Methods for early characterisation of the severity and dynamics of SARS-CoV-2 variants: a population-based time series analysis in South Africa

Emily Reichert, Beau Schaeffer*, Shae Gantt, Eva Rumpler, Nevashan Govender, Richard Welch, Andronica Moipone Shonhiwa, Chidozie Declan Iwu, Teresa Mashudu Lamola, Itumeleng Moema-Matiea, Darren Muganhiri, William Hanage, Mauricio Santillana, Waasila Jassat, Cheryl Cohen, David Swerdlow*

Summary

Background Assessment of disease severity associated with a novel pathogen or variant provides crucial information needed by public health agencies and governments to develop appropriate responses. The SARS-CoV-2 omicron variant of concern (VOC) spread rapidly through populations worldwide before robust epidemiological and laboratory data were available to investigate its relative severity. Here we develop a set of methods that make use of non-linked, aggregate data to promptly estimate the severity of a novel variant, compare its characteristics with those of previous VOCs, and inform data-driven public health responses.

Methods Using daily population-level surveillance data from the National Institute for Communicable Diseases in South Africa (March 2, 2020, to Jan 28, 2022), we determined lag intervals most consistent with time from case ascertainment to hospital admission and within-hospital death through optimisation of the distance correlation coefficient in a time series analysis. We then used these intervals to estimate and compare age-stratified case-hospitalisation and case-fatality ratios across the four epidemic waves that South Africa has faced, each dominated by a different variant.

Findings A total of 3 569 621 cases, 494 186 hospitalisations, and 99 954 deaths attributable to COVID-19 were included in the analyses. We found that lag intervals and disease severity were dependent on age and variant. At an aggregate level, fluctuations in cases were generally followed by a similar trend in hospitalisations within 7 days and deaths within 15 days. We noted a marked reduction in disease severity throughout the omicron period relative to previous waves (age-standardised case-fatality ratios were consistently reduced by >50%), most substantial for age strata with individuals 50 years or older.

Interpretation This population-level time series analysis method, which calculates an optimal lag interval that is then used to inform the numerator of severity metrics including the case-hospitalisation and case-fatality ratio, provides useful and timely estimates of the relative effects of novel SARS-CoV-2 VOCs, especially for application in settings where resources are limited.

Funding National Institute for Communicable Diseases of South Africa, South African National Government.

Lancet Microbe 2022; 3: e753–61

Published **Online**
August 31, 2022
https://doi.org/10.1016/S2666-5247(22)00182-3

*Contributed equally

Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, Boston, MA, USA (E Reichert BS, B Schaeffer MS, S Gantt MS, E Rumpler MS, W Hanage PhD, M Santillana PhD, D Swerdlow MD); **Division of Public Health Surveillance and Response** (N Govender MPH MSc, R Welch BSc, A M Shonhiwa MPH MSc, C D Iwu MPH, T M Lamola MPH, I Moema-Matiea MPH, D Muganhiri MPH, W Jassat FCPHM) **and Centre for Respiratory Diseases and Meningitis** (C Cohen PhD), **National Institute for Communicable Diseases, National Health Laboratory Service, Johannesburg, South Africa; School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa** (C Cohen)

Correspondence to:
Emily Reichert, Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
ereichert@hsph.harvard.edu

Introduction

The SARS-CoV-2 omicron (B.1.1.529) lineage (now BA.1, BA.1.1, and BA.2, among others) was first detected in Botswana and South Africa in specimens collected during early November, 2021.[1–3] In less than 2 months, omicron became the most prevalent variant globally, probably due to its increased growth rate and divergent spike protein structure relative to the delta (B.1.617.2) variant of concern (VOC). Rapid rates of case growth suggested that the doubling time for omicron, now estimated in the range of 1–2 days, is substantially shorter than for previous VOCs.[4] Assessing the extent to which growth rate will translate into public health impact depends upon multiple parameters, namely the magnitude and timing of severe disease following case ascertainment.

A surge in confirmed COVID-19 cases is followed temporally by one in hospital admissions and later, infection-associated deaths, as individuals take time to develop severe illness and seek care. For this reason, COVID-19-related hospital admissions and deaths are often referred to as lagging indicators of epidemic progression. Parameters describing the individual-level progression of the disease are often used to project the timing of these lags at the population level. Global estimates of the average time from symptom onset to hospitalisation range from 2 days to 10 days, and estimates of hospital length of stay for patients who die from infection are between 5 days and 19 days.[5–8] While multiple studies have explored the effect of age on these intervals, they have produced conflicting results,[5,8,9]

Research in context

Evidence before this study

We searched Google Scholar and medRxiv for articles published in English from database inception to Jan 1, 2022, with the search term "COVID-19" and additional terms such as "lag" and "time from cases to hospitalizations" or "time from cases to deaths". Although some studies have attempted to quantify this lag interval, few have done so stratified by age and variant, and none used lag intervals to approximate population-level severity metrics.

