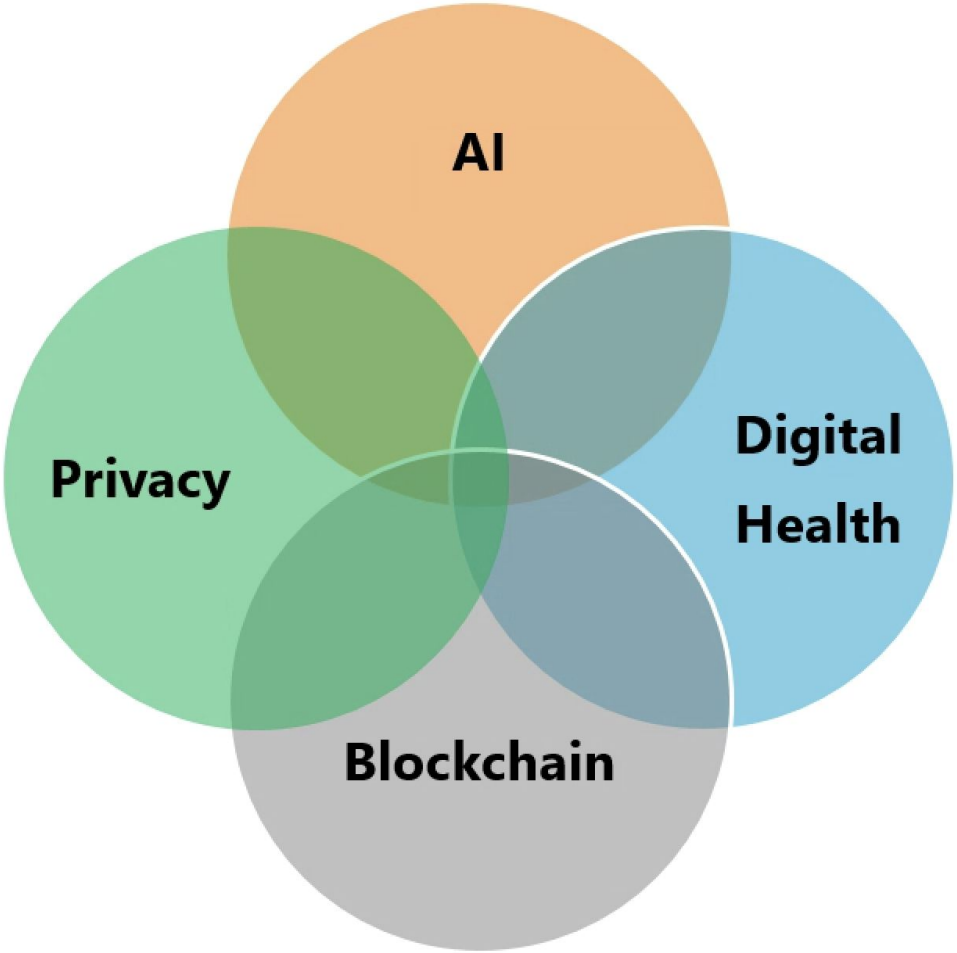
A3.AI
Covid-19
Project Overview

02/2021

- About Us
- Aims
- Data
- Approach
- Current Work
- Plan

Nonprofit Applied R&D



We are volunteers:

Data Scientists
AI Practitioners
Physicians
IT Engineers
Privacy Specialists

A3.AI

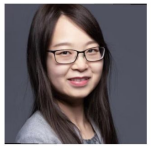
COVID-19 Project Team

Practitioners

Changrong Ji

Dr. Mahesh Shukla

Dr. David Patton



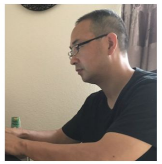
Dr. Xue Yang

Antonio Linari

James Scott

Premdutt Gaur

Katya Vert-Wong



Dr. Xingguo Zhang

Vance Degen

Advisors

Mark Cullen, MD

Greg Ewing, JD/MPH

Ben Carroll

Dr. Annastasiah Mhaka

Research Aims

Current

- COVID-19 Severe Outcome Risk Prediction with ML
  - Hospitalization
  - Death
  - ICU
- Descriptive analytics on social and clinical factors
- Preliminary work on Drug Repurposing
- Privacy-preserving Machine Learning

