Journal of Data Science 19 (1), 142–150
January 2021
DOI: 10.6339/21-JDS1005
Data Science in Action

Five Critical Genes Related to Seven COVID-19 Subtypes: A Data Science Discovery

Zhengjun Zhang[1,*]

[1]*Department of Statistics, University of Wisconsin, Madison, WI 53706, USA*

Abstract

Since the first confirmed case of COVID-19 was identified in December 2019, the total COVID-19 patients are up to 80,675,745, and the number of deaths is 1,764,185 as of December 27, 2020. The problem is that researchers are still learning about it, and new variants of SARS-CoV-2 are not stopping. For medical treatment, essential and informative genes can lead to accurate tests of whether an individual has contracted COVID-19 and help develop highly efficient vaccines, antiviral drugs, and treatments. As a result, identifying critical genes related to COVID-19 has been an urgent task for medical researchers. We conducted a competing risk analysis using the max-linear logistic regression model to analyze 126 blood samples from COVID-19-positive and COVID-19-negative patients. Our research led to a competing COVID-19 risk classifier derived from 19,472 genes and their differential expression values. The final classifier model only involves five critical genes, ABCB6, KIAA1614, MND1, SMG1, RIPK3, which led to 100% sensitivity and 100% specificity of the 126 samples. Given their 100% accuracy in predicting COVID-19 positive or negative status, these five genes can be critical in developing proper, focused, and accurate COVID-19 testing procedures, guiding the second-generation vaccine development, studying antiviral drugs and treatments. It is expected that these five genes can motivate numerous new COVID-19 researches.

Keywords *classification; competing risk; COVID-19 test; COVID-19 treatment; COVID-19 vaccine; gene-gene interaction*

1 Introduction

COVID-19 pandemic is a serious global health threat. Its impact on the whole world is tremendous. Many countries have put significant efforts and measures in preventing its spread, developing test procedures, vaccines, antiviral drugs, and treatments to battle this super severe disease. However, new variants of SARS-CoV-2 are not stopping. It is still unknown when the COVID-19 pandemic can be controlled, and the people's life can go back to normal.

There have been many research results published. These results help researchers, administrators, and ordinary people understand COVID-19 better. At the early stage, finding the origins of SARS-CoV-2 was fundamental for researchers to know how the virus was developed and spread. For example, in the genomic characterization and epidemiology of COVID-19, research results include the study of implications for virus origins and receptor binding (Lu et al., 2020), the proximal origin of SARS-CoV-2 (Andersen et al., 2020), and decoding the evolution and transmissions from the animal using whole genomic data (Yu et al., 2020), amongst many others.

The pharmaceutical industry has responded swiftly to the COVID-19 pandemic. Testing procedures and kits have been developed and widely applied. Potentially effective antiviral drugs

* Email: zjz@stat.wisc.edu.

Received December, 2020; Accepted January, 2021

have been studied in large-scale randomized controlled clinical trials. Most recently, randomized controlled vaccine trials have brought good news and hope to the public, and several countries have started vaccine processes.

For controlling the fast spread of COVID-19, rapid and accurate testing methods are in demand. On the one hand, rapid antigen tests are designed to tell whether or not someone has contracted the disease in a few minutes. A natural question will be whether or not they are accurate (Guglielmi, 2020). On the other hand, a chest x-ray radiograph (CXR) may be more reliable. For radiologists to differentiate SARS-CoV-2 infected pneumonia from different known pneumonia types on CXR, a trained deep neural network, CV19-Net, is introduced (Zhang et al., 2020). The performance of CV19-Net exceeds that of experienced thoracic radiologists. However, these tests' accuracy is about 80%, with some tests being better and up to 90% of accuracy.

In terms of potential effective antiviral drugs, published COVID-19 studies have shown no clear evidence of the clinical benefits of using antiviral drugs to treat patients (The-RECOVERY-Collaborative-Group, 2020). Recent work suggests that hydroxychloroquine can benefit some groups of people through a relative treatment effect design by Teng and Zhang (2020) which is a generalization of a proportional covariate effect model in Xie et al. (2019).

