Prognostic and Predictive Classification Approaches for Disease Prediction Modeling

Chetan Paul

Dr. Ravichandran Sarangan

February 28, 2023

EC #23-Leidos-0222-26020

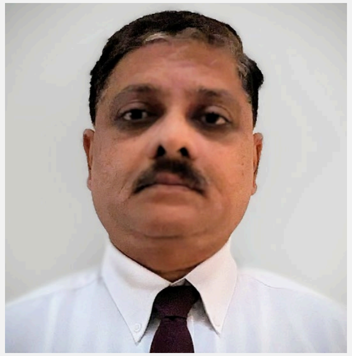
Introductions

leidos

Chetan Paul
Vice President – Technology & Innovation
Leidos Civilian Health Services

Highlights

- Advance Mission-solutions, research and outcomes for federal healthcare agencies (FDA, NIH, CDC, and CMS) through application of Emerging Technologies and HPC.
- Practitioner in Deep Learning, image processing, parallel computing GPU architectures for scientific and regulatory applications.
- Education: Strategic Leadership (Harvard), AI/ML Nanodegree (Stanford University), Masters Degree in Computers Science and Engineering
- Awards: “SASE” Technical Achievement, ACT-IAC Blockchain Innovation, G2Xchange Disruptive Tech, ACT-IAC Innovation, FDA Scientific Computing



Dr. Ravichandran Sarangan
Bioinformatics & Data Science Lead
Leidos Civilian Health Services

Highlights

- Bioinformatician and data scientist with extensive computing experience in analyzing and modeling public health and biomedical sciences data.
- Expert in developing statistical, Machine-Learning, and deep-learning models for high-dimensional Omics and health-focused (Real-world) data. Extensive experience collaborating and managing biomedical, and genetic diseases projects;
- Authored 54 peer-reviewed publications and a patent.
- Education: Ph.D. Computational Chemistry
- Awards: FNLCR Annual achievement team award; Annual performance award, ABCC; FDA Scientific Computing

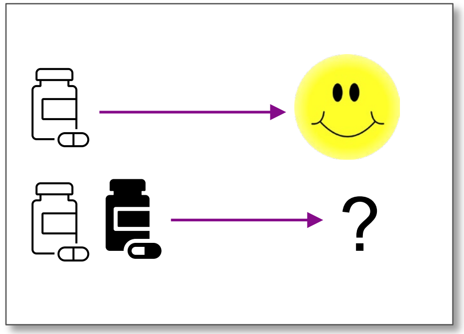
The Quest to Predict the "Future"

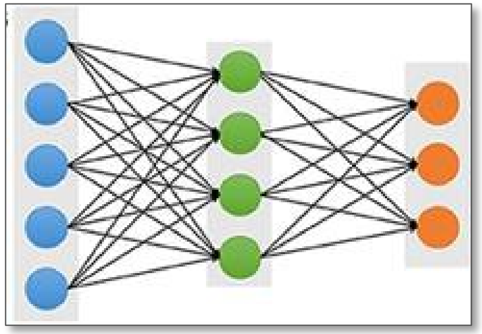
The ***Palantíri***, also known as the **Seven Seeing-stones**, were used for intelligence gathering and could show visions of the future

Prognostic and Predictive Modeling methods and approaches allow an estimation of future events which one can make by incorporating and casting forward data related to the past in a pre-determined and systematic manner

Agenda

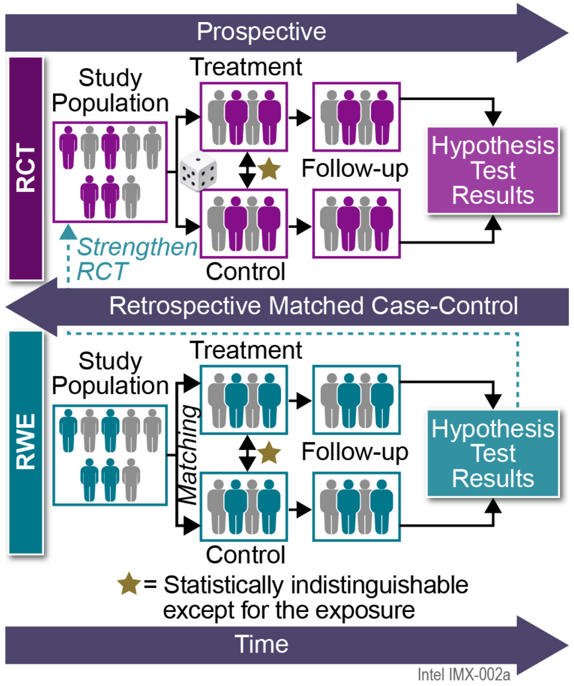
- **Causal Inference (Drug Repurposing Study)**
  - Population-based; sub-population-based
- **Prognostic Modeling (Long COVID Study)**
  - Population (group) based: Survival Analysis
  - Patient focused: Hazard Modeling; Random Survival Forest
- **Data Harmonization/Quality**
- **Predictive modeling pipeline**
- **Synthetic Data**
- **Deep-Learning Approaches and Transfer Learning**
- **AI Trust and Explainability**





Causal Inference

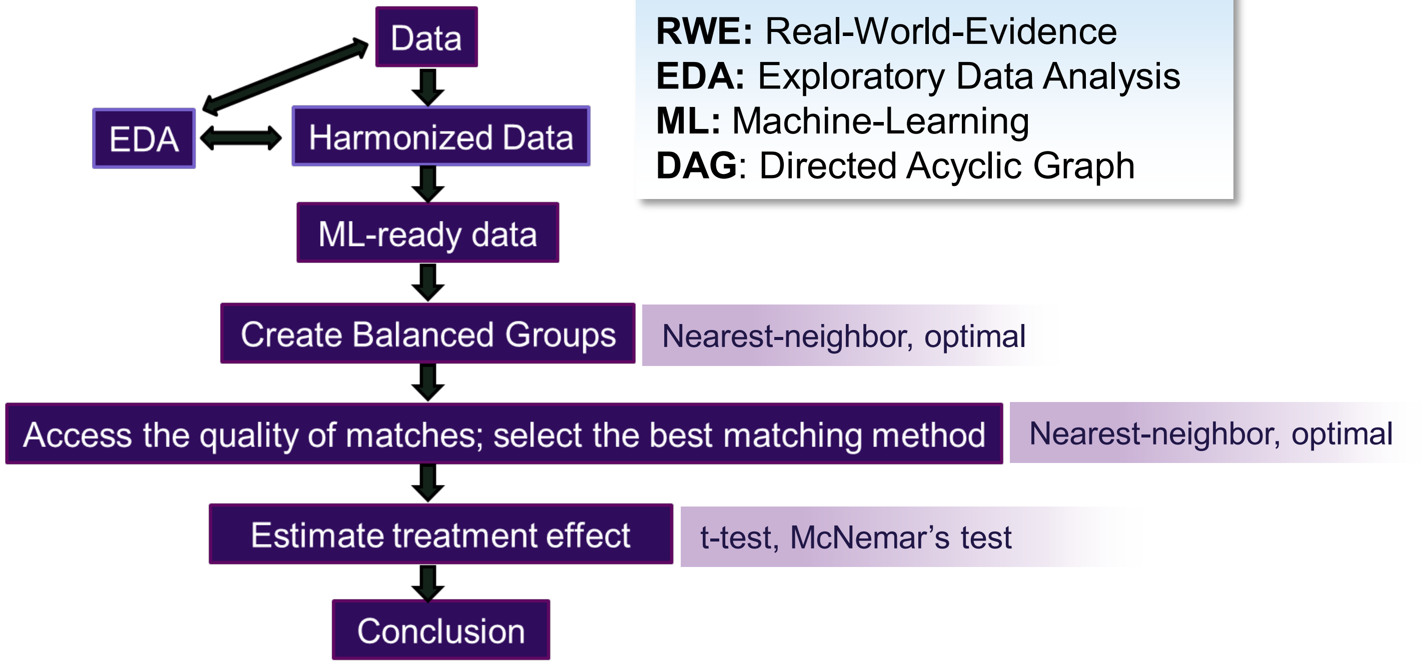
Using Real-World-Data (RWD), develop in-silico models for rapidly identifying repurposed drugs that can lower the risk of death due to disease/infection.



