National Institute of
Diabetes and Digestive
and Kidney Diseases

Conducting AI-Research for Health Data

Summer Rankin, PhD, Booz Allen Hamilton

NIDDK-CR Office Hours: December 29, 2023 through January 5, 2024

Speaker Introduction

Summer Rankin, PhD, is a computational neuroscientist who investigates the boundaries of AI and drives data science solutions for federal government clients.

She has a doctorate in complex systems and brain sciences and works as a senior lead data scientist at Booz Allen Hamilton's Honolulu Chief Technology Office.

She leads projects that involve a range of machine learning techniques including: deep learning, natural language processing, anomaly detection, and performance measurement.

She serves as an artificial intelligence subject matter expert for Indo-Pacific defense and health projects with recent publications modeling mortality rates in chronic kidney disease (ONC) and adverse event detection from EHRs (FDA).

She holds a PhD in Complex Systems and Brain Sciences and completed a postdoctoral fellowship with Charles Limb, MD at Johns Hopkins School of Medicine.

She has multiple peer-reviewed publications, public software releases, and conference presentations in the fields of AI, data science and neuroscience.

National Institute of
Diabetes and Digestive
and Kidney Diseases

Agenda

Overview of AI-Assisted Research

Bias in AI

Example of ML model for chronic kidney disease (CKD)

Q&A

What questions can be answered with AI?

AI is an outcome*—the ability of machines to perform tasks that typically require human-level intelligence*

	perception	**notification**	**suggestion**	**automation**	**prediction**	**prevention**	**situational awareness**
	Describe and understand surroundings	*Provide alerts, reminders, etc.*	*Build on past preferences and modify over time*	*Follow routine steps to accomplish an objective*	*Forecast the likelihood of future events based on past events*	*Apply cognitive reckoning to identify potential threats*	*Summarize the current, and likely future, environment*
Key Questions Answered	What's happening now?	What do I need to know?	What do you recommend?	What should I do?	What can I expect to happen?	What can/should I avoid?	What do I need to do now?

THE CURRENT ROLE OF AI:

Curator — Recommender — Orchestrator

NOT THE ROLE OF AI:

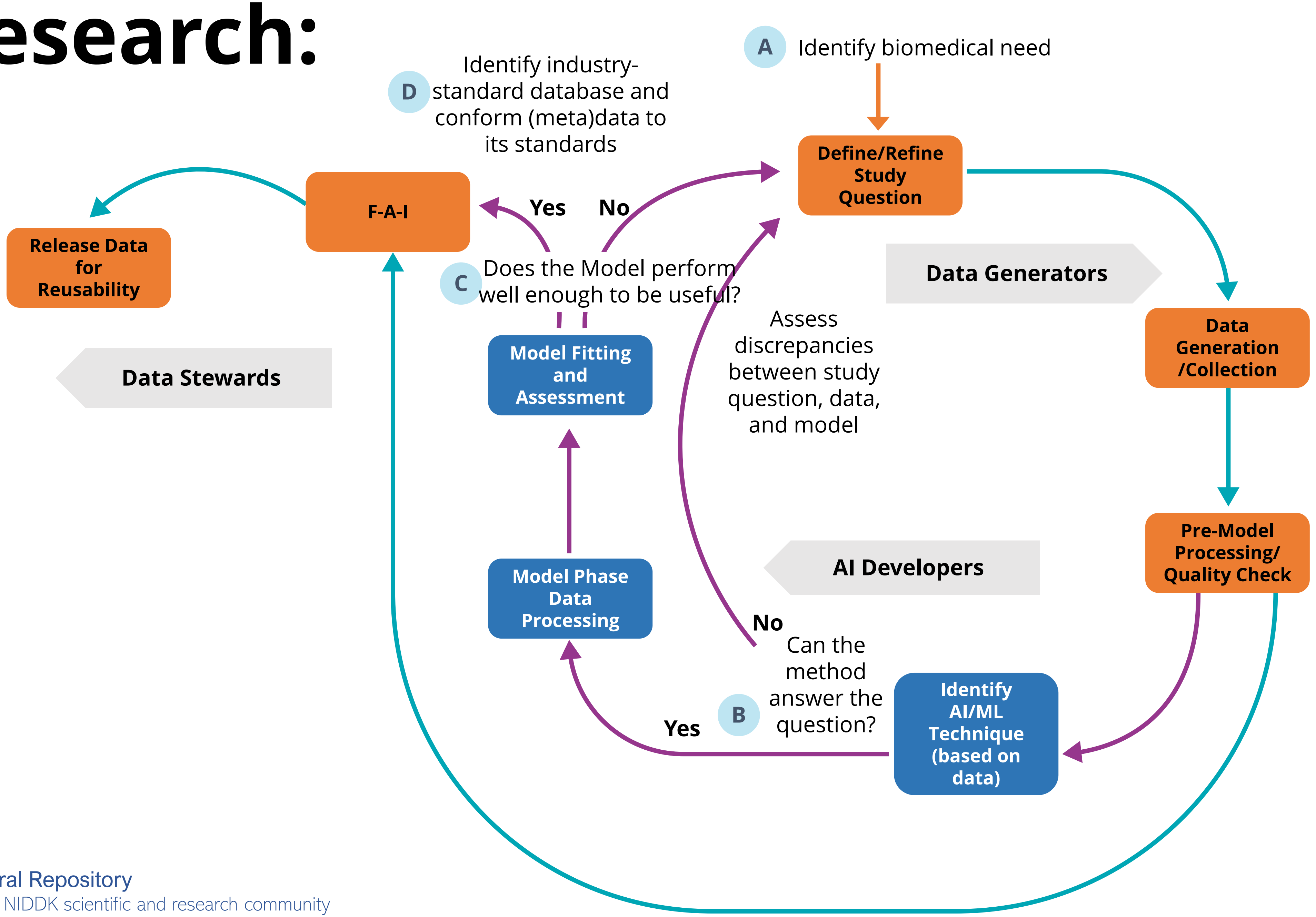
Critical Thinker — Decision Maker

What is an AI-ready dataset?

AI-readiness refers to data that are *machine-readable, reliable, accurate, explainable, predictive,* and *accessible for future AI applications*

- An AI-ready dataset consists of:
  - Data that is reflective of the population from which it was drawn
  - Data that is well documented and FAIR (findable, accessible, interoperable, and reusable)
  - Data that is model-agnostic
- AI-readiness will include:
  - ✓ **pre-processing steps** such as addressing errant values,
  - ✓ **handling of missing values**,
  - ✓ **relabeling and recoding** of data elements (aka columns, variables, features, or attributes) and values during harmonization to ensure consistency and standardized formatting
  - ✓ **documentation** of all data handling steps, all variables, and the dataset itself
- When possible,
  - attempt to **retain as much information as possible** by creating new data elements that are transforms of existing elements without deleting or overwriting existing elements.

AI Research:



A Identify biomedical need

Define/Refine Study Question

Data Generators

Data Generation /Collection

Pre-Model Processing/ Quality Check

AI Developers

Identify AI/ML Technique (based on data)

B Can the method answer the question?

Yes

No

Assess discrepancies between study question, data, and model

Model Phase Data Processing

Model Fitting and Assessment

C Does the Model perform well enough to be useful?

Yes

No

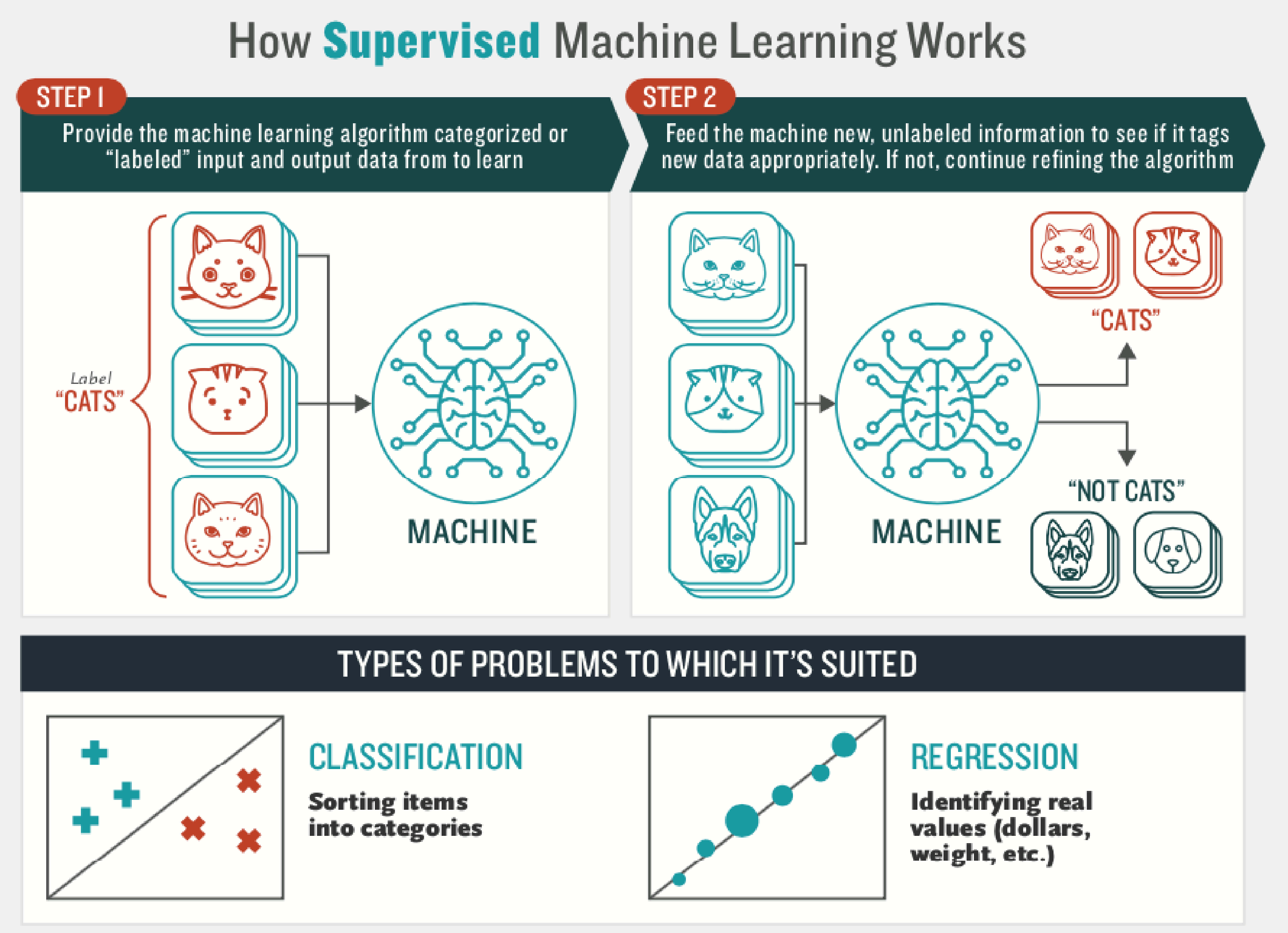
F-A-I

D Identify industry-standard database and conform (meta)data to its standards

Release Data for Reusability

Data Stewards

Supervised Learning



How Supervised Machine Learning Works

STEP 1

Provide the machine learning algorithm categorized or "labeled" input and output data from to learn

STEP 2

Feed the machine new, unlabeled information to see if it tags new data appropriately. If not, continue refining the algorithm

