Statistical Design and Analysis of Diagnostic Tests for Mutating Viruses

Yang Han[1,*], Yujia Sun[1], Jason C. Hsu[2,*], Thomas House[1,3,4], Nick Gent[5], and Ian Hall[1,4,5]

[1]Department of Mathematics, The University of Manchester, Manchester, UK, [2]Department of Statistics, The Ohio State University, Columbus, Ohio, USA, [3]IBM Research, Hartree Centre, Daresbury, UK, [4]The Alan Turing Institute for Data Science and Artificial Intelligence, London, UK, [5]Public Health England, UK.

*To whom correspondence should be addressed.

Abstract:

Mutations in the SARS-CoV-2 virus has given rise to concerns how diagnostic tests, treatments, and vaccines are affected. This article shows the presence of mutations in the Spike gene (S gene), which is prone to mutation, can be detected not only at the epidemiological (aggregate) level but in fact at the patient level using standard TaqPath PCR (polymerase chain reaction) test. However, only sequencing can reveal precise mutations at the amino acids level. Since PCR testing capacity far exceeds sequencing capability, patient level calling by our algorithm offers an efficient path toward targeted sequencing of cases infected by mutants, for discovering the actual mutations. Overcoming patient-to-patient variability by conditioning on an internal control, our statistical methodology classifies each patient as infected by wildtype or by some mutant strain(s) with reasonable accuracy. Besides adding a new tool for tracking emerging mutations epidemiologically, patient level calling is also important for treatment purpose, as there is evidence that some antibody treatments are less active in neutralizing the SARS-CoV-2 virus against certain mutant strains. Our algorithm can be continuously retrained as the SARS-CoV-2 virus evolves.

Key words: SARS-CoV-2, PCR, mutation tracking, S gene target failure, treatment prescription, statistical method, conditioning.

Contact: Yang.Han@manchester.ac.uk; Hsu.1@osu.edu.

1 Introduction

SARS-CoV-2, being a single stranded RNA virus, mutates easily. These mutations give rise to so-called variants of concern (VoC). A variant of concern (VoC) is one that increases transmissibility, causes the disease to be more severe or diagnostic detection to fail, or reduces effectiveness of treatments or vaccines. In the UK, the current variants of concern are Kent (B.1.1.7, from South East England), South Africa (B.1.351), Brazil (P.1) and B.1.1.7 with E484K [PHE]. In the U.S., the current variants of concern are Kent/UK (called just the UK variant in the U.S. media), Brazil, South Africa,

and two Californian variants (B.1.427, B.1.429) [CDC.1]. Variants that evidently impact on Medical Countermeasures (MCMs) such as diagnostic tests and/or therapeutics would be designated as Variant of High Consequence by the U.S. CDC. This article proposes a statistical method which turns potential failure of a popular diagnostic test around so that it not only can track emerging variants at the epidemiological level but also guide effective treatment at the patient level.

PCRs are popular diagnostic tests for SARS-CoV-2. In the UK, the TaqPath PCR test is the principal analytic platform for pillar 2 which consists of key workers in the NHS, social care and other sectors. Pillar 2 is also the diagnostic pathway for community cases. Such tests typically have multiple probes targeting different regions of the SARS-CoV-2 genome to bind to. Outcome of a PCR test is a Cycle threshold (Ct) for each target region, the thermal cycle number at which the fluorescent signal exceeds that of the background. Ct is a semi-quantitative value that can broadly categorize whether the amount of the target genome sequence in the sample is high or low. While a low Ct indicates substantial presence of the target, a high Ct may be due to either the amount of the target in the sample being low, or a mutation has caused a decrease in binding of the probe to the target (as the probe is built on the original genome sequence of SARS-CoV-2).

To be specific, within the SARS-CoV-2 virus, the S gene which codes the spike protein is especially prone to mutation. TaqPath targets the N (Nucleocapsid) gene, ORF1ab (Open Reading Frame 1a and 1b), and the S gene, for example. It is known that the 69-70del mutation in the S gene, a 6 base deletion that codes for two amino acids, causes a decrease in the signal from the S gene probe of TaqPath. This is referred to as the *S-dropouts* [1]. On the one hand, one wonders whether S-dropouts might cause an increase in false negative rates. On the other hand, this article shows S-dropouts can be used to advantage for epidemiological tracking of emerging variants as well as to guide prescription of treatments.