Plan

- COVID Long Term Effects
- Drug Safety and Repurposing
- **Gain access to richer data**
- **Collaborate with clinical researchers**

Current Data

- Medical Claims
  - 106 million patients
  - 7 years of medical claims history
  - 4 billion claim lines
  - ~900,000 COVID patients
  - Source: a claims clearinghouse
- Social Data
  - 240 million people, 80 attributes
  - People (demographics, finance, housing, jobs,etc)
  - Behaviors (health, lifestyle,, politics, etc)
  - Source: AIQ
- Death Registry
  - 80% of US population
  - Source: Datavant
- 40 million people have Medical + Social Data
- Electronic Health Record
  - Primarily Outpatient
  - 36 million patients
  - 7 million have Rx
  - 210,000 COVID Patients
  - Source: HealthJump

1000s lines of feature engineering code

Data characteristics:

Pros:

- Broad
- Multiple years
- Social + medical + mortality

Cons:

- Incomplete medical history
- Latency
- Sicker than general population
- Uneven regional distribution

Techniques

- Machine Learning
- Deep Learning
- Knowledge Graphs, NLP
- Handcrafted features
- Clinical Concept Embeddings
- Descriptive Analytics
- Causal Models
- Fixed Effects Models

Data Distribution and Labels

Presenter: Dr. Xue Yang

Hospitalization	15.7%
ICU	6.5%
Ventilation	0.4%
ECMO (heart-lung bypass)	0.01%
Death, etc	3.1%

