An Analysis of Variables Influencing COVID-19 Incidence and Mortality

Jon Javor
University at Buffalo Biostatistics Department

August 26th, 2022

Introduction

The virus SARS-CoV-2, now all too familiar to the world by its resultant disease, COVID-19 (or, colloquially, simply "corona", an abbreviation of "Coronavirus 2019"), has become a seemingly interminable and ubiquitous presence in the global social, economic, intellectual and cultural consciousness since it so abruptly appeared in the very last days of 2019.

Believing to have originated from bats, and to have passed up the food chain to humans via small predatory mammals in China's Wuhan Province, by the end of the first quarter of 2020 - a year that will now live forever in infamy - "COVID" and "Corona" were becoming household names as countries went into shutdown, desperately attempting to stop the rapid spread of this new virus. What would follow was an ever-changing status quo: a completely different global landscape on every level down to personal as individuals endured months of quarantine, social distancing, remote work, masking - to speak nothing of the sickness itself and its accompanying maladies - but, perhaps above all, a blanket uncertainty about everyday decisions that were once so commonplace as to be unnoticeable; without a doubt, a defining event for generations, the after-effects of which the world will be feeling for more to come.

However, as long as the adversities of the world have presented human beings with uncertainty, we have looked around and observed, out to find causes, and answers. It is from this fundamental instinct - mathematized and codified over the millennia since Aristotle - that the science of statistics (and its here highly topical sub-discipline, epidemiology) originated, and allowed human beings to develop to the stage of progress at which we currently reside: we, as a species, had the ability to fight even this. Within mere months of the outbreak, tests began to be distributed to help target and quell the spread, and - with astonishing speed - within less than a year, the first dose of the COVID-19 vaccination was administered, the beginning of our march out of the darkness that we are still undergoing today, as the virus fights for its life, mutating seemingly ceaselessly, and we fight each other. How much caution is too much, who is telling the truth and who is denying it, and, importantly, the question the analysis in the following pages will attempt to answer: What made it worse, or better?

Table of Contents

1 Abstract

In the war against the coronavirus, information on the new and largely unknown virus provided a distinct tactical advantage, especially in the earlier stages of the pandemic. Knowing who was at highest risk, and why, was vital to saving lives and understanding the nature of the threat. In the following analysis, various factors from several countries are fit to models in pre- and post-vaccination time frames, in order to see which of those population health or public health measures made case and death counts better, worse, or simply had no effect at all. In order to do this, the count data was scaled by country population, and had fit to it a Generalizing Estimation Equations (GEE) Negative Binomial model (with offset). Evaluated for appropriate parameter inclusion by minimization of Quasi-Likelihood under the Independence Criterion (QIC), models of case and death counts pre- and post-vaccination were fit, and the results interpreted.

2 Methods

In the following analysis, a repeated measures generalized linear model, known as generalized estimating equations, is fit under a Negative Binomial distribution:

$$f(k; r, p) \equiv \Pr(X = k) = \binom{k + r - 1}{r - 1}(1 - p)^k p^r$$

in order to account for the variance of the count variable in question being greater than the mean (a phenomenon known as overdispersion). This model is further fit using a distribution specific link function, notated and used also to determine means and variances as follows (respectively):

$$g(\mu_i) = \eta_i,$$

$$\mu_i = g^{-1}(\eta_i).$$

$$Var(Y_i) = \phi \cdot v(g^{-1}(\eta_i))$$

where

$$\eta_i = \boldsymbol{X}_i\boldsymbol{\beta} = \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}$$

(It is worth noting that in the case at hand of the Negative Binomial distribution, the link function is the natural log).

In addition, an offset will be used (a parameter with value treated as known throughout the model fitting process) in order to appropriately scale case and death counts by respective country populations.

The parameters for the aforementioned model will be determined using software (SAS), under a process known as quasi-likelihood estimation, which are the solution(s) to

$$\sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)' \frac{Y_i - \mu_i}{\phi v(\mu_i)} = \mathbf{0}$$

and a related quantity - the Quasi-Likelihood under the Independence Model Criterion (QIC) - which will be used to assess goodness of model fit relative to models with more/less covariates.

3 Data Filtration

3.1 Cleaning the Data for Read-In

Once a data set suitable (that is, accounting for enough potential covariates) for analytical use had been found[3], the next natural step was to read said data into the appropriate statistical software. Therein, a complication immediately presented itself: with COVID-19 data from 246 countries over 900 days (approximately; start dates of documentation were not identical by country), the row count of this data ended up just shy of 200,000. SAS's import wizard was capable of successfully reading in the data; however, when asked to print it (an essential function for "reality checking" the efficacy of data cleaning procedures), nearly six minutes of computing time was insufficient to execute, until the attempted "break" command caused a complete terminal crash. Turning to R, a similar problem presented itself: R could successfully read in the data, but values in the more extreme rows/columns were corrupted to "TRUE"s and "FALSE"s, despite being originally numeric. This specific stumbling point is indubitably common in professional statistical settings - especially epidemiological, where data sets can incidentally be prohibitively large - but has minimal academic application, except for in classes specifically devoted to the subject. Data is seldom so raw as here in the university context, presented instead in formats curated for the sake of clear conceptual demonstration.

The solution, as it turns out, would be as simple as employing a tertiary platform for the preliminary analytical stages: where SAS and R both failed to present the data in a manner useful for repeated viewing, Microsoft Excel was the ungainly .xlsx file's native environs, and had no such problems. So then, it was from this springboard that the early stages of data cleaning were carried out: predominantly, the data set was reduced down to only a select set of under a dozen countries "of interest", with the intent of drastically scaling down the number of rows so that more advanced statistical computational environments (namely R, where the finer points of the data cleansing would be completed due to its inherently vectorized nature) could better exhibit, and as such edit, the slimmed down set.

It is worth noting, from an experimental design/model-building perspective, that in this particular context, the designation of a country "of interest" was devised from a mixed consideration of graphics from the original data source[1], and countries from which a new variant of the virus are thought to have originated[2]. It is in the context of (an eventual subset of) these filtered locations that the analysis will be conducted.

3.2 Data Cleaning in R

With this newly refined data (including only around 15,000 rows of the initial 196,000) having been at last read intact into R, the next major step of cleaning was underway: namely, the combination of 900-odd day measurements into respective month-year bins, so that various repeated measures analyses could be more illustratively and efficiently run. In order to accomplish this, every remaining column variable (some having been beforehand strained out due either to an excess of missing values within a country, or to correlative redundancy with other covariates of more paramount relevance) would have to be either summed or averaged into a monthly measurement. Which arithmetic operation was performed was contingent on the context of the variable in question: for example, "New Cases" was combined into a monthly sum of all new case occurrences, whereas "Stringency Index" - a metric representing the severity of a country's COVID-19 preventative and precautionary protocols - was better represented as a monthly mean.

Once this process, and various other minutiae of data cleaning (among them of note, the exclusion of China and Russia, countries with an excess of clearly misrepresented "New Cases" and "New Deaths" dependent variables - specifically, a conspicuous abundance of improbable zeroes) had been resolved, the next significant step involved the generation of a new categorical covariate - one which was noticeably amiss from the original data, but is very likely of confounding significance: the predominant variant of the virus in the world at the time. By December 2020, the original strain of COVID-19 had been replaced in prevalence by its first well-known variant: Alpha. By July 2021, Alpha had in turn been overshadowed by the Delta variant, and by Christmas of 2021, the current prevailing variant - Omicron - was the driving force of global COVID-19 infection, as it and its sub-variants remain to this day[2].