Added value of this study

The current study uses a simple, correlational time series analysis rather than a more complex mathematical model for ease of implementation during emergence of novel SARS-CoV-2 variants. It proposes quantifying COVID-19 case-to-hospitalisation and case-to-death lag intervals using surveillance data not only for research purposes, but also for timely and feasible estimation of the relative severity of novel variants, by age or other stratifying factors. This data-informed approach will allow for more accurate severity estimates before individual-level epidemiological studies are available, while also providing information regarding age-specific dynamics and the anticipated timing of hospitalisations and deaths following a surge in COVID-19 cases.

Implications of all the available evidence

An understanding of the morbidity and mortality associated with a novel SARS-CoV-2 variant is crucial to informing the public health response. Our proposed methodology is an accessible tool that can be added to currently deployed approaches for attaining a comprehensive understanding of the threat of a novel variant.

although some variation can be expected due to contextual differences in testing policies and health-seeking behaviour. Quantifying the relative intrinsic severity of a novel VOC is an even more complex process. Such estimates require extensive individual-level information that takes weeks or months to collect and often requires data linkage. Despite the challenges of estimating these parameters, they are paramount to developing an effective, tailored response to a novel variant.

More feasible to conduct are population-level analyses that can be performed with deidentified surveillance data. To our knowledge, very few studies have been conducted on the population-level lag intervals between detected COVID-19 cases, hospitalisations, and deaths, and those that have are not consistent, reporting that deaths lag COVID-19 cases by as little as 8 days and up to 4–6 weeks.[9–12] Estimates for this lag interval disseminating from media outlets are similarly unclear and unsubstantiated, with major sources reporting that deaths lag cases by "at least 3 weeks".[13] Clear, data-driven estimates of these lag intervals would allow for more accurate estimation of population-level disease severity metrics.

Here we develop a method to quantify and compare variant-specific disease severity using non-linked, population-level surveillance data from South Africa. We first perform a correlational analysis to quantify the lag interval between the confirmed COVID-19 case burden and hospital admissions, as well as within-hospital deaths. Using these estimates, we calculate age-specific case-hospitalisation ratios (CHRs) and case-fatality ratios (CFRs) for each variant-dominated wave. South Africa's robust nationwide surveillance system for confirmed COVID-19 cases, hospital admissions, deaths, and genomic surveillance is well suited to this analysis.

Methods

Data source

Data used in this study were collected as part of national COVID-19 surveillance efforts by the National Institute for Communicable Diseases (NICD) in South Africa. Deidentified data summarising daily confirmed COVID-19 cases, hospital admissions, and within-hospital deaths for individuals with COVID-19 were shared through a data-sharing agreement between the Harvard T.H. Chan School of Public Health study authors and NICD. The DATCOV surveillance system captures all patients with a positive COVID-19 test result who are admitted or die, regardless of primary diagnosis or cause of death. Age-stratified data for the duration of the pandemic to date (March 2, 2020, to Jan 28, 2022) were obtained. Although age was provided in single-year increments, we categorised age into broader, clinically relevant categories (0–17, 18–29, 30–39, 40–49, 50–64, 65–74, and ≥75 years) to limit strata with small case numbers and facilitate comparison with severity estimates from other settings. Cases, hospitalisations, and deaths with missing age data were excluded. If no specimen collection date was available for case data, the specimen received date was used as a proxy. Confirmed COVID-19 cases include individuals with a positive laboratory-reported PCR or, beginning in November, 2020, antigen test for SARS-CoV-2; except for the integration of antigen tests, the COVID-19 case definition has remained constant throughout the pandemic in South Africa.

No ethical approval was required for this analysis, which uses only aggregate, deidentified surveillance data.

