Letter to the Editor

Testing-Related and Geo-Demographic Indicators Strongly Predict COVID-19 Deaths in the United States during March of 2020

James B. Hittner[1], Folorunso O. Fasina[2,#], Almira L. Hoogesteijn[3], Renata Piccinini[4], Dawid Maciorowski[5], Prakasha Kempaiah[5], Stephen D. Smith[6], and Ariel L. Rivas[7]

The COVID-19 pandemic has wreaked havoc around the globe and caused significant disruptions across multiple domains[1]. Moreover, different countries have been differentially impacted by COVID-19 — a phenomenon that is due to a multitude of complex and often interacting determinants[2]. Understanding such complexity and interacting factors requires both compelling theory and appropriate data analytic techniques. Regarding data analysis, one question that arises is how to analyze extremely non-normal data, such as those variables evidencing L-shaped distributions. A second question concerns the appropriate selection of a predictive modelling technique when the predictors derive from multiple domains (e.g., testing-related variables, population density), and both main effects and interactions are examined.

To address these questions, we propose a novel statistical approach for analyzing and understanding complex data interactions. Using data collected in the USA during the first month in which COVID-19 testing was performed (March of 2020 Supplementary Table S1 available in www.besjournal.com), we examined the following six predictors of COVID-19 related deaths: (i) the proportion of all tests conducted during the first week of testing; (ii) the cumulative number of (test-positive) cases through 3-31-2020; (iii) the number of tests performed/million inhabitants; (iv) the cumulative number of inhabitants tested; (v) the number of cases/million inhabitants (cases/mill inh); and (vi) the number of diagnostic tests performed in week one of testing/million inhabitants/state-specific population density (w1DT/MI/PD), where "population density" is defined as the number of inhabitants per square kilometer.

The purpose of this study was to examine the ability of the six variables to predict COVID-19 related deaths in the United States during March of 2020. We ran the predictive model twice, once for each dependent variable: mortality count (overall number of deaths), and deaths per million inhabitants. Because our model (a) uses predictors that leverage information from multiple domains, (b) captures both nationwide and state-specific dimensions, and (c) examines two different mortality-related outcomes, the results are expected to have relevance for policy-makers.

All data used in this study were obtained from three sources in the public domain: Worldometer (https://www.worldometers.info/coronavirus/), World Population Review (https://worldpopulationreview.com/states), and Covidtracking (https://covidtracking.com/). The data were processed and analyzed using IBM SPSS, Minitab, and *R*. Univariate skewness and kurtosis values indicated that all predictors and outcomes were non-normally distributed, with a few variables evidencing L-shaped distributions. The L-shaped variables were normalized using the rank-based inverse normal (RIN) transformation[3]. For extremely non-normal data, the RIN method is a highly effective normalizing transformation[3].

The prediction models were first examined using linear multiple regression, with the RIN-transformed versions of all variables used in the regressions. Because the homoscedasticity assumption (i.e.,

doi: 10.3967/bes2021.102

1. Department of Psychology, College of Charleston, Charleston, South Carolina, United States of America, 29424; 2. Food and Agriculture Organization, Dar es Salam, Tanzania & Department of Veterinary Tropical Diseases, University of Pretoria, South Africa, 0010; 3. Human Ecology, Centro de Investigación y de Estudios Avanzados (CINVESTAV), Mérida, México, 97133; 4. Department of Veterinary Medicine, University of Milan, Milan, Italy, 20122; 5. Loyola University Medical Center, Chicago, Illinois, United States of America, 60153; 6. Institute for Resource Information Science, College of Agriculture, Cornell University, Ithaca, United States of America, 14853; 7. Center for Global Health, Department of Internal Medicine, Medical School, University of New Mexico, Albuquerque, New Mexico, United States of America, 87131

constant variance of the predicted Y-values) was not met, we re-ran the prediction models using a non-parametric approach known as Kernel Regularized Least Squares (KRLS) Regression[4]. KRLS is an appropriate method to use when the assumptions of linear regression are not met and the precise functional forms between the predictors and outcomes are unknown. All KRLS regressions used the RIN-transformed variables and all analyses were performed using the KRLS package for *R.* The use of non-parametric, machine learning-based methods such as KRLS is consistent with recent calls to place greater reliance on artificial intelligence systems for understanding the causes and consequences of the COVID-19 pandemic[5].

