Title. Risk factors and prediction models of adverse outcomes in Veterans with COVID-19
PI: George N. Ioannou. Facility: Puget Sound (663)

Background and Significance.

The majority of patients infected with SARS-CoV-2 develop relatively mild, self-limiting, flu-like symptoms that do not require hospitalization. However, a small proportion develop severe viral pneumonia and hypoxemia, which may progress to respiratory failure, ARDS, cytokine release syndrome, multiorgan failure and death. These patients need to be hospitalized, frequently require admission to the intensive care unit and mechanical ventilation and have a high mortality.

The COVID-19 pandemic has been evolving so quickly that it is still unclear what proportion of infected patients experience adverse outcomes such as hospitalization, ICU admission, mechanical ventilation and death. When screening rates were very low and limited to patients with severe symptoms, hospitalization rates were reported to be high. These hospitalization rates have declined as screening programs expanded, identifying patients with milder disease or even asymptomatic contacts of infected patients. Early studies from China reported high mortality among confirmed cases (1023 out of 44672 or 2.3%)[1] and very high mortality among hospitalized patients (54 out of 191 patients or 28%)[2]. The mortality rate among US patients in general or VA patients in particular with SARS-CoV-2 infection is unknown.

It is also unclear why a small minority develop severe, life-threatening COVID-19 while the majority have mild or even asymptomatic presentation. Early published and anecdotal reports suggest multiple risk factors of adverse outcomes in COVID-19, including reversible factors and factors that appear to be related to COVID-19 pathophysiology. For example, hypertension, diabetes, older age, low albumin, low lymphocyte count, and high IL-6 and d-dimers were reported as risk factors for adverse outcomes including mortality[2-4], while the impact of medications such as angiotensin-converting enzyme inhibitors (ACEI) and angiotensin receptor blockers (ARB) remain controversial[5]. However, none of these risk factors have been confirmed in representative samples of US patients (such as US Veterans) and none of them are used for risk prediction or modification. It is important to cast a wide net in identifying such risk factors as we propose here, since COVID-19 is still such a novel disease with many unknowns, in addition to other studies pursuing specific “hypothesis-driven” associations. We propose to use national VA electronic data to extract the electronic health records of all VA patients who tested positive for SARS-CoV-2, determine the rate of adverse outcomes, identify independent predictors and develop innovative, state-of-the-art prediction models of adverse outcomes that can be used immediately in clinical practice.

SIGNIFICANCE: It is critical to identify the risk factors for and predict the development of adverse outcomes (hospitalization, ICU admission, mechanical ventilation, death) in patients with COVID-19 for three reasons.

- Some of these risk factors may be reversible or modifiable, such that eliminating them might be a strategy for reducing the mortality of COVID-19.
- Recognizing these risk factors, or, ideally combining them into accurate prediction models, will identify the high-risk patients early, focus limited resources on them, and target them for early interventions (e.g. close monitoring at home versus hospitalization, ICU admission or mechanical ventilation) in order to reduce their mortality.
- Prominent risk factors might provide clues as to the pathogenesis of severe, life-threatening COVID-19.

Specific Aims and Approach.

1. Identify all VA patients with SARS-CoV-2 infection and determine the risk of the following four adverse outcomes: a. Hospitalization, b. ICU admission, c. Mechanical ventilation, d. Death.

2. Identify independent risk factors for each of the four adverse outcomes, starting from a broad list of predictor variables, either previously described or novel.

3. Develop and internally validate prediction models to accurately estimate the risk of each of the four adverse outcomes using multivariable logistic regression as well as state-of-the-art machine learning models (ensemble tree-based, gradient-boosted algorithms, implemented by Extreme Gradient Boosting (XGBoost) in R.