TYPES OF PROBLEMS TO WHICH IT'S SUITED

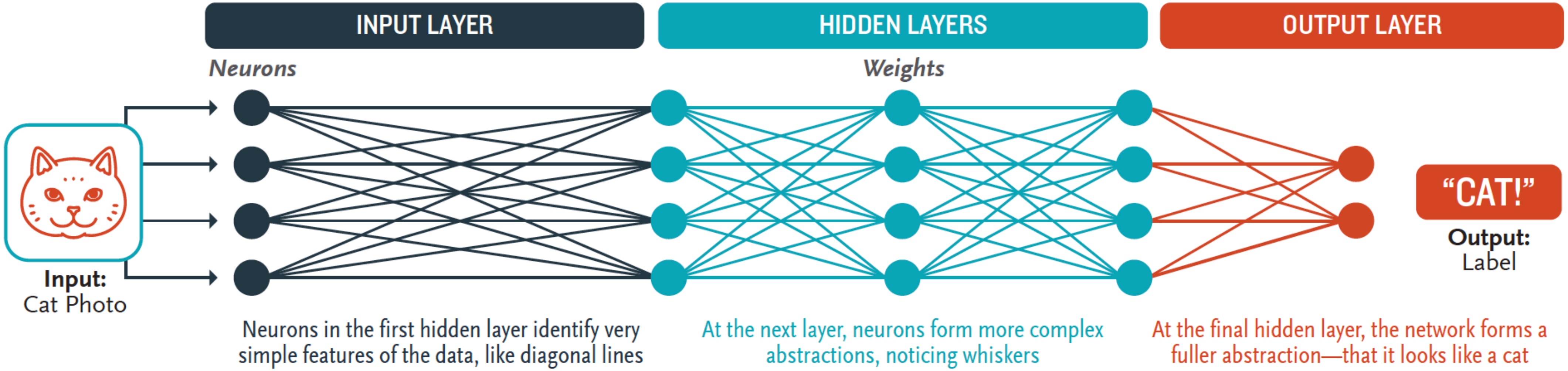
CLASSIFICATION

Sorting items into categories

REGRESSION

Identifying real values (dollars, weight, etc.)

Deep Learning

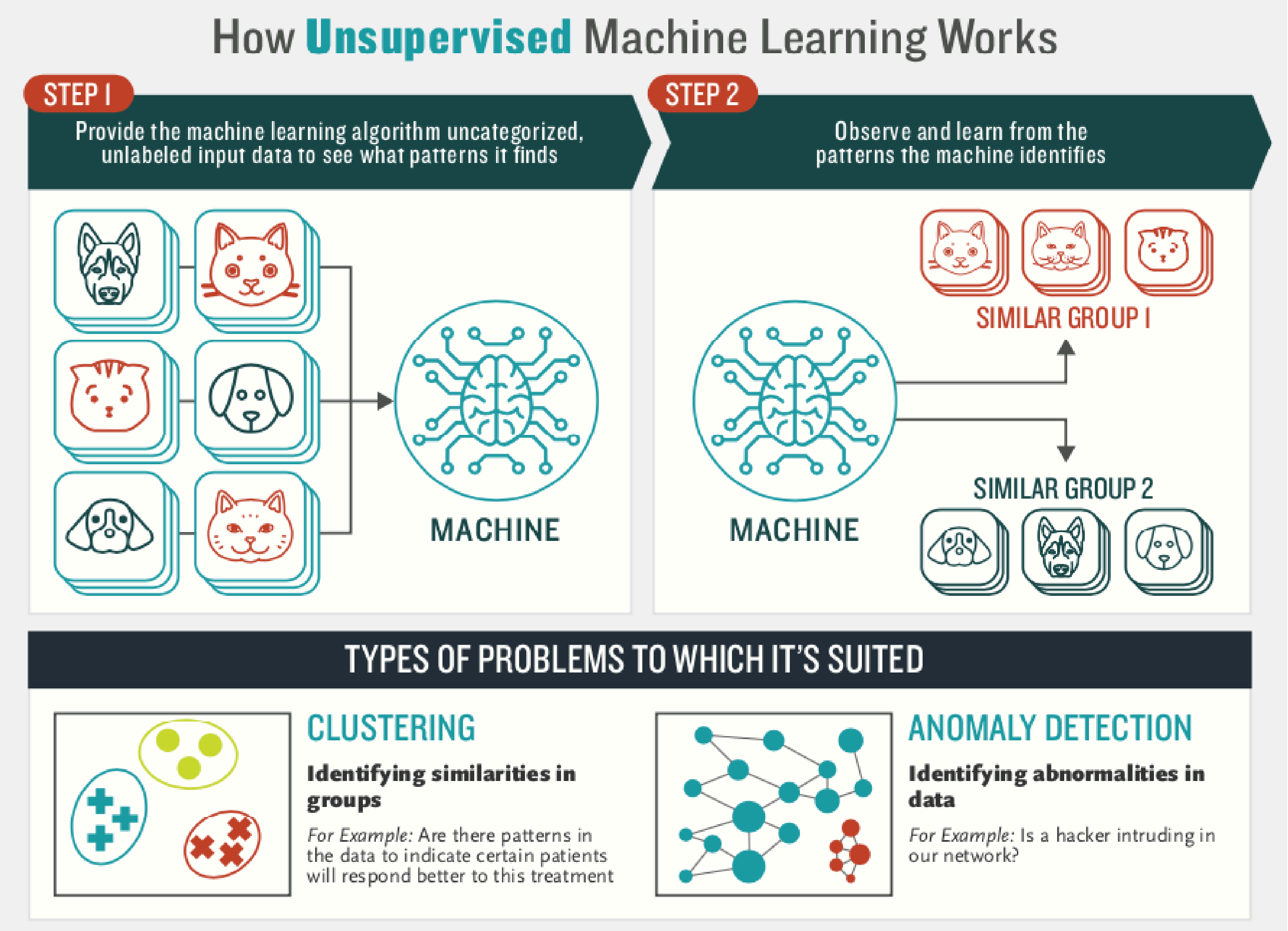


Neurons in the first hidden layer identify very simple features of the data, like diagonal lines

At the next layer, neurons form more complex abstractions, noticing whiskers

At the final hidden layer, the network forms a fuller abstraction—that it looks like a cat

Unsupervised Learning



How Unsupervised Machine Learning Works

STEP 1

Provide the machine learning algorithm uncategorized, unlabeled input data to see what patterns it finds

STEP 2

Observe and learn from the patterns the machine identifies

TYPES OF PROBLEMS TO WHICH IT'S SUITED

CLUSTERING

Identifying similarities in groups

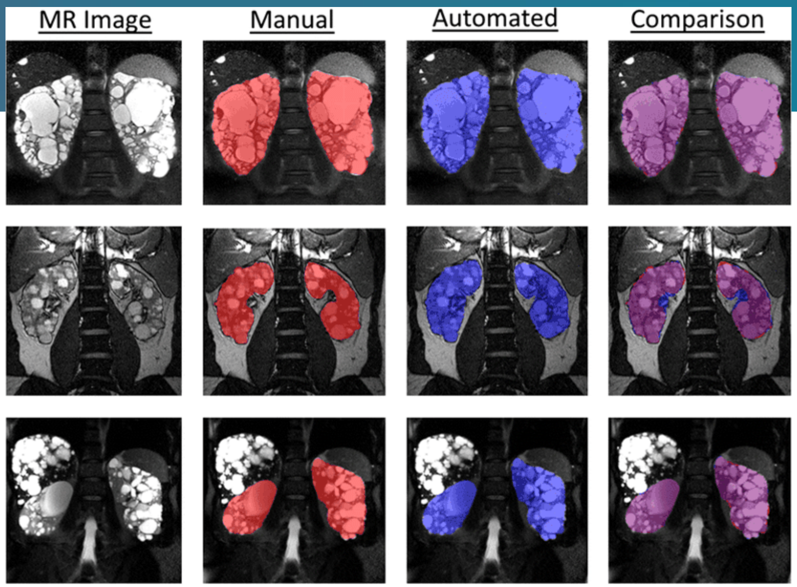
For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment

ANOMALY DETECTION

Identifying abnormalities in data

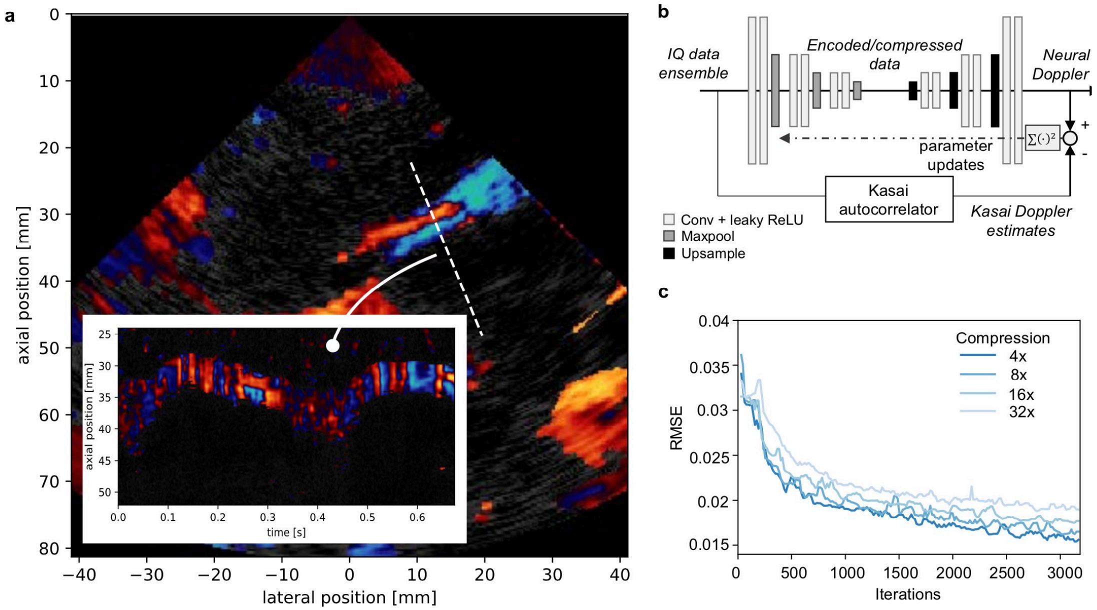
For Example: Is a hacker intruding in our network?

