VHA INNOVATION Ecosystem

FDA U.S. FOOD & DRUG ADMINISTRATION

precisionFDA

Leveraging precisionFDA and Synthetic Data to Improve Veteran Healthcare

VHA COVID-19 Risk Factor Challenge

Amanda Purnell, Clinical Data Specialist
Veterans Health Administration Innovation Ecosystem

VHA INNOVATION Ecosystem

U.S. FOOD & DRUG ADMINISTRATION

precisionFDA

Challenge Motivations, Objectives, and Data

On March 11, 2020, the World Health Organization (WHO) declared the outbreak of the novel coronavirus disease 2019 (COVID-19) a global pandemic.

Challenge Motivations:

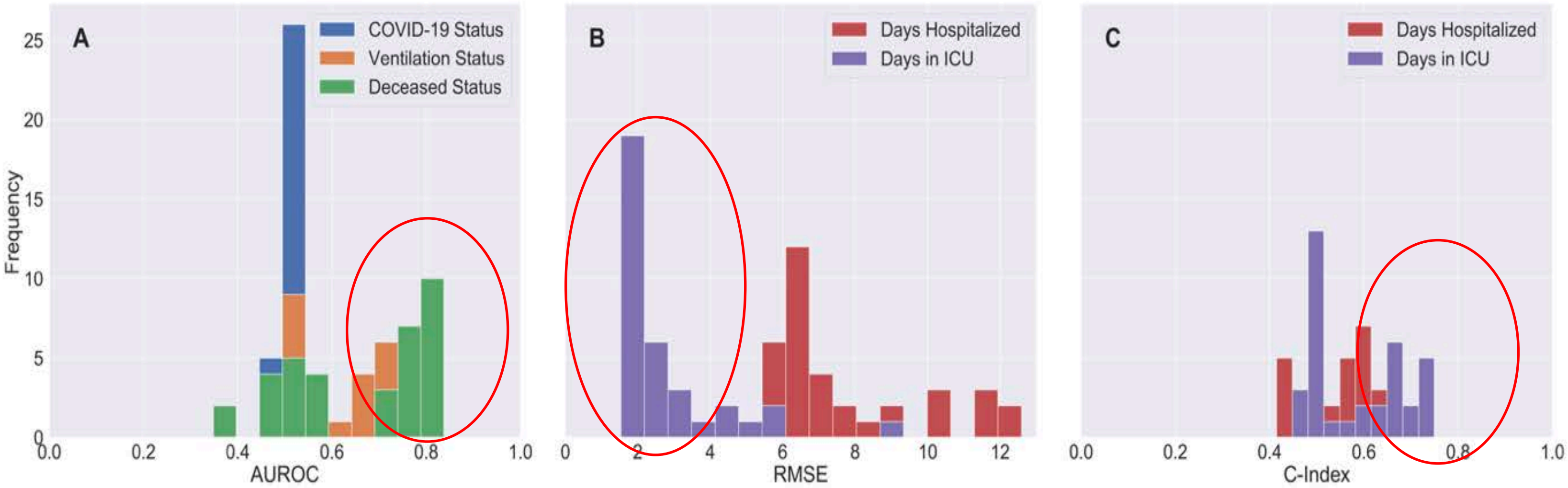
- To better understand COVID-19's impact on the Veteran population, which has a higher prevalence of several known risk factors for severe COVID-19 illness
- Identify key factors associated with COVID-19 outcomes
- Assess the usefulness of using synthetic data for Machine Learning modeling of a real-world problem

	Phase 1: Model Development	**Phase 2: Model Validation & Synthetic Data Assessment**
Objective	• Development of Machine Learning (ML) models to predict COVID-19 related health outcomes: ◦ COVID-19 status, Alive status, Ventilation status, Days in ICU, Days in Hospital	• Validated top performing models on additional synthetic EHRs and de-identified real EHRs • Assessed whether synthetic data is a reliable resource for model development
Data	• 147,451 Synthea-generated synthetic patients, including: ◦ patient information, conditions, care plans, observations, medications, procedures, providers, etc.	• 75,418 GAN-generated synthetic patients with medical history data aligning to the Synthea-generated dataset • >658,000 Veterans from the VA Covid Repository database with medical history data aligning to the Synthea-generated dataset