The KRLS regression results are presented in Table 1. For number of deaths, the six predictors accounted for 98.8% of the variance. Five of the predictors were statistically significant (*P*-values ≤ 0.002). Two of the significant predictors (i.e., number of test-positive cases, Cohen's *d* = 2.3; and cases per million inhabitants, Cohen's *d* = 1.3) represent different ways of quantifying the illness burden due to SARS-CoV-2 infection. The ratio of the two *d* values indicated that the predictive strength of number of test-positive cases was 77% greater than was cases per million inhabitants. Regarding the second dependent variable, the six predictors accounted for 92.6% of the variance in deaths per million inhabitants. Five of the predictors were significant (*P*-values ≤ 0.03). For this regression analysis, the number of test-positive cases (*d* = 1.1) and cases per million inhabitants (*d* = 1.4) were similar in predictive strength.

In addition to number of test-positive cases and cases per million inhabitants, another interesting predictor was our geo-demographic variable (i.e., the number of diagnostic tests/million inhabitants/ population density performed in week one of testing, or w1DT/MI/PD). This predictor was significantly associated with both dependent variables. Because w1DT/MI/PD is a complex, ratio-based predictor, discerning the precise nature of its predictive association from a single regression

Table 1. KRLS regression of potential predictors of COVID-19 related mortality

Items	**Estimate**	**Std. Error**	***t* value**	***P*-value**
Predictors of number of deaths				
Totaltests RIN	0.111	0.033	3.326	0.002
Testedpermil RIN	−0.153	0.026	−5.782	< 0.001
Wkonepropalltests RIN	0.044	0.030	1.452	0.153
Wkonepermilcitperpopden RIN	0.169	0.032	5.262	< 0.001
Confircases RIN	0.568	0.035	16.340	< 0.001
Casespermil RIN	0.215	0.023	9.185	< 0.001
Predictors of deaths per million inhabitants				
Totaltests RIN	−0.138	0.058	−2.352	0.023
Testedpermil RIN	0.004	0.048	0.091	0.928
Wkonepropalltests RIN	0.136	0.061	2.234	0.031
Wkonepermilcitperpopden RIN	0.161	0.063	2.570	0.014
Confircases RIN	0.408	0.055	7.353	< 0.001
Casespermil RIN	0.441	0.045	9.748	< 0.001

Note. All predictors were normalized using the rank-based inverse normal (RIN) transformation. Estimates are sample-average partial derivatives. The set of predictors accounted for 98.8% of the variance in number of deaths (R^2 = 0.9875). For deaths per million citizens, the predictors accounted for 92.6% of the variance (R^2 = 0.9264). Description of predictors: totaltests = number of tests performed in March of 2020; testedpermil = number of all tests conducted per million inhabitants, in March of 2020; wkonepropalltests = all tests conducted during the first week of testing, expressed as the percentage of all tests performed in March 2020; wkonepermilcitperpopden = the number of tests performed during week one per million inhabitants, divided by state-specific population density; confircases = total number of test-positive individuals, in March of 2020; casespermil = number of test-positive individuals per million inhabitants, in March of 2020.

estimate alone is challenging. To further enhance the interpretation of this variable, we created two scatterplots showing the association between w1DT/MI/PD and each dependent variable. Both scatterplots include a best fitting linear regression line and a *lowess* line (with accompanying 95% confidence interval). *Lowess* stands for locally weighted scatterplot smoothing. The *lowess* line is the best fitting non-linear curve that tracks the data points in the scatterplot. The *lowess* curves allow us to make inferences about COVID-19 related deaths at low and high levels of w1DT/MI/PD. Such inferences are tantamount to examining COVID-19 related deaths for U.S. states scoring low versus high on the geo-demographic predictor variable. The scatterplots were created using the *car* package for *R.*