Descriptive Analytics

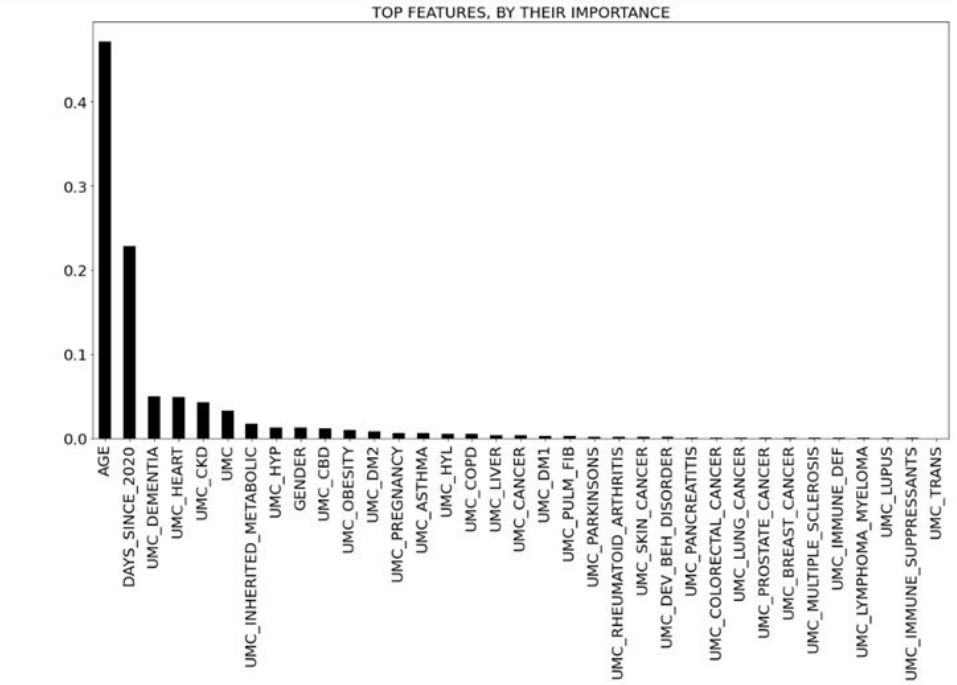
Demo

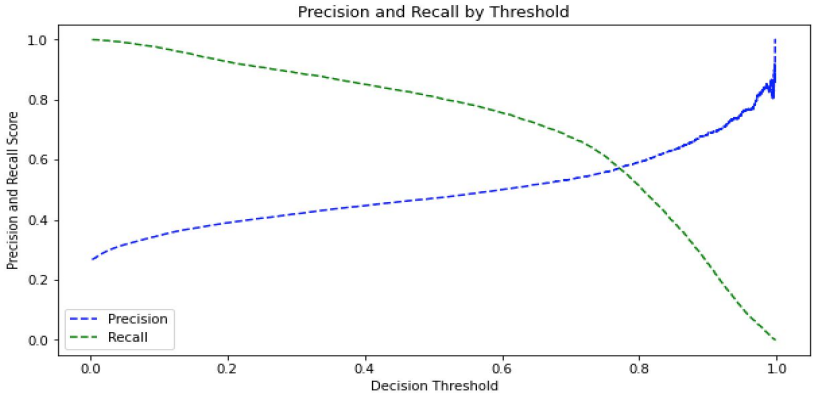
COVID Risk Prediction AI Models

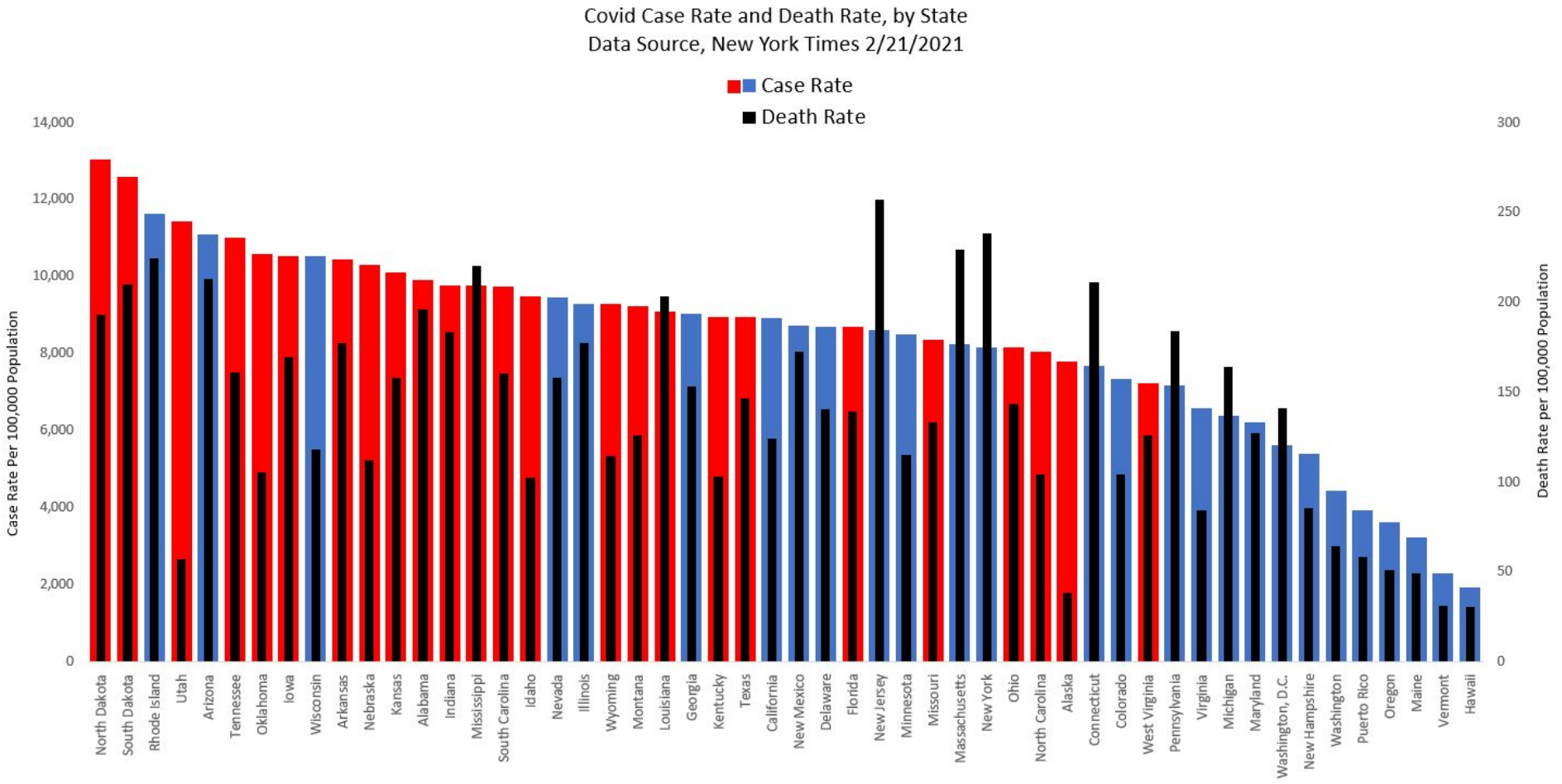
- Based on clinical and social features
- Predict risk of severe outcomes if infected with COVID
  - Hospitalization
  - ICU
  - Death
  - Specific complications