Areas of focus	Solution
Identify study population, randomization, and removing confounding	Inclusion-exclusion, matching, and DAG

Leidos process of Simulating in-silico Randomized Clinical Trials Using Real-World-Data (RWD) to Model Causal Effect of Repurposed Drugs

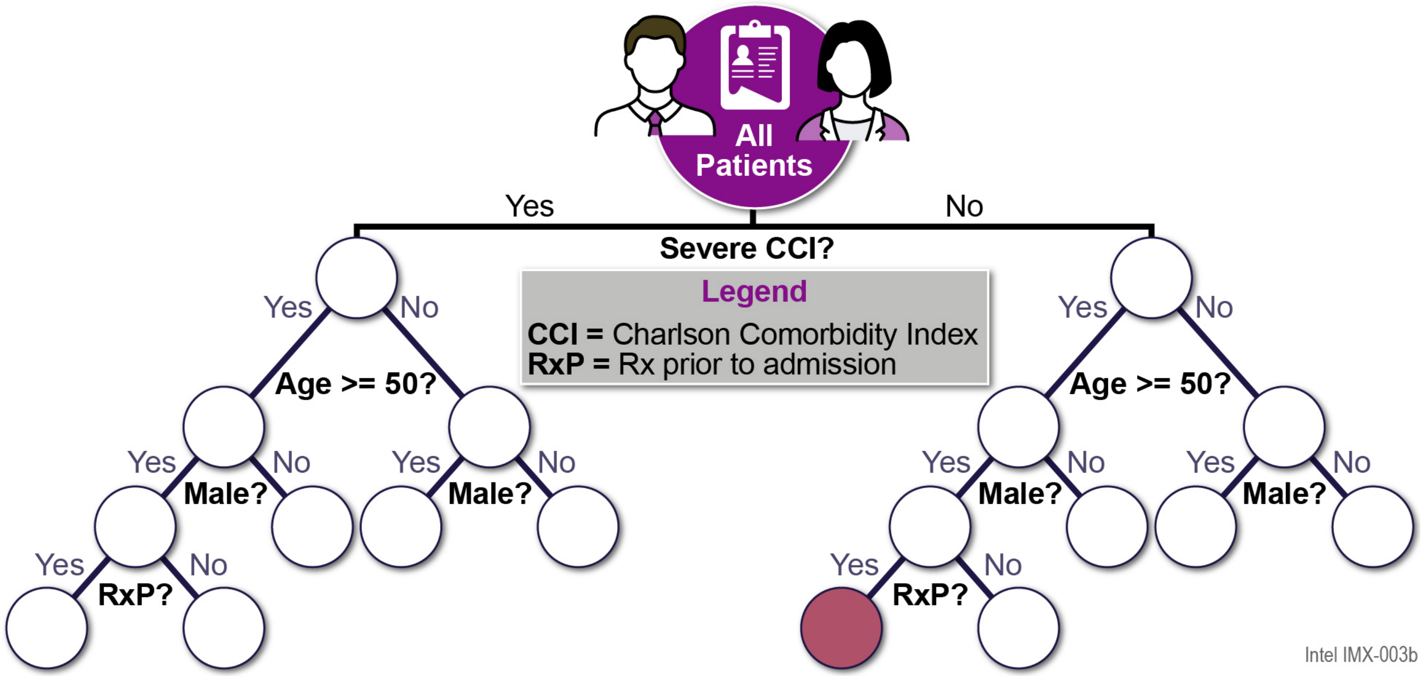
RWD: Real-World-Data
RWE: Real-World-Evidence
EDA: Exploratory Data Analysis
ML: Machine-Learning
DAG: Directed Acyclic Graph



Subpopulation Based Association

Using Real-World-Data (RWD), develop in-silico models for rapidly the efficacy of drugs in subpopulations

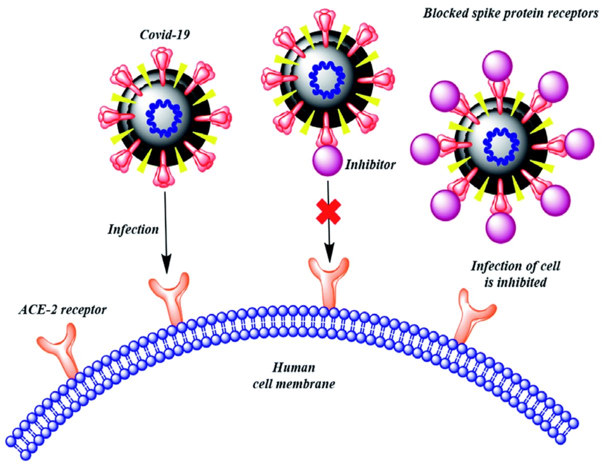
At each node of the tree, the population was split into two groups, and a two-sample proportion test was performed to determine whether any significant differences exist between patients who took medication A versus those who took medication B.

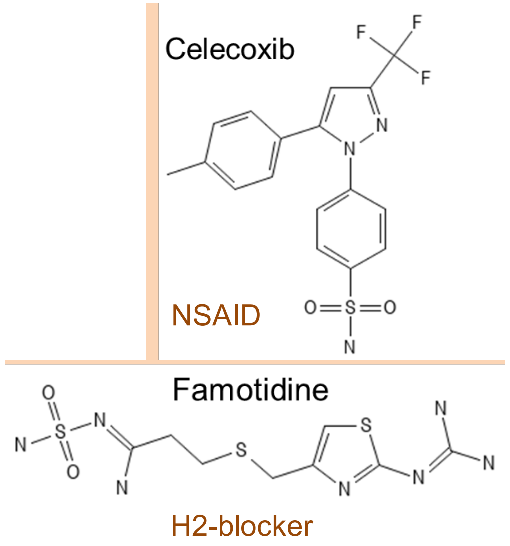


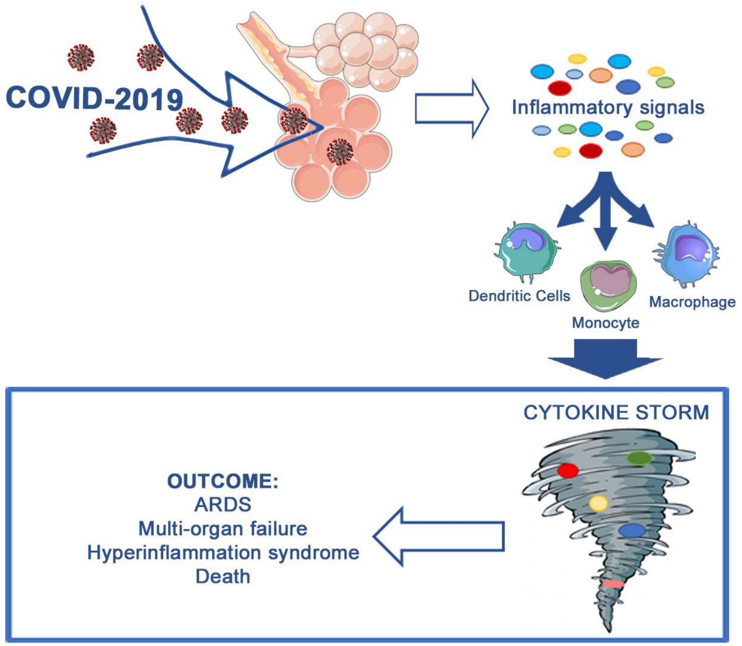
In-silico Simulated Randomized Clinical Trials Using Real-World-Data (RWD) to Model Causal Effect of Repurposed Drugs Study

- Using Real-World-Data (RWD), develop in-silico models for rapidly identifying repurposed drugs that can lower the risk of death due to Sars-CoV-2 infection

- Risk of death due to COVID-19 is predominantly due to hyperactive host inflammatory responses resulting from infection







COVID-19 Real World Data (RWD)

Charge Data Master (CDM)

CDM In-patient and out-patient having COVID19 diagnoses from 6/1/2020 - 1/31/2021

CDMVisitLocationFact
DISCHG_MONTH_ID
VST_ID
DAY_IN_VST_NBR
LOCATION_ID
LOCATION_DESC

CDMFacility
FCLT_ID
REGION_NM
RURAL_URBAN_CD
TCHG_IND
BED_SIZE_DESC

CDMVisitFact
DISCHG_MONTH_ID
VST_ID
PATIENT_ID
ADMT_CATG_DESC
ADMT_DIAG_CD
ADMIT_DIAG_VERS_TYP_ID
DISCHG_STATUS_DESC
TOTAL_CRG_AMT
ADMT_DT
FCLT_ID
IP_OP_IND
PAY_TYP_DESC
DISCHG_DT