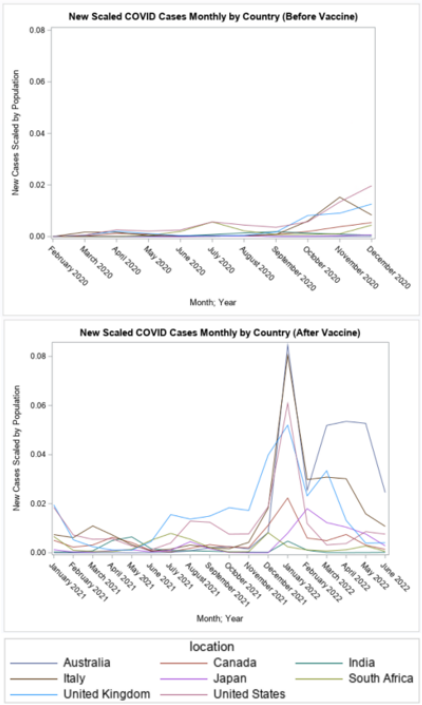
Finally, before the data could be passed over to SAS to begin the analytical stage properly, one more task had to be fulfilled: a subdivision of the data in accordance with a covariate with high likelihood of statistical significance: the vaccine. Since it would be unrepresentative to treat the "Full Vaccinations" variable as simply zero for all the time before a vaccine had even been developed - conveying the erroneous implication that countries simply distributed no doses of the vaccine, when in fact none existed at all - the data was split into two data sets: one pre-vaccine, and one post-vaccine. The analysis will be consequently conducted twofold, along that bifurcation.

4 Data Visualization

4.1 COVID-19 Case Summary

When faced with such a multifaceted data set, it is standard statistical recourse to reduce it down to summary measures, such as the mean and variance, and to present the data in a way where possible trends - or just general structural components - are more easily accessible, visually. In this case, the latter was accomplished by plotting monthly mean COVID-19 cases (adjusted for country population), before and after the advent of vaccination (see below left), and the former by a straightforward call to SAS's mean procedure, summarizing each country's overall mean and variance of case count (unscaled this time), also before and after vaccination became widely available (see below right).

The rightmost figures below seem to indicate a much higher case rate post vaccination, with a dramatic spike peaking around January 2022 (and a comparably much more subdued crest in the winter months of 2020). This may seem counter-intuitive to the clear efficacy of vaccination, but these plots, while illustrative of overall longitudinal trends, do not account for confounding variables such as current variant of the virus. It is of note that the January 2022 apex is highly correlated with the global emergence of the incredibly contagious Omicron variant, around the same time.



COVID-19 Case Summary by Country and Vaccine Status

The MEANS Procedure

	Analysis Variable : new_cases_monthly			
location	vaccine_status	N Obs	Mean	Variance
United St	Before	11	1835573.55	3.9021988E12
	After	18	3698128.00	1.9882844E13
United Ki	Before	11	226526.18	91422237690
	After	18	1073687.39	933342824214
South Afr	Before	11	96105.55	12845423624
	After	18	161906.11	26246038079
Japan	Before	11	21430.36	680827408.45
	After	18	497717.28	431821046380
Italy	Before	11	191573.82	85947087197
	After	18	886914.89	1.3608189E12
India	Before	11	935155.27	846388549808
	After	18	1837532.50	7.3697632E12
Canada	Before	11	53658.64	4300822770.1
	After	18	186396.06	37672098659
Australia	Before	11	2583.27	11549511.22
	After	18	441485.06	451877185242

In order to attempt to at least partially compensate for the aforementioned conflating factor, another analytical measure deemed sufficiently of-interest to be explored was the division of mean COVID-19 cases by viral variant. It should be noted that the output below, while scaled again for population, is not however scaled for duration of variant. Globally, the original strain of COVID-19 (coded as "None") was dominant for 10 months, Alpha variant for 7, Delta for 5, and Omicron for 7. Scaling for this confounding consideration, the Omicron variant of COVID-19 still - perhaps predictably - ends up as the strain with the highest (doubly) scaled mean case rate, at 0.002399.

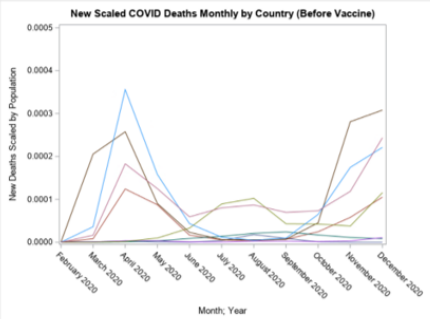
COVID-19 Scaled Case Summary by Variant (Unscaled for Time)

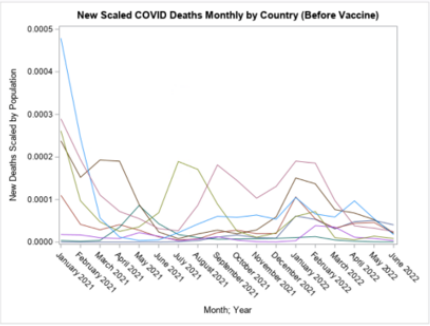
The MEANS Procedure

Analysis Variable : scaled_new_cases			
variant	N Obs	Mean	Std Dev
Alpha	56	0.003901	0.004747
Delta	40	0.004514	0.005363
None	79	0.001565	0.002808
Omicron	56	0.016791	0.020459

4.2 COVID-19 Death Summary

Besides COVID-19 case rates, another response variable of particular interest is COVID-19 death rates. In a similar fashion to the previously discussed cases, deaths are summarized and displayed below. Interestingly, the distribution of deaths is rather different than that of cases, with a significant crest around April and winter 2020, with countries individually spiking and fluctuating - although not as high - post vaccination. This is a likely indication of the vaccine's ability to prevent more serious illness, leading to death. The peaks in both graphs (with the exception of April 2020) still correspond roughly to the emergence of new global variants





COVID-19 Death Summary by Country and Vaccine Status

The MEANS Procedure

Analysis Variable : new_deaths_monthly

location	vaccine_status	N Obs	Mean	Variance
United St	Before	11	31912.64	544911973
	After	18	36906.06	618123902
United Ki	Before	11	6688.09	61665846.69
	After	18	5688.89	58927864.34
South Afr	Before	11	2588.09	6323961.09
	After	18	4050.67	18702201.06
Japan	Before	11	317.4545455	138533.07
	After	18	1533.17	1876605.21
Italy	Before	11	6744.55	56715971.87
	After	18	5211.56	18929307.79
India	Before	11	13545.00	137226963
	After	18	20477.00	872495461
Canada	Before	11	1516.82	3029929.76
	After	18	1416.06	1227563.35
Australia	Before	11	82.7272727	20184.02
	After	18	483.4444444	329350.85

It is also worth noting for future consideration that the country variances presented in both case and death rates are far from equal to their means. This will prove relevant in Section 4.

Again similarly to previous case consideration, and in light of the aforementioned peak coincidence in the previous longitudinal figures, deaths by variant were also briefly examined. Scaling for duration once again, the Alpha variant actually ends up being the deadliest variant. This is likely due to the fact that Alpha was significantly more dangerous and contagious than the original strain, and emerged with a comfortable enough time to wreak havoc before vaccination was widely available to the general population.