Combing through the genome, researchers have tied COVID-19 to some genes associated with the immune system's response. For example, genome-wide association analysis may allow for identifying potential genetic factors involved in the development of COVID-19 (The-Severe-Covid-19-GWAS-Group, 2020). A large-scale multi-omic study of COVID-19 severity illuminated the unique COVID-19 phenotype, and systems analysis revealed strong biomolecule associations with COVID-19 status and severity (Overmyer et al., 2020). Upper airway gene expression differentiated COVID-19 from other acute respiratory illnesses and showed the suppression of innate immune responses by SARS-CoV-2 (Mick et al., 2020), amongst much other research work.

Still, many uncertainties remain in testing procedures, vaccine, and antiviral drug developments (Rowland, 2020). The reported genes in the published work are positively associated with some parts of the immune system. However, it is not clear whether they are critical, i.e., they may be close to COVID-19 but may not be the actual cause of COVID-19. On the other hand, the number of reported genes is not small. As a result, these genes' inter-relationships will make the inference difficult and may mislead to wrong directions of vaccine and antiviral drug developments.

2 Statistical Methodology

The max-linear competing factor models (Cui and Zhang, 2018), the max-linear regression models (Cui et al., 2020), and the max-linear logistic models (Xu, 2019) have an advantage over existing models in a large class of research problems, e.g., nonlinear predictions and classifications. These models are different from the random forest, support vector machine, group lasso based machine learning methods, and deep learning methods. Max-linear models are not only interpretable but also outperform existing methods. In the literature, theoretical statistical and probabilistic foundations related to the competing risk factor models have been established (Cui et al., 2020; Cui and Zhang, 2018; Malinowski et al., 2016; Xu, 2019; Cao and Zhang, 2020; Zhang, 2005, 2020). The difference between the max-linear logistic regression and the classical logistic regression is that the original linear combination of predictors is replaced by the maximum of several linear combinations of predictors, called competing factors or competing-risk factors. We apply max-linear logistic regression in this study.

Suppose $(Y_i,\ X_i),\ i = 1, \ldots, n$, are the characteristics of n persons with Y_i corresponding to the ith individual' infected status ($Y_i = 0$ for not infected, $Y_i = 1$ for infected) and $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ being the gene expression values with $p = 19472$ in this study. Using a logit link (or probit link, Gumbel link), we can model the risk probability p_i of the ith person's infection status as:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + X_i\beta \tag{1}$$

or alternatively, we write

$$p_i = \frac{\exp(\beta_0 + X_i\beta)}{1 + \exp(\beta_0 + X_i\beta)}$$

where β_0 is an intercept, X_i is a $1 \times p$ observed vector, and β is a $p \times 1$ coefficient vector which characterizes the contribution of each predictor (gene in this study) to the risk. Given the ultra-high dimension of the predictor vector in many gene related studies, parameter penalization methods have to be implemented. This work is not going to discuss penalization, and the related work is referred to the most recent monograph by Fan et al. (2020).

There are at least three major problems applying the classical logistic classifier (1) to disease classifications. The first is that the number of genes selected is still not small. As a result, gene-gene interactions can hardly be interpretable, and hence the selected genes can not be directly used in drug development and treatment design. The second is that the classical logistic classifier cannot provide additional information about how genes interact with different disease subtypes. The third is that even with a relative non-small number of genes in the classical logistic classifier, the accuracy is not high enough, often just 80%.

There is one crucial factor, competing (risk) factors, that has not been considered in many existing statistical models, i.e., the existing classifiers do not distinguish the causes and the subtypes of the disease. In scientific studies, competing factors exist in many scenarios (Malinowski et al., 2016). The cause/regulation of each subtype of the disease can be different, i.e., each subtype of the disease can result from one factor or multiple factors. For example, in a system, e.g., a human body, all parts compete for resources to succeed. In terms of diseases (rare or non-rare), all subtype diseases also compete for resources. The dominant one wins all and will be diagnosed first. This study considers competing factors to be linear combinations of a set of predictors.

