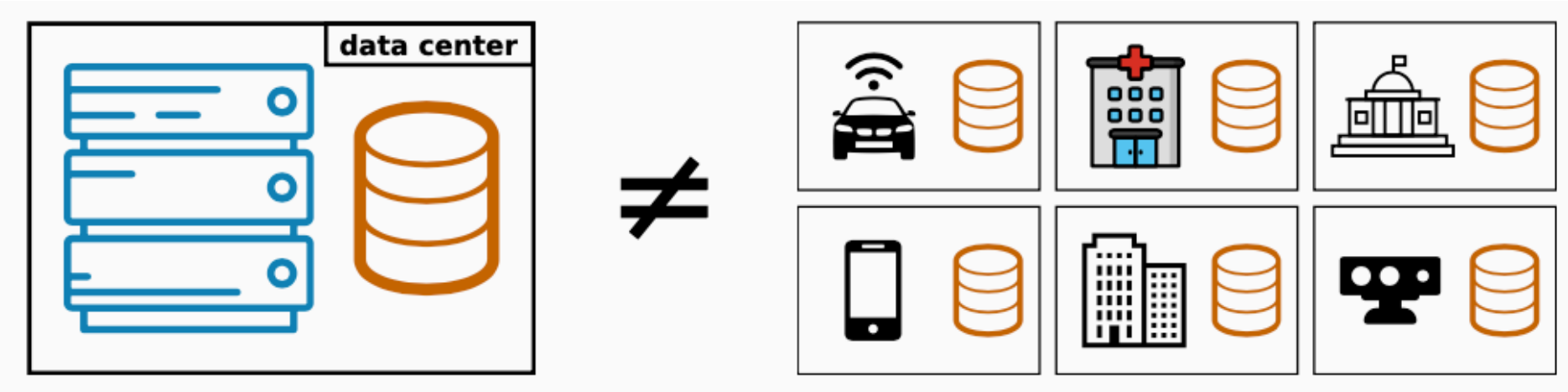
Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

Outline for today's class

1. Highlights of ML on EHR data:
   - Polypharmacy and adverse drug events
   - Modeling disease progression
   - Mortality and critical event prediction
2. Federated learning in healthcare
   - What is federated learning?
   - Why is it important for medical data?
3. What's next for clinical informatics research

A shift of paradigm: From centralized to decentralized data

- The standard setting in ML considers a centralized dataset processed in a tightly integrated system
- But in real-world data is often **decentralized across many parties**



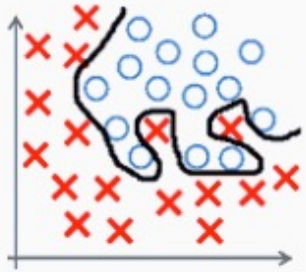
Why can't we just centralize the data?

1. Sending the data may be **too costly**
   - Self-driving cars are expected to generate several TBs of data a day
   - Some wireless devices have limited bandwidth/power

2. Data may be considered **too sensitive**
   - We see a growing public awareness and regulations on data privacy
   - Keeping control of data can give a competitive advantage in business and research

How about each party learning on its own?

1. The local dataset might be **too small**
   - Sub-par predictive performance (e.g., due to overfitting)
   - Non-statistically significant results (e.g., medical studies)

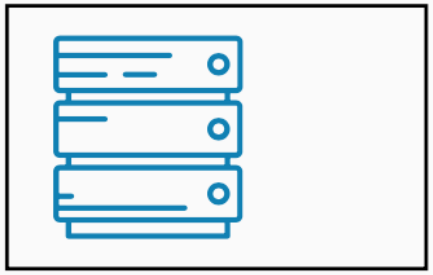


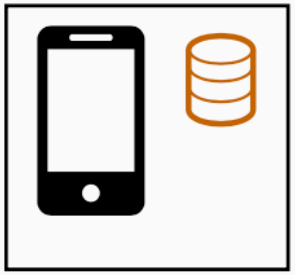
2. The local dataset might be **biased**
   - Not representative of the target distribution or patient population

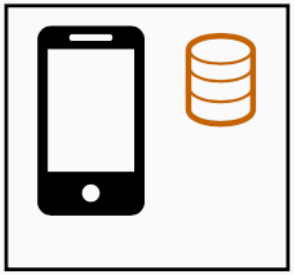
Broad definition of federated learning

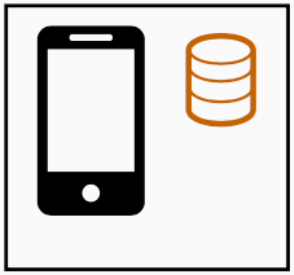
- Enable ML researchers to work productively with decentralized data with privacy by default
- It aims to **collaboratively train** ML models while **keeping the data centralized**
  - Federated learning allows for the decentralized refinement of independently built ML models via the iterative exchange of model parameters with a central aggregator, without sharing raw data
- We would like the final model to be as good as the centralized solution (ideally), or at least better than what each party can learn on its own

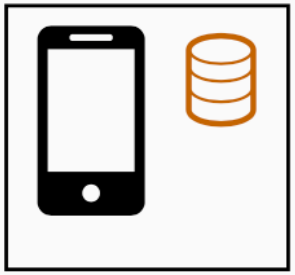
Broad definition of federated learning





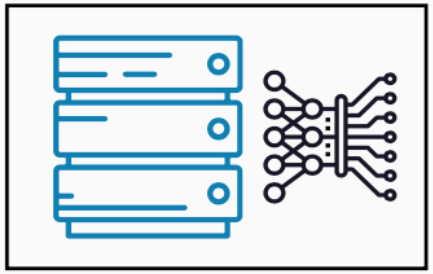


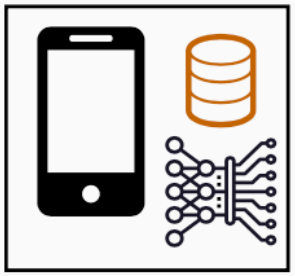


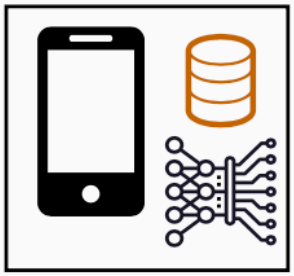


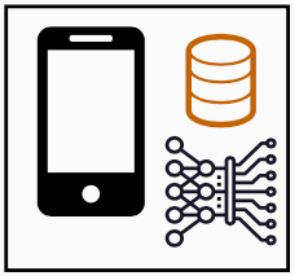
Broad definition of federated learning

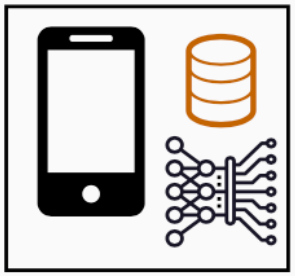
initialize model





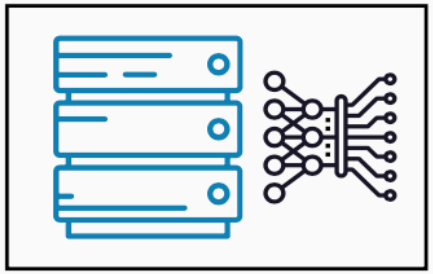


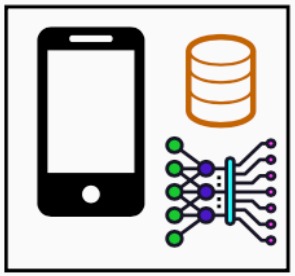


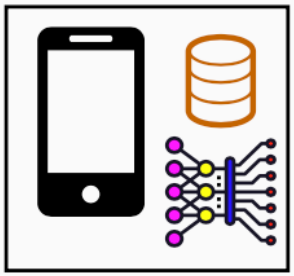


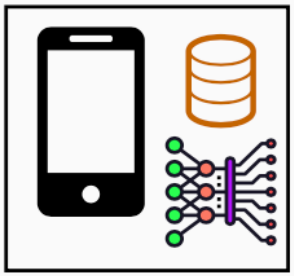
Broad definition of federated learning

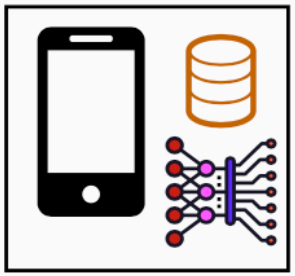
each party makes an update using its local dataset