Modeling

  - Deep Neural Net
  - Random Forest
  - XBGoost
  - Causal
  - Fixed Effects
  - Clinical Concept Embeddings



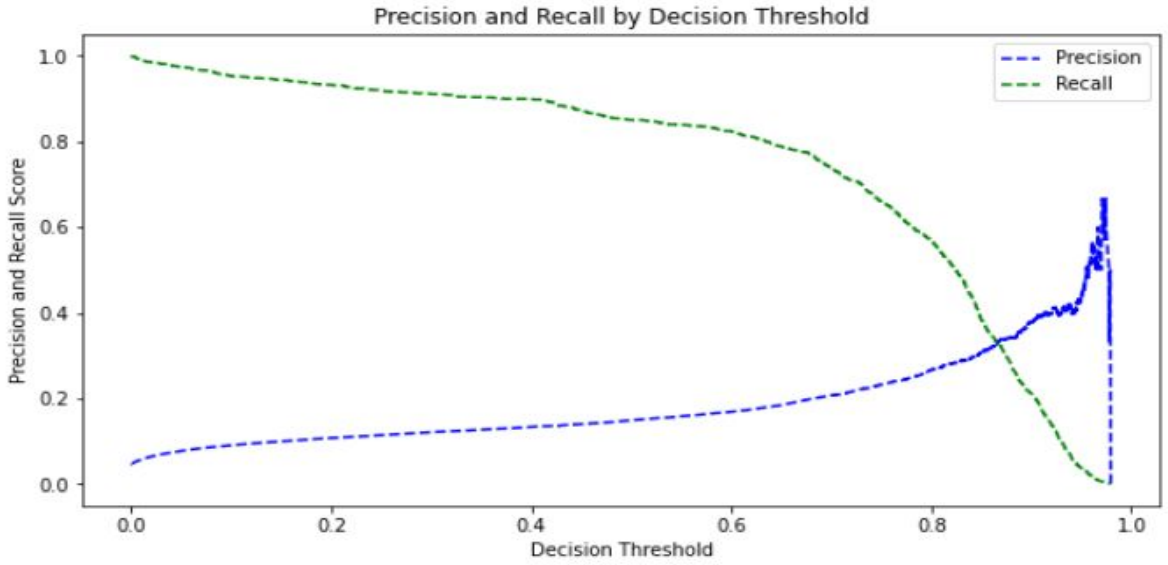



Covid Case Rate and Death Rate, by State
Data Source, New York Times 2/21/2021
Case Rate
Death Rate
Case Rate Per 100,000 Population
Death Rate per 100,000 Population
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
300
250
200
150
100
50
0
North Dakota
South Dakota
Rhode Island
Utah
Arizona
Tennessee
Oklahoma
Iowa
Wisconsin
Arkansas
Nebraska
Kansas
Alabama
Indiana
Mississippi
South Carolina
Idaho
Nevada
Illinois
Wyoming
Montana
Louisiana
Georgia
Kentucky
Texas
California
New Mexico
Delaware
Florida
New Jersey
Minnesota
Missouri
Massachusetts
New York
Ohio
North Carolina
Alaska
Connecticut
Colorado
West Virginia
Pennsylvania
Virginia
Michigan
Maryland
Washington, D.C.
New Hampshire
Washington
Puerto Rico
Oregon
Maine
Vermont
Hawaii