As the *lowess* curve in the top panel of Figure 1 indicates, at higher and medium levels of w1DT/MI/PD, the association between the geo-demographic predictor and death count was strongly negative and moderately negative, respectively. In contrast, at lower levels of w1DT/MI/PD, there was little if any association between the geo-demographic variable and number of fatalities. The bottom panel of Figure 1 indicates that at lower levels of w1DT/MI/PD, the association between the geo-demographic variable and deaths per million inhabitants was moderately positive. At medium levels of w1DT/MI/PD, there was little if any association between the two variables. Finally, at higher levels of w1DT/MI/PD, there was a moderately strong negative association between the geo-demographic variable and deaths per million inhabitants.

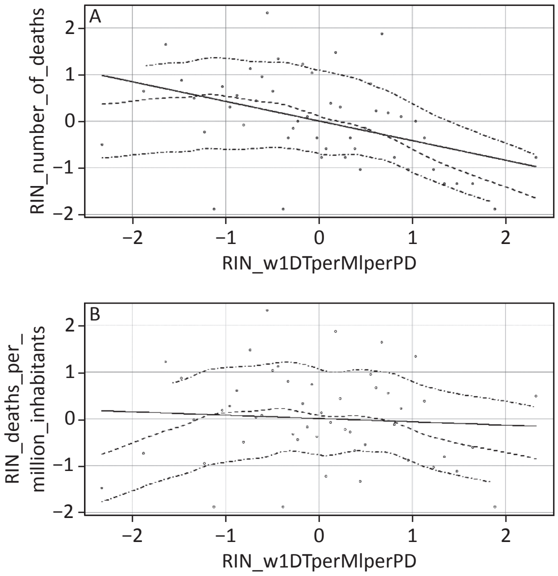

Figure 1. Scatterplots depicting *lowess* curves (the middle dashed lines) and accompanying 95% confidence intervals (top and bottom dashed lines) for the association between number of tests during week 1/million inhabitants/population density and (A) number of COVID-19 related deaths (top panel) and (B) number of COVID-19 related deaths per million inhabitants (bottom panel). All variables were normalized using the rank-based inverse normal (RIN) transformation.

In constructing our geo-demographic predictor variable, we controlled for population density because it is an important factor associated with disease transmission[6]. Moreover, because there typically is a lag time of several weeks or more between being infected with SARS-CoV-2 and showing disease-related symptoms, the association between population density and disease-related deaths should strengthen over time. To highlight this point, Figure 2 presents scatterplots showing the Pearson correlations between population density and cumulative COVID-19 related deaths per million inhabitants through March 31st and June 17th, 2020, respectively. The correlations were as follow: March 31st (r = 0.228, P > 0.05); June 17th (r = 0.800, P < 0.01). The difference between the two statistically dependent correlations was evaluated using Hittner, May and Silver's modification of Dunn and Clark's z test[7]. The two correlations were significantly different (z = 5.85, P < 0.0001), thereby supporting the prediction that the association between population density and COVID-19 related deaths will strengthen over time.

To the best of our knowledge, this is the first study that examines testing-, case count- and geo-demographic variables as predictors of COVID-19 related deaths. Using a flexible, machine learning-based approach (KRLS regression), we found that our predictors accounted for very high percentages of outcome variance (98.8% and 92.6% for number of deaths and deaths per million inhabitants, respectively). Furthermore, with very few exceptions, our predictors were both statistically significant and practically important.