AI in Health



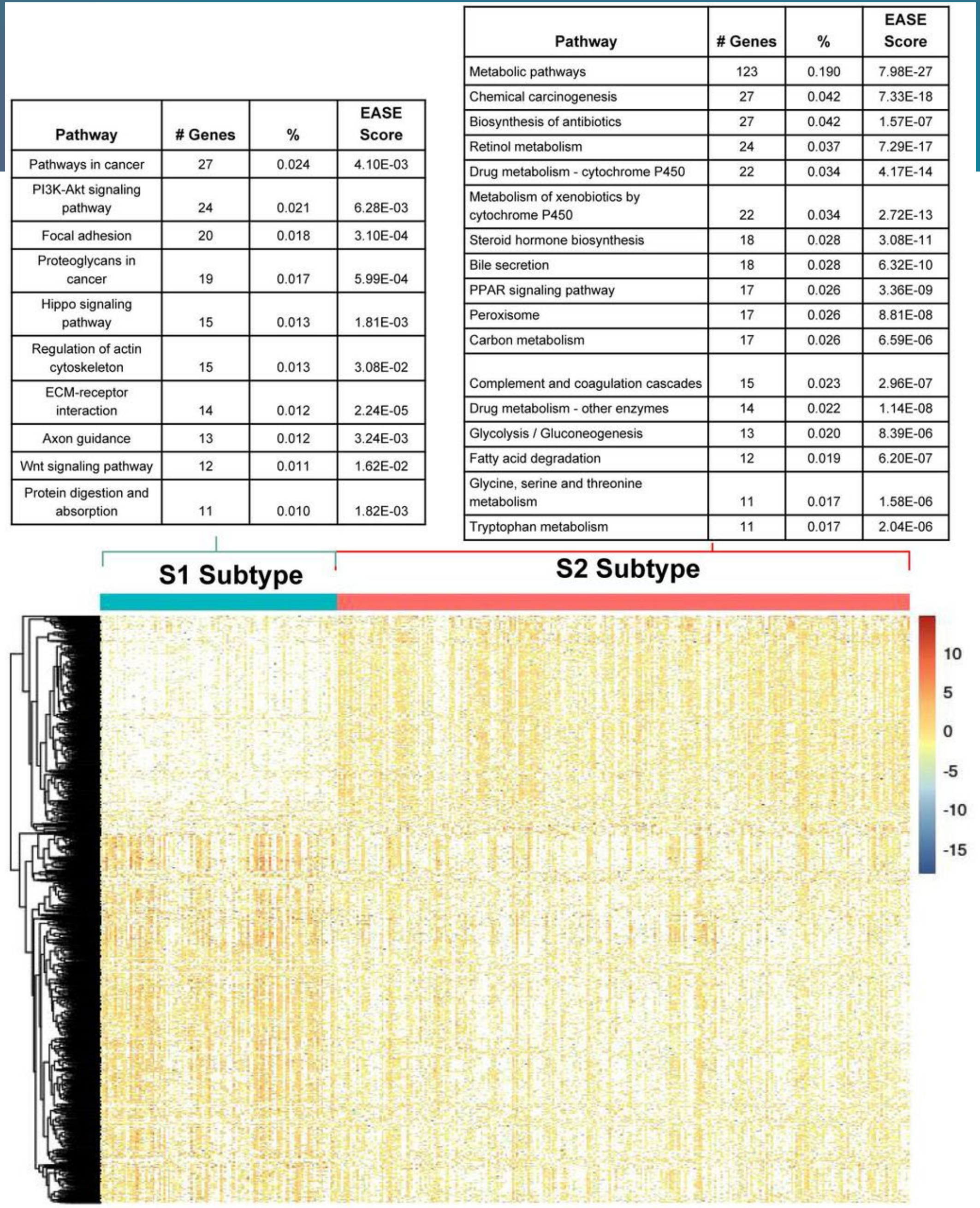
Labeled, annotated images

- Feature Extraction - Image segmentation (US, CT, MRI)
- Deep Learning - Learn important low-level and high-level features
  - *Image Augmentation*
  - *Transfer learning*
  - *Architectures for Deep Learning*
    - *Convolutional Neural Nets (CNN)*
    - *Autoencoders (AE)*
    - *Recurrent Neural Networks (RNN)*
    - *Deep Belief Network (DBN)*
  - Voxel-wise classification



AI in Health

- -omic sequence data is treated like a sequence and/or language
- Deep Learning Architectures
  - *Transfer learning from pre-trained models*
  - *Convolutional Neural Nets (CNN) - treat a window of the sequence as an image*
  - *Variational Autoencoders (VAE)*
  - *Recurrent Neural Networks (RNN)*
  - *Long Short-Term Memory (LSTM)*
    - *GENOMIC-ULMFiT – from FAST AI*
  - *Bi-directional Transformer models (BERT)*



Pathway	# Genes	%	EASE Score
Pathways in cancer	27	0.024	4.10E-03
PI3K-Akt signaling pathway	24	0.021	6.28E-03
Focal adhesion	20	0.018	3.10E-04
Proteoglycans in cancer	19	0.017	5.99E-04
Hippo signaling pathway	15	0.013	1.81E-03
Regulation of actin cytoskeleton	15	0.013	3.08E-02
ECM-receptor interaction	14	0.012	2.24E-05
Axon guidance	13	0.012	3.24E-03
Wnt signaling pathway	12	0.011	1.62E-02
Protein digestion and absorption	11	0.010	1.82E-03

Pathway	# Genes	%	EASE Score
Metabolic pathways	123	0.190	7.98E-27
Chemical carcinogenesis	27	0.042	7.33E-18
Biosynthesis of antibiotics	27	0.042	1.57E-07
Retinol metabolism	24	0.037	7.29E-17
Drug metabolism - cytochrome P450	22	0.034	4.17E-14
Metabolism of xenobiotics by cytochrome P450	22	0.034	2.72E-13
Steroid hormone biosynthesis	18	0.028	3.08E-11
Bile secretion	18	0.028	6.32E-10
PPAR signaling pathway	17	0.026	3.36E-09
Peroxisome	17	0.026	8.81E-08
Carbon metabolism	17	0.026	6.59E-06
Complement and coagulation cascades	15	0.023	2.96E-07
Drug metabolism - other enzymes	14	0.022	1.14E-08
Glycolysis / Gluconeogenesis	13	0.020	8.39E-06
Fatty acid degradation	12	0.019	6.20E-07
Glycine, serine and threonine metabolism	11	0.017	1.58E-06
Tryptophan metabolism	11	0.017	2.04E-06

Research Design

- Develop and define a systematic plan to study a scientific problem.
- Identify the type of study (e.g., descriptive, review, experimental), research question, hypothesis, variables, design, data collection, and subsequent statistical analysis plan.
- **Identify the data required to study this question: especially demographic details**

- Types of data that can support outcomes research:
  - Clinical Data – doctors' notes, prescription records, lab images and notes, insurance (claims) data, electronic health record (EHR) data
  - Patient-Sourced Data – sensors, survey measures, social media posts, preferences, wearables data

DATA CONSIDERATIONS

- Domain experts needed to inform data-use assumptions
- **Data source and details need to represent the population of interest**
- All algorithms inherently involve assumptions, some of which are *not* verifiable by the data
- Unmeasured, random variation mitigated by design/replication
- Non-random or systematic variation, more commonly encountered with "found" data (selection/confounding bias)[1]
- The learning 'target' (prediction, estimation) must guide chosen priorities in data considerations

Research Design

Use Case: Predict mortality for chronic kidney disease patients in the first 90 days of dialysis.

- The first 90 days following initiation of chronic dialysis represent a high-risk period for adverse outcomes, including mortality
- While the sudden and unplanned start of dialysis is a known risk factor, other factors leading to poor outcomes during this early period have not been fully delineated
- Tools to identify patients at highest-risk for poor outcomes during this early period are lacking

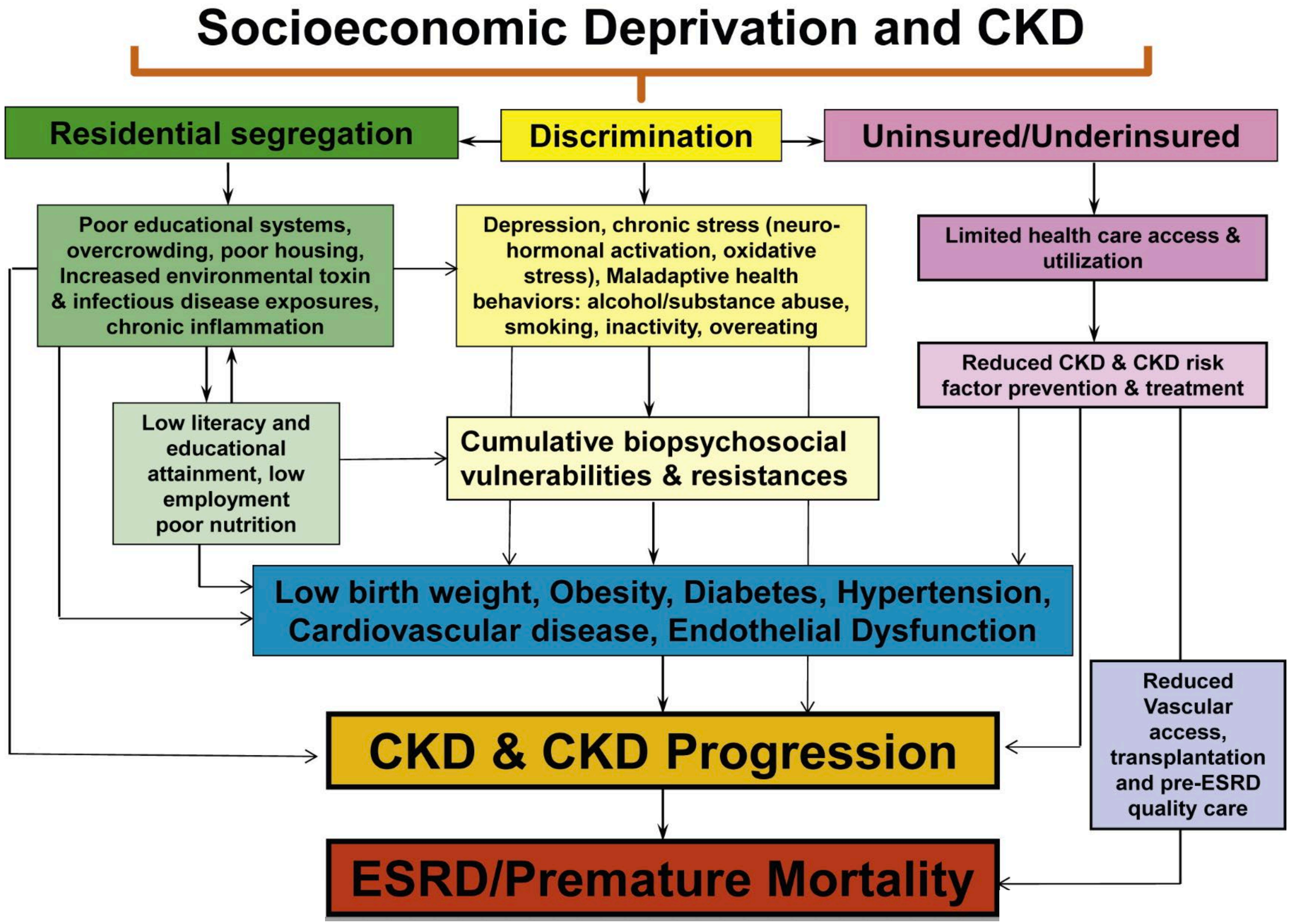
POTENTIAL DATA SOURCES
EHR data from any health system (e.g., VA, Optum)
Health claims data from Medicare/Medicaid and Payers
Vital statistics databases
Disease registries (e.g., USRDS, SEER)

The Office of the National Coordinator for Health Information Technology℠

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

Bias (socioeconomic)

- Many of the determinants of chronic kidney disease, such as obesity, diabetes, hypertension, chronic inflammation, neurohormonal activation, and oxidative stress may be related to socioeconomic disparities.
- Factors include substandard living conditions, limited quality health care to the uninsured or underinsured, and limited health literacy.