Analytical approach

Each of the four epidemic waves seen in the country over the past 2 years has been quite homogeneous in terms of the primary variant driving infections, making between-wave comparisons appropriate. We defined the beginning (t_0) of each epidemic wave as day 1 (Sunday) of the first epidemiological week in which case growth (using a 7-day rolling average) exceeded a 5% increase from the previous week for at least 3 weeks. We defined the end (t_1) of each epidemic wave as day 7 of the epidemiological week that was 4 weeks beyond the point at which case counts had stabilised or returned to t_0 levels, in an effort to capture all wave-associated hospital admissions and deaths. For omicron we used all available data, as cases have neither returned to t_0 levels nor stabilised at the time of writing. Using these definitions, we obtained the date bounds for each wave (appendix p 2). Monthly national genomic surveillance data were referenced to support these date bounds; 5761 (88%) of 6547 sequences in November, 2020, to March, 2021, were beta (B.1.351), 15 276 (79%) of 19 337 sequences in May to October, 2021, were delta, and 9514 (97%) of 9808 sequences in November, 2021, to January, 2022, were omicron. Robust sequencing data were not available from early 2020 to evaluate the proportion of sequences that were D614G.

We calculated the distance correlation coefficient value for a range of case-hospitalisation (1–20 days), case-death (1–25 days), and hospitalisation-death (1–25 days) lag intervals deemed reasonable on the basis of the existing literature. The distance correlation coefficient (dCor) of two vectors *X* and *Y* takes the general form below:

$$\mathrm{dCor}(X,Y) = \frac{\mathrm{dCov}(X,Y)}{\sqrt{\mathrm{dVar}(X)\mathrm{dVar}(Y)}}$$

Full equations for deriving the distance correlation and its component functions, the distance covariance (dCov) and variance (dVar), have been described previously.[14] Methods for calculating p values with a corresponding t-test of independence (here, $\alpha=0\cdot05$ significance level) have also been described previously.[15] Lag intervals with the maximum value of dCor were defined as optimal, as values closest to 1 indicate the highest dependency between leading and lagging indicators. Distance correlation was used instead of the Pearson product-moment correlation coefficient to allow for non-linear relations, ideal for time series data in which dependencies might exist in arbitrary dimensions.

Two interpretable measures of disease severity, the CHR and CFR, were then calculated for a given day by dividing the number of COVID-19-related hospital admissions or deaths *D* days in the future by the current 7-day average COVID-19 case count, where *D* represents a lag interval ranging from 1 day to 25 days. Calculations involving daily case counts used a 7-day average to smooth out substantial declines in case-reporting on weekends. Furthermore, a within-hospital CFR was calculated by dividing the number of COVID-19-related deaths *D* days in the future by the current daily hospital admission count. CHRs and CFRs are often justly criticised for their reliance on case counts, a metric most vulnerable to under-reporting; a within-hospital CFR provides a useful severity measure for comparison since it relies only on hospitalisation and death counts. The population-wide CFR was included alongside the within-hospital estimate so that comparisons of severity inclusive of all individuals, not just those at highest risk of hospitalisation, could be made. CHRs and CFRs are not intrinsic measures of disease virulence, but rather effective, context-specific measures of a disease's severity.

CHRs, CFRs, and within-hospital CFRs were calculated for each day using the optimal lag interval, stratified by age and variant. Results were then summarised across waves by taking the geometric mean of the daily estimates for days between the t_0 and t_1 thresholds. Geometric means were chosen to dampen the effect of outlying values, and geometric standard deviations—a dimensionless, multiplicative factor—were used to quantify spread. Finally, age-standardised severity