COVID Death Risk with DNN & Future Diagnosis Prediction

Presenter: James Scott

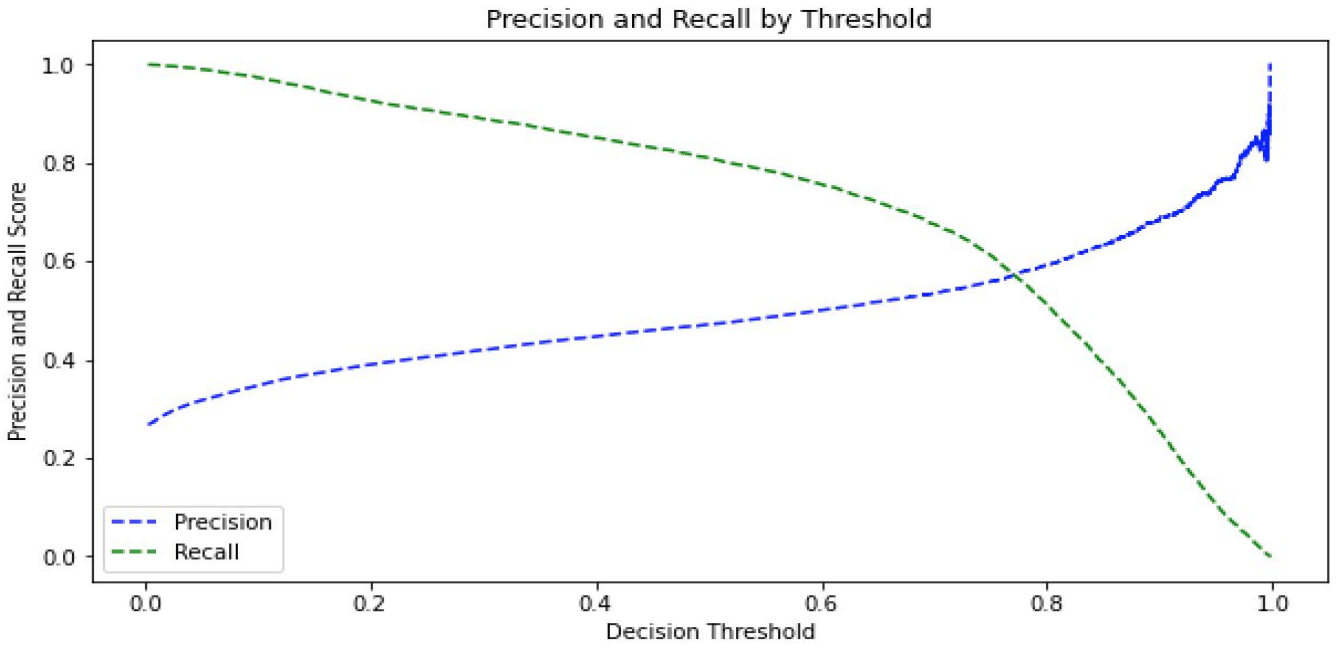
- Machine Learning vs Deep Learning
  - DNN Layers - 1 Flatten, 9 Dense, 3 Dropout Layers
- Used ML techniques in application with DNN architecture to produce better PR Curve than ML Models
- Results
  - AUROC: 0.8804, 20% Precision at 79% Recall
- Future State: Using conditions before and after COVID to determine future health complications using AI



ICU Risks of COVID Patients

Presenter: Dr. David Patton

- Preprocessing to build claim history and data cleaning
- Large DNN (over 850k parameters)
- Learned Embedding of diagnostic history
- Results
  - AUROC: 0.8154, AUPRC: 0.5921 48% Precision at 80% Recall