Approach. Using national VA electronic health record data from VINCI-CDW, we will obtain the results of all nasopharyngeal or oropharyngeal swabs performed for SARS-CoV-2 PCR as of approximately June-July 2020 and identify positive (detected) versus negative results. Potential predictor variables will be extracted as of the

time of testing for SARS-CoV-2 (for risk factors/prediction of hospitalization) and also as of the time of hospitalization (for risk factors/prediction of ICU admission, mechanical ventilation or death). Predictor variables will be derived from electronic data, which we are confident we can extract from VINCI-CDW, as we have done in prior similar studies. We will identify the following potential predictors/risk factors, which are suspected to be associated with adverse outcomes in COVID-19, deliberately casting a wide net since little is known definitively yet:
Demographic factors: age, race, ethnicity, sex.
Vital signs: temperature, blood pressure, heart rate and oxygen saturation.
Geographic factors: state, county, rural/urban.
Comorbidities: diabetes, hypertension, COPD, asthma, history of solid organ transplantation, chronic kidney disease, cirrhosis, obesity (BMI), congestive heart failure, stroke, cardiovascular disease, HIV.
Viral coinfection: influenza A or B, respiratory syncytial virus, Bacterial coinfection,
Laboratory tests: CBC with differential, low lymphocyte count and high neutrophil to lymphocyte ratio (implicated in early studies as predictor), comprehensive metabolic panel [electrolytes and liver function tests], albumin (low albumin implicated as adverse predictor), BNP, PT/INR, CRP, pro-calcitonin, magnesium, troponin and IL-6[4] (if available, implicated as marker of cytokine storm).
Medications that may worsen outcomes: immunosuppressive, immunomodulators, steroids, and ACEI, ARB, NSAIDS and thiazolidinediones which may increase the ability of the virus to enter the cells by increasing the expression of ACE2 receptors.
Supportive medications: vasopressors, inotropes, antibiotics.
Medications potentially used to treat COVID-19: remdesivir (antiviral), tocilizumab (anti-IL-6), lopinavir/ritonavir (no benefit in early studies)[6], chloroquine/hydroxychloroquine, azithromycin, ribavirin, interferon-alpha.
We will also identify receipt of imaging studies (e.g. chest X-rays and CT scans), echocardiograms and EKGs, but their findings cannot be extracted directly from electronic records and will be considered for extraction in future studies.

LOGISTIC REGRESSION MODELS
We will develop multivariable logistic regression models for each of the four adverse outcomes to identify independent predictors among those listed above, with medical center as a random effect. Since there are many candidate predictor variables and not every one of them is necessarily related with the outcome conditional on the existence of other predictors, we will use the Least Absolute Shrinkage and Selection Operator (LASSO) procedure. LASSO simultaneously fits models and performs variable selection and regularization, thereby balancing the goals of predictive accuracy and parsimony.

GRADIENT BOOSTED MACHINE LEARNING MODELS
Additionally, we will develop gradient-boosted machine learning models, which might be expected to outperform the logistic regression models. Gradient-boosted models are some of the most widely used and best performing, state-of-the-art ML models and new high-performance implementations are an active area of research[7-9]. XGBoost models are particularly powerful when features are on different scales as is commonly found in electronic health records. Additionally, XGBoost models handle missing data and do not require transforming or categorizing of continuous measures. Tree-based models are prone to overfitting, so particular care will be applied to the tuning of the tree (depth, child weight, subsampling) and regularization (L1 and L2) parameters, aiming for conservatively tuning regularization parameters. Parameter tuning will be achieved through a 5-fold cross-validation over a grid of possible values. The gradient loss function will be chosen to account for the imbalance in groups (i.e. only ~2% experience death within 30 days). Loss functions based on precision and recall (e.g. F1 score, Brier score, and precision-recall area-under-the curve) instead of accuracy will be considered. When multiple values are found, the one with the more conservative regularization value will be chosen.
Internal validation of selected models will be performed by bootstrapping in order to more efficiently use the data.