At the population level, in the UK, S-dropouts have already been used to advantage *for epidemiological tracking*, to discover and locate outbreaks from the Kent variant [3] by using the frequency of S-gene target negatives among PCR positives as a proxy for the frequency of the mutant strain. In the U.S., where the UK variant (B.1.1.7, called the Kent variant in UK) is not yet dominant [CDC.2] but may become so, this article provides a new technique, a statistical method, of tracking of the spread of B.1.1.7.

We also show in this article that S-dropouts can potentially also be used to advantage *at the patient level*, to differentiate patients infected by wildtype from patients infected by mutant variant(s). Being able to make such patient level calls is useful for treatment purpose. For example, there is evidence

that the antibody treatments bamlanivimab made by Eli Lilly (LY-CoV555) and casirivimab made by Regeneron (REGN10933), while retaining efficacy in neutralizing the B.1.1.7 variant [4], are less active in neutralization of the B.1.351 variant [5].

Methodologically the difference between the two uses is while epidemiological tracking can be done at the aggregate level, wildtype versus mutated infection calling at the *patient level* must overcome patient-to-patient variability, which appears to be considerable in the case of PCR testing for SARS-CoV-2. Guided by biology and statistical principles, we develop a simple yet effective method that can be deployed at point-of-care. Our proof-of-concept method is trained on a large Public Health England (PHE) data set and validated on confirmed Kent variant cases.

2 Training and Validation Data Sets

The training data set consists of all records created from 28 May 2020 to 2 November 2020 [2]. Total sample size n = 1,048,575. Among those samples, n = 500,440 are Covid-19 Positive by PCR or culturing. Among those confirmed Covid-19 Positive samples, n = 309,139 have complete Ct values for the N gene, the S gene, and ORF1ab. As the Kent variant, termed the Variant of Concern in [3], was first identified on 20 September 2020, this data set likely contained both patients infected by the original SARS-CoV-2 strain as well as those infected by the Kent variant.

The validation data set is Variants of Concern data under investigation created from 1 November 2020 up to 3 February 2021, from the COG-UK dataset, PHE Second Generation Surveillance System and the PHE Rapid Investigation Team Kent investigation [3]. Total sample size n = 7,582,696. Among the n = 2,646,907 confirmed Covid-19 cases (called *Positive* by PCR or culturing), n = 2,561,279 are missing sequencing test results. (Sequencing results are available only on a subset of the samples of positive cases as PCR testing capacity exceeds the national sequencing capability.) Among those sequenced, n = 38,612 are called as Kent variant, n = 116 are called as South African variant, with n = 46,900 called as Wildtype. Confirmation of being a Kent variant is by phylogenetic tree matching analysis with the B.1.1.7 strain. Wildtype is defined as sequenced samples other than Kent variants and South African variants. Within those called Kent variant by sequencing, n = 34,018 are further classified as Confirmed Kent variant, n = 4,536 as Probably Kent variant, and n = 58 as Low-quality Genome. Among the Confirmed and Probable Kent variants, only n = 76 have complete Ct values for the N gene, the S gene, and ORF1ab. Those 76 cases with complete Ct values are the ones we use for validation.

3 Observation from Preliminary Analysis and Rationale for Our Conditional Control Approach

We first tried simply plotting the distribution of the S gene's Ct in the training set, to see if the density indicates it is a mixture of distributions, one for patients infected by wildtype, one or more others infected by mutant strains, see Figure 1. Neither the histogram nor its smoothed version indicates anything definitive. Apparently, there is too much person-to-person variability in the S gene's Ct within those infected by wildtype, and within those infected by some mutant strain, that their distributions overlap too much for us to tell which patients are infected by wildtype and which are infected by some mutated variant. Sources of such person-to-person Ct variability include varying stage of the illness when cases are tested. Initial viral load changes roughly as a gamma distribution from the late prodromal stage until the early recovery period, so having a mixture of screening, asymptomatic, and symptomatic tests will cause the Ct distribution to be complicated.