Procedure
PRC_CD
PRC_VERS_TYP_ID
PRC_TYP_CD
PRC_SHORT_DESC
PRC_DESC

Diagnosis
DIAG_CD
DIAG_VERS_TYP_ID
DIAG_SHORT_DESC
DIAG_DESC

CDMVisitResourceFact
DISCHG_MONTH
VST_ID
SVC_DT
REV_CD
BILLG_DESC
PRC_CD
PRC_VERS_TYP_ID
RSRC_QTY
RSRC_TOTAL_CRG_AMT

Patient
PATIENT_ID
PAT_BRTH_YR_NBR
PAT_GENDER_CD

CDMVisitDiagnosisFact
DISCHG_MONTH_ID
VST_ID
DIAG_CD
DIAG_VERS_TYP_ID
PRI_DIAG_IND

Longitudinal Rx (LRx)

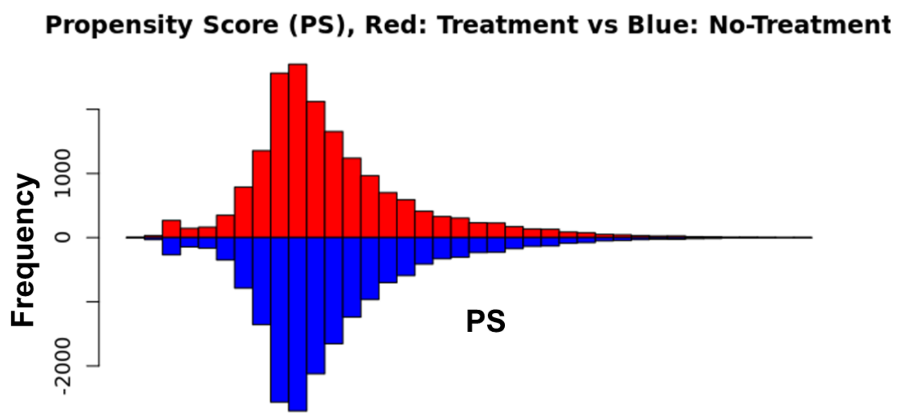
Rx data for 120-day lookback from each patient’s earliest CDM visit with a COVID-19 diagnosis

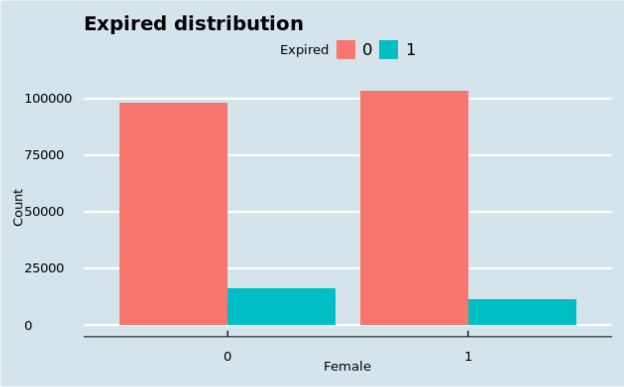
RxFact
MONTH_ID
SVC_DT
PATIENT_ID
CHNL_CD
CLAIM_ID
RX_TYP_CD
PROVIDER_ID
PAY_TYP_DESC
PRODUCT_ID
AUTH_RFLL_NBR
DSPNSD_QTY
DAYS_SUPPLY_CNT

RxProduct
PRODUCT_ID
NDC_CD
MKTED_PROD_NM
STRNT_DESC
DOSAGE_FORM_NM
LBLER_NM
LBLER_TYP_CD

CDM and LRx Data Source: IQVIA Inc

Results and Findings





Results shown for Famotidine; Celecoxib results are similar

$$\text{SMD} = \left(\frac{\bar{x}_{treatment} - \bar{x}_{\text{control}}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}} \right)$$

McNemar's Exact Test Results

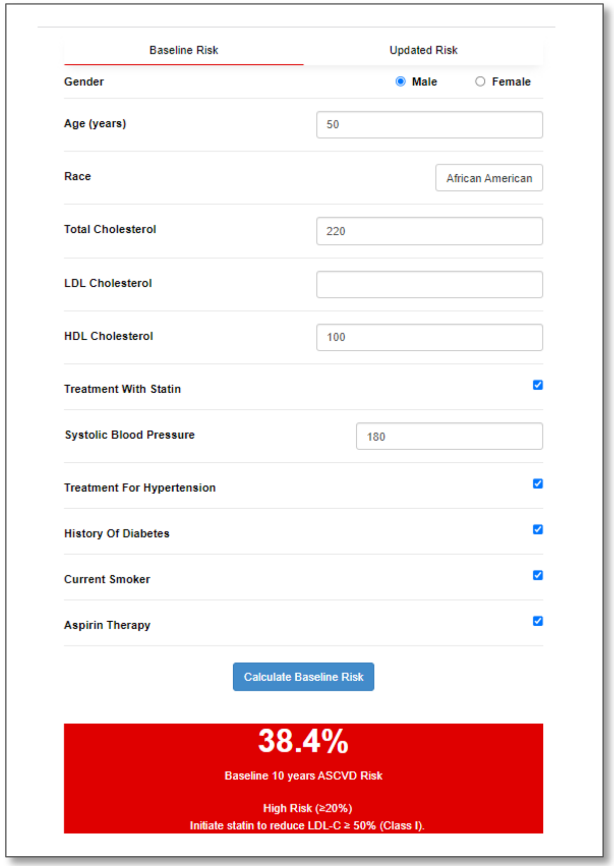
	Celecoxib				Famotidine			
Run	**N**	**OR**	**CI (95%)**	**P-value**	**N**	**OR**	**CI (95%)**	**P-value**
1	1013	2.3870	1.5498, 3.7573	3.276e-05	17916	2.400	2.2254, 2.5898	< 2.2e-16
2	999	4.5882	2.6903, 8.2730	1.642e-10	17892	2.5143	2.3304, 2.7145	< 2.2e-16
3	1019	2.0000	1.3148, 3.0927	8.200e-04	17622	2.5978	2.4045, 2.8085	< 2.2e-16
4	1026	2.3636	1.5545, 3.6669	2.326e-05	17897	2.4851	2.3029, 2.6833	< 2.2e-16
5	1046	2.4838	1.6175, 3.9002	1.115e-05	17916	2.5967	2.4056, 2.8050	< 2.2e-16

Our matched case-control study results for both treatment options Celecoxib and Famotidine show Odds Ratio OR > 1 indicating that the Famotidine and Celecoxib did not provide protective effects for COVID-19 patients. Please note that conclusions need more strengthening with follow-up studies

Prognostic Modeling?

- **What is Prognostic Modeling?**
  - Predicting risk of a future event
    - Heart attack, death
- **Applications:**
  - Useful for finding the survival with a disease (ex., brain tumor)
  - What is the risk of a disease?
  - Effective for creating treatment guidance
    - Who is eligible for end-of-life care

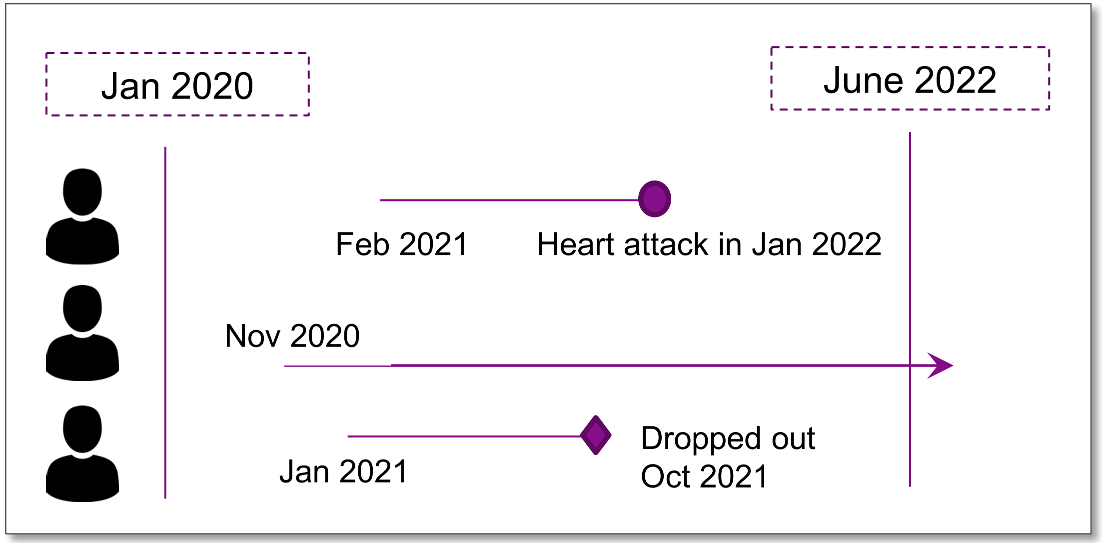
2018 Prevention Guidelines Tool CV risk calculator (American Heart Association)