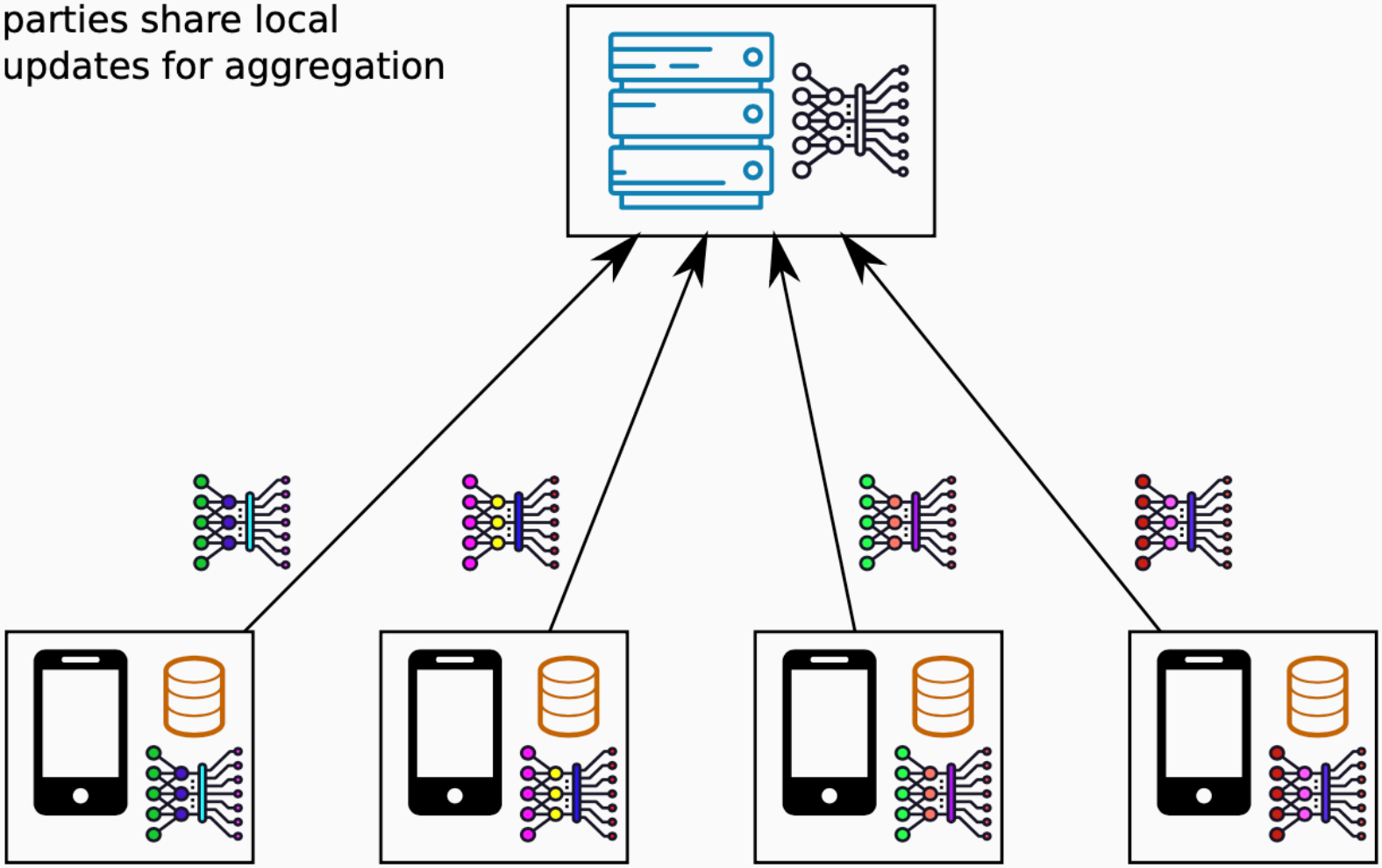




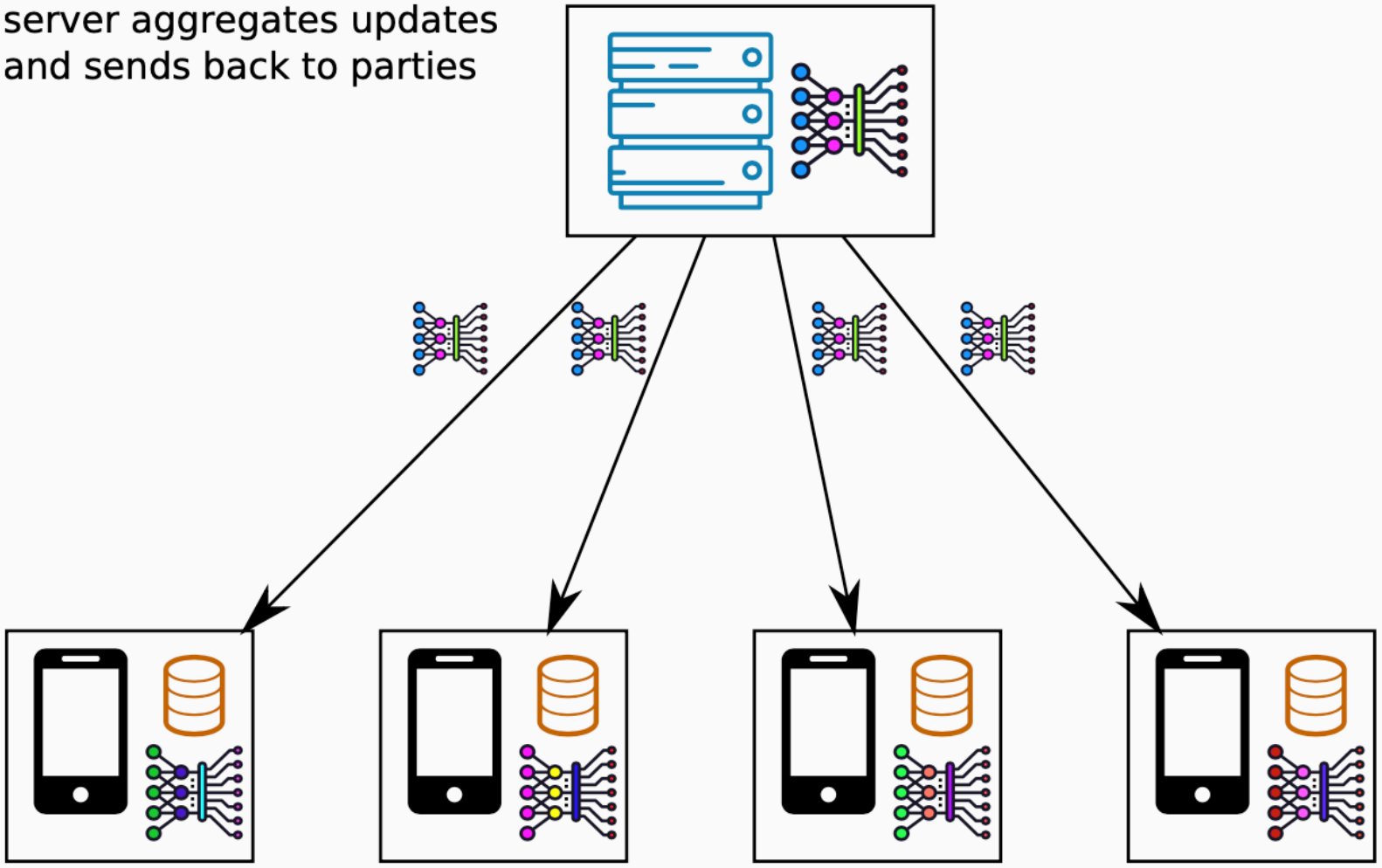




Broad definition of federated learning

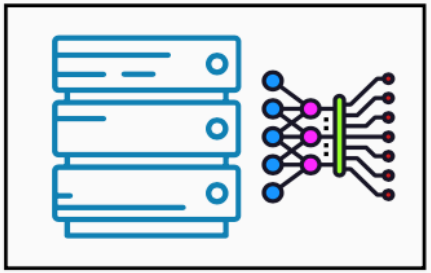


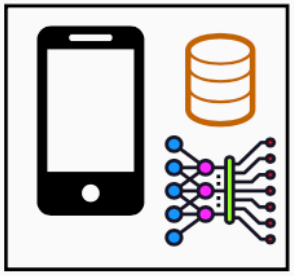
Broad definition of federated learning

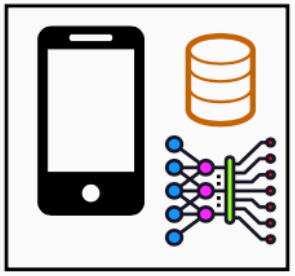


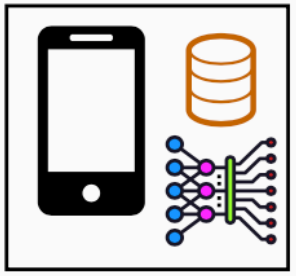
Broad definition of federated learning

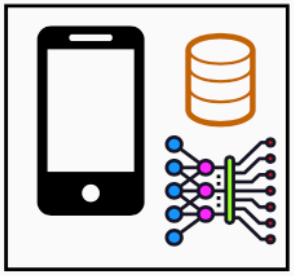
parties update their copy of the model and iterate









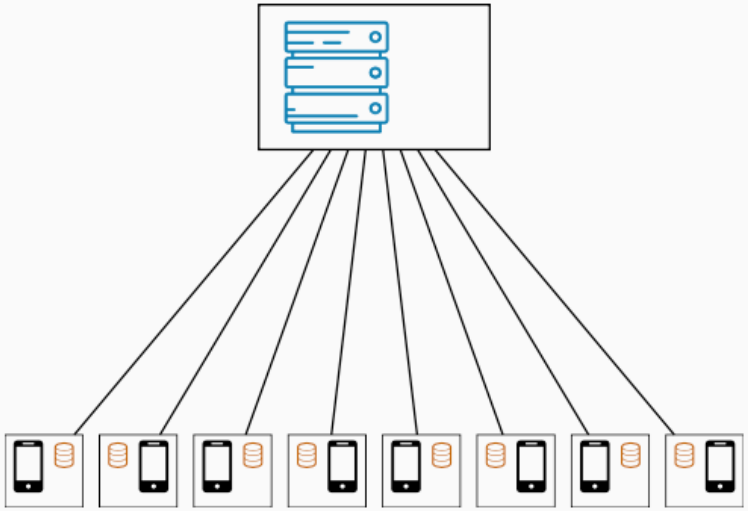


Key differences with distributed learning

- In **distributed learning**, data is centrally stored (e.g., in a data center)
  - The main goal is just to train faster
  - We control how data is distributed across parties: usually, it is distributed uniformly at random across parties