We thus examine the possibility of controlling for person-to-person variability using one or more *covariates* of the target gene, specifically Ct's of the N gene and ORF1ab. We view them simply as variables that might be used to "baseline" each patient, not particularly as quantitative measurements of initial viral loads per se. For our purpose of differentiating patients with S-dropouts from those without, an ideal baseliner should be stable relative to mutation in the S gene, as if there were no mutation in that gene.

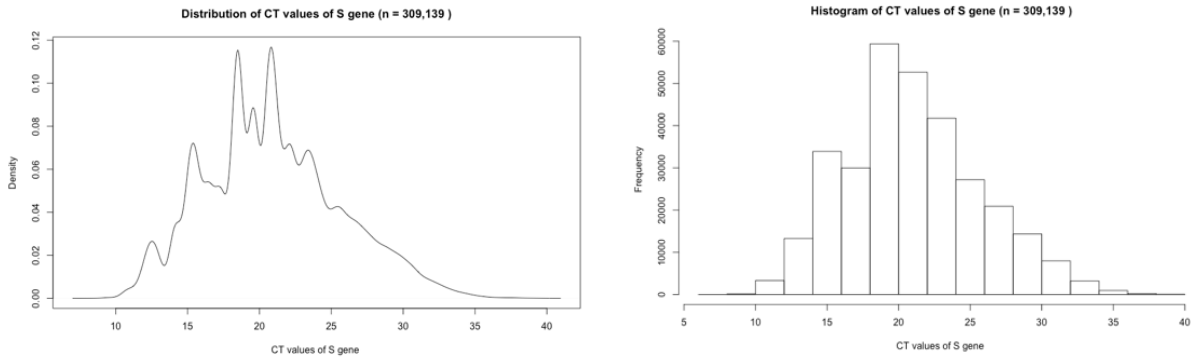

Fig. 1. Distributions of the S gene's Ct in the training data set.

4 Method and Result

Conditioning is a tried and true principle of statistical inference. Idea is that, for each patient, Ct's of the N gene and ORF1ab may be correlated with the Ct of the S gene (high or low in the same direction), so we might use the Ct of the N gene or Ct of ORF1ab or both to baseline each patient. Once a variable is found to suitably reduce S gene's Ct variability conditionally given its value, we calculate its predicted value for S gene's Ct disregarding potential mutation. Then hopefully, in most cases, patients infected by wildtype will have an observed S gene Ct lower than its predicted value, while patients infected by a mutant strain will either experience S-dropout (and have no Ct value for the S gene) or have an observed S gene Ct higher than its predicted value.

We first try conditioning on a single variable, either the Ct of the N gene, or Ct of ORF1ab. The conditioning technique we use, the simplest, is linear regression. Raw value of the S gene's Ct is then replaced by its *residual*, its observed S gene Ct value minus its predicted S gene Ct value (given the corresponding Ct value of the conditioning variable). We call such residuals S-residuals. The hope is these S-residuals will form clusters, of patients infected by wildtype and of patients infected by mutant(s). We do this on the logarithmic scale which seems to work better than on the raw Ct scale. Ct of ORF1ab works well as a conditioning variable. However, Ct of the N gene does not work well. While we speculate that there may be differential sensitivity/specificity of the receptor targets used in the PCR system, these relative differences need further exploration.

Figure 2 shows the result of this regression from the training data set. Labeling log(Ct) of the S gene as *log(S)* and log(Ct) of ORF1ab as *log(ORF1ab)* on the axes, each dot represents a patient's log(ORF1ab) and log(S) values. Due to the large size of the training data set, for clarify of message, the dots in Figure 2 represent only 10,000 randomly selected patients. The regression line *fitted based on the entire training set data* is the red dashed line. For a patient with a particular log(ORF1ab) value, its S-residual from this regression is the log(S) value of the dot representing the patient minus that patient's log(S) value as predicted by the red dashed line. So a dot above the red line will have a positive S-residual, while a dot below the red line will have a negative S-residual.