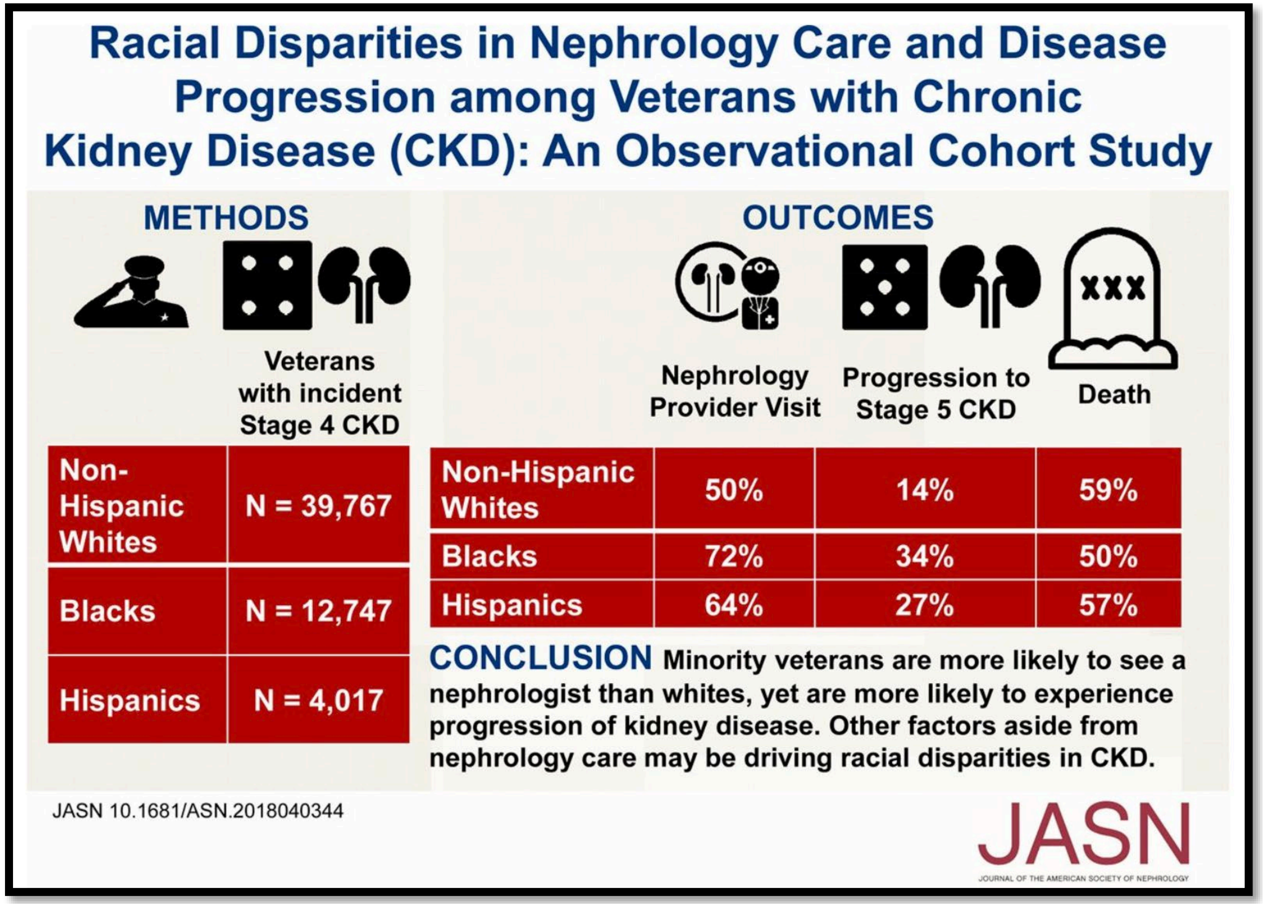

Socioeconomic Deprivation and CKD

Residential segregation
Discrimination
Uninsured/Underinsured
Poor educational systems, overcrowding, poor housing, Increased environmental toxin & infectious disease exposures, chronic inflammation
Depression, chronic stress (neuro-hormonal activation, oxidative stress), Maladaptive health behaviors: alcohol/substance abuse, smoking, inactivity, overeating
Limited health care access & utilization
Reduced CKD & CKD risk factor prevention & treatment
Low literacy and educational attainment, low employment poor nutrition
Cumulative biopsychosocial vulnerabilities & resistances
Low birth weight, Obesity, Diabetes, Hypertension, Cardiovascular disease, Endothelial Dysfunction
Reduced Vascular access, transplantation and pre-ESRD quality care
CKD & CKD Progression
ESRD/Premature Mortality

Source: https://www.sciencedirect.com/science/article/pii/S1548559514001086

Bias (Racial)

- Despite being more likely to receive nephrology consultation, black patients with stage 4 chronic kidney disease (CKD) were 62% more likely to develop end-stage renal disease (ESRD) after adjustment for comorbidities and socioeconomic factors.
- These findings suggest that biologic or environmental factors drive ESRD progression through mechanisms that nephrologists cannot currently treat.



Racial Disparities in Nephrology Care and Disease Progression among Veterans with Chronic Kidney Disease (CKD): An Observational Cohort Study

METHODS

	Veterans with incident Stage 4 CKD
Non-Hispanic Whites	N = 39,767
Blacks	N = 12,747
Hispanics	N = 4,017

OUTCOMES

	Nephrology Provider Visit	Progression to Stage 5 CKD	Death
Non-Hispanic Whites	50%	14%	59%
Blacks	72%	34%	50%
Hispanics	64%	27%	57%

CONCLUSION Minority veterans are more likely to see a nephrologist than whites, yet are more likely to experience progression of kidney disease. Other factors aside from nephrology care may be driving racial disparities in CKD.

JASN 10.1681/ASN.2018040344

JASN

JOURNAL OF THE AMERICAN SOCIETY OF NEPHROLOGY

Source: https://jasn.asnjournals.org/content/29/10/2563

Bias in AI

- Advances in AI offer the potential to provide personalized care by taking into account individual differences[1]
- **At the same time, because machine learning algorithms aggregate and assess large volumes of real-world data, AI can reinforce bias in data, potentially reinforcing existing patterns of discrimination**
- Machine learning algorithms may work well for one patient group, but results may not be appropriate for others

SOURCES OF BIAS

- Missing data – patients without consistent care at a single institution and/or lower health literacy
- Sample size – certain subgroups of patients may not exist in sufficient numbers, leading to uninformative predictions
- Misclassification or measurement error – implicit bias leads to disparities in care, teaching clinics (where patients of low socioeconomic status may be seen) may have less accurate data input[2]

Sources: 1. https://journalofethics.ama-assn.org/article/can-ai-help-reduce-disparities-general-medical-and-mental-health-care/2019-02; 2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/#!po=15.6250

Bias in AI

POTENTIAL CHALLENGES	RECOMMENDED SOLUTIONS
Data diversity due to limited population representation	• Assess the limitations • Identify the strategy for mitigating a lack of diversity as part of the research design
Overreliance on machine learning solutions	• Ensure interdisciplinary approach and continuous human involvement • Conduct follow-up studies to ensure results are meaningful
Algorithms based on biased data	• Identify the target population and select training and testing sets accordingly • Build and test algorithms in socioeconomically diverse health care systems • Ensure that key variables that are related to race, gender, etc. are being captured and included in algorithms where appropriate • Test algorithms for potential discriminatory behavior throughout processing • Develop feedback loops to monitor and verify output and validity
Non-clinically meaningful algorithms	• Focus on clinically important improvements in relevant outcomes rather than strict performance measures • Impose human values in algorithms at the cost of efficiency

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/#!po=15.6250

Bias in AI

- Preventing algorithms from making biased decisions is challenging and there is often a tradeoff between fairness and accuracy
- Three main strategies for reducing bias:
  - Eliminating sources of unfairness in the data before training a machine learning algorithm
  - Making fairness adjustments as part of the process by which the algorithm is constructed
  - Adjusting performance after an algorithm is applied to make it fairer

WHY IS IT SO DIFFICULT TO ELIMIATE UNFAIRNESS?

- There is a lack of agreement among researchers about which definition of fairness is the most appropriate[1]
- Removing sensitive information from data, such as race, age, and gender, may not result in unbiased outcomes since non-sensitive attributes and outcome variables are often statistically dependent on sensitive information[2,3,4]
- A user's judgment about a model feature may change after learning how the use of the feature impacts decision outcomes[5]

Sources: 1. https://journalofethics.ama-assn.org/article/can-ai-help-reduce-disparities-general-medical-and-mental-health-care/2019-02; 2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/#!po=15.6250

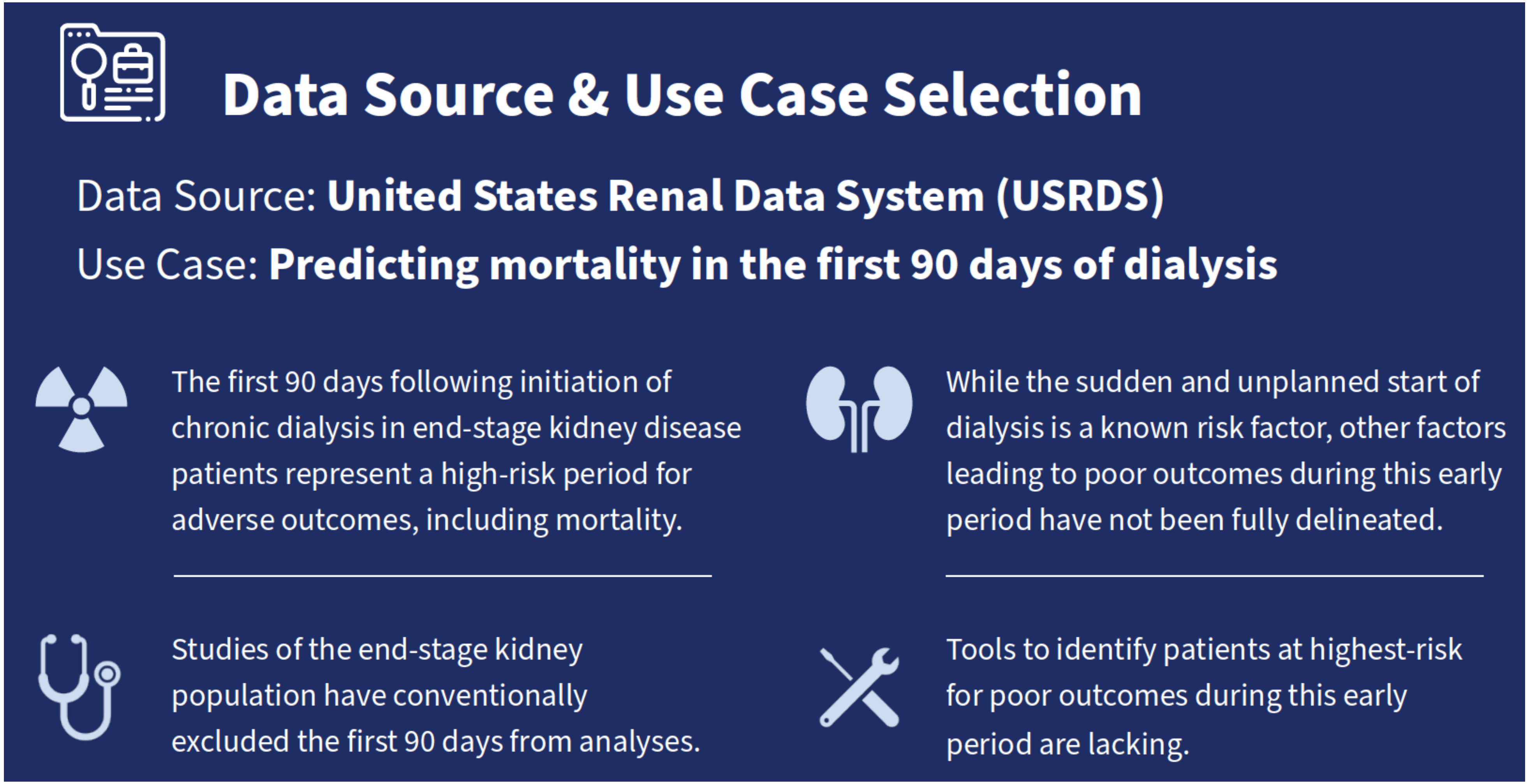
Quiz (Type Answers in Chat)