- In **federated learning**, data is naturally distributed and generated locally
  - Data is not independent and identically distributed (non-i.i.d.), and it is imbalanced
  - Additional challenges arise in federated learning:
    - Enforcing privacy constraints
    - Dealing with possibly limited reliability/availability of parties
    - Achieving robustness against malicious parties

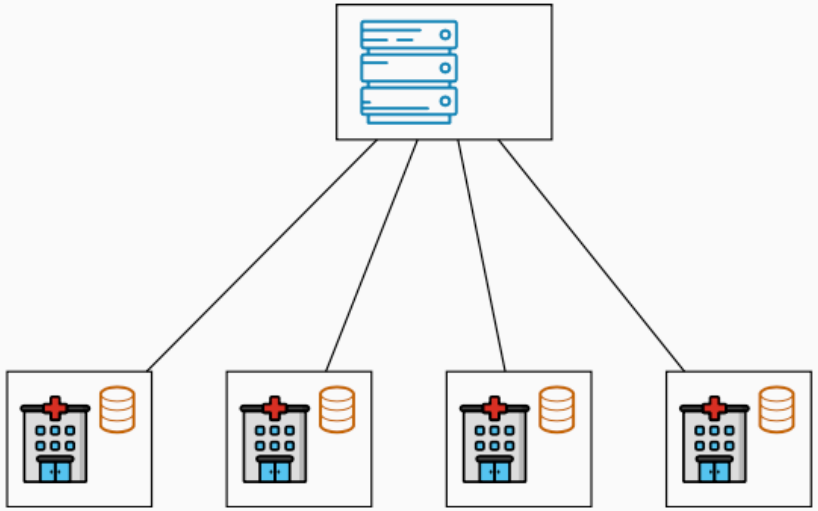
Cross-device vs. cross-silo

Cross-device federated learning



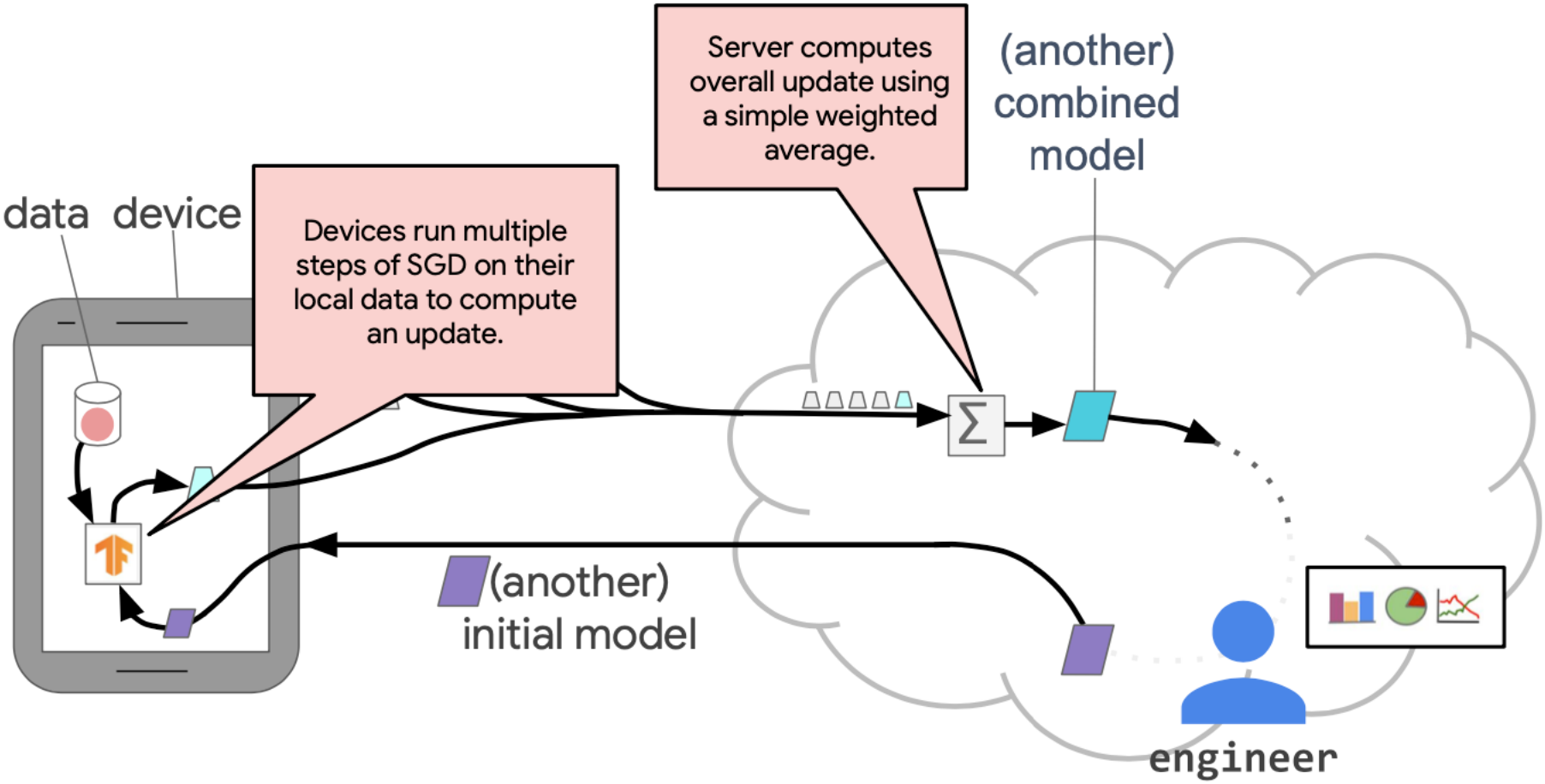
- Massive number of parties (up to 10^10)
- Small dataset per party (could be size 1)
- Limited availability and reliability
- Some parties might be malicious