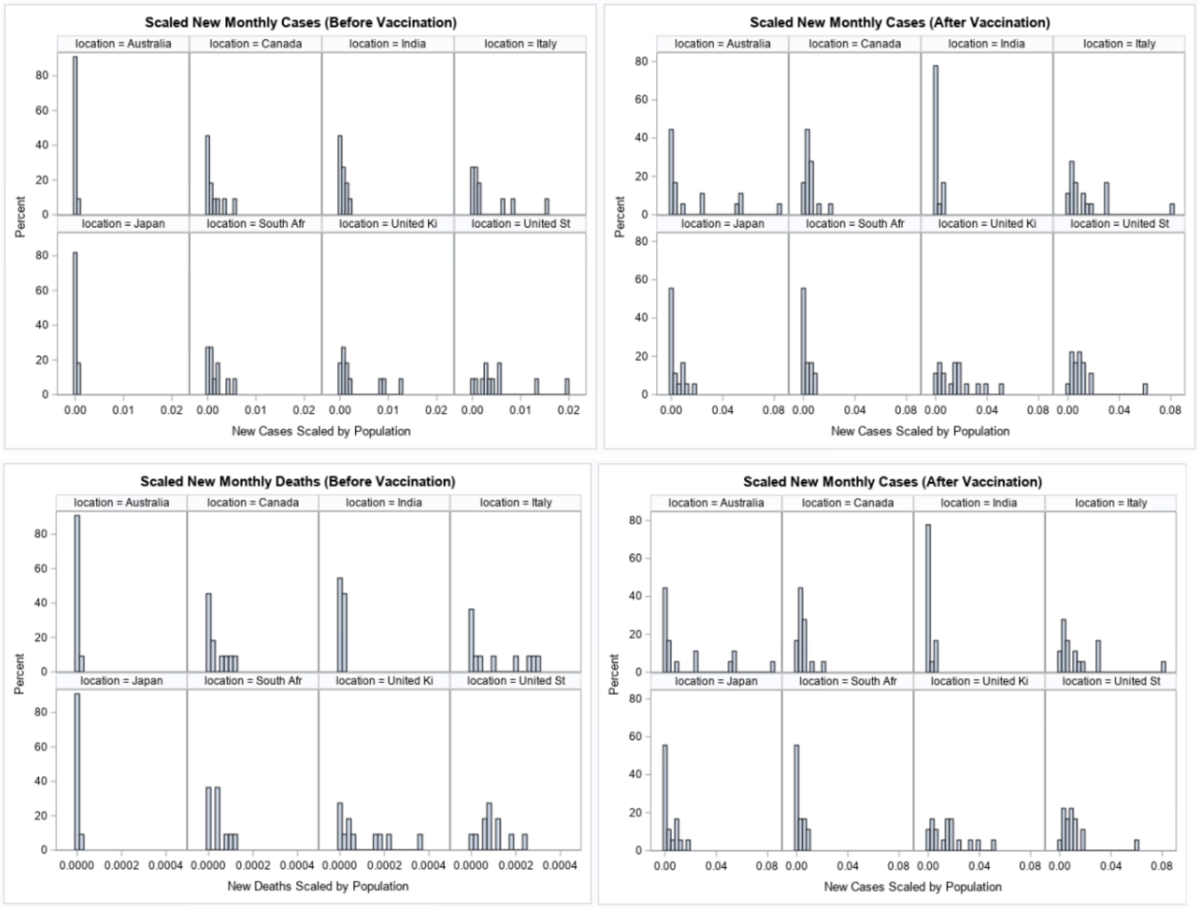
COVID-19 Scaled Death Summary by Variant (Unscaled for Time)

The MEANS Procedure

Analysis Variable : scaled_new_deaths			
variant	N Obs	Mean	Std Dev
Alpha	56	0.00008058	0.00010242
Delta	40	0.00003959	0.00005131
None	79	0.00004345	0.00007012
Omicron	56	0.00004770	0.00004544

5 Finding the Right Model

In order to get a broad sense of what distributional family of model should be fit to this data (or, perhaps more accurately, what if any family of model must be excluded), a series of country-specific histograms were generated, both before and after widespread vaccination, for (population-scaled) cases and deaths:



As expected of count data (or, in the case of the above figures, proportional data based directly on counts), the countrywide distributions of interest appear Poisson. With the widespread and commonly used linear regression and ANOVA families of analysis out of the question - based as they are around the assumption of response normality, such began a search for the appropriate modeling scheme for this somewhat unruly data.

Further hurdles were still at hand, however. One of the main drawbacks of any Poisson process is a phenomenon known as overdispersion, a distributional quality in which the data mean and its variance are not equal (specifically, the variance exceeds the mean). This will result in the need for a distribution with more descriptive parameters than just one lambda (the Poisson parameter, which describes both its mean and variance at once). The previous summary measures explored in Section 3 were quite clear in the conclusion that overdispersion was, in fact, present here. So, moving forward, Negative Binomial alternative models were to be considered.

In addition, the very nature of the data itself proved problematic, in particular the response variables of interest. Cases and deaths were indeed well-described by counts when consideration was confined to one country, but in a comparison against multiple, the vastly differing populations of Australia versus India, for example, would be a powerful confounding variable. In order to correct for this, the case and death counts were scaled for population via simple count proportion (as had been alluded to in Section 3, and displayed in the figures on the preceding page). While this allowed for cases and deaths to be compared more genuinely among countries, it immediately lent itself to further complication: Negative Binomial models are specifically tailored to count data, not proportional data. This would necessitate using an algorithm that could account for the various populations of the countries involved in the analysis.

Negative Binomial Generalized Estimating Equations (GEE) were decided upon as the best methodology for modeling this thickly layered data set. This technique for modeling repeated measures data (especially count data, as is present here) is more thoroughly mathematically expounded upon in Section 2. Conceptually, not only does GEE have within its assumptions the handy prerequisite that observations are correlated, but also can easily account for repeated measures (in context, the multiple observations from the same country over time).

6 Fitting the Model(s)

In order to fit a count model on proportion data, an offset would have to be used. An offset is standardly used for a covariate in Poisson or Negative Binomial regression with known slope, to hold it fixed and account for that knowledge in the process of model building[4]. In this context, the offset in question would account for a difference in population between countries, turning the analysis from a count to a rate over time analysis. This offset would be specified in SAS's GENMOD procedure under a "Weight" statement, transformed by the link function of the Negative Binomial GEE being fit: the natural log. Consequently, parameter estimates would need to be exponentiated with natural base to translate to proper real-world interpretation.

The variables to be considered for inclusion in the model were Average Median Population Age, Average Diabetes Prevalence, Average Population Density, Average Number of Hospital Beds per Thousand, Concentration of Smokers, and Average Human Development Index ("average" here referring to monthly averages of each variable by country). It is from among these options, descriptive of general livelihood health within each country, that predictors will be chosen.

Models of some variation of the following will be fit, with number of additive terms depending on the number of the aforementioned covariates deemed prudent to include:

$$\eta_i = \beta_1 + \beta_2 \log(X_{i2}) + \beta_3 \log(X_{i3}) + \beta_4 \log(X_{i4}) + \beta_5 \log(X_{i5}) + \beta_6 \log(X_{i6})$$

where β_2 is the parameter corresponding to Average Median Population Age, β_3 to Average Diabetes Prevalence, etc.

It is worth noting, firstly, that for each of these four models, SAS produced the following:

Algorithm converged.