1. Select a potential source of bias for an electronic health record data set.
   a) Sample size (not enough representation of all subgroups)
   b) Measurement error
   c) Equipment choice
   d) Missing context
   e) All of the above
2. The type of model that can be used if you have a set of labeled data
   a) Unsupervised Learning
   b) Supervised Learning
   c) Independent Learning
   d) Observation Learning
3. An AI-ready dataset does not need to be documented fully because the model will do it automatically.
   a) True
   b) False

Quiz (Type Answers in Chat)

- Properly handling self-reported demographic data is an emerging field of interest. What are your thoughts?
- Some points to consider:
  - *Free response is the most accurate, but how do you analyze this?*
  - *Offering many categories can lead to "small n", where few observations are recorded in some categories. Is it then okay to combine categories?*
  - *Is "race" or "sex" or "gender" just a proxy for something else within your population? Are there variables you should be recording instead that are more related to exposure or outcome?*

CKD Example - Research Design



Data Source & Use Case Selection

Data Source: **United States Renal Data System (USRDS)**

Use Case: **Predicting mortality in the first 90 days of dialysis**

The first 90 days following initiation of chronic dialysis in end-stage kidney disease patients represent a high-risk period for adverse outcomes, including mortality.

While the sudden and unplanned start of dialysis is a known risk factor, other factors leading to poor outcomes during this early period have not been fully delineated.

Studies of the end-stage kidney population have conventionally excluded the first 90 days from analyses.

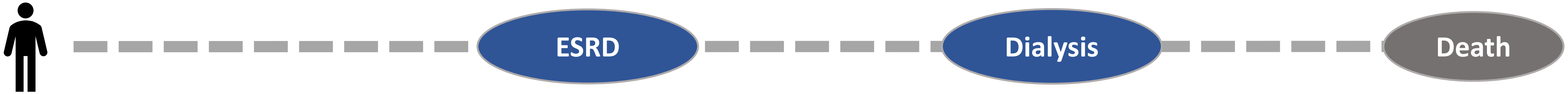
Tools to identify patients at highest-risk for poor outcomes during this early period are lacking.

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – USRDS Data Mapping to Use Case

CKD Patient

Selected use case: *Predicting mortality in the first 90 days of dialysis*



ESRD

Dialysis

Death

1. **CMS Pre-ESRD Claims Datasets**
   - Parts A and B claims prior to ESRD diagnosis
   - Used to build features, such as prior nephrology care

2. **ESRD Medical Evidence Report (MEDEVID) (CMS 2728)/ PATIENTS Dataset**
   - Form is completed when a patient is diagnosed as ESRD and receives their first chronic dialysis treatment(s) or transplant
   - Used to build features such as patient demographics, comorbid conditions, primary cause of renal failure, and laboratory values

2A. PATIENTS Dataset
- Provides basic demographic and ESRD-related data
- Used to obtain dialysis start date and modality
- Used in conjunction with MEDEVID to build demographic features such as age, sex, race, etc.

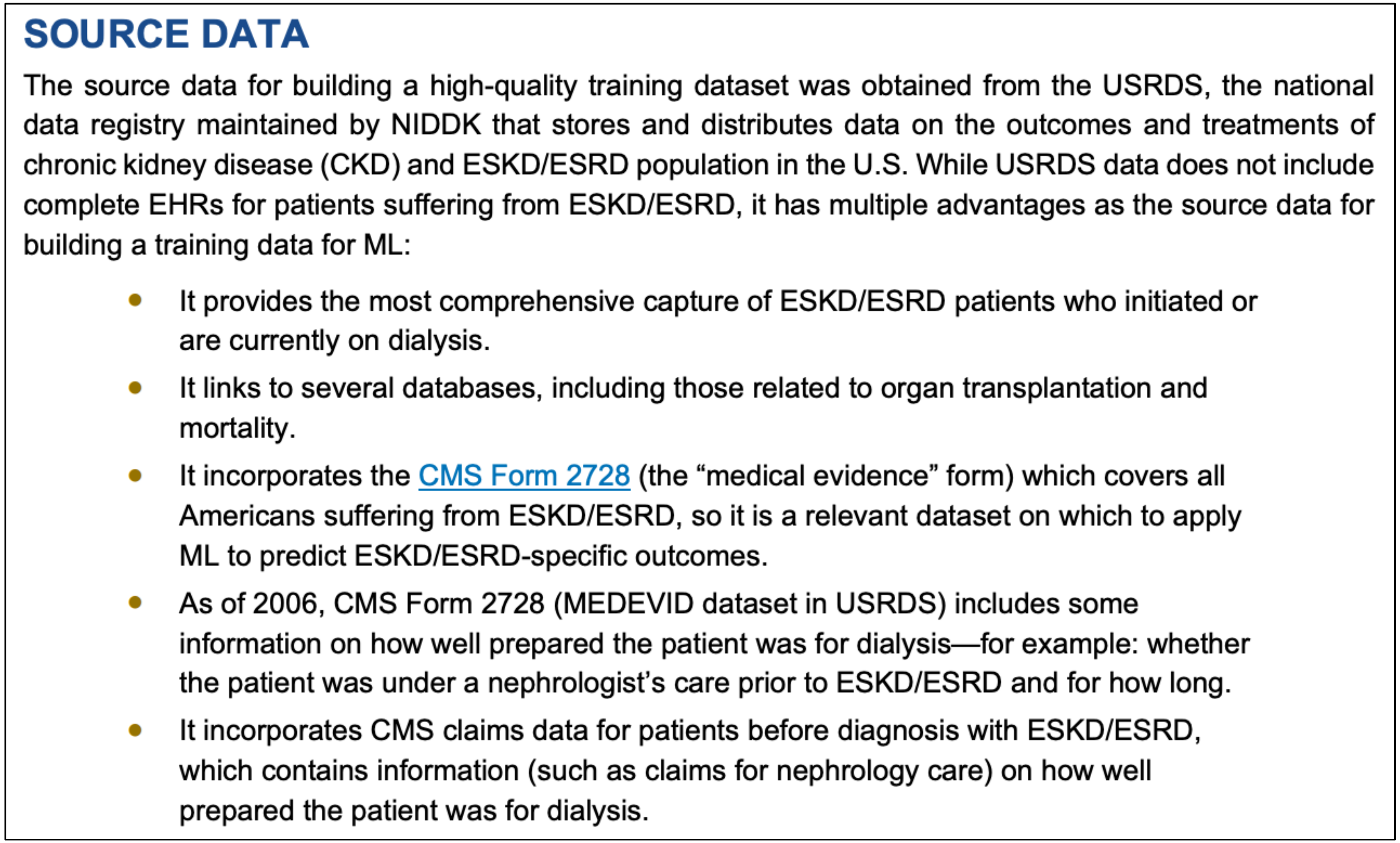
2B. Transplant Dataset (TX)
- Provides information on kidney transplants such as list date/data on eligibility pre-dialysis
- Used to build features such as transplant waitlist status

3. **PATIENTS Dataset/ DEATH Dataset (CMS ESRD Death Notification Form 2726)**
   - Used to determine if a patient died in the first 90 days after dialysis start

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – Data Documentation

Documentation of source for dataset(s)



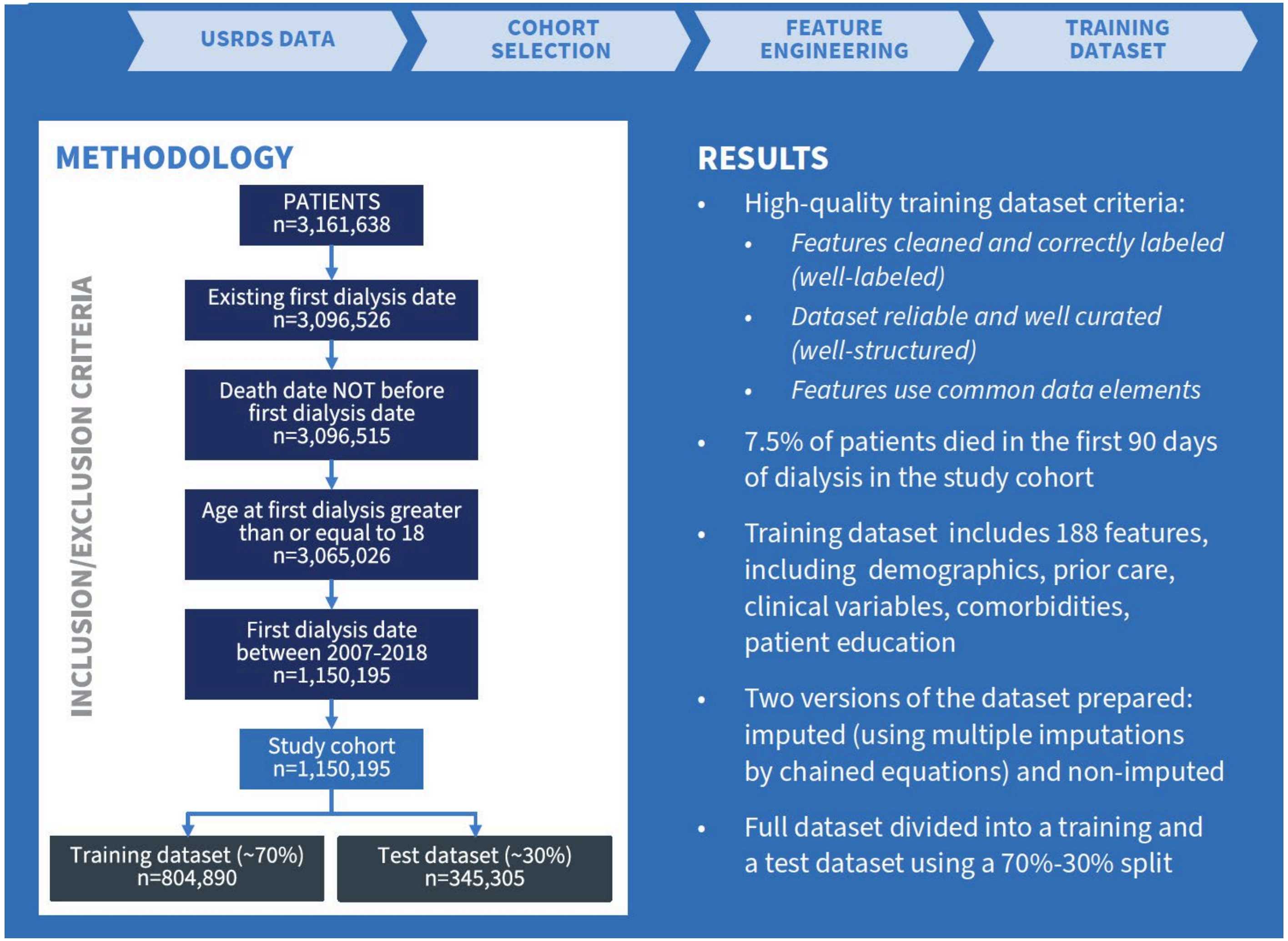
SOURCE DATA

The source data for building a high-quality training dataset was obtained from the USRDS, the national data registry maintained by NIDDK that stores and distributes data on the outcomes and treatments of chronic kidney disease (CKD) and ESKD/ESRD population in the U.S. While USRDS data does not include complete EHRs for patients suffering from ESKD/ESRD, it has multiple advantages as the source data for building a training data for ML:

- It provides the most comprehensive capture of ESKD/ESRD patients who initiated or are currently on dialysis.
- It links to several databases, including those related to organ transplantation and mortality.
- It incorporates the CMS Form 2728 (the "medical evidence" form) which covers all Americans suffering from ESKD/ESRD, so it is a relevant dataset on which to apply ML to predict ESKD/ESRD-specific outcomes.
- As of 2006, CMS Form 2728 (MEDEVID dataset in USRDS) includes some information on how well prepared the patient was for dialysis—for example: whether the patient was under a nephrologist's care prior to ESKD/ESRD and for how long.
- It incorporates CMS claims data for patients before diagnosis with ESKD/ESRD, which contains information (such as claims for nephrology care) on how well prepared the patient was for dialysis.