http://static.heart.org/riskcalc/app/index.html#!/baseline-risk

Prognostic Modeling

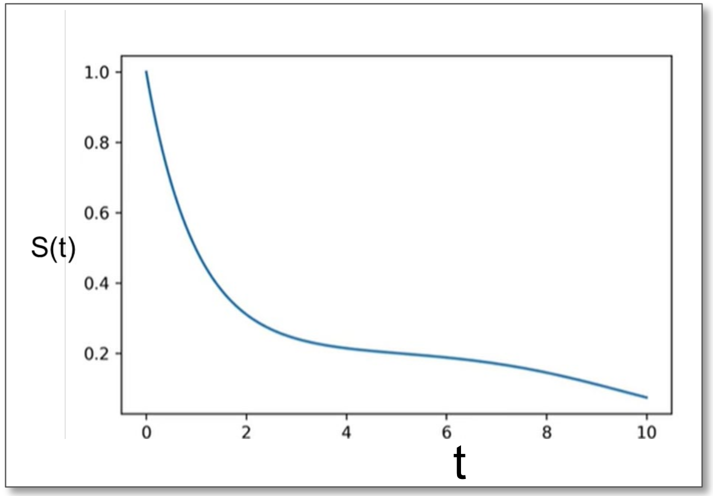
- **Survival Models**
  - Computes the probability of survival past any time ‘t’
  - Group based analysis
    - Ex., stage-1 cancer vs stage-4 cancer patients' survival
    - Ex., P(time to death > 2 years) = 0.8
- **Evaluating models**
  - C-index (Concordance index)



+ indicates censored patients

i	T_i	Event
1	10	1
2	55+	0
3	20	1
4	15+	0
5	30+	0
...	...	...

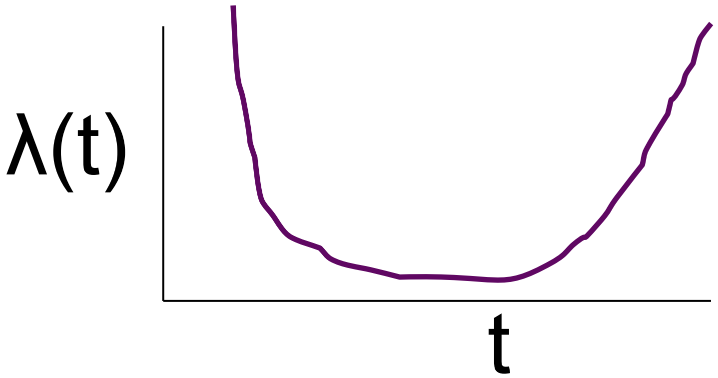
***S(t) = Pr(T>t);* T is the time to an event**



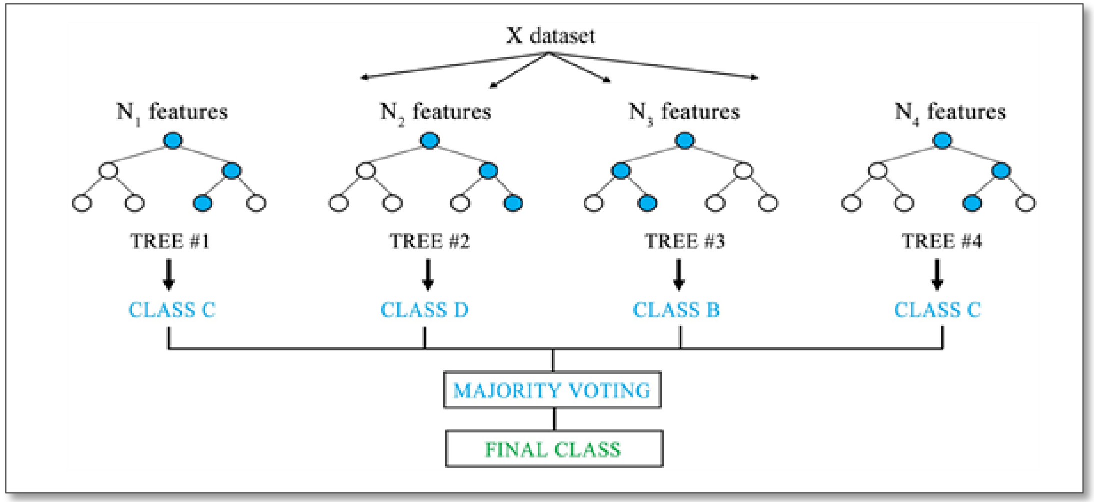
Prognostic Modeling

- **Cox Proportional Hazard Model:**
  - Popular choice for right-censored time-to-event data
  - Individual patient prediction

Hazard Function: H(t) or λ(t) = Pr(T=t | T >= t); T is the time to an event



- **Random Survival Forest:**
  - Alternative approach to Cox Proportional Hazard Models
  - Tree-based method for analysis of right censored time-to-event data
  - Extension of Random Forest
  - Approach to model subpopulations



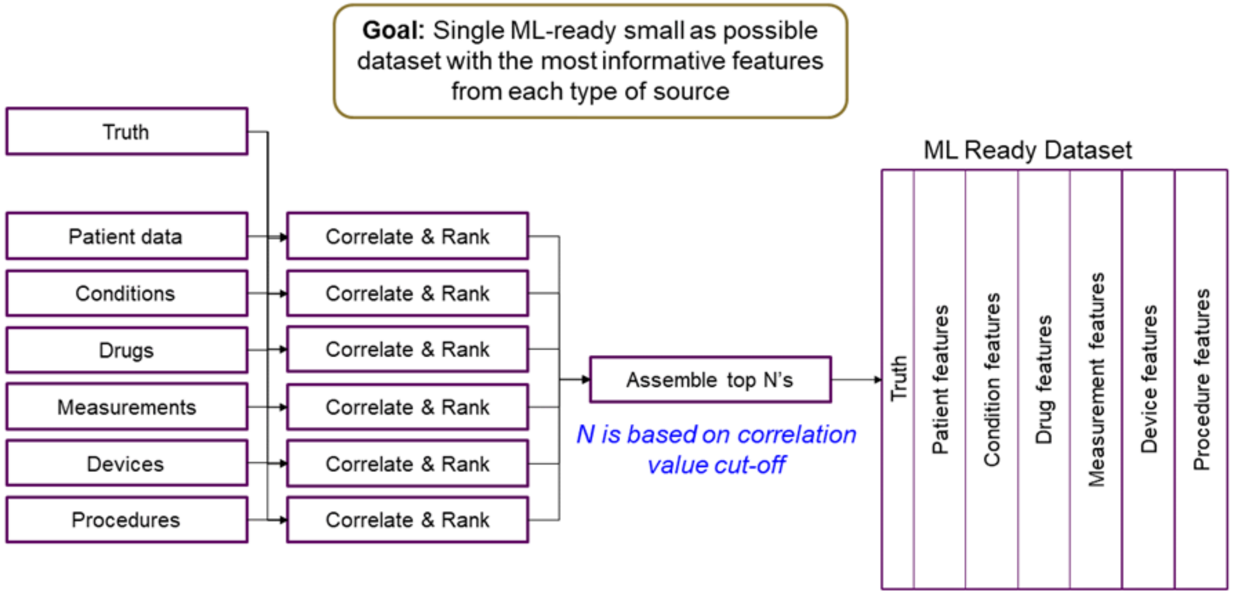
Prognostic Modeling: Long COVID Study

Predicting the likelihood of developing Long COVID is essential to identify at-risk patients and to provide timely treatment options

- Data: The EHR data contained COVID-19 infected patients with information such as demographics, procedures, medical conditions, physical measurements, lab results, and many more factors. The dataset contains more than 15 million patients, including more than 5 million COVID-19 infected patients, and over 17.5 billion rows of raw data (Source: N3C)
- Assumptions: Data missing ‘time-to-event’ Ti for censored patients. Couldn’t determine if these patients left the study before the end or survived until the end of the study. Because majority of the population was censored (> 80%), chose to retain by assigning a maximum value for time-to-event Ti.