Research Team and Relevant Experience

The PI, Dr. George Ioannou, is the Director of Hepatology at VAPSHCS and Professor of Medicine at UW. He has had continuous CSR&D funding since 2010 performing a large number of high-impact, epidemiological studies in hepatitis C virus using national VA data from CDW-VINCI. He is an expert in extracting relevant analytic variables from CDW, investigating independent risk factors for clinically meaningful outcomes and developing clinical prediction tools that are being used in the VA in clinical practice (such as the HCC Risk Calculator). His research has led to the development and national VA implementation of healthcare tools and

clinical strategies for HCC screening or preventing HBV reactivation. The Data Analyst (Pamela Green, PhD) and Biostatistician (Kristin Berry, PhD) have worked with Dr. Ioannou on national VA data analyses for more than 10 years with >50 relevant publications, and already have access to VINCI-CDW data required to complete the study very quickly. Vincent Fan (MD, MPH), co-investigator, is an expert pulmonologist and intensivist with first-hand experience managing COVID-19 patients. He is a funded VA investigator with a track record of health services research and implementations. Javeed Shah (MD), co-investigator, is the Section Chief of Infectious Diseases at VAPSHCS and will provide expertise in SARS-CoV-2 virology to inform our selection and modeling of risk factors. Wesley Van Voorhis (MD), consultant, former Chief Division of Allergy and Infectious Diseases, UW, is an expert Virologist actively involved in the development of treatments and serological testing for SARS-CoV-2. He will advise us on the optimal selection and modeling of risk factors. Emily Locke, research coordinator, has been at VAPSHCS for many years working with the study co-investigators. She will facilitate the necessary rapid regulatory approvals and coordination between investigators.

Data Sources

We have been able to identify using national VA electronic health records all nasopharyngeal swabs tested for COVID-19 PCR in the VA nationally as of 3/22/2020 (n=2560) and among them those that tested positive (n=251) as proof of feasibility. We confirmed that in a single facility the number of positive patients that we identified was exactly the same as the number reported by the facility, suggesting high accuracy of case identification. Given current exponential growth patterns, the number of positive tests would be expected to exceed 5,000 by June-July, which is when we propose to extract data for analysis. We propose to identify all SARS-CoV-2 tests performed in the VA, identify positive cases and extract their risk factors listed above using VINCI-CDW data, with which we are very familiar having completed >30 studies using these data.

We are aware that VINCI is working with other partners (NIH, CSP, PH Surveillance, PBM) to define a common set of data extracted from CDW (the "COVID-19 National Surveillance Tool" or COVID-NST) that would allow investigators to examine questions of risk and outcomes, such as the ones we propose. We would be delighted to work with the COVID-NST if was available or to contribute to its creation and have expressed this interest to Dr. Nicholas Smith, Director of the Seattle CSP ERIC.

Timeline: Gantt Chart

We anticipate the study will be approved by the R&D and IRD committees at VAPSHCS in an expedited manner as minimal risk, since it involves only secondary analysis of existing VA electronic data.

	STUDY MONTH								
	1	2	3	4	5	6	7	8	9
Obtain IRB and R&D Approvals									
Identify all VA patients with SARS-CoV-2 test and confirm those positive									
Extract analytic variables on baseline risk factors and predictor									
Extract analytic variables of the 4 outcomes: hospitalization, ICU, ventilation, death									
Develop logistic regression and XGBoost Prediction Models									
Publish, Disseminate, Implement Findings									