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – Data



USRDS DATA → COHORT SELECTION → FEATURE ENGINEERING → TRAINING DATASET

METHODOLOGY

INCLUSION/EXCLUSION CRITERIA

PATIENTS
n=3,161,638

Existing first dialysis date
n=3,096,526

Death date NOT before first dialysis date
n=3,096,515

Age at first dialysis greater than or equal to 18
n=3,065,026

First dialysis date between 2007-2018
n=1,150,195

Study cohort
n=1,150,195

Training dataset (~70%)
n=804,890

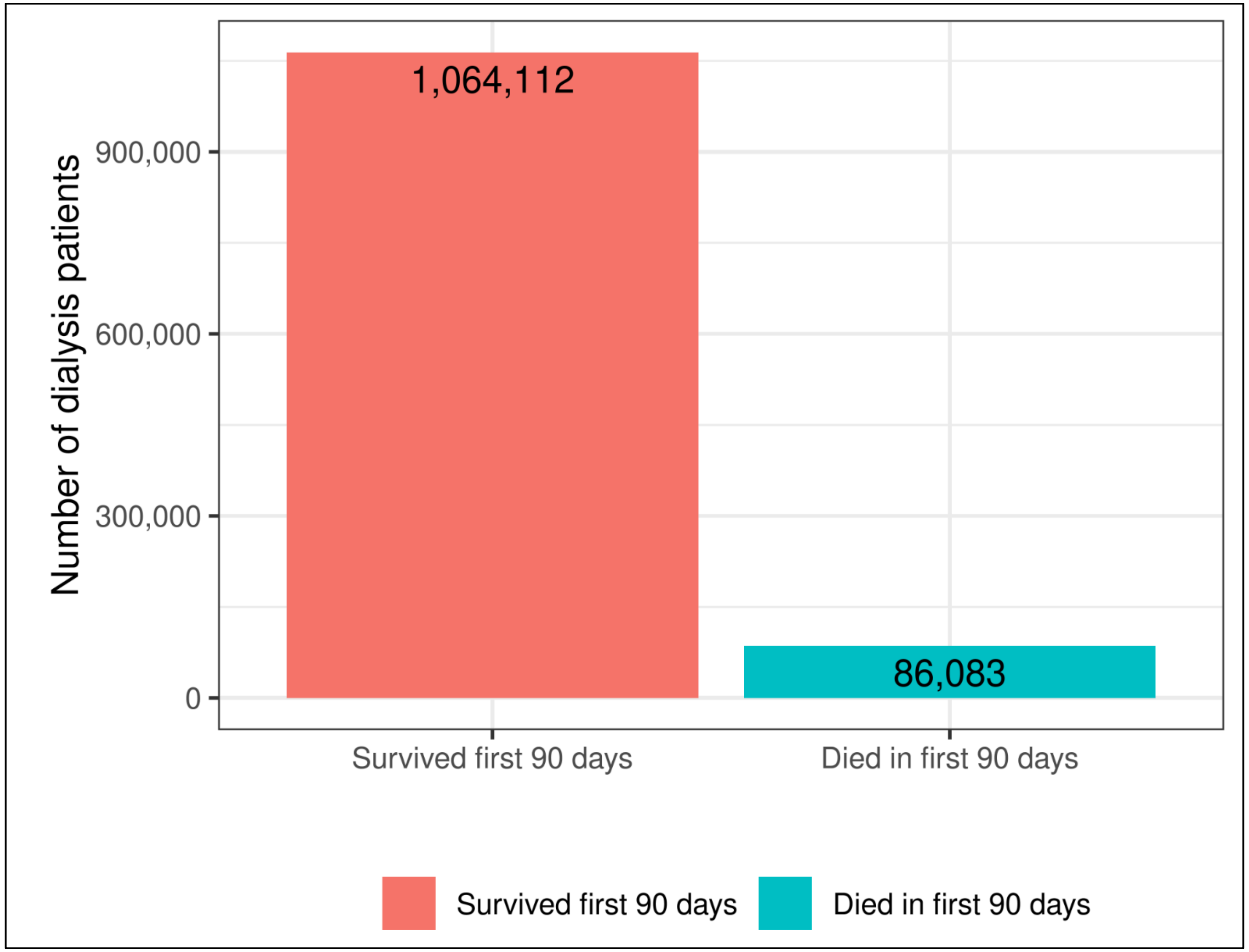
Test dataset (~30%)
n=345,305

RESULTS

- High-quality training dataset criteria:
  - *Features cleaned and correctly labeled (well-labeled)*
  - *Dataset reliable and well curated (well-structured)*
  - *Features use common data elements*
- 7.5% of patients died in the first 90 days of dialysis in the study cohort
- Training dataset includes 188 features, including demographics, prior care, clinical variables, comorbidities, patient education
- Two versions of the dataset prepared: imputed (using multiple imputations by chained equations) and non-imputed
- Full dataset divided into a training and a test dataset using a 70%-30% split

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

Imbalanced Data



https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

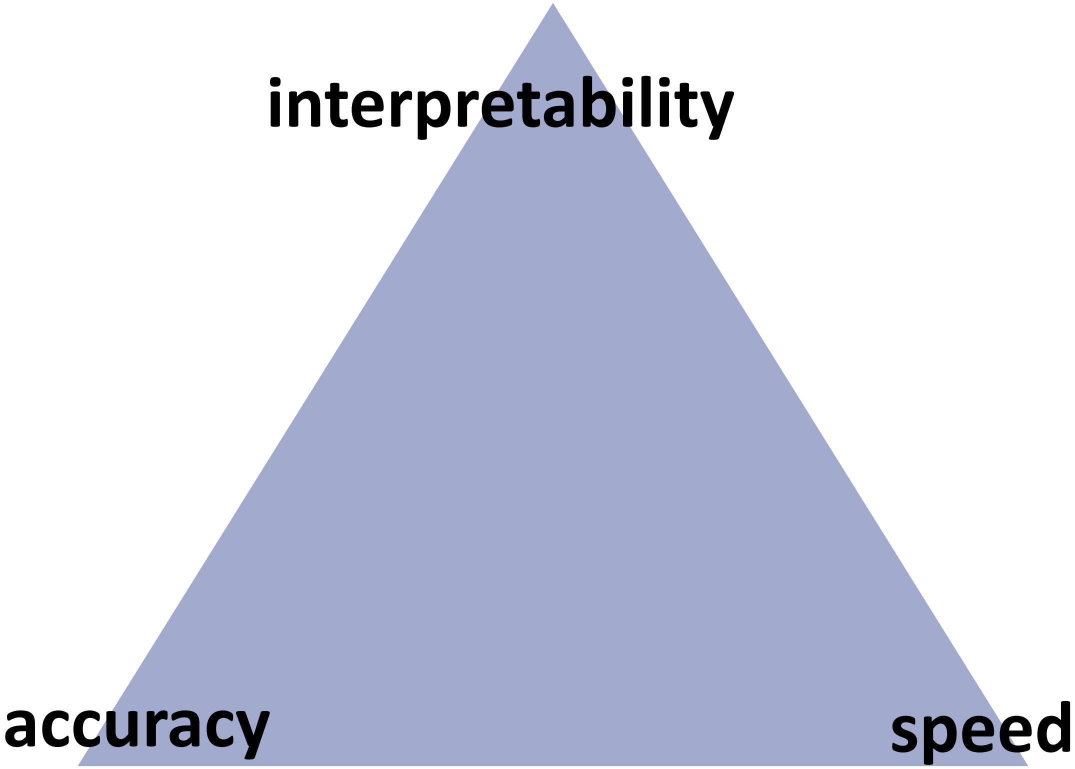
Model Selection

Labeled Data

- Supervised learning

Unlabeled Data

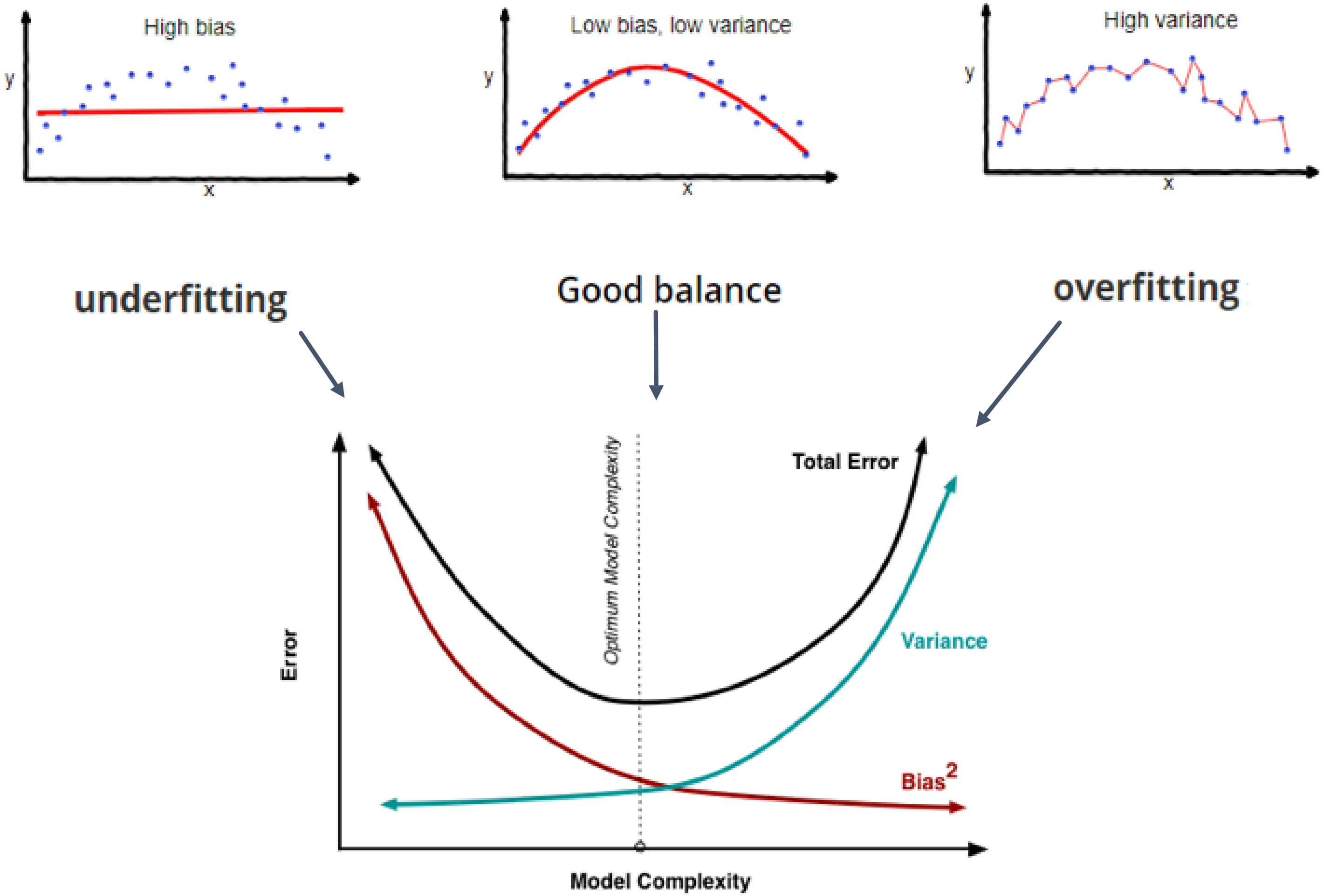
- Unsupervised Learning
- Dimensionality Reduction