Above that barebones requirement, the measure of model fit itself will be quantified using a metric provided through SAS: the Quasi-Likelihood under the Independence model Criterion, or QIC. Similar to the familiar AIC, the aim is the minimization of this quantity: the smaller QIC, the better the model. It was under this framework that an ad hoc forward selection process was carried out, initializing the model with a single of the aforementioned predictors, and then cautiously incorporating others, until software output a model with the maximal number of predictors but smallest QIC (see Code Appendix Section 4.2 for an example of this process, carried out for the Post Vaccination Case Count model).

6.1 Pre-Vaccination COVID-19 Case Model

The model fitting process and procedures resulted in the series of output over the following pages, with a model for each case and death count within Pre- and Post-Vaccination time periods:

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	7.9179	2.0944	3.8129	12.0229	3.78	0.0002
avg_diabetes_prevale	0.5006	0.1109	0.2832	0.7179	4.51	<.0001
avg_median_age	-0.0873	0.0432	-0.1719	-0.0026	-2.02	0.0433
smokers_monthly	0.1130	0.0315	0.0512	0.1747	3.58	0.0003

The first model with minimum QIC identified was the above pre-vaccination case model, with (not yet exponentiated) parameter estimates and accompanying significance tests thereof all displayed in the above table (parameters within the "Estimate" column, and significance tests in the "Pr> $|Z|$" column). According to those tests, all predictors included in the model were significant predictors of countrywide COVID-19 case count: that is, Average Diabetes Prevalence, Average Median Age, and Smoker Concentration, all of which had associated p-values of less than a standard alpha-level significance of 0.05. In terms of parameter estimates, both Average Diabetes Prevalence and Smoker Concentration had positive estimates, indicating that a one unit increase those predictor values will result in an $e^{0.5006} = 1.6497108$ multiplicative increase in case count, and an $e^{0.1130} = 1.11963193$ likewise increase in a country's expected case count (respectively).

However, with somewhat more nuanced interpretation, it can be seen that Average Median Age (the predictor with the greatest p-value, and as such the most tenuous predictive significance), has a negative parameter estimate, resulting in the interpretation that a one unit increase in Average Median Age in a country will yield an $e^{-0.0873} = 0.9164$ multiplicative decrease in case count. While on the surface this appears counterintuitive, it is worth considering that, within the data itself, Average Median Age by country is held constant over time, which is likely not a fully accurate representation of actual trends. Likely census data was used at the initialization of the data set, and simply not updated. But COVID-19 is well known for having a more profound effect the elderly populations within countries – especially within

nursing homes – so holding such a dynamic variable constant is a simple setback of the data, and must be taken into account for a representative analysis. Furthermore, this model was produced for all countries in general, over any given time period. Due to the reduction of elderly populations over time, the mortality rate may have actually had a significant enough effect to bring down overall case count, due to there being significantly fewer surviving individuals within the country's age group.

Overall, in the pre-vaccination era, case rate was predominantly determined by a country's Average Diabetes Prevalence and Concentration of Smokers (referred to at the time, all too familiar now, as "pre-existing conditions"), with an effect of Median Age that was dynamic over time, unadjusted as the model is for elderly death rates.

6.2 Post-Vaccination COVID-19 Case Model

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	8.0216	1.5777	4.9294	11.1137	5.08	<.0001
avg_median_age	0.0592	0.0321	-0.0037	0.1221	1.85	0.0650
avg_diabetes_prevale	0.3008	0.0704	0.1629	0.4387	4.28	<.0001
avg_hosp_bed_per_tho	-0.1298	0.0354	-0.1993	-0.0604	-3.66	0.0002
smokers_monthly	0.0368	0.0289	-0.0199	0.0934	1.27	0.2033
avg_population_densi	0.0019	0.0011	-0.0002	0.0040	1.80	0.0722

In the post-vaccination era, the preceding model output is the combination of predictors (nearly all of them) that successfully minimized QIC. Within this model, however, only two of the predictors have significant p-values: Average Diabetes Prevalence, and Average Hospital Bed per Thousand (although it is worth noting that Average Median Age and Average Population Density have p-values very close to the 0.05 threshold for significance).

Average Diabetes Prevalence having a significant effect (per one unit increase, an $e^{.3008} = 1.35094$ multiplication on case count) is due likely to similar "pre-existing conditions" as discussed in the previous model. Average Hospital Bed per Thousand's significant effect is notably negative, indicating an $e^{-0.1298} = 0.87827$ factor decrease in case count per one unit increase, probably attributable to the confounding influence of more hospital beds in a country accompanying better healthcare and public health systems in general.

6.3 Pre-Vaccination COVID-19 Death Model

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	4.1372	2.0375	0.1438	8.1305	2.03	0.0423
avg_diabetes_prevale	0.4131	0.1047	0.2078	0.6183	3.94	<.0001
avg_median_age	-0.0682	0.0420	-0.1504	0.0141	-1.62	0.1043
smokers_monthly	0.1107	0.0297	0.0525	0.1688	3.73	0.0002

In the above minimal QIC model for pre-vaccination deaths, there are only two significant terms (excluding Intercept): Average Diabetes Prevalence and Concentration of Smokers, both with associated p-values of less than a standard alpha level significance of 0.05. Average Median Age appears statistically insignificant here (likely for the secondary reasoning detailed in the pre-vaccination case model, wherein it is acknowledged that the above analysis is pooled over time, with significant proportions of countries' elderly populations incurring fatal infections in the early months of the pandemic, leading to their fatality incidence decreasing due to diminished population count and thusly leveling over the whole time period). With positive parameter estimates, it appears as if a one unit increase in Average Diabetes Prevalence by country will result in an $e^{0.4131} = 1.5115$ increase in mean predicted case count, and a similar increase in Concentration of Smokers will result in an $e^{0.1107} = 1.1171$increase in that same quantity.

In light of the above interpretations, it seems the indicators of overall countrywide lifestyle health – as opposed to merely age – were the most significant indicators of a country's COVID-19 death incidence, pre-vaccination.

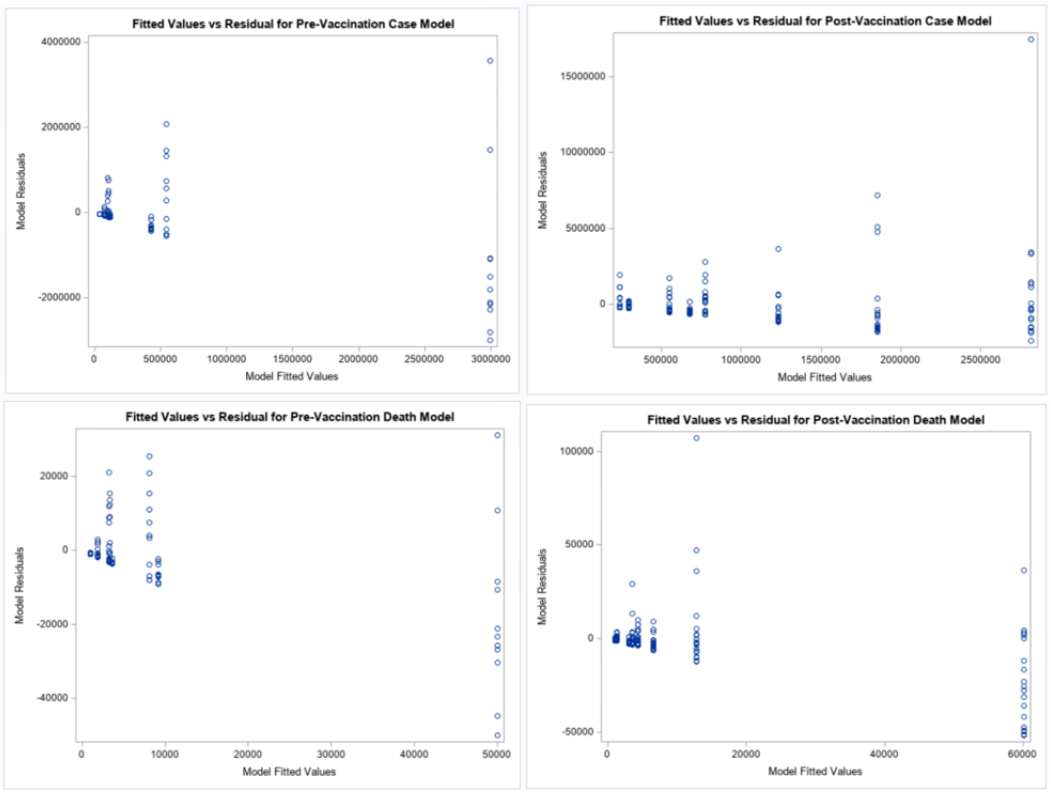
6.4 Post-Vaccination COVID-19 Death Model