References

1. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention. JAMA 2020.
2. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X, Guan L, Wei Y, Li H, Wu X, Xu J, Tu S, Zhang Y, Chen H, Cao B. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020;395:1054-1062.
3. Wu C, Chen X, Cai Y, Xia J, Zhou X, Xu S, Huang H, Zhang L, Zhou X, Du C, Zhang Y, Song J, Wang S, Chao Y, Yang Z, Xu J, Zhou X, Chen D, Xiong W, Xu L, Zhou F, Jiang J, Bai C, Zheng J, Song Y. Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. JAMA Intern Med 2020.
4. MerdRxiv perprint: Level of IL-6 predicts respiratory failure in hospitalized symptomatic COVID-19 patients. Tobias Herold III, Vindi Jurinovic, Chiara Arnreich, Johannes C Hellmuth, Michael Bergwelt-Baildon, Matthias Klein, Tobias Weinberger. Available at https://www.medrxiv.org/content/10.1101/2020.04.01.20047381v1 Last accessed 04/06/2020.
5. Patel AB, Verma A. COVID-19 and Angiotensin-Converting Enzyme Inhibitors and Angiotensin Receptor Blockers: What Is the Evidence? JAMA 2020.
6. Cao B, Wang Y, Wen D, Liu W, Wang J, Fan G, Ruan L, Song B, Cai Y, Wei M, Li X, Xia J, Chen N, Xiang J, Yu T, Bai T, Xie X, Zhang L, Li C, Yuan Y, Chen H, Li H, Huang H, Tu S, Gong F, Liu Y, Wei Y, Dong C, Zhou F, Gu X, Xu J, Liu Z, Zhang Y, Li H, Shang L, Wang K, Li K, Zhou X, Dong X, Qu Z, Lu S, Hu X, Ruan S, Luo S, Wu J, Peng L, Cheng F, Pan L, Zou J, Jia C, Wang J, Liu X, Wang S, Wu X, Ge Q, He J, Zhan H, Qiu F, Guo L, Huang C, Jaki T, Hayden FG, Horby PW, Zhang D, Wang C. A Trial of Lopinavir-Ritonavir in Adults Hospitalized with Severe Covid-19. N Engl J Med 2020.
7. Tianqi Chen and Carlos Guestrin. “XGBoost: A scalable tree boosting system”. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2016, pp. 785–794.
8. Guolin Ke et al. “Lightgbm: A highly efficient gradient boosting decision tree”. In: Advances in Neural Information Processing Systems. 2017, pp. 3146–3154.
9. Liudmila Prokhorenkova et al. “CatBoost: unbiased boosting with categorical features”. In: Advances in Neural Information Processing Systems. 2018, pp. 6637–6647.

Budget

The total budget for this 9-month proposal is $100,000

SUMMARY BUDGET WORKSHEET - Single Site ver 6.30.17

Expense Category	Budget Period 1	Budget Period 2	Budget Period 3	Budget Period 4	Budget Period 5		
Primary Site:	VA Puget Sound						
VA PI ONLY (Section A, top line)						Yr 1 Effort (Cal Mo)	Degree
PI Salary	donated	0	0	0	0	2	MD
PI Fringe	donated	0	0	0	0		
Other VA Personnel (include Sr/Key VA personnel, exclude IPAs)						Total Yr 1 Effort (Cal Mo)	# unique staff
Hired - Salary	73,507	0	0	0	0	9	5
Hired - Fringe	26,094	0	0	0	0		
TBH - Salary		0	0	0	0		
TBH - Fringe		0	0	0	0		
Total Other VA Personnel Salary	73,507	0	0	0	0	9	5
Fringe (Section B, last line)	26,094	0	0	0	0		
Total Personnel (Total Salary, Wages and Fringe Benefits A+B)	99,601	0	0	0	0		
Equipment/Start-up (total-do not itemize)	0	0	0	0	0		
Travel	0	0	0	0	0		
Other Direct Costs (List any subcategory over $5000, lump remainder)						Total Yr 1 Effort (Cal Mo)	# unique persons
IPAs		0	0	0	0		
Consultants		0	0	0	0		
add…		0	0	0	0		
add…		0	0	0	0		
add…		0	0	0	0		
add…		0	0	0	0		
Other direct costs		0	0	0	0		
Subtotal Other Direct	0	0	0	0	0		
Subtotal Non-Personnel (Equipment, Travel, Other Direct) **(Section F, Line 8)**	0	0	0	0	0		
Total Project	**99,061**	**0**	**0**	**0**	**0**		**99,061**

BUDGET JUSTIFICATION

A. SENIOR/KEY PERSONNEL

George N. Ioannou, MD, MS (Epi) Principal Investigator (GS 15/1, 8/8th VA: 2 CM, YR 1, no salary requested) will serve as the Principal Investigator of the project. Dr. Ioannou is the Director of Hepatology at VAPSHCS and a Professor of Medicine at the University of Washington (UW). He is co-Director of the VA's VISN 20 "Hepatic Innovation Team" (HIT). He is the Chair of the Practice Guidelines Committee of the American Association for the Study of Liver Diseases (AASLD).