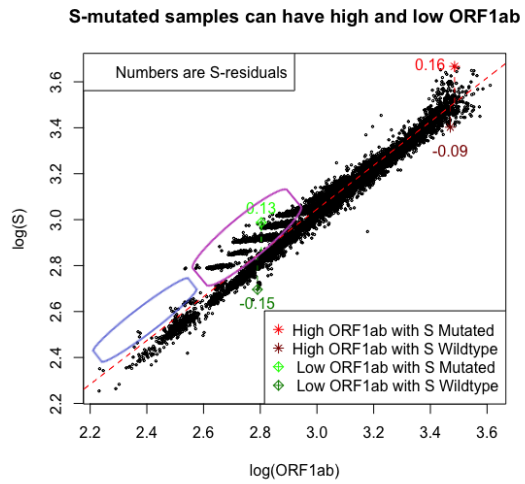

Fig. 2. Regression from the training data set.

Figure 3 shows a density plot (a smoothed histogram) of the S-residuals from the entire training data set. This plot suggests it is a mixture of densities, one centered to the left of zero while the others to the right of zero. The relatively flat area represents where the densities blend. In the training data, about 58% of the patients have S-residuals less than zero, 42% have S-residuals greater than zero. Theorizing that patients infected by wildtype without S-dropout would tend to have lower S gene Ct values than patients infected by mutant(s) and thus experiencing S-dropouts, we suggest simply calling a patient as infected by wildtype if his/her S-residual is negative, and calling a patient as infected by mutant(s) if his/her S-residual is positive. In other words, call a case wildtype if its dot is below the red line, and call it mutated if its dot is above the red line.

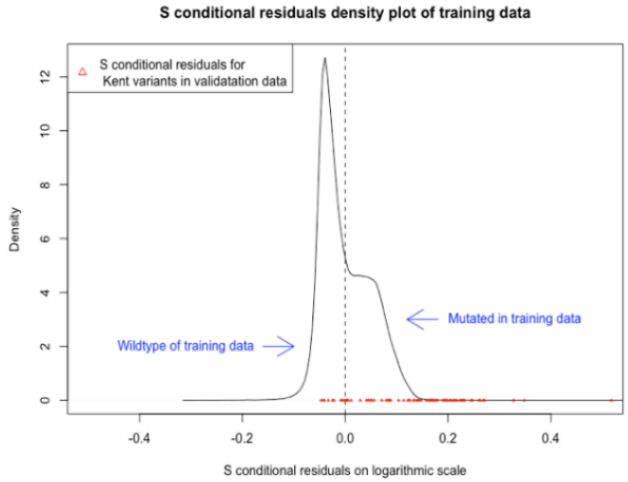

Fig. 3. Density plot of the S-residuals from the entire training data set.

We tested this algorithm using confirmed and probable Kent variant cases in the validation data set (confirmed by sequencing). Represented by the red triangles in Figure 3, out of the 76 confirmed and probable Kent variant cases, 66 (about 87%) have positive S-residuals while 10 (about 13%) have negative S-residuals. So we are cautiously optimistic that our method which we call the Conditional S-residual Method has potential for calling patients either as infected by wildtype or by mutant(s) with some accuracy.

4.1 Requirement for Our Method to Be Applicable

Concerns with mutations include diagnostic testing failure, reduced anti-body neutralization, increased severity of the disease, and increased transmissibility. These are separate issues. For example, the 69-70del deletion causes S gene target failure with TaqPath PCR, while the N501Y substitution increases transmissibility. Medical definition of a variant is not based on a single concern but takes all these considerations into account. Currently, PHE's definition of a Confirmed B.1.1.7 case in [6] is it includes 13 lineage defining mutations (out of 29 known mutations), while a Probable B.1.1.7 case is it includes at least 5 of the lineage defining mutations (and all other positions are either unknown or mixed bases). Deletions such as 69-70del in fact are not currently included in the B.1.1.7 definitions.