One novel contribution of this study was our examination of a complex, ratio-based geo-demographic predictor variable. This variable—the number of diagnostic tests performed in week one of testing/million inhabitants/state-specific population density (w1DT/MI/PD)—significantly predicted COVID-19 related deaths, but did so differently depending on where, along the continuum of geo-demographic values, the predictive association was examined. At the lower end of the geo-demographic predictor, more tests during week one per million inhabitants, normalized by population density, were associated with more deaths per million citizens. In contrast, at the higher end of the geo-demographic predictor, more tests during week one per million inhabitants, normalized by population density, were associated with fewer deaths per million inhabitants. These different quantitative patterns could reflect different qualitative situations. In the first case (lower values on the geo-demographic variable, where more tests are associated with more deaths), testing seems to pursue a *confirmatory* purpose. In contrast, for the second case (higher values on the geo-demographic variable, where more tests are associated with fewer deaths), *diagnostic* testing appears to be emphasized[8]. One implication of these findings is that when examining our geo-demographic variable as a predictor of deaths, the inflection points along the *lowess* curves (the positions where the slope rises and falls) can serve as approximate cut-points demarcating three types of testing: confirmatory, diagnostic, and other.

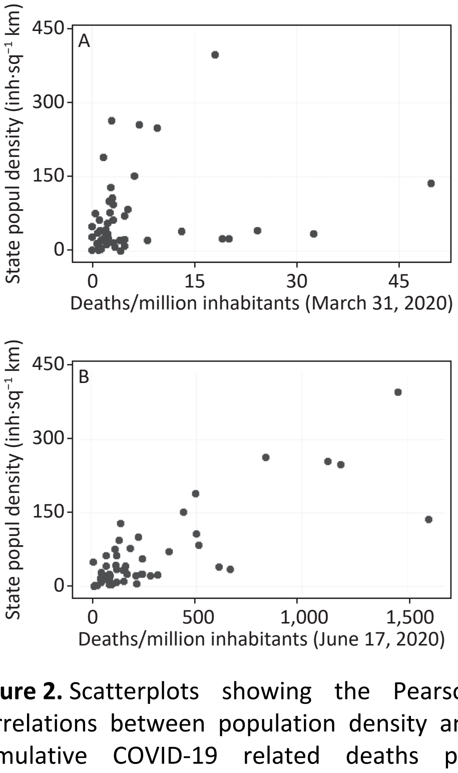

Figure 2. Scatterplots showing the Pearson correlations between population density and cumulative COVID-19 related deaths per million inhabitants through (A) March 31, 2020, top panel (r = 0.228, 95% *CI*: −0.054, 0.476) and (B) June 17, 2020, bottom panel (r = 0.80, 95% *CI*: 0.671, 0.882).

When testing prioritizes symptomatic cases, it is expected that most tested individuals will result in positive results (infection will be confirmed). Because deaths will occur within a subset of infected individuals, when testing is confirmatory (when only symptomatic patients are tested), more tests will be associated with more deaths. In contrast, when asymptomatic individuals are also tested, more tests, conducted earlier, will allow clinicians to detect, treat, and isolate infections earlier and prevent further viral dissemination which, in turn, will result in fewer deaths/million inhabitants. Our findings thus support an important recommendation from the World Health Organization, which is that early and frequent testing helps to prevent deaths[9].

In addition to the contributions described above, we performed supplemental analyses examining the association between population density and COVID-19 related deaths. The role of population density in predicting epidemic dispersal and epidemic-related deaths is receiving increased research attention[10]. To the best of our knowledge, the present study is the first to demonstrate that the magnitude of association between population density and COVID-19 related deaths *strengthens* as the time since first infection increases. Understanding how factors such as testing frequency, the relative proportion of confirmatory versus diagnostic testing, and sociodemographic composition influence the temporal association between population density and COVID-19 related deaths is an important priority for future research.