Dr. Ioannou is an expert in the analysis of national VA data to address liver-related outcomes, including antiviral treatment for hepatitis C infection, cirrhosis and hepatocellular carcinoma. He is also an expert in the development of clinically useful prediction tools, such as prediction models for the development of hepatocellular carcinoma in patients with cirrhosis or hepatitis C infection (www.hccrisk.com). He is an experienced clinical hepatologist. He has had prior experience leading VA REAP and Merit Review programs as well as NIH R01 programs.

As a VA staff physician, he is not requesting salary support.

Vincent Fan, Co-Investigator (GS 15/1, 8/8th VA: 2 CM, YR 1, no salary requested). Dr. Fan is a Staff Physician at VAPSHCS and an Associate Professor of Medicine at the University of Washington in the Division of Pulmonary and Critical Care. He is an expert pulmonologist and intensivist with first-hand experience managing COVID-19 patients. He is a funded VA investigator with a track record of health services research and implementations.

As a VA staff physician, he is not requesting salary support.

Kristin Berry (Wyatt), PhD, Co-investigator (GS 13/3, 8/8th VA: 3 CM $26,651 salary, $11,726 fringe) is a **Biostatistician** at the VAPSHCS, who has been working very closely with Dr. Ioannou over the last 10 years co-authoring 26 original research papers, including many with analytic aims similar to the ones described in the current proposal using very similar national VA data. She will work with Dr. Ioannou to create the logistic regression and gradient-boosted machine learning models for adverse outcomes in patients with SARS-CoV-2 infection. She has previously developed our HCC risk prediction models using similar VA data (www.hccrisk.com). She will work closely with our analyst, Dr. Green, who will be extracting the analytic variables from VINCI-CDW data.

Javeed Shah, MD, PhD, Co-investigator (GS 15/1, 8/8th VA: 0.5 CM, YR 1, no salary requested). He is the Section Chief of Infectious Diseases at VAPSHCS. He will provide expertise in SARS-CoV-2 virology and risk factors.

As a VA staff physician, he is not requesting salary support.

Wesley Van Voorhis, MD, Consultant. He is Professor and former Chief Division of Allergy and Infectious Diseases at the University of Washington (UW). He is actively involved in the development of treatments and serological testing for SARS-CoV-2 at the UW. He will be consulted on the choice and modeling of predictors and risk factors based on the most-up-to-date information available.

B. OTHER PERSONNEL

Pamela Green, PhD, MPH, Analyst (GS 12/6, 8/8th VA: 3 CM, $23,708/yr salary, $7,586/yr fringe Year 1) is an extremely experienced analyst who has been working with Dr. Ioannou on national VA VINCI-CDW data for the last 10 years co-authoring 25 original research papers. She will work under the supervision of Dr. Ioannou to extract analytic variables from VINCI-CDW. She will work closely with our biostatistician, Dr. Berry (Wyatt), to ensure correct interpretation of all analytic variables and incorporation into multivariable prediction models.

Emily Locke, MPH, Research Coordinator ((GS 11/1, 8/8th VA: 3 CM, $17,334 salary, $4680 fringe) is a Research Health Science Specialist at the VAPSHCS HSR&D COIN who has worked on VA research projects for the past 10 years. She will prepare and expedite all regulatory applications (IRB and R&D) working with Dr. Ioannou and she will coordinate all research activities.

C. EQUIPMENT

N/A

D. TRAVEL

N/A

E. OTHER DIRECT COSTS

N/A