Cross-silo federated learning



- 2-100 parties
- Medium to large dataset per party
- Reliable parties, almost always available
- Parties are typically honest

Federated averaging (FedAvg)



McMahan, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data. AISTATS 2017

FedAvg: Notation

- We consider a set of *K* parties (clients)
- Each party *k* holds a dataset D_k of n_k points
- Let $D = D_1 \cup \cdots \cup D_K$ be the joint dataset and $n = \sum_k n_k$ the total number of points
- We want to solve problems of the form $\min_{\theta \in \mathbb{R}^p} F(\theta, D)$ where:

$$F(\theta, D) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\theta, D_k) \text{ and } F(\theta; D_k) = \sum_{d \in D_k} f(\theta, d)$$

- $\theta \in \mathbb{R}^p$ are model parameters (e.g., weights of a logistic regression or a neural network)
- This covers a broad class of ML problems that can be formulated as empirical risk minimization problems

McMahan, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data. AISTATS 2017

FedAvg: Algorithm

Algorithm FedAvg (server-side)

Parameters: client sampling rate ρ

initialize θ
for each round $t = 0, 1, \dots$ **do**
  $\mathcal{S}_t \leftarrow$ random set of $m = \lceil \rho K \rceil$ clients
  for each client $k \in \mathcal{S}_t$ in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \sum_{k \in \mathcal{S}_t} \frac{n_k}{n} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: batch size B, number of local steps L, learning rate η

for each local step $1, \dots, L$ **do**
  $\mathcal{B} \leftarrow$ mini-batch of B examples from $\mathcal{D}_k$
  $\theta \leftarrow \theta - \frac{n_k}{B} \eta \sum_{d \in \mathcal{B}} \nabla f(\theta; d)$
send θ to server

- For $L = 1$ and $\rho = 1$, it is equivalent to classic parallel learning: updates are aggregated, and the model synchronized at each step
- For $L > 1$: each client performs multiple local updates to the model before communicating

McMahan, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data. AISTATS 2017

FedAvg: Algorithm

Server

Until Converged:

1. Select a random subset of clients

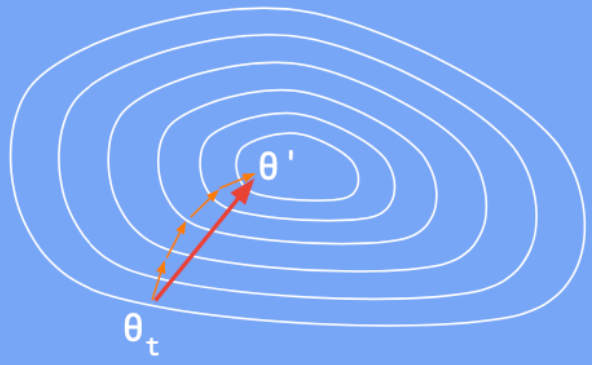
2. In parallel, send current parameters θ_t to those clients

Selected Client k

1. Receive θ_t from server.

2. Run some number of minibatch SGD steps, producing θ'

3. Return $\theta' - \theta_t$ to server.

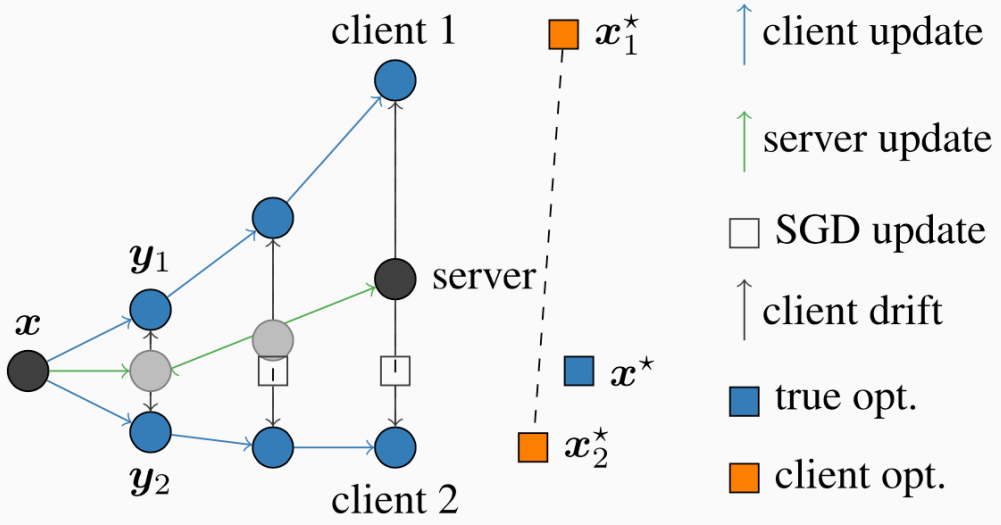


3. $\theta_{t+1} = \theta_t +$ data-weighted average of client updates

H. B. McMahan, *et al*. **Communication-Efficient Learning of Deep Networks from Decentralized Data.** AISTATS 2017

Challenge in federated learning: dealing with non i.i.d. data

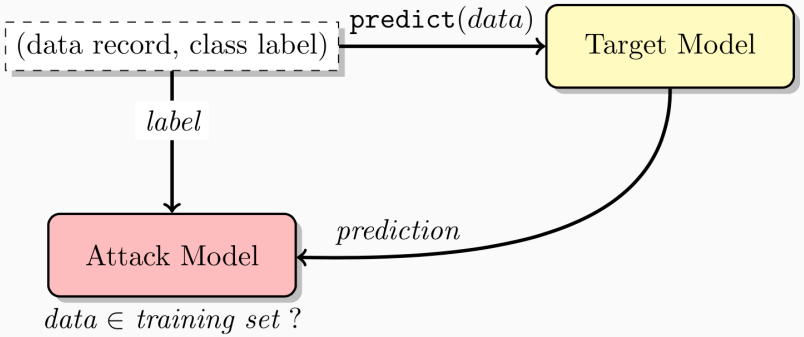
- **Problem:** Client drift, when local datasets are non-i.i.d., FedAvg suffers from client drift
- **Solution:** To avoid this drift, one must use fewer local updates and/or smaller learning rates, which can hurt convergence of the model