Suppose a disease (e.g., COVID-19) may be related to G groups of genes

$$\Phi_{ij} = (X_{i,j_1}, X_{i,j_2}, \ldots, X_{i,j_{g_j}}),\ j = 1, \ldots, G,\ g_j \geqslant 0 \tag{2}$$

where i is the ith individual in the sample, g_j is the number of genes in jth group. The competing (risk) factor classifier is defined as

$$\log\left(\frac{p_i}{1-p_i}\right) = \max(\beta_{01} + \Phi_{i1}\beta_1,\ \beta_{02} + \Phi_{i2}\beta_2, \ldots, \beta_{0G} + \Phi_{iG}\beta_G) \tag{3}$$

where β_{0j}'s are intercepts, Φ_{ij} is a $1 \times g_j$ observed vector, β_j is a $g_j \times 1$ coefficient vector which characterizes the contribution of each predictor in the jth group to the risk.

Remark 1. In the definition of the competing risk factor classifier, the number (G) of competing factors is unknown, predictors (genes) in each competing factor are unknown. They may be solved simultaneously using penalization approach together with conditional likelihood (or composite likelihood) method. We refer the readers to Xu (2019) for theoretical results.

Remark 2. In this study, we will set $G = 3$. In each competing factor, we set the number of genes to be three to have the genes in the competing factors interpretable and to avoid computational complexity.

Remark 3. After the final model is fitted, each $\beta_{0j} + \Phi_{ij}\beta_j$ can be used as a classifier and the risk probabilities can be computed using Equation (1).

In practice, we have to choose a threshold probability value to decide a patient's class label. Following the general trend in the literature, we set the threshold to be 0.5. As such, if $p_i \leqslant 0.5$, the ith individual is classified as disease free, otherwise the individual is classified to have the disease.

The remaining problem is to identify the genes in each competing factor, which can be implemented in the following optimization problem:

$$(\hat{\beta}, \hat{S}) = \arg\min_{\beta,\ S_j \subset S,\ j=1,2,\ldots,G} \sum_{i=1}^{n} \left(I(p_i \leqslant 0.5) I(Y_i = 1) + I(p_i > 0.5) I(Y_i = 0) \right) \tag{4}$$

where $I(.)$ is an indicate function, p_i is defined in Equation (3), $S = \{1, 2, \ldots, 19472\}$ is the index set of all genes, $S_j = \{j_{j1}, \ldots, j_{j,g_j}\}$, $j = 1, \ldots, G$ are index sets corresponding to (2), and $\hat{S} = \{j_{j1}, \ldots, j_{j,g_j},\ j = 1, \ldots, G\}$ is the final gene set selected in the final classifiers.

Remark 4. A perfect classifier (100% sensitivity and 100% specificity) will have $\sum_{i=1}^{n} \left(I(p_i \leqslant 0.5) I(Y_i = 1) + I(p_i > 0.5) I(Y_i = 0) \right) = 0$ in Equation (4), which is the case in our study.

Remark 5. We note that the optimization procedure in Equation (4) is different from existing approaches, e.g., likelihood method and composite likelihood. $\sum_{i=1}^{n} \left(I(p_i \leqslant 0.5) I(Y_i = 1) + I(p_i > 0.5) I(Y_i = 0) \right)$ is a newly introduced loss function in this study. It takes values 0 (the best), $1, 2, \ldots, n$ (the worst). The approach applied in this study is more like a machine learning approach. Nevertheless, the final competing risk classifier is interpretable and gene-gene interactions are expressed.

The optimization problem (4) is a combination of combinatorial optimization and continuous variable optimization. As a result, its algorithm complexity is extremely high. In this study, we adopt a simple approach to find some feasible solution. The following is the procedure.

1. Randomly draw three sets of genes with each set having three genes;
2. Use any optimization procedures (e.g., Nelder–Mead method, genetic algorithm, simulated annealing) to solve (4);
3. Repeat the above two steps until an acceptable solution is reached.