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	3.5782	1.3358	0.9600	6.1963	2.68	0.0074
avg_diabetes_prevale	0.4454	0.0767	0.2951	0.5957	5.81	<.0001
avg_median_age	-0.0266	0.0269	-0.0794	0.0262	-0.99	0.3236
smokers_monthly	0.0921	0.0242	0.0446	0.1396	3.80	0.0001
avg_hosp_bed_per_tho	-0.1397	0.0359	-0.2100	-0.0694	-3.90	<.0001

Finally, the minimum QIC post-vaccination death model is summarized above. With three significant predictors, it still appears the factors of Average Diabetes Prevalence and Concentration of Smokers, as well as the newly pertinent Average Hospital Beds per Thousand (excluded from the pre-vaccination model due to a significant increase in QIC upon its inclusion), are all significant factors. Average Median Age still has a p-value too high to be significant, due likely to previously discussed reasoning. Interpretations require natural exponentiation, as before, and it can be observed that the Average Hospital Bed per Thousand variable has a negative parameter estimate, suggesting that an increase in that predictor will result, commonsensically, in a reduction of COVID-19 deaths.

7 Assessment of Fitted Values vs Residuals

Scatterplots of model fitted values versus model residual values for all four models are provided below:



All residual plots are quite clearly centered at/around zero.

It is also worth noting that heteroskedasticity is quite clearly present in all four models, which is only further validation of the analytical decision towards a brand of generalized linear model: methodologies like ANOVA or MANOVA have model assumptions involving not only distributional normality, but also constancy of variance among residuals must be met. Clearly, it is not here - all the more reason to use the analytical methods previously implemented. A further longitudinal decomposition (perhaps into six- or even three-month intervals) may be prudent to examine that dependence structure.

8 Conclusions

After an extensive and exhaustive data cleaning stage in Excel and R, filtering the data down to a usable format and combining time measurement variables into monthly bins, the data was read into SAS for analysis. Following the computation of summary statistics on COVID-19 case and death counts and a time series plot of both variables by country through time, and an assessment of time independent distributional structure – suggesting the Poisson methods characteristic to count data, but then being refined to Negative Binomial to account for evident overdispersion – lead after some misdirection and deliberation to the decision to implement a Negative Binomial GEE in order to most accurately model this complicated data. As a tradeoff to gaining a model that could account for the clearly present dependence in time-based observations by country, any time-dependent potential covariates had to be excluded from analysis, resulting in models with minimized QIC structured around some combination of variables Median Age, Concentration of Smokers, Population Density, and Average Hospital Bed per Thousand.

From the four models, built on cases and deaths in the pre- and post-vaccination era, certain overall conclusions became evident: Average Diabetes Prevalence was unilaterally significant in every case, a strong indication of the damage that particular pre-existing condition can do to both the transmission of the virus, and its fatality rate; when Average Hospital Bed per Thousand was deemed fit to appear in the model, it was also always significant, generally telling as it was of the impact a country's healthcare and public health resources had on viral impact; Concentration of Smokers seemed to minimally influence a country's case rate (especially pre-vaccination), but seemed to quite strongly influence a country's death rate, suggesting that such behaviors harmful to a person's lungs can make them much more susceptible to serious consequence from the illness, should they contract it. The other aforementioned variables of Median Age and Population proved not consistently significant to either metric, likely due to a homogenizing effect of time on a country's median age corresponding to early death rates, or a similar effect of rural and urban locations balancing each other's incidence.

In conclusion, this exercise as both a data familiarity and analysis process has yielded variables with significant predictive value in COVID-19 case and death rates. It is these and other analyses that have given us, and will continue to give us, the capability to combat such epidemiological threats as a species, and have taken us to back to the functioning "new normal" we have currently achieved in our fight against the coronavirus, saving lives and penetrating the fog of unknown data with no structure through which the last nearly 3 years have been spent walking.

9 Citations

1. Covid-19 Data Explorer. Our World in Data. (n.d.). Retrieved August 22, 2022, from

https://ourworldindata.org/explorers/coronavirus-data-explorer?zoomToSelection=true&time=2020-03-01..latest&facet=none&pickerSort=asc&pickerMetric=location&Metric=Confirmed%2Bcases&Interval=7-day%2Brolling%2Baverage&Relative%2Bto%2BPopulation=true&Color%2Bby%2Btest%2Bpositivity=false&country=USA~GBR~CAN~DEU~ITA~IND

2. Katella, K. (2022, July 5). Omicron, Delta, Alpha, and more: What to know about the coronavirus variants. Yale Medicine. Retrieved August 22, 2022, from

https://www.yalemedicine.org/news/covid-19-variants-of-concern-omicron#:~:text=1.1.as%20a%20variant%20of%20concern.