Clinical Concept Embedding & Patient Fingerprint

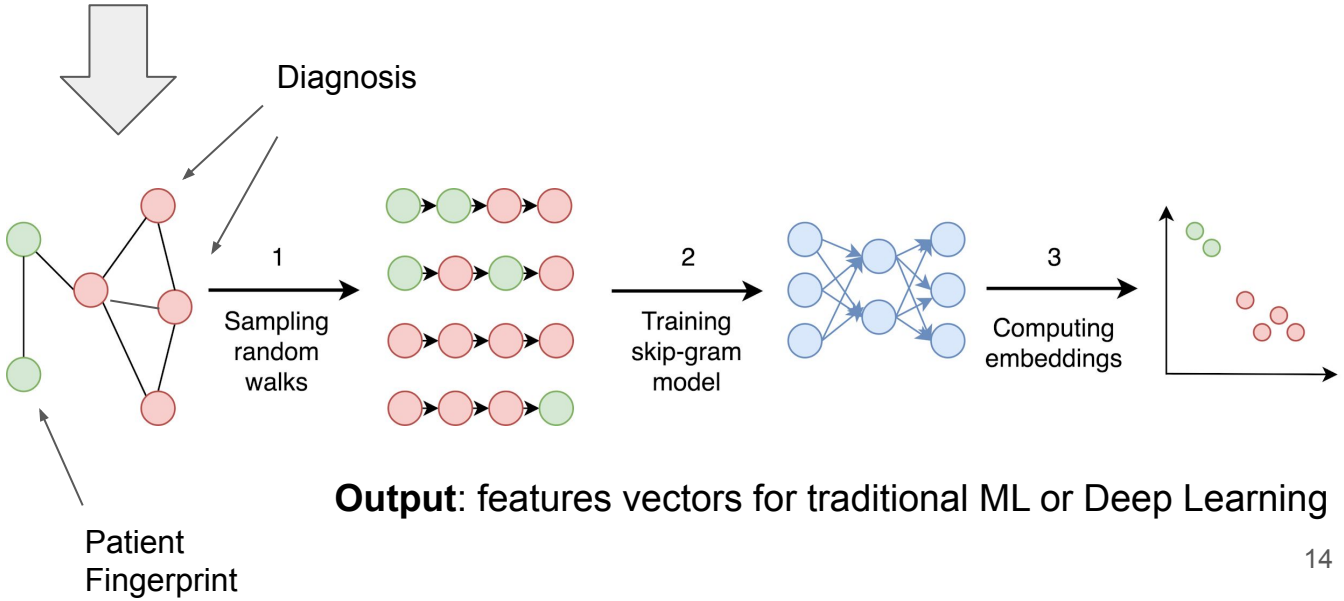
Presenter: Antonio Linari

DESYNPUF_ID	BENE_BIRTH_	BENE_DEA	BENE_SEX	BENE_RAC	BENE_ESR	SP_STATE_	BENE_COU
00013D2EFD8E45D1	19230501		1	1	0	26	950
00016F745862898F	19430101		1	1	0	39	230
0001FDD721E223DC	19360901		2	1	0	39	280
00021CA6FF03E670	19410601		1	5	0	6	290
00024B3D2352D2D0	19360801		1	1	0	52	590
0002DAE1C81CC70D	19431001		1	2	0	33	400
0002F28CE057345B	19220701		1	1	0	39	270
000308435E3E5B76	19350901		1	1	0	24	680
000345A39D4157C9	19760901		2	1	0	23	810

DESYNPUF_ID	CLM_ID	CLM_FRO	CLM_THRU	ICD9_DGN	ICD9_DGN	ICD9_DGN	ICD9_DGN
00013D2EFD8E45D1	887733386680966	20090725	20090725	7245	7244	6272	
00013D2EFD8E45D1	887213386947664	20091014	20091014	3598	27541		
00013D2EFD8E45D1	887243388666441	20100401	20100401	29606			
00013D2EFD8E45D1	887893388307089	20100817	20100817	8410	8472	8409	
00013D2EFD8E45D1	887463387476539	20101105	20101105	29521			
00016F745862898F	887523386668713	20081004	20081004	82311	3558	V0481	
00016F745862898F	887793388522804	20081008	20081008	78963	4010	78909	78701
00016F745862898F	887783385091632	20081009	20081018	2514	30000	2449	2859
00016F745862898F	887953385848149	20081020	20081020	59971			
00016F745862898F	887423388326030	20081209	20081209	87362	49390		
00016F745862898F	887383388842050	20090105	20090105	1628			
00016F745862898F	887783384850257	20090113	20090113	V5867			