Remark 6. We have done an extensive Monte Carlo search to find our final competing classifier. A Matlab® demo code for solving Equation (4) is available online. However, we have experienced quite a few times man-machine interactions to reduce the dimensions from 19472 to 5. As such, we don't have a well-documented algorithm for solving Equation (4). It will be a future project as it is an algorithm problem, i.e., not a methodological problem. As the number of genes is big, the first step may not be efficient. Dimension reduction can be helpful. In our man-machine interactions, to train our program, we first allowed the loss function to take a value around 12, i.e., 10% of error rate. We recorded some sets of genes that performed better than other sets of genes, and to form a new set of genes, then repeated the above procedure to get the final classifier. We were able to find an optimal solution to have a loss function taking the value zero. The dimension reduction procedure we used is ad hoc. Other dimension reduction procedures may be useful and worthy of further investigation. Besides some well-documented dimension

reduction algorithms, we have used quotient correlation coefficients for dimension reduction in other projects (Zhang, 2008; Zhang and Ma X, 2011; Zhang et al., 2017).

Remark 7. Given that we used Monte Carlo method in this study, we have set a seed number (just the day we started the project) in our Matlab programs. The seed number can help, but not sure for final results as we had quite a few steps man-machine interactions, i.e., the seed number might not have an effect.

Remark 8. Given the objective function in Equation (4) is heavily flat (taking integer values), non-smooth, and non-convex, there may be multiple optimal solutions that exist. Our final solution is a global optimal. We have obtained some different sets of estimated coefficients, but the conclusions remain the same.

3 Data Descriptions, Results and Interpretations

The data used in this analysis are publicly available: Large-scale Multi-omic Analysis of COVID-19 Severity (Overmyer et al., 2020) Public on August 29, 2020, the link and the summary are https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157103. The experiment type is "Expression profiling by high throughput sequencing." One hundred twenty-six samples were analyzed in total, with 100 COVID-19 patients and 26 non-COVID-19. There are two types of datasets available. One type is TPM (Transcripts Per Million), while another type is expected counts. We used TPM data in this study.

The goal is to select a sparse (single digit) number of genes with high performance. We start with three competing factors in logistic regression models, with each factor having only three genes randomly drawing from 19472 genes. A Monte Carlo method with extensive computation is used to find the final model with the best performance of sensitivity and specificity and the smallest number of genes. We refer the details to Remark 6.

Using a probability higher than 50% as the threshold, we identify five critical genes (ABCB6, KIAA1614. MND1, SMG1, RIPK3), which lead to a 100% precision of classifying all 126 patients in their respective groups. Our final classifier is a combined classifier of three competing factor (CF) classifiers expressed as:

$$\begin{aligned}
&\text{CF1)} \quad -0.330340967+3.415275789 \times \text{KIAA1614}-0.124771579 \times \text{SMG1}+0.21769849 \times \text{MND1},\\
&\text{CF2)} \quad -0.737841461-0.462005922 \times \text{ABCB6}+0.0653607995 \times \text{SMG1}+0.909277249 \times \text{MND1},\\
&\text{CF3)} \quad 6.928277138 - 0.392092650 \times \text{RIPK3}. \qquad (5)
\end{aligned}$$

The final classifier CFmax is the maximum of (CF1, CF2, CF3), i.e., CFmax = max(CF1, CF2, CF3).

Table 1 (in an online supplementary file) lists all of the final selected genes and their differentiated expression values from 126 plasma and leukocyte samples. Columns CF1–3 are computed from the formulas listed above. Predictive probabilities (Columns P1, P2, P3, Pmax) are computed using Columns of classifiers (CF1, CF2, CF3), and CFmax (the combined max-classifier), respectively using the classical logistic regression function.