For our method to apply to a particular PCR test, mutations between strains that one is trying to differentiate need to be different at some target sequences of the probes for bindings to have differentials. From a diagnostic test's point of view, SGTF (an extreme differential) indicates B.1.1.7 because it is known that many B.1.1.7 patients harbor the 69-70del and it causes SGTF. However, a case can be a confirmed or probable B.1.1.7 without the 69-70del, which we believe many of the cases with S-residual values greater than zero are. In principle, our method applies so long as there is sufficient differential in probe bindings.

4.2 Some Observations

As indicated by the blue oval in Figure 2, for low ORF1ab Ct value cases, almost no S-gene mutated cases show up in this plot of the training data set. Those missing observations correspond to mutated cases in the training data set with missing S Ct values. S-dropouts are caused by a failure of the qPCR probe to bind due to the 69-70del mutation (page 7 of [3]). When probes of the N gene and ORF1ab successfully bind to amplified targets but probe of the S gene fails to bind, this is termed S gene target failure (SGTF). So those missing cases likely are patients infected by strains with the 69-70del (especially B.1.1.7). In Figure 2, there are mutated cases with higher ORF1ab Ct values that are plotted.

Those are cases for which primer for the S-gene retained some ability to bind, though at the limit of the technical capabilities of the PCR systems.

As indicated by the magenta oval in Figure 2, we observe five horizontal bands of log(S) values for mutated cases in the 2.5 to 3.0 range of log(ORF1ab). These bands correspond to five distinct S gene Ct values from 16.4 to 20.8, spaced about 1,1 Ct units apart. Uncertain of the reason why, it is our hope that observing these bands sheds some light on S dropouts.

Comparing the scatterplot of log(S) vs log(N) with the scatterplot of log(S) vs log(ORF1ab) in Figure 4, we can see somewhat that log(ORF1ab) can predict log(S) value better than log(N) can. To some extent, this explains why the N gene does not work well as a Control for the S gene.

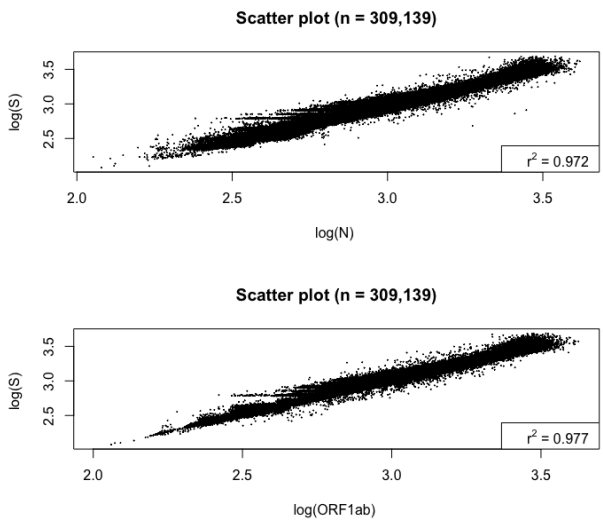

Fig. 4. Scatterplot of log(S) vs log(N) with the scatterplot of log(S) vs log(ORF1ab).

4.3 Continuous Re-training

The original training data set contained samples up to November 2020. Since the Kent variant was first discovered in September 2020, the original data set almost surely contained patients infected by the original strain and by the Kent and other mutated variants. But since the Kent variant and other mutated strains likely constituted a minority in the training data set, in our analysis of the training data set, the original strain is called "wildtype".

However, aligning with the definition of wildtype in [3], in our analysis of the validation data set which targets identification of Kent variant samples, wildtype became all strains that are not the Kent variant (i.e., the *complement* of the Kent variant).

Viruses mutate over time and, in the case of SARS-CoV-2, by now the Kent variant has become the dominant strain in the UK. Thus, assume at any given one time there is a dominant (> 50%) strain

which we will call “wildtype”, and there is one or more emerging strains which we will call “mutants”. It is important that the conditional S-residual regression equation (or N-residual if surveillance of mutation is on the N gene) be continuously re-trained each time there is a meaningfully updated data set, separating “wildtype” from “mutants”, to not only guide prescription of treatments to patients, but also to monitor which mutant strain is evolving to become dominant.