Overall, our findings highlight the importance of considering predictor variables from multiple domains. When ratio-based predictors such as our geo-demographic variable are analyzed, we recommend examining *lowess* curves as a visual interpretational aid for explicating the (often) complex non-linear associations between such ratio-based predictors and various outcomes of interest. An important direction for future research on epidemic dissemination and potential control is to examine both ratio-based composite variables—such as our geo-demographic measure—and traditional multiplicative interaction terms (created as linear products of two or more variables). The joint examination of both types of complex variables might result in greater predictive power and/or might foster additional insights into the dynamics of infectious diseases, such as COVID-19.

This work was previously released as a preprint by J. B. Hittner, F. O. Fasina, A. L. Hoogesteijn, R. Piccinini, P. Kempaiah, S. D. Smith, and A. L. Rivas, with the title '*Early and massive testing saves lives: COVID-19 related infections and deaths in the United States during March of 2020*' medRxiv 2020.05.14.20102483; https://doi.org/10.1101/2020.05.14.20102483.

Acknowledgements The authors appreciate the data gathering efforts of those citizens who contributed to COVID-19 tracking (https://covidtracking.com). FOF is currently funded by the United States Agency for International Development (USAID) grant to the Food and Agriculture Organization of the United Nations, (Global Health Security Agenda- Zoonotic Diseases and Animal Health in Africa). The views and opinions expressed in this paper are those of the authors and not necessarily the views and opinions of the United States Agency for International Development and the Food and Agriculture Organization of the United Nations.

Financial Support This research received no specific grant from any funding agency, commercial or not-for-profit entity.

Conflict of Interest The authors declare that they have no conflict of interest.

Author Contributions JBH and ALR designed the study. DM curated the original data. ALH, RP, PK, SDS, JBH, and FOF reviewed the literature and extracted and filtered the available data from online repositories. JBH conducted the statistical analyses. All authors contributed to writing of the report.

[#]Correspondence should be addressed to Folorunso O. Fasina, E-mail: Folorunso.fasina@fao.org, Tel: 255-686-132-852.

Biographical note of the first author: James B. Hittner, male, born in 1965, PhD Degree, majoring in clinical and applied psychology, risky behavior, statistical methodology, and infectious disease dynamics.

Received: August 10, 2020;

Accepted: January 18, 2021

REFERENCES

1. Haleem A, Javaid M, Vaishya R. Effects of COVID-19 pandemic in daily life. Curr Med Res Prac, 2020; 10, 78−9.
2. Holmes EA, O'Connor RC, Perry VH, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. Lancet Psychiatry, 2020; 7, 547−60.
3. Bishara AJ, Hittner JB. Testing the significance of a correlation with nonnormal data: comparison of Pearson, spearman, transformation, and resampling approaches. Psychol Methods, 2012; 17, 399−417.
4. Hainmueller J, Hazlett C. Kernel Regularized Least Squares: reducing misspecification bias with a flexible and interpretable machine learning approach. Pol Anal, 2014; 22, 143−68.
5. Vaishya R, Javaid M, Haleem Khan I, et al. Artificial intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr: Clin Res Rev, 2020; 14, 337−9.
6. Rivas AL, Fasina FO, Hoogesteyn AL, et al. Connecting network properties of rapidly disseminating epizoonotics. PLoS One, 2012; 7, e39778.
7. Hittner JB, May K, Silver NC. A Monte Carlo evaluation of tests for comparing dependent correlations. J Gen Psychol, 2003; 130, 149−68.
8. Padula WV. Why only test symptomatic patients? Consider random screening for COVID-19. Appl Health Econ Health Policy, 2020; 18, 333−4.
9. World Health Organization. COVID 19: Public Health Emergency of International Concern (PHEIC). Global research and innovation forum: towards a research roadmap. https://www.who.int/blueprint/priority-diseases/key-action/Global_Research_Forum_FINAL_VERSION_for_web_14_feb_2020.pdf?ua=1. [2020-04-04].
10. Rocklöv J, Sjödin H. High population densities catalyse the spread of COVID-19. J Travel Med, 2020; 27, taaa038.