In Table 1 in the #ID column, C1-103 stands for patients with COVID-19, NC1-26 stands for patients without COVID-19. Figure 1 illustrates the summarized results. Note that in Column Pmax, a probability of 0.50 is due to truncation, i.e., its actual value is less than 0.50. From Column Pmax, we can see that the combined max-classifier correctly classified all patients into

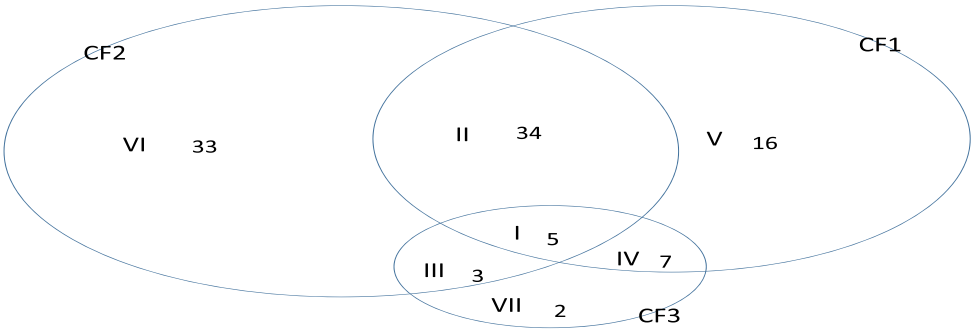

Figure 1: The performance of each individual classifier and the combined classifiers.

their corresponding COVID-19 groups, respectively. Classifiers CF1, CF2, CF3 alone correctly classified 62 (I, II, IV, V in Figure 1), 75 (I, II, III, VI), 17 (I, III, IV, VII) COVID-19 patients out of 100 patients into the COVID-19 group respectively. Classifiers CF1 and CF2 together correctly classified 98 (except VII) COVID-19 patients out of 100 patients into the COVID-19 group. Classifiers CF1 and CF3 together correctly classified 67 (except VI) COVID-19 patients out of 100 patients into the COVID-19 group. Classifiers CF2 and CF3 together correctly classified 84 (except V) COVID-19 patients out of 100 patients into the COVID-19 group. There are 16 (V), 33 (VI), 2 (VII) COVID-19 patients out of 100 being classified correctly by only one classifier of CF1, CF2, CF3, respectively, i.e., only one classifier works for those patients. The numbers of correctly classified among 100 COVID-19 patients by (CF1 and CF2 simultaneously, not CF3), (CF1 and CF3 simultaneously, not CF2), (CF2 and CF3 simultaneously, not CF1), are 34 (II), 7 (IV), 3 (III), respectively. The number of correctly classified among 100 COVID-19 patients by CF1, CF2, CF3 simultaneously (note: not the max-classifier) is 5 (I).

Based on the classifiers' performance in Figure 1, we can see that there are at least seven subtypes (I–VII) of patients as they have different characteristics, e.g., competing factors and their combinations. Of course, these seven subtypes can further be divided into subtypes based on the interactions (coefficients in (5)). It is clear that these five genes can be used to test and predict which type of COVID-19 diseases a patient may have.

We further notice that the signs of gene SMG1 are reversed in competing classifiers CF1 and CF2, respectively. This observation indicates that the diversified types of COVID-19 patients need different antiviral drugs and vaccines for treatment. It also raises one question whether or not one type of the first generation vaccines can be protective for all subtypes of SARS-CoV-2.

Remark 9. Reproducibility of all results in Table 1 is guaranteed as the readers can directly work on the downloaded data set from the link mentioned earlier to calculate the risk probabilities using the coefficients listed in Equation (5). Table 1 includes the original gene expression values. If the original published data has updated values which can happen at NCBI database, we cannot guarantee the coefficients in Equation (5) will lead to the same conclusions.

Remark 10. The conclusion depends on data quality. The data used in this study are plasma and leukocyte samples. Other types of gene expression data may not work. Also, the data is not a random sample, and then its representativeness may not be guaranteed. For data descriptions, we refer readers to Overmyer et al. (2020). As our objective functions (3) and (4) are new in the literature, their statistical properties are unknown. The uncertainty quantification of the estimators needs further investigation. In this study, our approach can be regarded as a man-guided machine learning data science discovery.