Model Selection

Overfitting: Model captures too much of the noise along with the signal or pattern of the data and cannot generalize to new data (i.e., data too noisy, not enough data). It has merely memorized the data it has seen before.

Underfitting: Model does not capture the signal/pattern of the data.

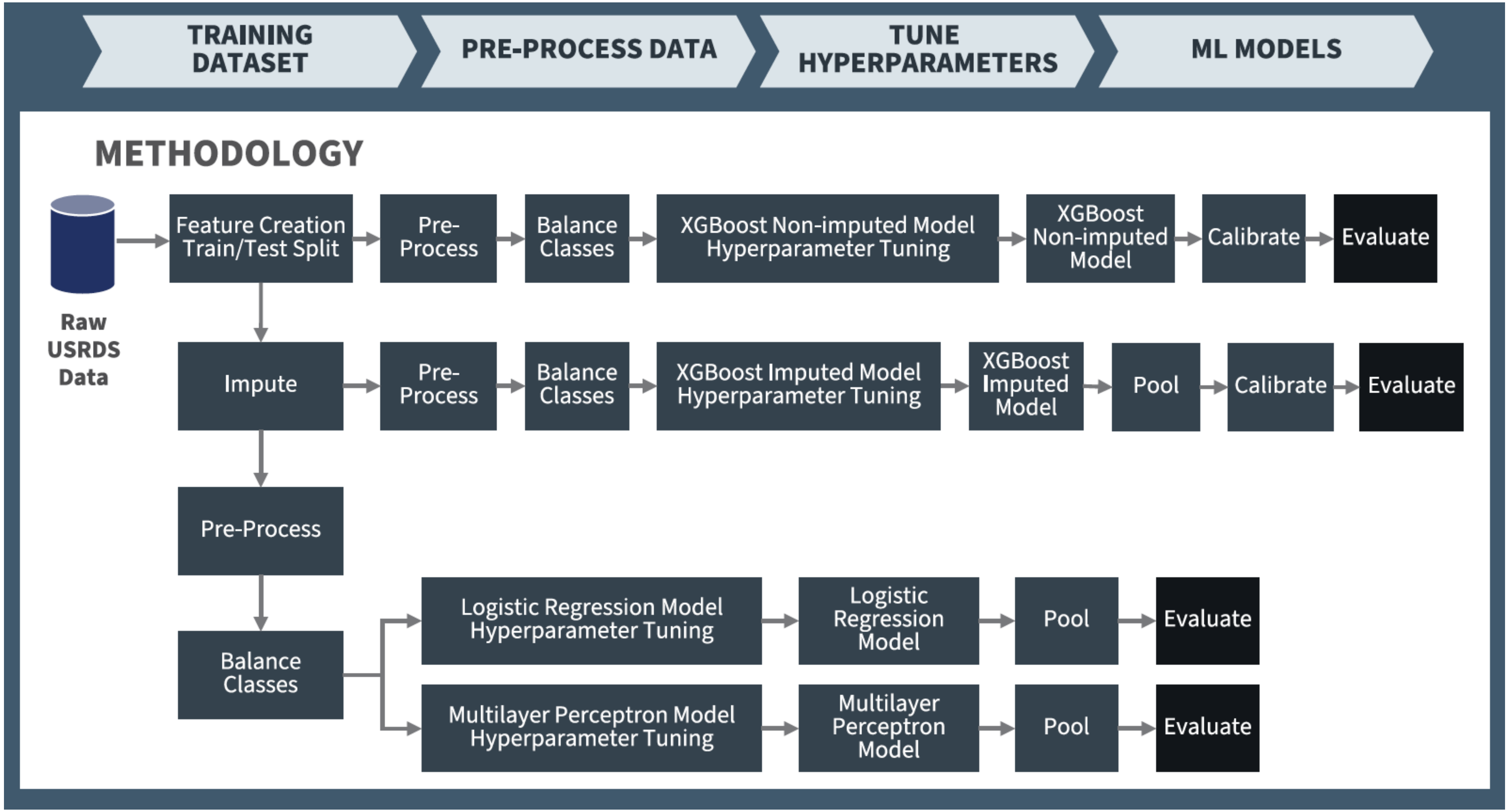


underfitting

Good balance

overfitting

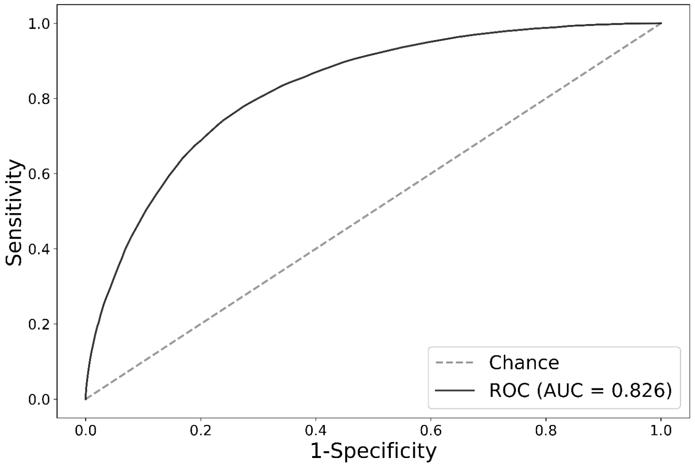
CKD Example – Model Selection


TRAINING DATASET
PRE-PROCESS DATA
TUNE HYPERPARAMETERS
ML MODELS
METHODOLOGY
Raw USRDS Data
Feature Creation Train/Test Split
Pre-Process
Balance Classes
XGBoost Non-imputed Model Hyperparameter Tuning
XGBoost Non-imputed Model
Calibrate
Evaluate
Impute
Pre-Process
Balance Classes
XGBoost Imputed Model Hyperparameter Tuning
XGBoost Imputed Model
Pool
Calibrate
Evaluate
Pre-Process
Balance Classes
Logistic Regression Model Hyperparameter Tuning
Logistic Regression Model
Pool
Evaluate
Multilayer Perceptron Model Hyperparameter Tuning
Multilayer Perceptron Model
Pool
Evaluate

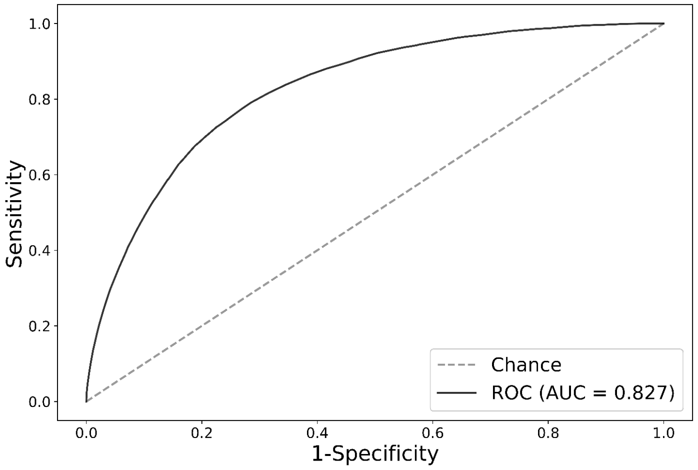
https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – Model Results

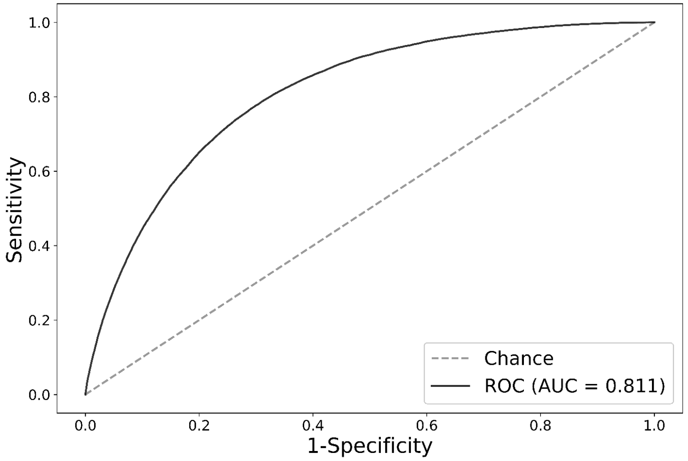
XGBoost Non-imputed



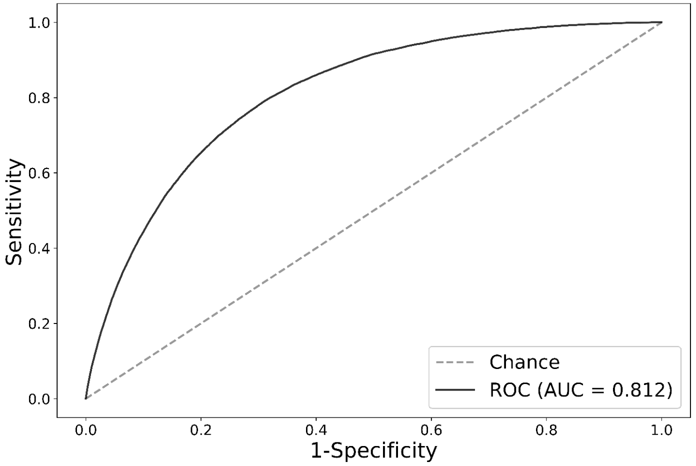
XGBoost Imputed



Logistic Regression



Multilayer Perceptron



https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – Model Interpretability