Phase 1 Results – Predictive Modeling with Synthetic Health Records

34 Total Submissions:

- Models use a wide array of ML techniques including Random Forest, Adaptive Boost (AdaB), Neural Network, and Ensemble approaches.
- As shown in Figure 1, model predictions are generally better for more severe outcomes like days in ICU
- COVID-19 status predictions were not better than chance (AUROC = 0.516)



Phase 2 Results – Assessing Synthetically Generated Datasets

- Compared prediction accuracy of models using Synthea and Generative Adversarial Network (GAN) generated synthetic datasets.
  - Submissions trained and tested on GAN-generated data scored significantly higher in predicting COVID-19 status
  - Model performance was similar on GAN and Synthea generated health data for all other outcomes
  - Both had strongest performance against more severe COVID-19 outcomes

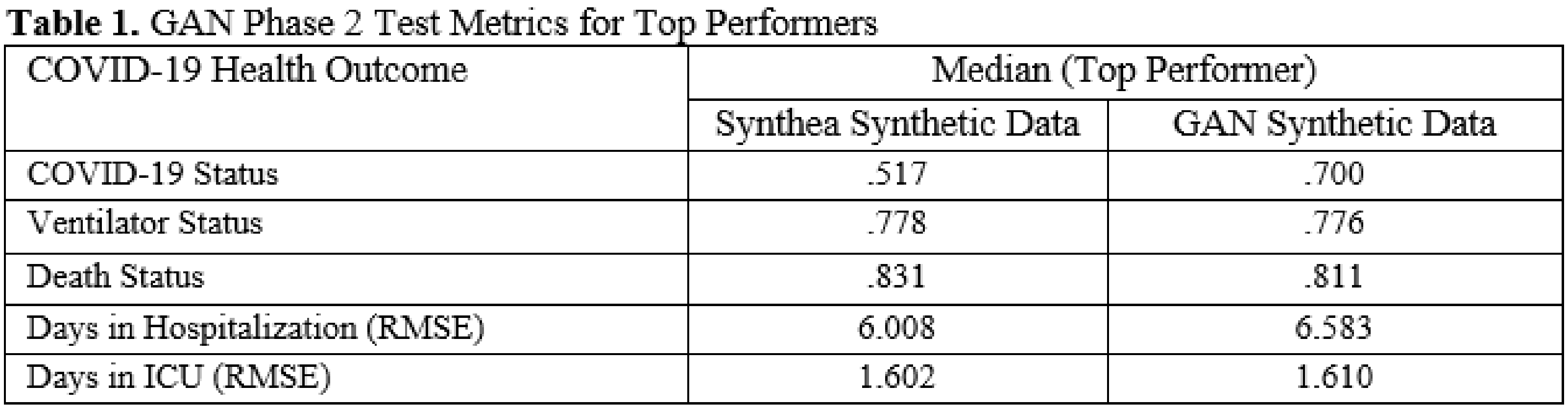


Table 1. GAN Phase 2 Test Metrics for Top Performers

COVID-19 Health Outcome	Median (Top Performer)	
	Synthea Synthetic Data	GAN Synthetic Data
COVID-19 Status	.517	.700
Ventilator Status	.778	.776
Death Status	.831	.811
Days in Hospitalization (RMSE)	6.008	6.583
Days in ICU (RMSE)	1.602	1.610

COVID-19 Risk Factor Modeling Challenge: Lessons Learned and Next Steps

What did we learn?

- Participant models performed better on patients with more severe outcomes (e.g., days in ICU versus days hospitalized)
- Top Phase 1 performer models highlighted age, smoking status, oxygen saturation, blood pressure and previous healthcare cost coverage as strong indicators of COVID-19 health outcomes
- Synthea synthetic data and GAN-generated data performed similarly, suggesting comparable efficacy

Next Steps

- Validate the top-performing models on de-identified Veteran data
- Explore methods to improve synthetic data quality
- Create a synthetic dataset to mimic VA data that non-VA researchers can access for modeling purposes