4 Discussions

This study is the first time in the medical literature that an infectious disease (and other diseases, cancer, flu, etc.) can be classified 100% correctly using only a few (five) genes. Many published results usually contain dozens of genes (e.g., 27 genes in Mick et al., 2020) for various purposes but still couldn't reach 100% correctness. In addition, we have also analyzed the expected counts data accompanied to TPM data in Overmyer et al. (2020), and again found our newly introduced competing classifier can 100% correctly classify the COVID-19 patients and non-COVID-19 patients, which shows that our approach is invariance preserving with different measures. The inference/analysis approach used in this study is robust and can shed new light on all gene-related research, i.e., not just the COVID-19 study. Researchers can apply max-linear type models in their studies.

The discovery of the five critical genes ABCB6 (ATP Binding Cassette Subfamily B Member 6 - Langereis Blood Group), KIAA1614 (Uncharacterized Protein), MND1 (Meiotic Nuclear Divisions 1), SMG1 (SMG1 Nonsense Mediated mRNA Decay Associated PI3K Related Kinase), RIPK3 (Receptor Interacting Serine/Threonine Kinase 3) can have an immediate application: design accurate test kit, and then determine a patient's COVID-19 subtype. The critical genes can save time and cost as researchers and drug companies can be more focusing on the targets. The discovery of the five critical genes can motivate many new research directions and laboratory experiments. For example, we can use these five genes as a starting point to conduct gene network analysis, test other reported genes, find the causal directions in various projects, find their connections directly or indirectly (through other genes) to conserved areas with critical functions in spike proteins, and find their connections to (directly or indirectly) other genes which can disrupt the life cycle of SARS-CoV-2. As a result, many other existing pieces of research can be enriched. It is important to notice that KIAA1614 is uncharacterized, and SMG1 is an mRNA type gene. They may need further explorations. It can also be hoped that new types of diseases and another set of critical genes can be discovered. We note that the five critical genes identified in this study were not reported in published work from our best knowledge. Ultimately, new testing procedures, vaccines, antiviral drugs, and treatments for COVID-19 can be designed.

The optimization problem (4) remains an open problem. In this study, setting $G = 3$ and $g_j = 3$ is purely ad hoc and for computational convenience and feasibility. We have tried $g_j = 2$, but the outcome was not satisfactory. A combination $G = 2$ and $g_j = 4$ can be tried. However, with $g_j = 4$, the interpretation of gene-gene interaction can become complicated. Finally, we note that the number of competing factors is smaller than the number of observed disease subtypes, which indicates that the competing factor classifier is not only workable but also interpretable.

On Overfitting and Underfitting Compared with the classical logistic regression model, the model complexity of model/equation (4) is higher, and hence Equation (4) can be over-fitting the data. On the one hand, note that by taking the intercept terms in the second competing factor and the third competing factor to be negative infinity or a substantial negative value, Equation (4) reduces to the classical logistic regression. As a result, Equation (4) won't cause an over-fitting problem than the classical logistic regression. On the other hand, compared with different classifiers with dozens of genes as predictors, the five-gene-based competing classifier is the most sparse gene-based classifier. Moreover, as discussed in Cui et al. (2020) that the classical linear regression is a particular case of the competing factor regression; the classical logistic model is a particular case of the competing factor classifier. In terms of the competing

factor classifier itself, if the number of competing factors or the number of gene predictors can be reduced, the five-gene-based competing factor classifier is overfitting the data. However, We were not able to reduce the number of genes to 4 or the number of competing factors to 2 in our extensive computation. More computational work may be needed. In summary, these five genes can potentially be drivers or messengers of COVID-19 disease.

Supplementary Material

Outcome Table 1 is in a supplementary file available online. A Matlab® demo code for solving Equation (4) is also available.

Acknowledgement

The author thanks Editor Jun Yan for constructive comments, careful reading, and guidance of the paper presentation and two anonymous referees for valuable comments. The author also thanks Dr. Yuqing Xu for reading an earlier version of the paper and commenting on some parts of the paper, Dr. Hao Yang Teng and Dr. Ye Zheng for their helpful discussions on COVID-19 studies.

Funding

The work was partially supported by NSF-DMS-2012298 (NSF).

References

Andersen K, Rambaut A, Lipkin W, et al. (2020). The proximal origin of SARS-COV-2. *Nature Medicine*, 26: 450–452.