Each new lineage of SARS-CoV-2 will be defined by a set of mutations. For example, PHE's previous definition of B.1.351 in [6] includes 13 lineage-defining amino acid changes. Depending on how many of these mutations a person has, that person may be classified as a Confirmed or a Probable B.1.351 case. To illustrate surveillance of mutations in a gene that do not cause the target probe to fail (which the 69-70del in the S gene does) but decreased binding, consider applying our method to the N gene using ORF1ab as the Control. We regressed logarithm of the Ct of the N gene on the logarithm of the Ct of ORF1ab in the randomly selected subset of sample size n = 10,000 of the training data set, obtaining the following regression equation

Sample size of subset	Intercept	Slope
10,000	0.23524	0.93901

To assess how changing prevalence of the dominant strain and training sample size might affect the regression equation, we also randomly selected subsets of the validation data set of sample sizes n ranging from 5,000 to 50,000. Regressing logarithm of the Ct of the N gene on the logarithm of the Ct of ORF1ab, the resulting regression equations were as follows.

Sample size of subset	Intercept	Slope
5,000	0.05507	0.98962
10,000	0.05595	0.98887
50,000	0.06558	0.98618

Note that the regression equation did change somewhat between our training data set and the validation data set, which is not surprising since the dominant strain had changed. However, within our validation data set, the regression equation did not change much as sample size ranged from 5,000 to 50,000; and a sample size of 10,000 certainly seems adequate. In contrast to Figure 3, density plot of the conditional residuals of the logarithm of Ct of the N gene of the training data set did not show the distribution

to be a mixture, while that of the validation data set showed just a hint that it might be a mixture, see Figure 5.

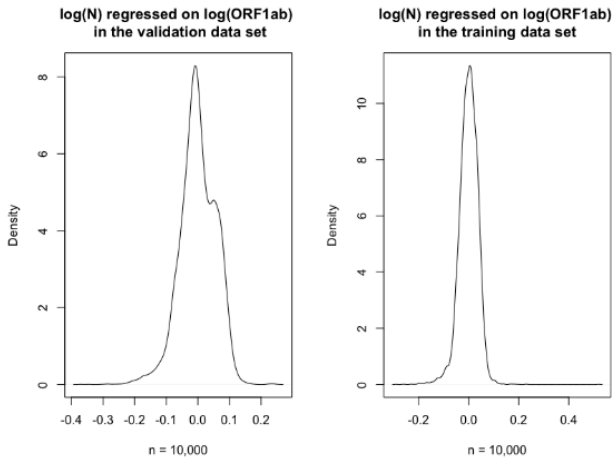

Fig. 5. Density plot of N-residuals from validation and training data sets

5 Recommended Design and Analysis Principles

We suggest the following principles for designing diagnostic tests for mutating viruses and for statistically analyzing the test results:

1. Target multiple genes and/or open reading frames
2. Choose probes as unique as possible relative to other organisms
3. Within the target virus, choose probes that are biologically as functionally independent as possible
4. Within the target virus, use simple methodology to make the probes statistically as independent (orthogonal) as possible

Our rationales are as follows.

Principle 1: While TaqPath targets two genes (S gene and N gene) and one open reading frame (ORF1ab), some other PCRs target only one gene and one open reading frame (the N gene and ORF1ab in the case of the Chinses CDC), or even just a single gene (the N gene in the case of the U.S. CDC). How a virus evolves is hard to anticipate, so building diverse targets into probe designs is safer, and turned out to be useful in the case of SARS-CoV-2.

Principle 2: To avoid non-specific bindings, rationale for the second principle is biologically self-evident and is standard practice.

Principle 3: Technically, making the probes statistically independent to classify patients with tends to work better when the probes are more biologically functionally independent. This is because, with

more biological independence, there is a higher chance that a suitable internal Control variable can be found to baseline the target (mutating) gene with.