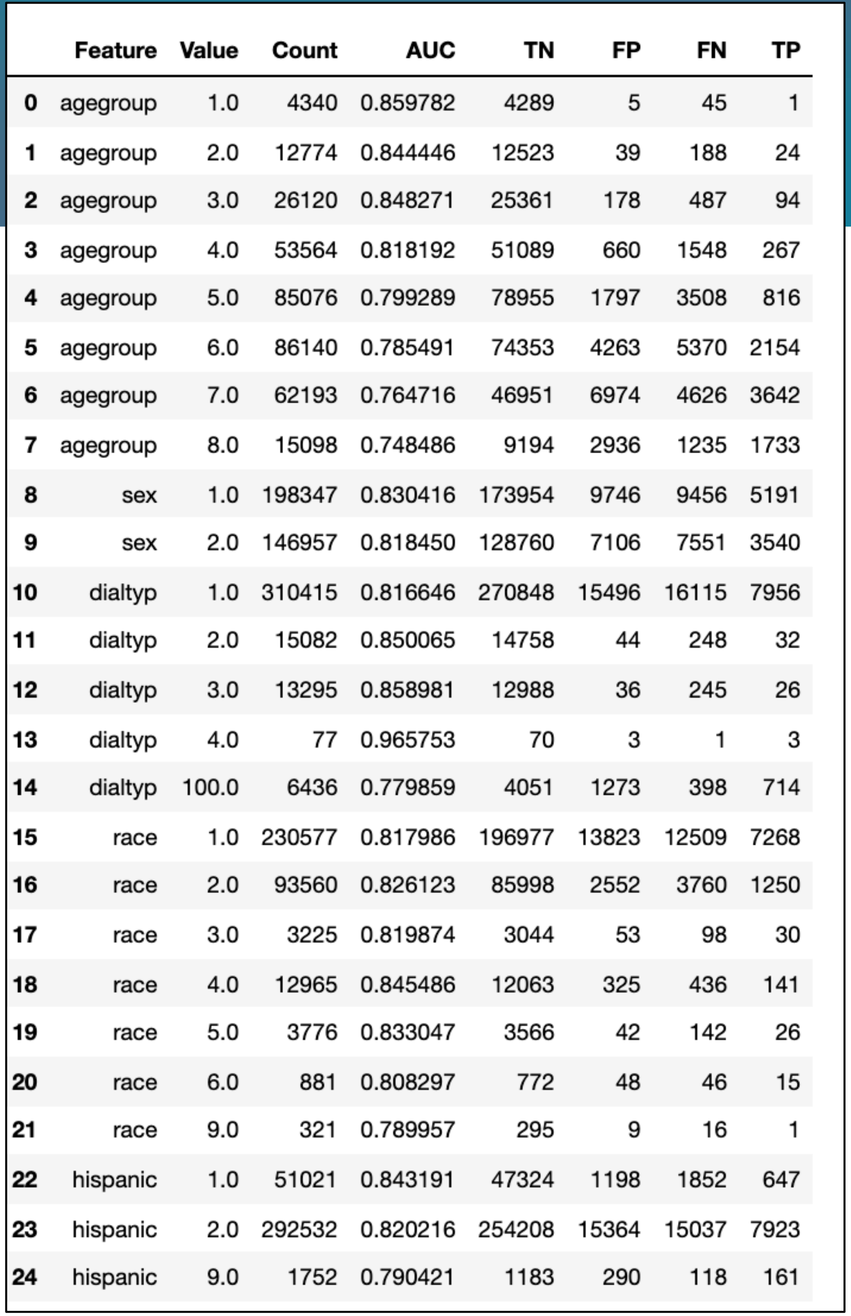
	Feature	Explanation
1.	**Age**	• Older age is associated with worse survival
2.	**Inpatient stays**	• Longer inpatient stays is more common in older and sicker patients and has been associated with early mortality
3.	**Received erythropoietin (EPO)**	• EPO hormone is produced by kidneys when it senses low oxygen levels in the blood; EPO triggers bone marrow to produce more red blood cells which raises blood oxygen • Patients on EPO typically have advanced CKD at the time of dialysis and are under the care of a nephrologist • Patients with kidney failure produces less EPO; therefore, are given EPO
4.	**Albumin**	• Albumin reflects the patient’s overall health status (including nutrition and inflammation) • Risk of death is increased by poor serum albumin levels reflecting inadequate nutrition
5.	**Arteriovenous Fistula (AVF)**	• The presence of a maturing AVF indicates prior nephrology care • Hemodialysis through AVF access is associated with reduced mortality

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – Fairness Assessment

- ML models can perform differently for different categories of patients, so the non-imputed XGBoost model was assessed for fairness, or how well the model performs for each category of interest (demographics—sex, race, and age—as well as initial dialysis modality). Age were binned into the following categories based on clinician input and an example in literature: 18-25, 26-35, 36-45, 46-55, 56-65, 66-75, 76-85, 86+. The USRDS predefined categories for race, sex, and dialysis modality were used for the fairness assessment.

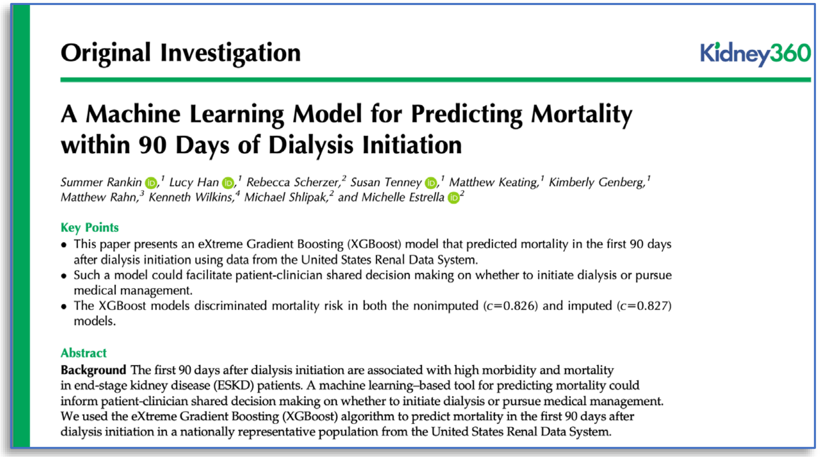
- Performing the fairness assessment on the categories of interest gives additional insight into how the model performs by different patient categories of interest (by demographics, etc.). Future researchers should perform fairness assessments to better evaluate model performance, especially for models that may be deployed in a clinical setting. Other methods of assessing fairness include evaluating true positives, sensitivity, positive predictive value, etc. at various threshold across the different groups of interest, which would allow selection of a threshold that balances model performance across the groups of interest.



	Feature	Value	Count	AUC	TN	FP	FN	TP
0	agegroup	1.0	4340	0.859782	4289	5	45	1
1	agegroup	2.0	12774	0.844446	12523	39	188	24
2	agegroup	3.0	26120	0.848271	25361	178	487	94
3	agegroup	4.0	53564	0.818192	51089	660	1548	267
4	agegroup	5.0	85076	0.799289	78955	1797	3508	816
5	agegroup	6.0	86140	0.785491	74353	4263	5370	2154
6	agegroup	7.0	62193	0.764716	46951	6974	4626	3642
7	agegroup	8.0	15098	0.748486	9194	2936	1235	1733
8	sex	1.0	198347	0.830416	173954	9746	9456	5191
9	sex	2.0	146957	0.818450	128760	7106	7551	3540
10	dialtyp	1.0	310415	0.816646	270848	15496	16115	7956
11	dialtyp	2.0	15082	0.850065	14758	44	248	32
12	dialtyp	3.0	13295	0.858981	12988	36	245	26
13	dialtyp	4.0	77	0.965753	70	3	1	3
14	dialtyp	100.0	6436	0.779859	4051	1273	398	714
15	race	1.0	230577	0.817986	196977	13823	12509	7268
16	race	2.0	93560	0.826123	85998	2552	3760	1250
17	race	3.0	3225	0.819874	3044	53	98	30
18	race	4.0	12965	0.845486	12063	325	436	141
19	race	5.0	3776	0.833047	3566	42	142	26
20	race	6.0	881	0.808297	772	48	46	15
21	race	9.0	321	0.789957	295	9	16	1
22	hispanic	1.0	51021	0.843191	47324	1198	1852	647
23	hispanic	2.0	292532	0.820216	254208	15364	15037	7923
24	hispanic	9.0	1752	0.790421	1183	290	118	161

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example - Project Resources



Original Investigation

Kidney360

A Machine Learning Model for Predicting Mortality within 90 Days of Dialysis Initiation

Summer Rankin,[1] Lucy Han,[1] Rebecca Scherzer,[2] Susan Tenney,[1] Matthew Keating,[1] Kimberly Genberg,[1] Matthew Rahn,[3] Kenneth Wilkins,[4] Michael Shlipak,[2] and Michelle Estrella[2]

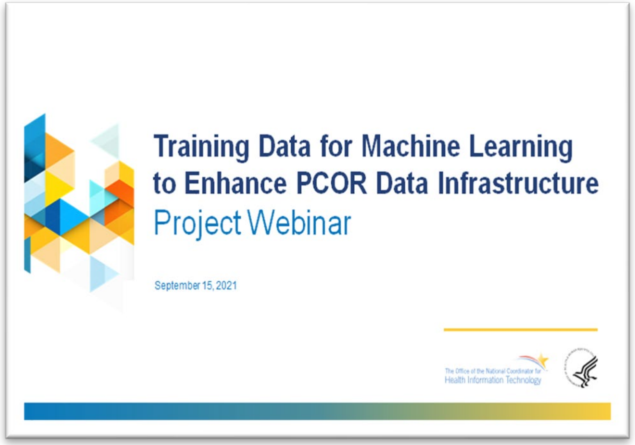
Key Points

- This paper presents an eXtreme Gradient Boosting (XGBoost) model that predicted mortality in the first 90 days after dialysis initiation using data from the United States Renal Data System.
- Such a model could facilitate patient-clinician shared decision making on whether to initiate dialysis or pursue medical management.
- The XGBoost models discriminated mortality risk in both the nonimputed (c=0.826) and imputed (c=0.827) models.

Abstract

Background The first 90 days after dialysis initiation are associated with high morbidity and mortality in end-stage kidney disease (ESKD) patients. A machine learning–based tool for predicting mortality could inform patient-clinician shared decision making on whether to initiate dialysis or pursue medical management. We used the eXtreme Gradient Boosting (XGBoost) algorithm to predict mortality in the first 90 days after dialysis initiation in a nationally representative population from the United States Renal Data System.

- Main:
  - https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning
- Blog Post:
  - https://www.healthit.gov/buzz-blog/health-it/the-application-of-machine-learning-to-address-kidney-disease
- Peer-reviewed publication
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9528387/
- Infographic
  - https://www.healthit.gov/sites/default/files/page/2021-09/ONC%20Training%20Data%20Project_Infographic-FINAL.pdf
- Code Repository:
  - https://github.com/onc-healthit/2021PCOR-ML-AI



Training Data for Machine Learning to Enhance PCOR Data Infrastructure

Project Webinar

September 15, 2021

Quiz (Type Answers in Chat)

1. Select the techniques that can be used to handle imbalanced data.
   a) Tiprapping
   b) Bootstrapping
   c) None. A model cannot be fit to imbalanced data
   d) Oversampling

2. Select the reason that interpretability in AI models is important for health domain.
   a) The weights can be compared to benchmarks
   b) The importance of the features can be analyzed
   c) A peer-reviewed paper can be published
   d) Trick question, it isn't important

Questions