(Figure taken from [Karimireddy et al., 2020])

Challenge in federated learning: preserving privacy

- **Problem:** ML models are susceptible to attacks on data privacy
  - Membership inference attacks try to infer the presence of a known individual in the training set, e.g., by exploiting the confidence in model predictions

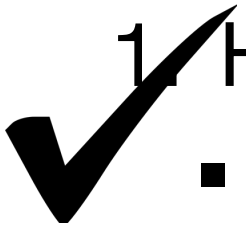


  - Reconstruction attacks try to infer some of the points used to train the model, e.g., by differencing attacks

- FedAvg offers an additional attack surface because server and clients observe model updates (not only the final model)

- **Solution:** Differential privacy (opt. reading: Sabater et al., 2020)

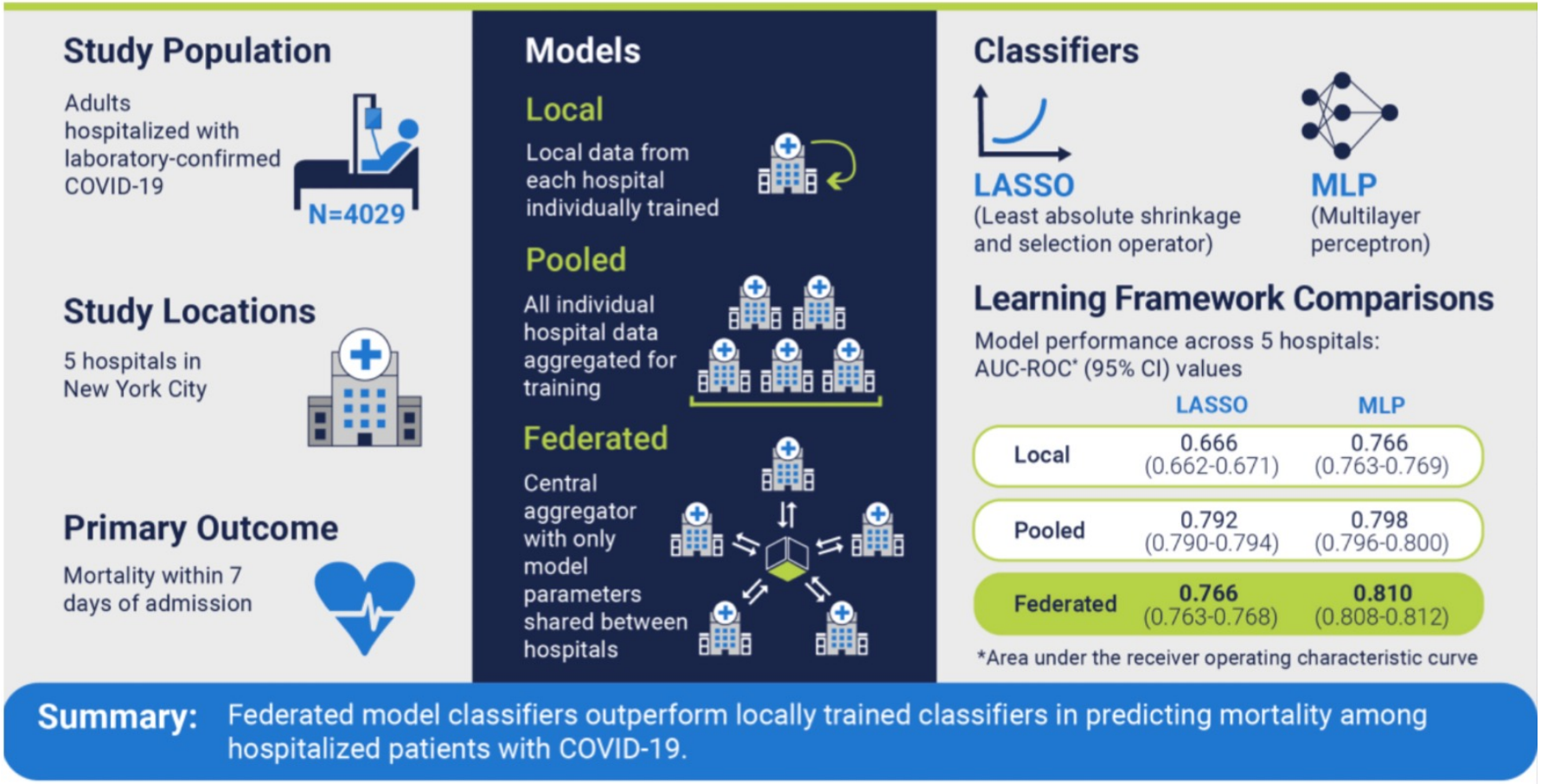
Outline for today's class



1. Highlights of ML on EHR data:
   - Polypharmacy and adverse drug events
   - Modeling disease progression
   - Mortality and critical event prediction
2. Federated learning in healthcare
   - What is federated learning?
   - Why is it important for medical data?

3. What's next for clinical informatics research

Example: Improving predictive capabilities with federated learning



Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

Overview of the study

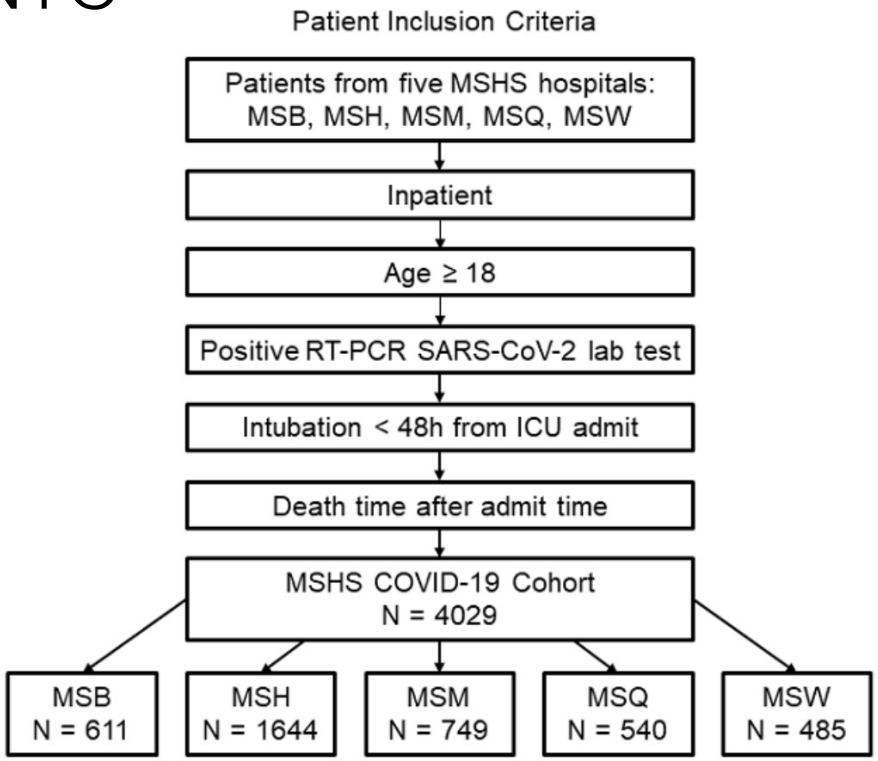
- **Motivation:** ML models require large datasets that may be siloed across different health care institutions
- **Objective:** Use federated learning to avoid aggregating clinical data across 5 hospitals, to predict mortality in hospitalized COVID-19 patients within 7 days

- **Approach:**
  - Patient data are collected from EHRs of 5 hospitals within Mount Sinai Health System
  - Logistic regression with L1 regularization/least absolute shrinkage and selection operator (**LASSO**) and multilayer perceptron (**MLP**) models are trained by using local data at each site
  - **Pooled model:** Combine data from all 5 sites
  - **Federated model:** Only share parameters with a central aggregator

Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

Data and study population (1/2)

Patients who tested positive for COVID-19 (N=4029) were derived from the EHRs of 5 Mount Sinai hospitals in NYC



Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

Data and study population (2/2)

- Study data included the demographics, past medical history, vital signs, lab test results, and outcomes of all patients
- Varying prevalence of COVID-19 across hospitals → assess multiple class balancing techniques

Table 1. Demographic characteristics of all hospitalized patients with COVID-19 included in this study (N=4029)[a].