Feature Engineering & Selection for High-Dimensional Real-World Data

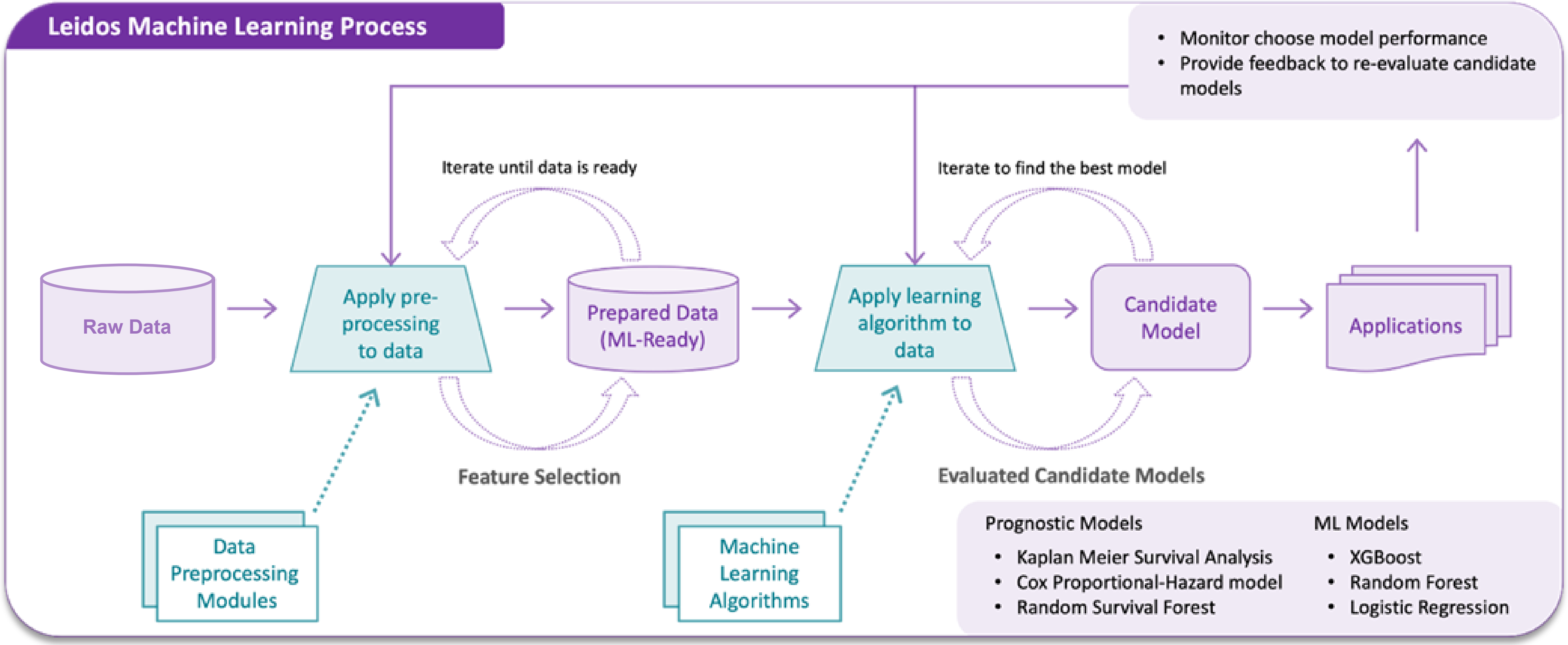
- Correlations of each feature with the presence of outcome were tabulated and ranked from most to least informative.
- Approach provided substantial improvement to prognostic and predictive classification models.
- Applications
  - Reduce dimensionality, especially for high-dimensional Real-World-Data



Correlation method (binary-binary)
- Tetrachoric correlation
- Based on Chi-squared
- Theil's U

Modeling Methodology Pipeline

- Includes Feature Selection and Modeling (Prognostic and Machine-Learning models)

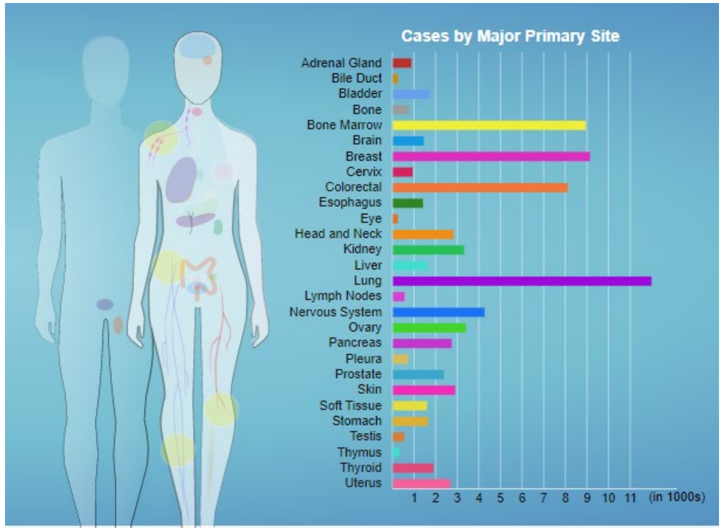


Prognostic Modeling: Long COVID prediction

- **Results and Findings**
  - Performance analysis of predictive classification models showed XGBoost as best with AUC = 0.93 and a balanced accuracy of 0.75.
  - Feature selection based on measuring feature power provided substantial improvement to prognostic and predictive classification models.
  - The most powerful features had correlations with Long COVID diagnoses as great as 0.45 and were substantially better than the correlations of 0.15 obtained on the original feature set

	Performance	
Models	**AUC**	**Other**
XGBoost	0.9260	-
XGBoost_upsampling	0.9236	-
XGBoost_downsampling	0.9104	-
RandomForest	0.9114	-
Logistic Regression	0.8500	-
Odds Ratio (exposed: gender)	-	OR(CI): 1.27 (1.21, 1.33)
Kaplan-Meier (log-rank test; gender group test)	-	p-value: 3.835e-23
Cox Proportional Hazard	-	Harrell C-index: 0.7939
SurvivalRandomForest	-	Harrell C-index: 0.8624