Patient Fingerprint, Diagnosis...
E.g. **1:5:1:101100000000**

```
gender = tokens[0]
age = tokens[1]
covid = tokens[2]
preconditions = tokens[3]
ckd = preconditions[0]
copd = preconditions[1]
obesity = preconditions[2]
heart = preconditions[3]
dm2 = preconditions[4]
asthma = preconditions[5]
cbd = preconditions[6]
hyp = preconditions[7]
dementia = preconditions[8]
liver = preconditions[9]
pregnancy = preconditions[10]
hyl = preconditions[11]
```



Output: features vectors for traditional ML or Deep Learning

Drug Repurposing - Goals

- Utilize Real World Evidence (RWE) and FDA's chemical and pharmacological data, identify and analyze patterns in COVID-19 disease to determine how existing prescription drug products and their active ingredients impact infection rate and disease progression.
- Validate existing hypotheses
- Discover new patterns

External Collaborators

- Dr. Larry Callahan (FDA)
- Dr. Frank Weichold (FDA)
- Dr. Ana Szarfman (FDA)
- Dr. Noel Southall (NCATS)

Example: Early Findings on Hydroxychloroquine

- In 04/ 2020, in collaboration with FDA, we conducted a preliminary analysis on millions of patients' Rx claims data on 16 drugs, including Hydroxychloroquine.
- Comparing the COVID-19 infection rate among patients taking Hydroxychloroquine for pre-existing (non COVID) conditions, against the general population.
- **Real-world-evidence revealed NO statistically significant preventive effect of Hydroxychloroquine against COVID-19 infection as early as 04/08/2020.**

Organization/ Member Count	Count of members with CDC's Covid surrogate diagnosis as of 3/31/2020	Distinct members on drug list for entire 7 months	Drug Name	2020/02	2020/01	2019/12	2019/11	2019/10	2019/09	Count of members with Covid diagnosis as of 3/31/2020 by Drug	Used Emergency Department	Admitted as inpatient	Admitted to ICU
			Hydroxychloroquine										
			chloroquine										
			Adalimumab										
			Benazepril HCl										
			Canakinumab										
members			Captopril										
			Enalapril Maleate										
			Fosinopril Sodium										
			Lisinopril										
members with claims in past 3 months scanned for COVID diagnosis			METHOTREXATE										
			Perindopril Erbumine										
			Quinapril HCl										
			Ramipril										
			Sarilumab										
			Tocilizumab										
			Trandolapril										

Drug Repurposing

Study Proposal for Data requires academic medical center

SYMPHONY

One of the largest existing repositories of patient-level integrated data, including retail pharmacy data, medical claims, and remittance data. They have vast coverage of the market, including: 92% of retail pharmacy claims, 71% of mail order pharmacy claims, 65% of specialty pharmacy activity, >280 million patients, >1.8 million prescribers and >16,000 health plans. Does not include data on costs.

CHANGE HEALTHCARE

Largest claims clearinghouse with a network of 900K providers, 5500 hospitals nationally, 2100 payers, 33K pharmacies, and 600 labs. Processes nearly 50% of all commercial claims (including Medicare Advantage but not Medicare FFS) in the U.S. Claims are available T+1 days from claim processing. Data held includes 1 billion distinct medical and remittance claims per year on 125 million patients, 800 million distinct claims per year on 33 million unique patients, and 75 million lab test results per year on 2 million unique patients.

VERADIGM

Veradigm is a health information, analytics and intervention solutions company representing the largest single source of de-identified patient EMR data in ambulatory care. It will provide de-identified outpatient records from two national EHR vendors, Practice Fusion and Allscripts.