3. Owid.(n.d.). Covid-19-data/public/data at master · owid/covid-19-DATA. GitHub. Retrieved August 22, 2022, from

https://github.com/owid/covid-19-data/tree/master/public/data

4. Patrick Breheny April 11 - University of Iowa. (n.d.). Retrieved August 23, 2022, from

https://myweb.uiowa.edu/pbreheny/uk/teaching/760-s13/notes/4-11.pdf

10 Appendix of Code

10.1 R Code (Data Cleaning)

```
library(readxl)
owid_covid_data_filtered_2=
read_excel("C:/Users/Jon Javor/Desktop/owid_covid_data_filtered_2.xlsx")
#View(owid_covid_data_filtered_2)

keep=c("location","date","new_cases","new_deaths","icu_patients",
       "hosp_patients", "hospital_beds_per_thousand",
       "stringency_index","population","median_age",
       "gdp_per_capita","extreme_poverty","diabetes_prevalence",
       "female_smokers","male_smokers","handwashing_facilities",
       "human_development_index","cardiovasc_death_rate","new_tests",
       "new_tests_smoothed", "people_fully_vaccinated", "new_vaccinations",
       "new_vaccinations_smoothed")
data_filtered_2=owid_covid_data_filtered_2[keep]
#View(data_filtered_3)

vax_vs_smooth_vax=cbind(data_filtered_2$new_vaccinations,
                        data_filtered_2$new_vaccinations_smoothed)

test_vs_smooth_test=cbind(data_filtered_2$new_tests,
                          data_filtered_2$new_tests_smoothed)

data_filtered_3=subset(data_filtered_2,
                       data_filtered_2$location!="Africa")
data_filtered_4=subset(data_filtered_3,
                       data_filtered_3$location!="World")

smokers=data_filtered_4$female_smokers+
        data_filtered_4$male_smokers
data_filtered_5=cbind(data_filtered_4,smokers)

data_filtered_6=subset(data_filtered_5,select=
                       -c(female_smokers,
                          male_smokers,
                          handwashing_facilities))
#View(data_filtered_6)
```

```
for(i in 1:nrow(data_filtered_6)){
    month_number=substr(data_filtered_6$date[i],5,8)
    data_filtered_6$month_number[i]=month_number
    year=substr(data_filtered_6$date[i],1,4)
    data_filtered_6$year[i]=year
    }

monthsnum=c("-01-","-02-","-03-","-04-","-05-","-06-","-07-","-08-","-09-",
            "-10-","-11-","-12-")
monthsname=c("January","February","March","April","May","June","July","August",
             "September","October","November","December")
data_filtered_6$month=factor(gsub("\\s*","",data_filtered_6$month_number),
                             levels=monthsnum,labels=monthsname)
data_filtered_6$bin=paste(data_filtered_6$month,data_filtered_6$year)

data_filtered_6point5=
transform(data_filtered_6, variants=
          ifelse(data_filtered_6$date>='2020-02-01'&
                 data_filtered_6$date<='2020-11-30',"None",
                 ifelse(data_filtered_6$date>='2020-12-01'&
                        data_filtered_6$date<='2021-06-30',"Alpha",
                        ifelse(data_filtered_6$date>='2021-07-01'&
                               data_filtered_6$date<='2021-11-30',"Delta",
                               ifelse(data_filtered_6$date>='2021-12-01'&
                               data_filtered_6$date<=as.Date('2022-06-24'),
                               "Omicron", "NULL")))))

data_filtered_7=subset(data_filtered_6point5,select=
                       -c(month_number,year,month))

#View(data_filtered_7)

data_filtered_8=subset(data_filtered_7,
                       data_filtered_7$bin!="January 2020")

#View(data_filtered_8)

#new variable definition#

#new_cases_monthly#
```

```
data_filtered_8$new_cases[is.na(data_filtered_8$new_cases)]=0

new_cases_monthly=aggregate(data_filtered_8$new_cases,
                            by=list(location=data_filtered_8$location,
                            bin=data_filtered_8$bin),FUN=sum)

names(new_cases_monthly)[names(new_cases_monthly)=='x']='new_cases_monthly'

#new_deaths_monthly#

data_filtered_8$new_deaths[is.na(data_filtered_8$new_deaths)]=0

new_deaths_monthly=aggregate(data_filtered_8$new_deaths,
                             by=list(location=data_filtered_8$location,
                             bin=data_filtered_8$bin),FUN=sum)

names(new_deaths_monthly)[names(new_deaths_monthly)=='x']='new_deaths_monthly'

#icu_patients_monthly_rounded#

icu_patients_monthly=aggregate(data_filtered_8$icu_patients,
                               by=list(location=data_filtered_8$location,
                               bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)

names(icu_patients_monthly)[names(icu_patients_monthly)=='x']=
     'icu_patients_monthly'

icu_patients_monthly$icu_patients_monthly_rounded=
                     round(icu_patients_monthly$icu_patients_monthly,digits=0)

#hosp_patients_monthly_rounded#

hosp_patients_monthly=aggregate(data_filtered_8$hosp_patients,
                                by=list(location=data_filtered_8$location,
                                bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)

names(hosp_patients_monthly)[names(hosp_patients_monthly)=='x']=
     'hosp_patients_monthly'

hosp_patients_monthly$hosp_patients_monthly_rounded=
                      round(hosp_patients_monthly$hosp_patients_monthly,digits=0)
```

```
#stringency_index_monthly#

stringency_index_monthly=aggregate(data_filtered_8$stringency_index,
                                   by=list(location=data_filtered_8$location,
                                   bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)

names(stringency_index_monthly)[names(stringency_index_monthly)=='x']=
     'stringency_index_monthly'

#avg_population#

avg_population=aggregate(data_filtered_8$population,
                                  by=list(location=data_filtered_8$location,
                                  bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)

names(avg_population)[names(avg_population)=='x']=
     'avg_population'

#avg_median_age#

avg_median_age=aggregate(data_filtered_8$median_age,
                         by=list(location=data_filtered_8$location,
                         bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)

names(avg_median_age)[names(avg_median_age)=='x']='avg_median_age'

#avg_gdp_per_capita#

avg_gdp_per_capita=aggregate(data_filtered_8$gdp_per_capita,
                             by=list(location=data_filtered_8$location,
                             bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)

names(avg_gdp_per_capita)[names(avg_gdp_per_capita)=='x']='avg_gdp_per_capita'

#avg_extreme_poverty#

avg_extreme_poverty=aggregate(data_filtered_8$extreme_poverty,
                              by=list(location=data_filtered_8$location,
                              bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)
```

```
names(avg_extreme_poverty)[names(avg_extreme_poverty)=='x']=
     'avg_extreme_poverty'

#avg_diabetes_prevalence#

avg_diabetes_prevalence=aggregate(data_filtered_8$diabetes_prevalence,
                                  by=list(location=data_filtered_8$location,
                                  bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)

names(avg_diabetes_prevalence)[names(avg_diabetes_prevalence)=='x']='
      avg_diabetes_prevalence'

#avg_human_development_index#

avg_human_development_index=aggregate(data_filtered_8$human_development_index,
                                      by=list(location=data_filtered_8$location,
                                      bin=data_filtered_8$bin),FUN=mean,
                                      na.rm=TRUE)

names(avg_human_development_index)[names(avg_human_development_index)=='x']=
     'avg_human_development_index'

#cardiovasc_death_rate_monthly#

cardiovasc_death_rate_monthly=aggregate(data_filtered_8$cardiovasc_death_rate,