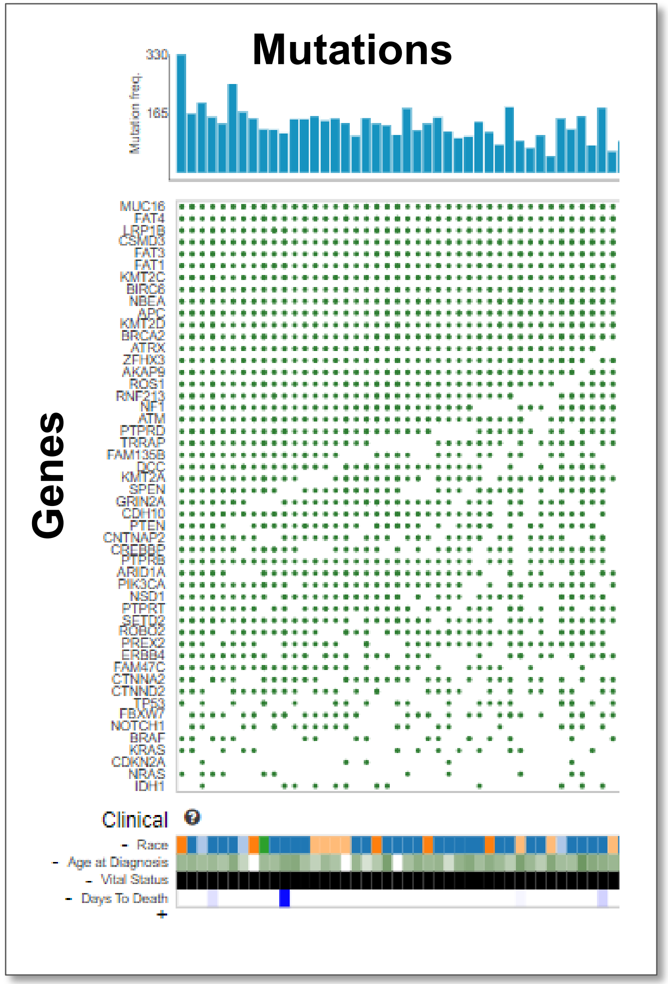
Genomic Expression Data Modeling Study

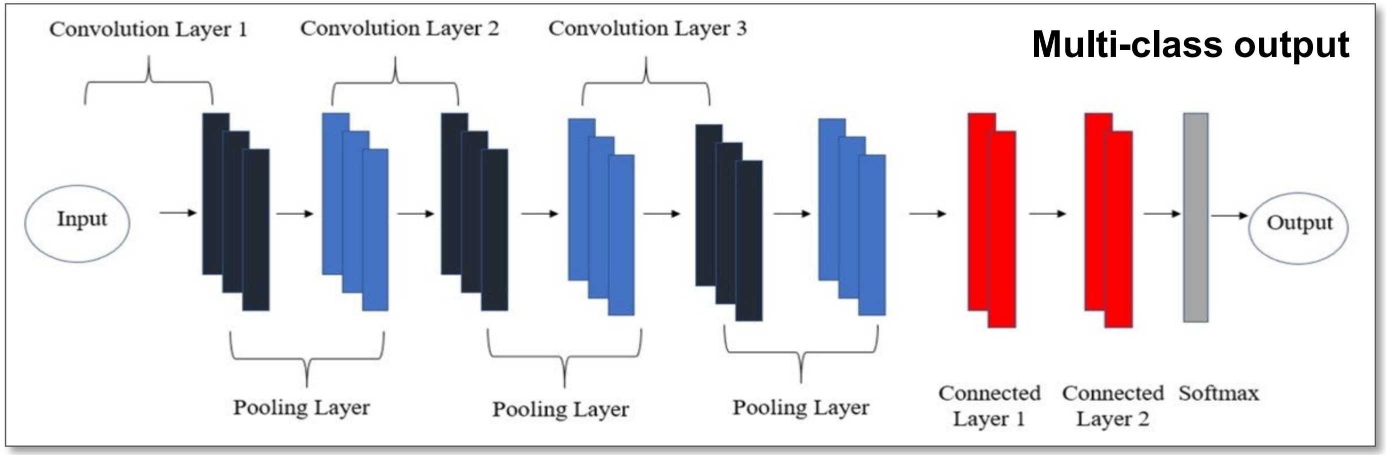


https://portal.gdc.cancer.gov/

Input

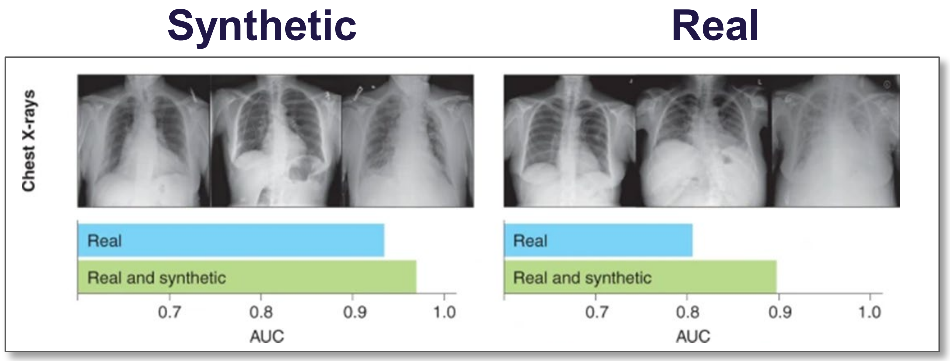
https://portal.gdc.cancer.gov/



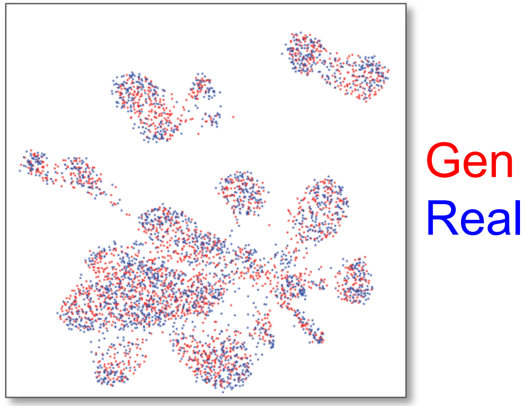


Synthetic Data & Synthea

- Having access to good-quality Healthcare data is a bottle-neck for research, training, and technology development
- The scientific community is struggling to strike a balance between the competing interests of protecting patient privacy/confidentiality and making data public
- Can we generate/share realistic synthetic healthcare data that statistically reflects the concerned population, and protecting patient privacy and confidentiality?



https://www.nature.com/articles/s41551-021-00751-8



doi: 10.1093/bioinformatics/btab035

Synthetic Data & Bias detection

- Synthetic data can be used as a test for Bias detection
- It is the data and not the algorithm that is biased
- If we want to eliminate bias from our AI systems, then we need to remove the bias from the data before we use it to build models
- Bias mitigation strategies:
  - Before training: rebalance the data by collecting representative datasets (not easy);
  - During training: Data augmentation, adversarial training
  - After training: Model outcomes can be post-processed based on sub-groups

https://www.nature.com/articles/s43856-021-00028-w;

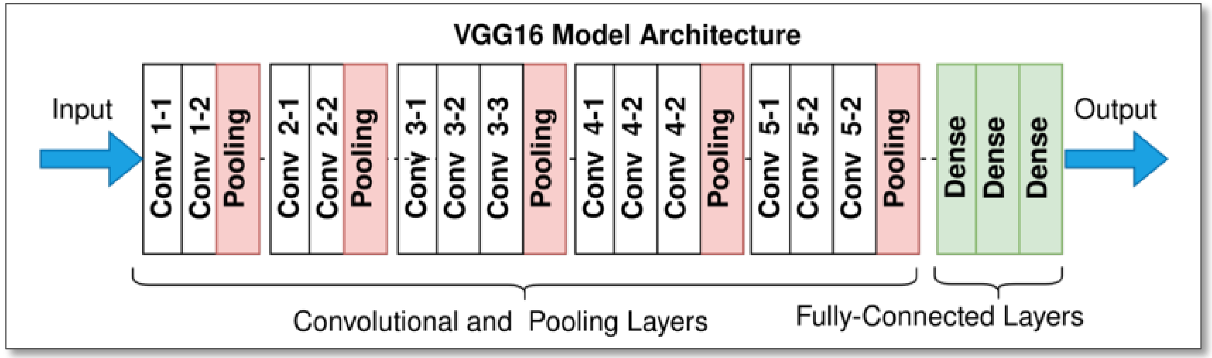
https://www.nature.com/articles/s41467-022-32186-3

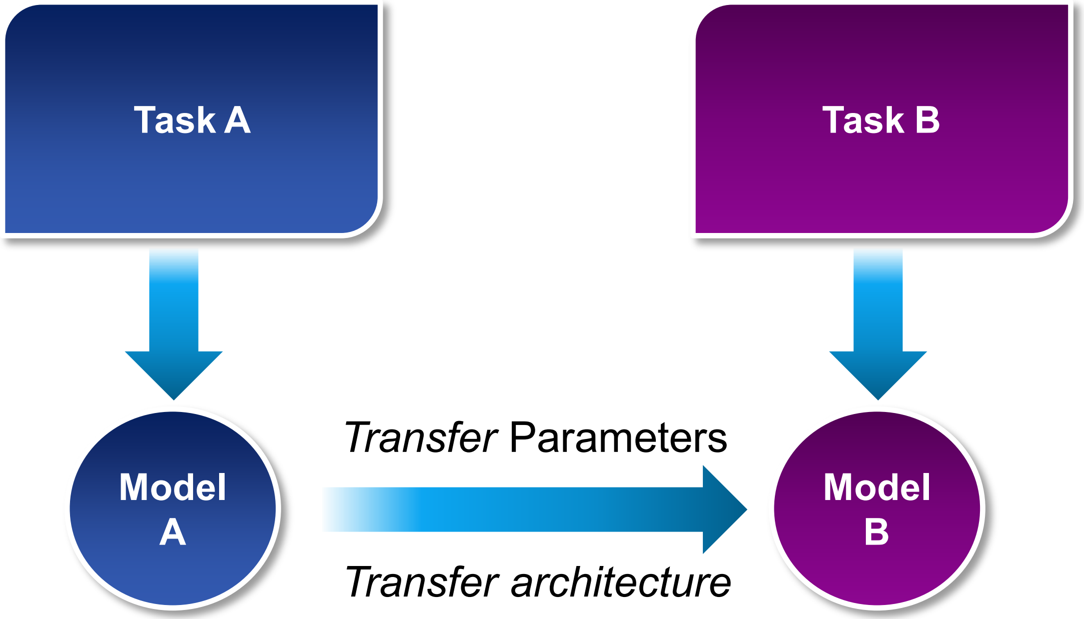
https://github.com/synthetichealth/synthea

Transfer Learning

- We could use image (Chest X-ray) dataset and train a CNN to classify a presence or absence of disease (ex., pneumonia).
- This model predictions can help human radiologist to speed up the predictions.
- One could take the technology and submit it to the FDA for 510(k) clearance as Software as a Medical Device.

CNN: Convolutional Neural Network





Current Interests

- How to handle confounding due to unobserved features? (causal modeling)
- Best method to combine RCTs and observational data? (causal modeling)
- Patient self-selection is a problem in observational studies. How to control this issue?
- Feature selection and dimensionality reduction in high-dimensional RWD
- Bias detection modeling/evaluation for healthcare data is complex. If the prevalence of the target disease is different between different groups, then what `selection fairness condition` would be appropriate?