Cao W, Zhang Z (2020). New extreme value theory for maxima of maxima. *Statistical Theory and Related Fields*. Forthcoming, https://doi.org/10.1080/24754269.2020.1846115.

Cui Q, Xu Y, Zhang Z, Chan V (2020). Max-linear regression models with regularization. *Journal of Econometrics*. Forthcoming, https://doi.org/10.1016/j.jeconom.2020.07.017.

Cui Q, Zhang Z (2018). Max-linear competing factor models. *Journal of Business & Economic Statistics*, 36(1): 62–74.

Fan J, Li R, Zhang CH, Zou H (2020). *Statistical Foundations of Data Science*. Chapman and Hall/CRC.

Guglielmi G (2020). Fast coronavirus tests: What they can and can't do. *Nature*, 585: 496–498.

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *The Lancet*, 395: 565–574.

Malinowski A, Schlather M, Zhang Z (2016). Intrinsically weighted means and non-ergodic marked point processes. *Annals of the Institute of Statistical Mathematics*, 68(1): 1–24.

Mick E, Kamm J, Pisco A, Ratnasiri K, et al. (2020). Upper airway gene expression reveals suppressed immune responses to SARS-COV-2 compared with other respiratory viruses. *Nature Communications*, 11: 5854.

Overmyer KA, Shishkova E, Miller IJ, Balnis J, Bernstein MN, Peters-Clarke TM, et al. (2020). Large-scale multi-omic analysis of COVID-19 severity. *Cell Systems*, 12(1): 23–40. https://doi.org/10.1016/j.cels.2020.10.003.

Rowland C (2020). Doctors and nurses want more data before championing vaccines to end the pandemic: Health systems are launching bids to assure their medical workers that vaccines will be safe and effective. *CNN*, November 21, 2020 at 6:00 a.m. CST.

Teng HY, Zhang Z (2020). Absolute and relative treatment effects in clinical trials: Models and applications in COVID-19 treatments. *Manuscript submitted*, University of Wisconsin.

The-RECOVERY-Collaborative-Group (2020). Effect of hydroxychloroquine in hospitalized patients with COVID-19. *The New England Journal of Medicine*, 383(21): 2030–2040.

The-Severe-Covid-19-GWAS-Group (2020). Genomewide association study of severe COVID-19 with respiratory failure. *The New England Journal of Medicine*, 383(16): 1522–1534. PMID: 32558485.

Xie Y, Zhang Z, Rathouz PJ, Barrett B (2019). Multivariate semi-continuous proportionally constrained two-part fixed effects models and applications. *Statistical Methods in Medical Research*, 28: 3516–3533.

Xu Y (2019). Regression models with max-linear structure, *PhD Dissertation*, University of Wisconsin.

Yu WB, Tang GD, Zhang L, Corlett RT (2020). Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2/HCoV-19) using whole genomic data. *Zoology Research*, 41(3): 247–257.

Zhang R, Tie X, Qi Z, Bevins NB, et al. (2020). Diagnosis of coronavirus disease 2019 pneumonia by using chest radiography: Value of artificial intelligence. *Radiology*, 298(2): E88–E97. Published Online: September 24, 2020.

Zhang Z (2005). A new class of tail-dependent time series models and its applications in financial time series. *Advances in Econometrics*, 20(B): 323–358.

Zhang Z (2008). Quotient correlation: A sample based alternative to Pearson's correlation. *The Annals of Statistics*, 36(2): 1007–1030.

Zhang Z (2020). On studying extreme values and systematic risks with nonlinear time series models and tail dependence measures (with discussions). *Statistical Theory and Related Fields*, Forthcoming, https://doi.org/10.1080/24754269.2020.1856590.

Zhang Z, Qi Y, Ma X (2011). Asymptotic independence of correlation coefficients with application to testing hypothesis of independence. *Electronic Journal of Statistics*, 5: 342–372.

Zhang Z, Zhang C, Cui Q (2017). Random threshold driven tail dependence measures with application to precipitation data analysis. *Statistica Sinica*, 27(2): 685–709.