                                        by=list(location=data_filtered_8$location,
                                        bin=data_filtered_8$bin),FUN=mean,
                                        na.rm=TRUE)

names(cardiovasc_death_rate_monthly)[names(cardiovasc_death_rate_monthly)=='x']=
     'cardiovasc_death_rate_monthly'

#smokers_monthly#

smokers_monthly=aggregate(data_filtered_8$smokers,
                          by=list(location=data_filtered_8$location,
                          bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)

names(smokers_monthly)[names(smokers_monthly)=='x']='smokers_monthly'
```

```
#start_of_month#

data_filtered_8$reformatted_date=format.Date(data_filtered_8$date, "%Y-%d-%m")
data_filtered_8$start_of_month_check=grepl("-01-",data_filtered_8$reformatted_date)
start_of_month_date=data_filtered_8[data_filtered_8$start_of_month_check==TRUE, ]

keep2=c("location","bin","reformatted_date")
start_of_month=start_of_month_date[keep2]

#avg_hosp_bed_per_thousand#

avg_hospital_bed=aggregate(data_filtered_8$hospital_beds_per_thousand,
                           by=list(location=data_filtered_8$location,
                           bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)

names(avg_hospital_bed)[names(avg_hospital_bed)=='x']='avg_hosp_bed_per_thousand'

#avg_population_density#

avg_population_dense=aggregate(data_filtered_8$population_density,
                               by=list(location=data_filtered_8$location,
                               bin=data_filtered_8$bin),FUN=mean, na.rm=TRUE)

names(avg_population_dense)[names(avg_population_dense)=='x']=
     'avg_population_density'

#total_new_tests#

data_filtered_8$new_tests[is.na(data_filtered_8$new_tests)]=0

new_tests_monthly=aggregate(data_filtered_8$new_tests,
                            by=list(location=data_filtered_8$location,
                            bin=data_filtered_8$bin),FUN=sum)

names(new_tests_monthly)[names(new_tests_monthly)=='x']='total_new_tests'

#total_vaccinations#

data_filtered_8$people_fully_vaccinated[is.na(
                                        data_filtered_8$people_fully_vaccinated)]=0
```

```
start_of_month_date$full_vaccinations=
subset(data_filtered_8$people_fully_vaccinated,
       data_filtered_8$start_of_month_check==TRUE)

keep3=c("location","bin","full_vaccinations")
start_of_month_vaccinations=start_of_month_date[keep3]
#View(start_of_month_vaccinations)

#data_filtered_8$new_vaccinations[is.na(data_filtered_8$new_vaccinations)]=0

#new_vaccinations_monthly=aggregate(data_filtered_8$new_vaccinations,
                                    by=list(location=data_filtered_8$location,
                                    bin=data_filtered_8$bin),FUN=sum)

#names(new_vaccinations_monthly)[names(new_vaccinations_monthly)=='x']=
      'total_vaccinations'
#View(new_vaccinations_monthly)

#data_filtered_8$total_tests[is.na(data_filtered_8$total_tests)]=0
#start_of_month_date$total_tests=subset(data_filtered_8$total_tests,
                                        data_filtered_8$start_of_month_check==TRUE)

#keep4=c("location","bin","total_tests")
#start_of_month_tests=start_of_month_date[keep4]
#View(start_of_month_tests)

#predominant_variant

start_of_month_date$variant=subset(data_filtered_8$variants,
                                   data_filtered_8$start_of_month_check==TRUE)

keep4=c("location","bin","variant")
variants_monthly=start_of_month_date[keep4]

#end of new variable definition#

#View(data_filtered_8)

install.packages("plyr")
library(plyr)
```

```
data_binned=join_all(list(start_of_month,new_cases_monthly,new_deaths_monthly,
                          icu_patients_monthly, hosp_patients_monthly,
                          stringency_index_monthly, avg_population,
                          avg_median_age,avg_gdp_per_capita,avg_extreme_poverty,
                          avg_diabetes_prevalence,avg_human_development_index,
                          cardiovasc_death_rate_monthly,smokers_monthly,
                          new_tests_monthly,start_of_month_vaccinations,
                          variants_monthly, avg_hospital_bed,avg_population_dense),
                          by=c('location','bin'), type='left')

data_binned_2=subset(data_binned,select=-c(icu_patients_monthly,
                     hosp_patients_monthly))

names(data_binned_2)[names(data_binned_2)=='icu_patients_monthly_rounded']=
     'icu_patients_monthly'

names(data_binned_2)[names(data_binned_2)=='hosp_patients_monthly_rounded']=
     'hosp_patients_monthly'

data_binned_3=subset(data_binned_2,
                     data_binned_2$location!="South America")

data_binned_4=subset(data_binned_3,select=-c(icu_patients_monthly,
                                             hosp_patients_monthly))

data_binned_5=subset(data_binned_4,select=-avg_extreme_poverty)

data_binned_6=subset(data_binned_5,
                       data_binned_5$location!="China")
#View(data_binned_6)

#check for missing data to correct#
#which(unlist(is.nan(data_binned_6$new_cases_monthly)))
#which(unlist(is.nan(data_binned_6$new_deaths_monthly)))
#which(unlist(is.nan(data_binned_6$stringency_index_monthly)))
#which(unlist(is.nan(data_binned_6$avg_population_density)))
#which(unlist(is.nan(data_binned_6$avg_median_age)))
#which(unlist(is.nan(data_binned_6$avg_gdp_per_capita)))
#which(unlist(is.nan(data_binned_6$avg_diabetes_prevalence)))
#which(unlist(is.nan(data_binned_6$avg_human_development_index)))
#which(unlist(is.nan(data_binned_6$cardiovasc_death_rate_monthly)))
```

```
#which(unlist(is.nan(data_binned_6$smokers_monthly)))

data_binned_6$stringency_index_monthly[is.nan(
                                        data_binned_6$stringency_index_monthly)]=
                                        39.46875

data_binned_6$reformatted_date=as.Date(data_binned_6$reformatted_date,
                                        format="%Y-%m-%d")

data_binned_7=subset(data_binned_6,
                     data_binned_6$location!='Russia')

pre_date_check=data_binned_7$reformatted_date<='2021-01-03'

data_binned_7$full_vaccinations_filtered=
              ifelse(pre_date_check, data_binned_7$full_vaccinations,
                     ifelse(data_binned_7$full_vaccinations==0, NA,
                            data_binned_7$full_vaccinations))

vaccine_date_check=data_binned_7$reformatted_date<'2021-01-01'

data_binned_7$vaccine_status=ifelse(vaccine_date_check,'Before','After')

data_binned_7$scaled_new_cases=
    data_binned_7$new_cases_monthly/data_binned_7$avg_population

data_binned_7$scaled_new_deaths=
    data_binned_7$new_deaths_monthly/data_binned_7$avg_population

safricafeb2020=c('South Africa','February 2020','2020-01-02',0,0,2.78,60041996,
                 27.3,12294.876,5.52,.709,200.38,NA,0,0,'None',2.32,46.754,0,
                 'Before',0,0)

transpose=(t(safricafeb2020))

colnames(transpose)=colnames(data_binned_7)

data_binned_7=rbind(data_binned_7,transpose)

data_binned_7=data_binned_7[order(data_binned_7$location,
              data_binned_7$reformatted_date),]
```

```
data_binned_7$time=rep(seq(1,29,1),8)

install.packages("writexl")
library(writexl)

write_xlsx(data_binned_7,"C:/Users/Jon Javor/Desktop/final_data.xlsx")

pre_vax_data=subset(data_binned_7,
                    data_binned_7$reformatted_date<'2021-01-01')

post_vax_data=subset(data_binned_7,
                     data_binned_7$reformatted_date>='2021-01-01')

post_vax_data=subset(post_vax_data,select=-c(full_vaccinations)
pre_vax_data=subset(pre_vax_data,select=-c(full_vaccinations)

write_xlsx(pre_vax_data,"C:/Users/Jon Javor/Desktop/pre_vax_data.xlsx")
write_xlsx(post_vax_data,"C:/Users/Jon Javor/Desktop/post_vax_data.xlsx")
```