Current Interests

- What is the best Deep-Learning architecture?
- Create Interpretable Neural Network models? (Explainability)
- Data normalizations for biological expression data?
- Model repositories, storage/retrieval? (Sharing Models, Outcomes, Collaboration)

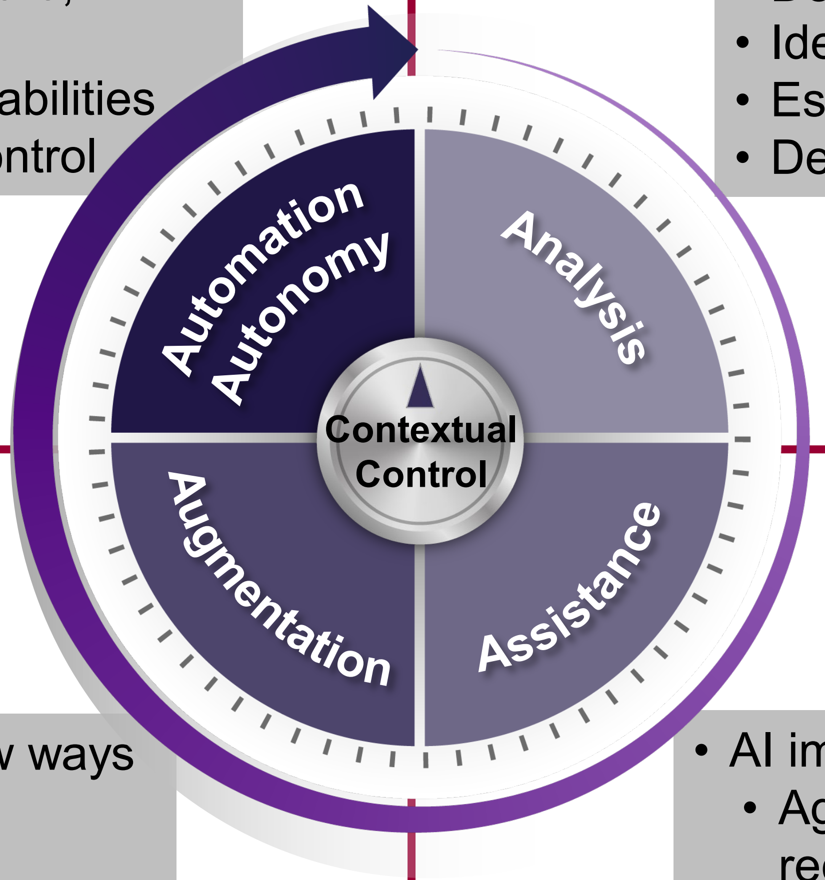
4A Methodology for AI Trust and Explainability

leidos

- Replace humans for complex tasks
  - Control, decide, act, navigate, collaborate, achieve goals
- Feedback/data drives AI capabilities
- Human on-the-loop (OTL) control

- Characterize initial data
- Design data strategy
- Identify / rank opportunities for AI
- Estimate ROI, risk
- Deliver Data Science capabilities

Collaboration



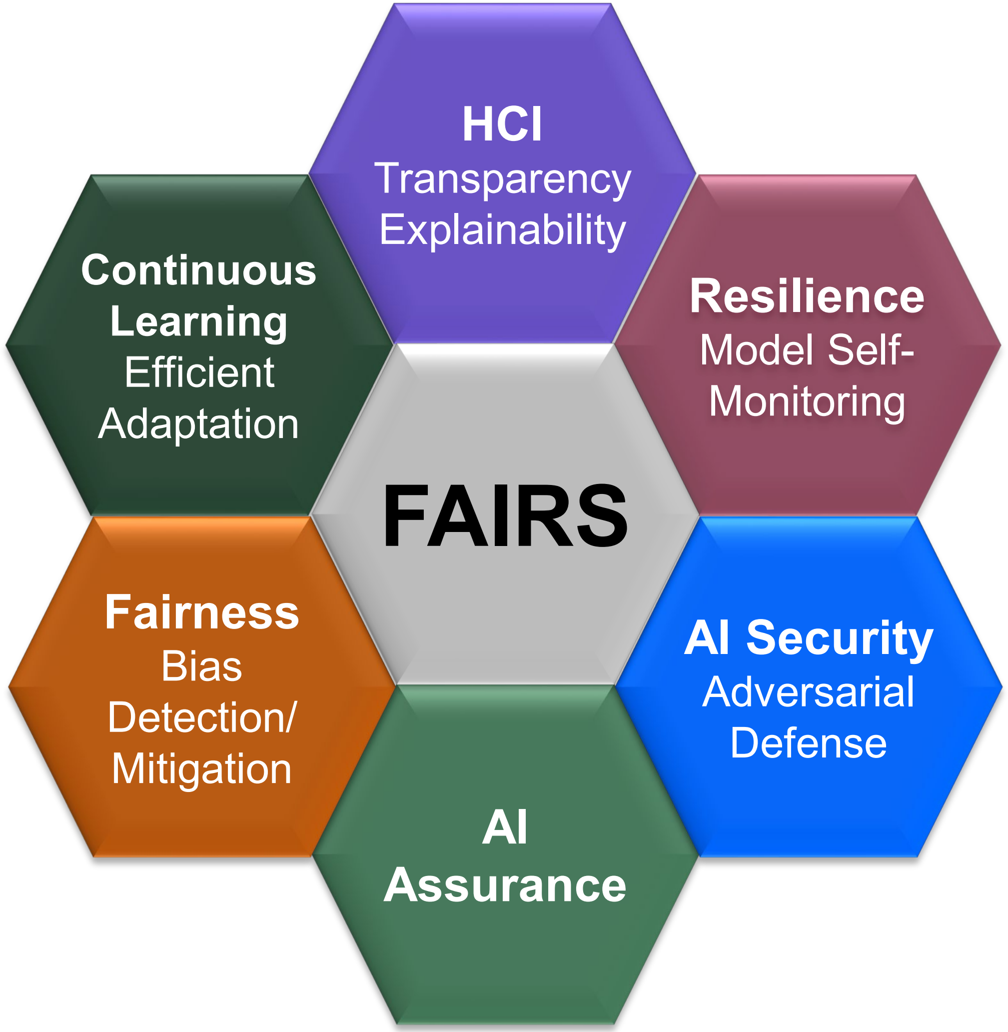
Open Science

- AI helps humans work in new ways
  - Track, transform, organize, predict, warn, hold state
- Feedback/data drives AI capabilities
- Human ITL control