Principle 4: Based on the statistical principle of *conditioning*, a key contribution of this article is to offer a simple yet effective statistical methodology for patient infection classification. Once this scientific problem has been explicitly formulated and a solution proven feasible, other techniques (such as statistical learning or machine learning) can potentially be used to classify patients. While not opposed to their use, we do feel simple methodologies such as ours with clear biological interpretations are preferable to black boxes.

6 Recommended Case Calling Algorithm and Future Research

Based on our current algorithm, in countries where the B.1.1.7 strain has not become dominant, for patients with positive N gene and ORF1ab Ct's, we recommend calling a patient as infected by a 69-70del strain (especially B.1.1.7) if the sample experiences SGTF (with its Ct for the S gene missing). Call a patient as "infected by mutants" (which may include B.1.1.7) if its Conditional S-residual value is greater than zero. Otherwise, call a patient as infected by wildtype.

A future research goal of ours is to assign statistical confidence levels to such calling. Once that has been achieved, an app based on Figure 3 can be made available. The current S residual density plot is trained on data with 390,139 cases from Public Health England (2020) which consists of records created from 28 May 2020 to 2 November 2020. At the moment, our proof-of-concept (for research only) app calls a case wildtype (as defined for the training data set) if its red dot is to the left of zero, and call a case as mutated if its red dot is to the right of zero. Being web-based, eventually such an app (if approved) can be deployed at patient point-of-care. Our algorithm for the app can be continuously retrained as the SARS-CoV-2 virus evolves.

7 Deeper Statistical Insight into Our Method

In traditional statistical thinking, the regression coefficients are of primary interest while the noises are nuisances. So in usual regression practice, we keep the estimated regression coefficients and the residuals are discarded once we have used them for variability estimation and regression diagnostics. It is the opposite with the Conditional S-residual method: after regression, we discard the estimated coefficients and look for signals of mutations in the residuals which we keep.

"The signals that we seek, are in the noises that we keep!"

Acknowledgements

In addition to thanking Liang-Chun Liu, Yushi Liu, and JJH for helpful biological discussions, and dedicating this article to the memory of Wu laoshi, we acknowledge sadness for the suffering caused by the Covid-19 pandemic, admiration for courage of the essential workers, and gratitude for the opportunity to contribute our effort.

Funding

The following funding sources are acknowledged as providing funding for the named authors:
IH is supported by the UK Research and Innovation Joint Universities Pandemic and Epidemiological Research (grant number MR/V038613/1) and the National Institute for Health Research Policy Research Programme in Operational Research (grant number PR-R17-0916-21001). TH is supported by the Royal Society (grant number INF/R2/180067), the UK Research and Innovation COVID-19 rolling scheme (grant numbers EP/V027468/1, MR/V028618/1 and MR/V038613/1) and the Alan Turing Institute for Data Science and Artificial Intelligence. YS is supported by the University of Manchester for PhD funding.

Conflict of Interest: none declared.

References

[1] Public Health Wales (2020), Technical Brief: Viral Variant VOC-202012/01, Public Health Wales, Cardiff, UK

[2] Public Health England (2020), Public Health England, London, UK.

[3] Public Health England (2021), Investigation of novel SARS-COV-2 variant: Variant of Concern 202012/01, Public Health England, London, UK.

[4] Wang, P., Nair, M.S., Liu, L. et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature (2021). https://doi.org/10.1038/s41586-021-03398-2

[5] Wibmer, C.K., Ayres, F., Hermanus, T. et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat Med* (2021). https://doi.org/10.1038/s41591-021-01285-x

[6] Public Health England (2021), Technical Briefing 6: Investigation of SARS-CoV-2 variants of concern in England, Public Health England, London, UK.

[CDC.1] Centers for Disease Control and Prevention (2021), SARS-CoV-2 variant Classifications and Definitions. Retrieved March 16, 2020 from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html#print.

[CDC.2] Centers for Disease Control and Prevention (2021), Variant Proportions in the U.S. Retrieved March 17, 2020 from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-proportions.html.