Characteristic	Mount Sinai Brooklyn	Mount Sinai Hospital	Mount Sinai Morningside	Mount Sinai Queens	Mount Sinai West	*P* value
Number of patients, n	611	1644	749	540	485	—[b]
Gender, n (%)						
Male	338 (55.3)	951 (57.8)	411 (54.9)	344 (63.7)	257 (53.0)	.004
Female	273 (44.7)	693 (42.2)	338 (45.1)	196 (36.3)	228 (47.0)	.004
Age (years), median (IQR)	72.5 (63.6-82.7)	63.3 (51.3-73.2)	69.8 (57.4-80.3)	68.1 (57.1-78.8)	66.3 (52.5-77.6)	<.001
Ethnicity, n (%)						
Hispanic	21 (3.4)	460 (28.0)	259 (34.6)	198 (36.7)	111 (22.9)	<.001
Non-Hispanic	416 (68.1)	892 (54.3)	452 (60.3)	287 (53.1)	349 (72.0)	<.001
Unknown	174 (28.5)	292 (17.8)	38 (5.1)	55 (10.2)	25 (5.2)	<.001
Race, n (%)						
Asian	13 (2.1)	83 (5.0)	16 (2.1)	56 (10.4)	27 (5.6)	<.001
Black/African American	323 (52.9)	388 (23.6)	266 (35.5)	64 (11.9)	109 (22.5)	<.001
Other	54 (8.8)	705 (42.9)	343 (45.8)	288 (53.3)	164 (33.8)	<.001
Unknown	27 (4.4)	87 (5.3)	25 (3.3)	14 (2.6)	14 (2.9)	<.001
White	194 (31.8)	381 (23.2)	99 (13.2)	118 (21.9)	171 (35.3)	<.001
Past medical history, n (%)						
Acute myocardial infarction	14 (2.3)	16 (1.0)	—	15 (2.8)	7 (1.4)	.006
Acute respiratory distress syndrome	—	28 (1.7)	—	—	—	<.001
Acute venous thromboembolism	—	11 (0.7)	—	—	—	.74
Asthma	—	100 (6.1)	39 (5.2)	19 (3.5)	27 (5.6)	<.001
Atrial fibrillation	23 (3.8)	113 (6.9)	44 (5.9)	49 (9.1)	28 (5.8)	.005
Cancer	22 (3.6)	190 (11.6)	47 (6.3)	21 (3.9)	41 (8.5)	<.001
Chronic kidney disease	46 (7.5)	208 (12.7)	75 (10.0)	81 (15.0)	33 (6.8)	<.001

Outcome: prediction target for ML models

Mortality within 7 days, n (%)	148 (24.2)	118 (7.2)	93 (12.4)	124 (23.0)	27 (5.6)	<.001

Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

Study design: 3 experiments

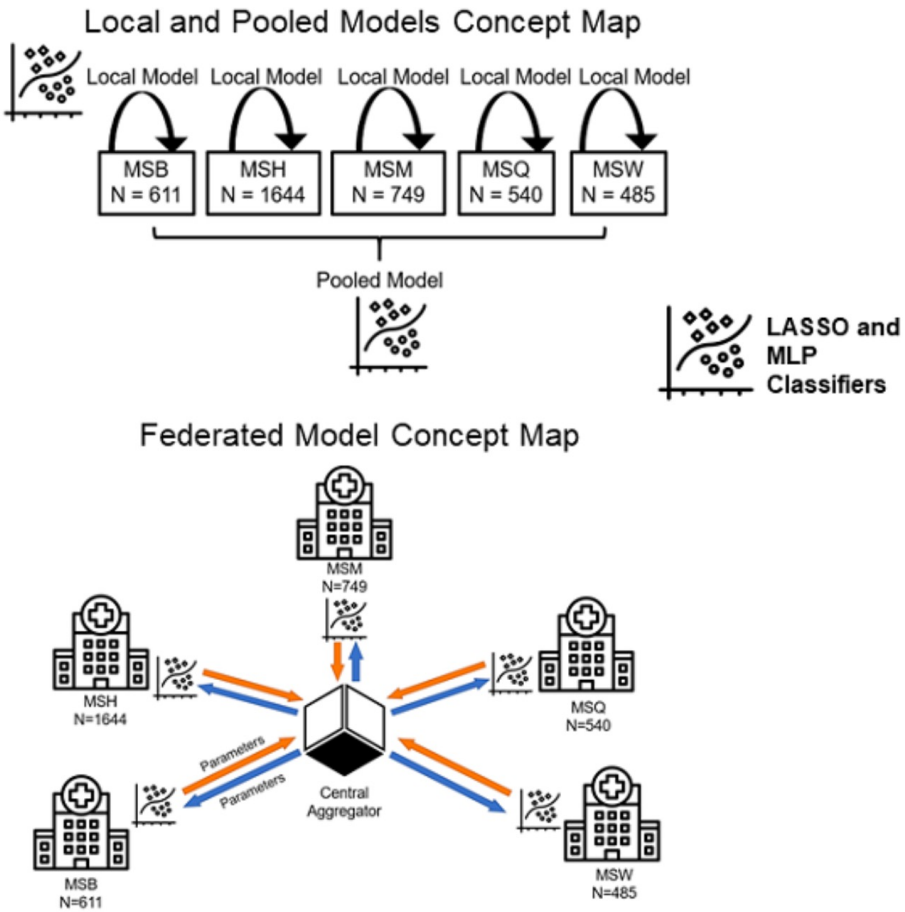
1. **Local model:** Classifiers that used and were tested on local data from each hospital separately
2. **Federated model:** Federated learning model by aggregating the model parameters of each individual hospital
3. **Pooled model:** Combine all individual hospital data into a superset to develop a pooled model that represents an ideal framework

Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

Study design: 3 experiments

- Local models only use data from the site itself
- Pooled model incorporates data from all sites
- Both the local and pooled MLP and LASSO models are used

- Federated model:
  - Parameters from a central aggregator are shared with each site
  - Sites do not have direct access to clinical data from other sites
  - After the models are locally trained at a site, parameters are sent back to the central aggregator
  - Aggregator updates federated model parameters



Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

ML model development

- **Primary outcome:** mortality within 7 days of admission
- **Baseline models:** 2 algorithms
  - Multilayer perceptron (MLP)
  - Logistic regression with LASSO regularization
- **Primary model of interest:** federated learning model
  1. Central aggregator is used to initialize the federated model with random parameters
  2. This model is sent to each site and trained for 1 epoch/iteration
  3. Afterward, model parameters are sent back to the central aggregator
  4. Federated averaging (**FedAvg**) is performed
     - Recall: FedAvg scales parameters of each site according to the number of available data points and sums all parameters by layer
  5. Updated parameters from the central aggregator are sent back to each site
  6. This cycle was repeated for multiple epochs/iterations.

Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

Overview of results

- Results:
  - Federated vs. local:
    - $LASSO_{federated}$ model outperformed $LASSO_{local}$ model at 3 hospitals
    - $MLP_{federated}$ model outperformed MLP_{local} model at 5 hospitals
  - Federated vs. pooled:
    - $LASSO_{pooled}$ model outperformed the $LASSO_{federated}$ model at all hospitals
    - $MLP_{federated}$ model outperformed the MLP_{pooled} model at 2 hospitals.
- Findings:
  - Federated learning of COVID-19 EHRs can produce robust ML models without compromising patient privacy

Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

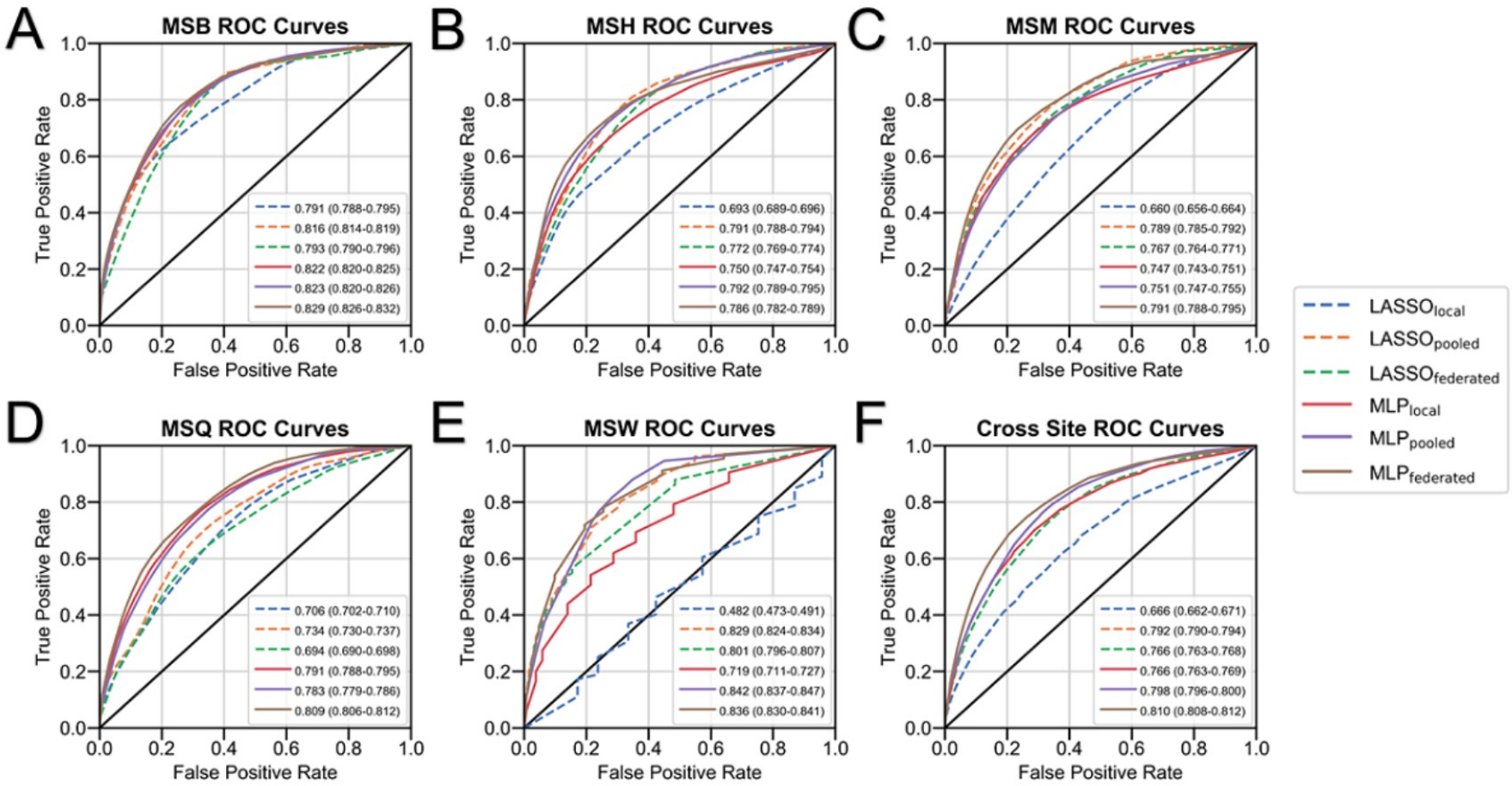
Results: Classifier performance

Table 2. Performance of the local, pooled, and federated LASSO[a] and MLP[b] models at each site, based on AUROCs[c] with 95% confidence intervals.

Model	Mount Sinai Brooklyn (n=611), AUROC (95% CI)	Mount Sinai Hospital (n=1644), AUROC (95% CI)	Mount Sinai Morningside (n=749), AUROC (95% CI)	Mount Sinai Queens (n=540), AUROC (95% CI)	Mount Sinai West (n=485), AUROC (95% CI)
LASSO model					
Local	0.791 (0.788-0.795)	0.693 (0.689-0.696)	0.66 (0.656-0.664)	0.706 (0.702-0.710)	0.482 (0.473-0.491)
Pooled	0.816 (0.814-0.819)	0.791 (0.788-0.794)	0.789 (0.785-0.792)	0.734 (0.730-0.737)	0.829 (0.824-0.834)
Federated	0.793 (0.790-0.796)	0.772 (0.769-0.774)	0.767 (0.764-0.771)	0.694 (0.690-0.698)	0.801 (0.796-0.807)
MLP model					
Local	0.822 (0.820-0.825)	0.750 (0.747-0.754)	0.747 (0.743-0.751)	0.791 (0.788 - 0.795)	0.719 (0.711-0.727)
Pooled	0.823 (0.820-0.826)	0.792 (0.789-0.795)	0.751 (0.747-0.755)	0.783 (0.779-0.786)	0.842 (0.837-0.847)
Federated (no noise	0.829 (0.826-0.832)	0.786 (0.782-0.789)	0.791 (0.788-0.795)	0.809 (0.806-0.812)	0.836 (0.83-0.841)

Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

Results: Classifier performance



Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

Key findings

- $MLP_{federated}$ and $LASSO_{federated}$ models outperform their respective local ML models at most hospitals
- Results show the potential of federated learning in overcoming the drawbacks of fragmented local ML models
- Scenarios in which federated models should either be approached with caution or favored:
  - Mount Sinai Queens hospital:
    - The only hospital where $LASSO_{federated}$ model performed worse than $LASSO_{local}$ model
    - This may have been attributed to the hospital having a smaller sample size (n=540) and higher mortality prevalence (23%) than other sites
  - Mount Sinai West hospital:
    - $LASSO_{local}$ model severely underperformed compared to $LASSOf_{ederated}$ model, with an AUROC difference of 0.319.
    - Mount Sinai West hospital had the lowest sample size (n=485) and the lowest COVID-19 mortality prevalence (5.6%) among all hospitals
  - This finding emphasizes the benefit of using federated learning for sites with small sample sizes and large class imbalances

Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

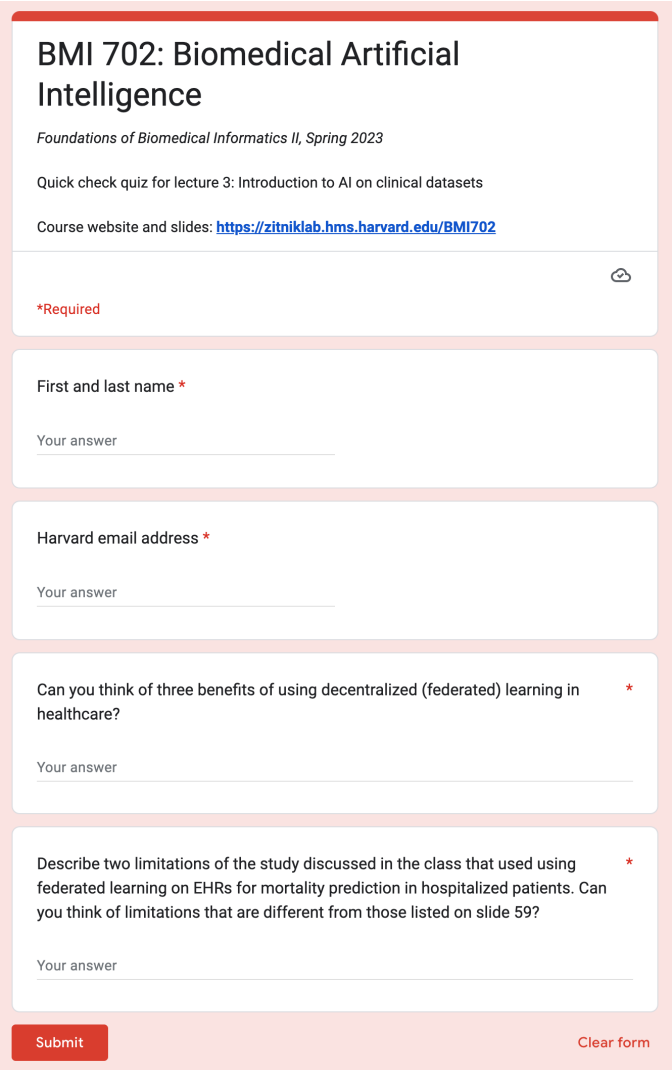
What are limitations of this study?

- Data collection was limited to Mount Sinai hospitals in NYC:
  - This may limit model generalizability to hospitals in other regions
- Focus on applying federated learning to the prediction of outcomes based on patient EHRs rather than creating an operational framework for immediate deployment
- Various aspects of federated learning are not addressed:
  - For example, load balancing, convergence, and scaling
- Only clinical data were included. Other data types could be incorporated
- The study implemented only two widely used classifiers:
  - Other ML algorithms may perform better (more on that in future lectures)
  - Although identical MLP architectures were used across all experiments, these architectures could have been further optimized
- Future work need to focus on model accessibility and comprehensive analysis of federated models to improve scalability, understand feature importance, and integrate additional data modalities

Vaid and Jaladanki et al., Journal of Medical Internet Research, 2020

Quick Check

https://forms.gle/iaYBgBRB7viwXfMa8



BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2023

Quick check quiz for lecture 3: Introduction to AI on clinical datasets

Course website and slides: https://zitniklab.hms.harvard.edu/BMI702

*Required

First and last name *

Your answer

Harvard email address *

Your answer

Can you think of three benefits of using decentralized (federated) learning in healthcare? *

Your answer

Describe two limitations of the study discussed in the class that used using federated learning on EHRs for mortality prediction in hospitalized patients. Can you think of limitations that are different from those listed on slide 59? *

Your answer

Submit

Clear form

Outline for today's class

1. Highlights of ML on EHR data:
   - Polypharmacy and adverse drug events
   - Modeling disease progression
   - Mortality and critical event prediction
2. Federated learning in healthcare
   - What is federated learning?
   - Why is it important for medical data?

3. What's next for clinical informatics research

Utility of ML for clinical datasets

- Must be **FAIR** (trained, assessed, and tuned to work on all populations)
- Must be **generalizable across health systems**
- Need to **overcome data limitations** (include non-traditional, multi-omic data)
- Must be tested prospectively in **proper trial framework (silent pilot first)**
- Must serve a specific use/clinically relevant task

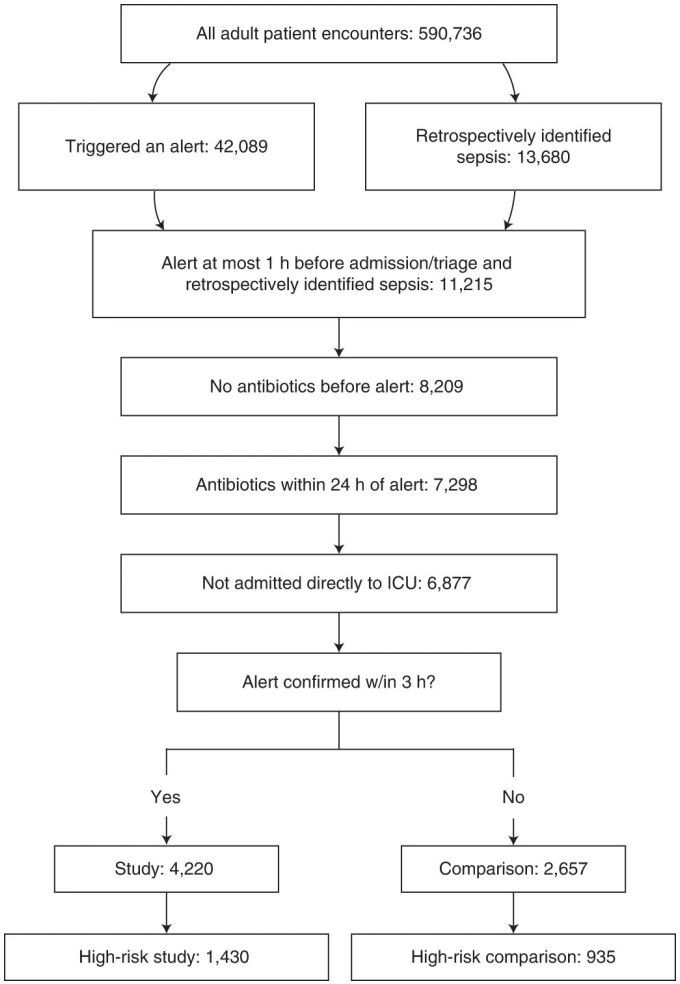
- They must be **implemented/operationalized within the health system**
  - How can we embed something in care process without disrupting it?
  - Who receives results? How are results presented? When are results presented?
  - How to give feedback? Must be iterative cycle

Example: Targeted real-time early warning system

Article | Published: 21 July 2022

Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis

Early recognition and treatment of sepsis are linked to improved patient outcomes. Machine learning-based early warning systems may reduce the time to recognition, but few systems have undergone clinical evaluation. In this prospective, multi-site cohort study, we examined the association between patient outcomes and provider interaction with a deployed sepsis alert system called the Targeted Real-time Early Warning System (TREWS). During the study, 590,736 patients were monitored by TREWS across five hospitals. We focused our analysis on 6,877 patients with sepsis who were identified by the alert before initiation of antibiotic therapy. Adjusting for patient presentation and severity, patients in this group whose alert was confirmed by a provider within 3 h of the alert had a reduced in-hospital mortality rate (3.3%, confidence interval (CI) 1.7, 5.1%, adjusted absolute reduction, and 18.7%, CI 9.4, 27.0%, adjusted relative reduction), organ failure and length of stay compared with patients whose alert was not confirmed by a provider within 3 h. Improvements in mortality rate (4.5%, CI 0.8, 8.3%, adjusted absolute reduction) and organ failure were larger among those patients who were additionally flagged as high risk. Our findings indicate that early warning systems have the potential to identify sepsis patients early and improve patient outcomes and that sepsis patients who would benefit the most from early treatment can be identified and prioritized at the time of the alert



Adams et al., Nature Medicine, 2022