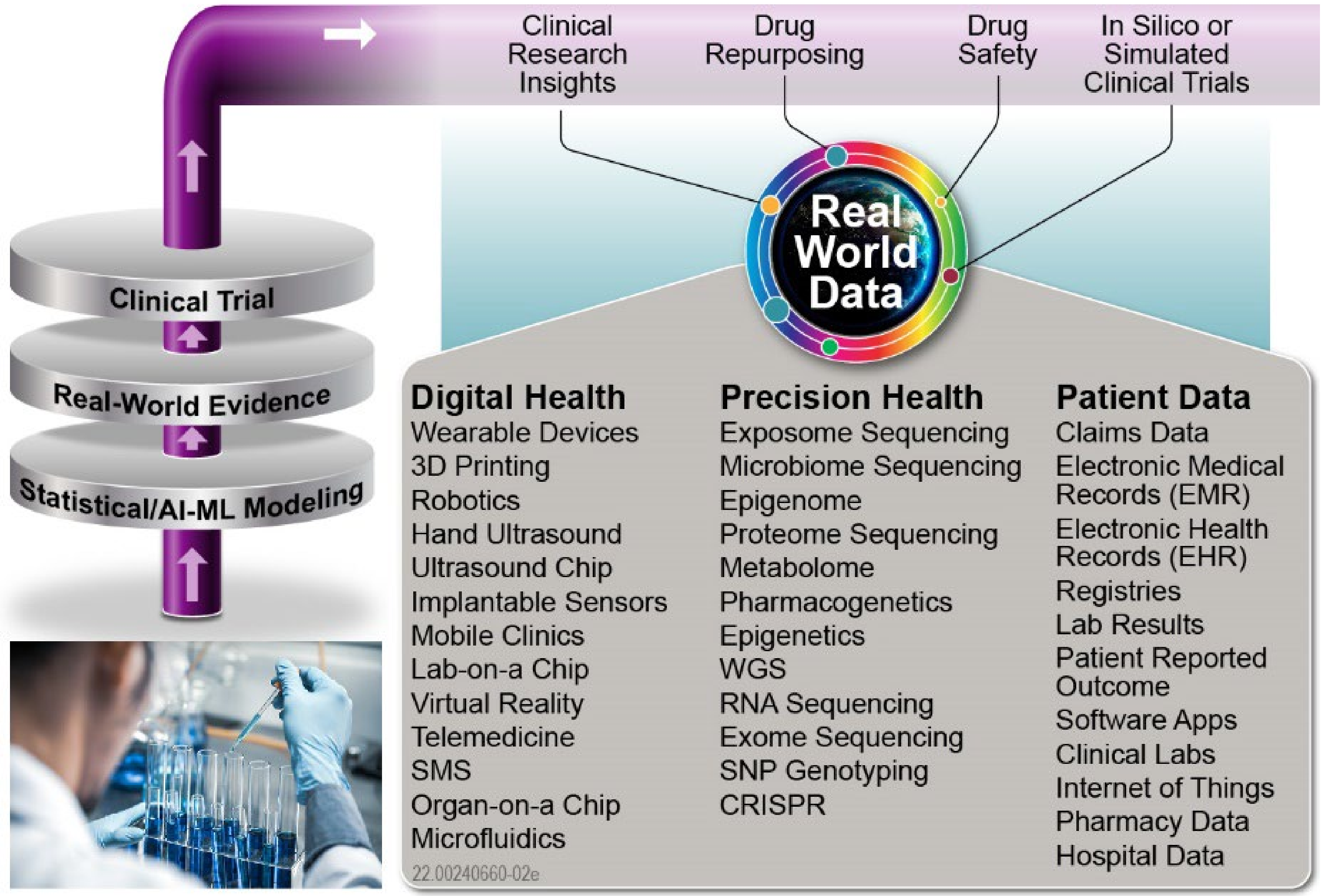
- AI improves efficiency of workflows
  - Aggregate, alert, enrich, recommend, visualize
- User feedback improves accuracy
- Human in-the-loop (ITL) control

Framework for AI Resilience and Security (FAIRS)

- **FAIRS** provides a broad set of complementary and mutually reinforcing capabilities for AI Trust in a unified framework
- **FAIRS** components packaged as microservices, allow flexible deployment, configuration with speed, scale and security



Future Directions



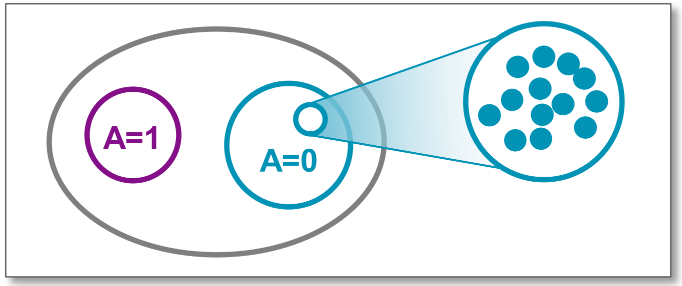
Contact Info

Chetan Paul
Chetan.Paul@leidos.com

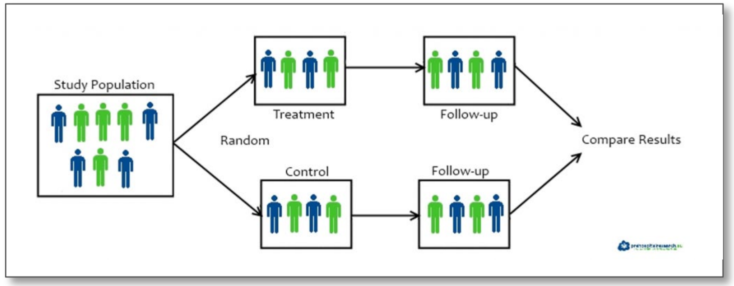
Dr. Ravichandran Sarangan
Sarangan.Ravichandran@leidos.com

Backup Slides

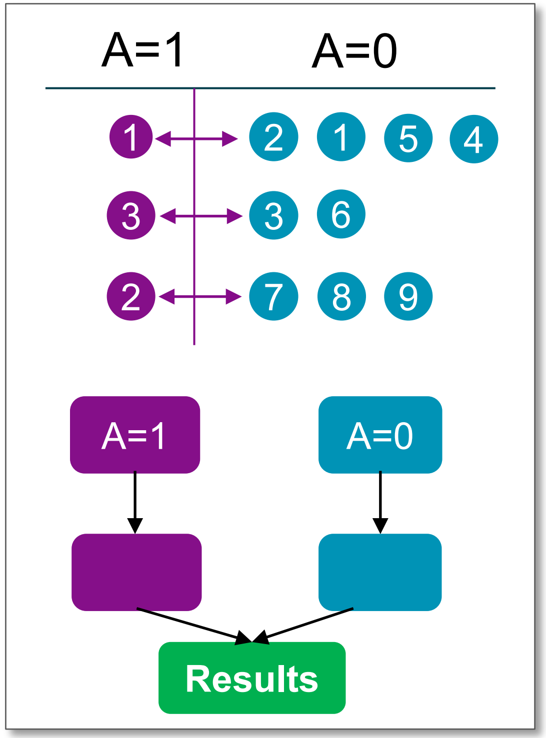
RCT vs RWE



Time



Time



Id	Age	Female	BC	CC
1	36	1	1	1
2	37	1	1	0
3	36	0	0	0

	1	2	3
1	0	$\sqrt{2}$	$\sqrt{3}$
2		0	$\sqrt{3}$
3			0

Distance → Matching using Nearest Neighbor

Treated

Weight:

$$\frac{1}{P(A=1|X=1)} = \frac{1}{0.1} = 10$$

Control

Weight:

$$\frac{1}{P(A=0|X=1)} = \frac{1}{0.9} = \frac{10}{9}$$

Inverse Probability of Treatment Weighting (IPTW)

Overview of RWE Modeling

leidos

EDA: Exploratory Data Analysis
RCT: Randomized Control Trial
IPTW: Inverse Probability Treatment Weighting

$$\text{smd} = \left(\frac{\bar{x}_{treatment} - \bar{x}_{\text{control}}}{\sqrt{\frac{s_{treatment}^2 + s_{control}^2}{2}}} \right)$$

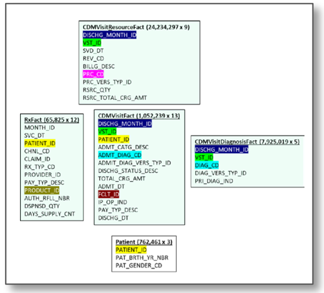
Data → Harmonized Data → ML-ready data → Explore distance metrics → Check Initial Imbalance → Try Matching methods → Access the quality of matches; select the best matching method → Estimate the treatment effect → Conclusion

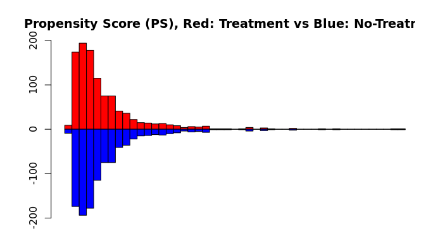
EDA ↔ Data; EDA ↔ Harmonized Data

Mahalanobis

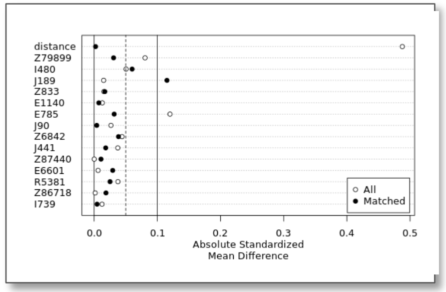
NN, full etc.

Use IPTW reduce confounding





Risk difference/ratio etc



Matching Methods

Hypothetical Data to Illustrate Matching Methods

Treated individuals		Comparison individuals	
Individual	Income (in $10,000)	Individual	Income (in $10,000)
A	42	a	44
B	35	b	42
C	24	c	37
D	22	d	34
		e	23

Dev Psychol. 2008 March ; 44(2): 395–406. doi:10.1037/0012-1649.44.2.395.

Method	Matched Pair	Global Distance
Greedy (or Nearest-Neighbor)	{Ab}, {Bd}, {Ce}, {Dc}	17 = (0 + 1 + 1 + 15)
Optimal (1:1) matching	{Ab}, {Bc}, {Cd}, {De}	13 = (0 + 2 + 10 + 1)
Full Matching	{Aab}, {Bcd}, {Cde}	7 = (2 + 0 + 2 + 1 + 1 + 1)