10.2 SAS Code (Data Analysis)

```
filename datain 'S:\windows\SASSY\pre_vax_data.csv';
proc import out=pre_vax_data
datafile=datain
dbms=csv replace;
getnames=Yes;
datarow=2;
run;

filename datain 'S:\windows\SASSY\post_vax_data.csv';
proc import out=post_vax_data
datafile=datain
dbms=csv replace;
getnames=Yes;
datarow=2;
run;

filename datain 'S:\windows\SASSY\final_data.csv';
proc import out=final_data
datafile=datain
dbms=csv replace;
getnames=Yes;
datarow=2;
run;

proc sgplot data=pre_vax_data noautolegend;
series x=bin y=scaled_new_cases / group=location;
title 'New Scaled COVID Cases Monthly by Country (Before Vaccine)';
xaxis label='Month; Year';
yaxis label='New Cases Scaled by Population' values=(0 to 0.08 by 0.02);
run;

proc sgplot data=post_vax_data noautolegend;
series x=bin y=scaled_new_cases / group=location;
title 'New Scaled COVID Cases Monthly by Country (After Vaccine)';
xaxis label='Month; Year';
yaxis label='New Cases Scaled by Population';
run;

proc sort data=final_data;
```

```
by vaccine_status;
run;

proc means data=final_data maxdec=2 mean var;
class location vaccine_status / descending;
var new_cases_monthly;
label new_cases_monthly=;
title 'COVID-19 Case Summary by Country and Vaccine Status';
run;

proc means data=final_data maxdec=6 mean std;
class variant;
var scaled_new_cases;
label scaled_new_cases =;
title 'COVID-19 Scaled Case Summary by Variant (Unscaled for Time)';
run;

proc sgplot data=pre_vax_data noautolegend;
series x=bin y=scaled_new_deaths / group=location;
title 'New Scaled COVID Deaths Monthly by Country (Before Vaccine)';
xaxis label='Month; Year';
yaxis label='New Deaths Scaled by Population' values=(0 to 0.0005 by 0.0001);
run;

proc sgplot data=post_vax_data noautolegend;
series x=bin y=scaled_new_deaths / group=location;
title 'New Scaled COVID Deaths Monthly by Country (Before Vaccine)';
xaxis label='Month; Year';
yaxis label='New Deaths Scaled by Population';
run;

data pre_vax_data_2;
set pre_vax_data;
scaled_new_tests=total_new_tests/avg_population;
log_population=log(avg_population);
run;

data post_vax_data_2;
set post_vax_data;
scaled_new_tests=total_new_tests/avg_population;
log_population=log(avg_population);
```

```
run;

proc means data=final_data mean var;
class location vaccine_status / descending;
var new_deaths_monthly;
label new_deaths_monthly=;
title 'COVID-19 Death Summary by Country and Vaccine Status';
run;

proc means data=final_data maxdec=8 mean std;
class variant;
var scaled_new_deaths;
label scaled_new_deaths =;
title 'COVID-19 Scaled Death Summary by Variant (Unscaled for Time)';
run;

proc sgpanel data=pre_vax_data;
panelby location / rows=2 columns=4;
histogram scaled_new_cases / binwidth=0.0007;
title 'Scaled New Monthly Cases (Before Vaccination)';
colaxis label='New Cases Scaled by Population';
run;

proc sgpanel data=post_vax_data;
panelby location / rows=2 columns=4;
histogram scaled_new_cases / binwidth=0.003;
title 'Scaled New Monthly Cases (After Vaccination)';
colaxis label='New Cases Scaled by Population';
run;

proc sgpanel data=pre_vax_data;
panelby location / rows=2 columns=4;
histogram scaled_new_deaths / binwidth=0.00002;
title 'Scaled New Monthly Deaths (Before Vaccination)';
colaxis label='New Deaths Scaled by Population';
run;

proc sgpanel data=post_vax_data;
panelby location / rows=2 columns=4;
histogram scaled_new_deaths / binwidth=0.00003;
title 'Scaled New Monthly Deaths (After Vaccination)';
```

```
colaxis label='New Deaths Scaled by Population';
run;

proc genmod data=pre_vax_data_2;
  class variant location time;
  model new_cases_monthly= avg_diabetes_prevalence avg_median_age
        smokers_monthly / type3 dist=negbin link=log;
  weight log_population;
  repeated subject=location / withinsubject=time;
  output out=out_pre_case predicted=fits resraw=residuals;
run;

proc sgplot data=out_pre_case noautolegend;
scatter x=fits y=residuals;
title 'Fitted Values vs Residual for Pre-Vaccination Case Model';
xaxis label='Model Fitted Values';
yaxis label='Model Residuals';
run;

proc genmod data=post_vax_data_2;
  class variant location time;
  model new_cases_monthly=avg_diabetes_prevalence / type3 dist=negbin link=log;
  weight log_population;
  repeated subject=location / withinsubject=time;
run;

proc genmod data=post_vax_data_2;
  class variant location time;
  model new_cases_monthly=avg_median_age avg_diabetes_prevalence
        avg_hosp_bed_per_thousand / type3 dist=negbin link=log;
  weight log_population;
  repeated subject=location / withinsubject=time;
run;

proc genmod data=post_vax_data_2;
  class variant location;
  model new_cases_monthly=avg_median_age
        avg_diabetes_prevalence avg_population_density
        avg_hosp_bed_per_thousand / type3 dist=negbin link=log;
  weight log_population;
  repeated subject=location / withinsubject=time;
```

```
run;

proc genmod data=post_vax_data_2;
  class variant location time;
  model new_cases_monthly=avg_median_age avg_diabetes_prevalence
        avg_hosp_bed_per_thousand smokers_monthly
        avg_population_density / type3 dist=negbin link=log;
  weight log_population;
  repeated subject=location / withinsubject=time;
  output out=out_post_case predicted=fits resraw=residuals;
run;

proc sgplot data=out_post_case noautolegend;
scatter x=fits y=residuals;
title 'Fitted Values vs Residual for Post-Vaccination Case Model';
xaxis label='Model Fitted Values';
yaxis label='Model Residuals';
run;

proc genmod data=pre_vax_data_2;
  class variant location time;
  model new_deaths_monthly=avg_diabetes_prevalence avg_median_age
        smokers_monthly / type3 dist=negbin link=log;
  weight log_population;
  repeated subject=location / withinsubject=time;
run;

proc sgplot data=out_pre_deaths noautolegend;
scatter x=fits y=residuals;
title 'Fitted Values vs Residual for Pre-Vaccination Death Model';
xaxis label='Model Fitted Values';
yaxis label='Model Residuals';
run;

proc genmod data=post_vax_data_2;
  class variant location time;
  model new_deaths_monthly=avg_diabetes_prevalence avg_median_age smokers_monthly
        avg_hosp_bed_per_thousand/ type3 dist=negbin link=log;
  weight log_population;
  repeated subject=location / withinsubject=time;
  output out=out_post_deaths predicted=fits resraw=residuals;
```

```
run;

proc sgplot data=out_post_deaths noautolegend;
scatter x=fits y=residuals;
title 'Fitted Values vs Residual for Post-Vaccination Death Model';
xaxis label='Model Fitted Values';
yaxis label='Model Residuals';
run;
```

11 Acknowledgements

I would like to thank Professor Joseph Consiglio for inspiring me to enter the statistical field, in a time when my way was less than clear to me, for being such a good professor for so many years of courses, and for both advising me on this project and sitting on my committee. I would also like to personally thank my other two committee members, Lili Tian and Changxing Ma, for taking the time to read over this paper and sit in on my presentation. To Professor Tian: thank you for all the wonderful conversations, and for believing in and encouraging my potential within the program, and supporting my decisions where I stumbled.

I would also like to thank Professor Dietrich Kuhlmann for being the professor that made me love statistics - my least favorite high school subject - from the first day of class, for serving as a reference more times than I can count, and for entrusting so many of his students to me in my time as a teaching assistant.

One final special thank you goes out to Samantha Brosius for her patient consultation on the methods within this project. In a chapter of life full of change, where I could never quite seem to get my footing, you gave me solid ground and clear eyes. Here's to the next.