See Online for appendix

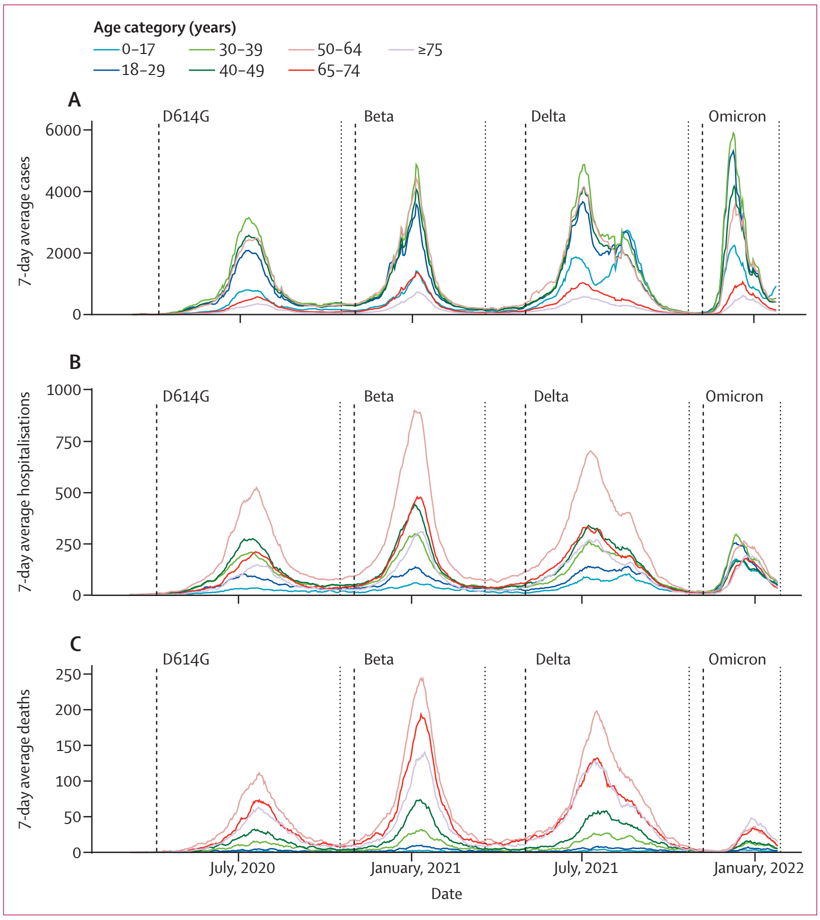

Figure 1: **Age stratified COVID-19 cases, hospitalisations, and deaths (7-day averages) in South Africa**
Data are displayed from the first detected presence of SARS-CoV-2 in South Africa (March 2, 2020) to Jan 28, 2022, by (A) specimen collection date, (B) hospital admission date, and (C) death date, and are aggregated across all nine South African provinces. Vertical lines indicate t_0 (dashed) and t_1 (dotted) dates used to define each variant-dominated wave.

measures were calculated by taking a weighted average of the age-specific estimates for each variant, weighted to the age distribution of confirmed COVID-19 cases throughout the entirety of the pandemic to date in South Africa (March 2, 2020, to Jan 28, 2022) to enable cross-variant comparisons. Cross-variant comparisons assume testing, hospitalisation protocols, and reporting quality remain similar over time.

Data cleaning and statistical analysis was performed using R 4.1.2 software. Distance correlation coefficients were calculated using the dcor function within the R package energy.[16] Code used to produce all analyses and figures is available online.

For the **code** see https://github.com/emreichert13/covid19lag

Sensitivity analysis

To evaluate the accuracy and timeliness of this method, we conducted a sensitivity analysis using what data would have been available at 5-day increments post-t_0 for the omicron and delta VOC-dominated waves. This retrospective way of evaluating the method's performance in real time assumes there is no reporting lag (ie, all outcomes that occur on day 7 are available on day 7). Age-stratified severity measures from our proposed methods were compared with a conventional method, defined simply by dividing the cumulative lagging indicator (here, deaths) by the cumulative number of cases available at each time point without taking a lag interval into account. Both methods were compared with the overall CFR estimates calculated using the conventional method between t_0 and t_1, once all outcomes had theoretically been allowed to accumulate.

Role of the funding source

The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

In total, 3 603 618 cases, 496 648 hospitalisations, and 100 150 deaths attributable to COVID-19 were detected in South Africa between March 2, 2020, and Jan 28, 2022. Missing age data led to the exclusion of a small fraction of confirmed cases (33 397 [0·9%]), hospitalisations (2462 [0·5%]), and deaths (196 [0·2%]), leaving a total of 3 569 621 cases, 494 186 hospitalisations, and 99 954 deaths. No specimen collection date was available for 1847 [0·05%] of 3 603 644 cases, and so the specimen received date was used as a proxy for these cases. The data were representative of all nine provinces and roughly evenly composed of cases captured by private sector (1 878 335 [52·6%] of 3 569 621) and public sector (1 691 286 [47·4%] of 3 569 621) laboratories. Four discernible epidemic waves, each dominated by a single variant—D614G, beta, delta, and omicron—account for most of the cases.[17] Age-stratified epidemic curves for 7-day averaged daily COVID-19 cases, hospital admissions, and deaths (figure 1) show the national burden of each variant. Specific t_0 and t_1 timepoints used to define each wave, as well as the total confirmed COVID-19 cases, hospital admissions, and deaths associated with each wave, can be found in the appendix (p 2).

The four distinct epidemic waves seen in South Africa make its national surveillance data well suited for

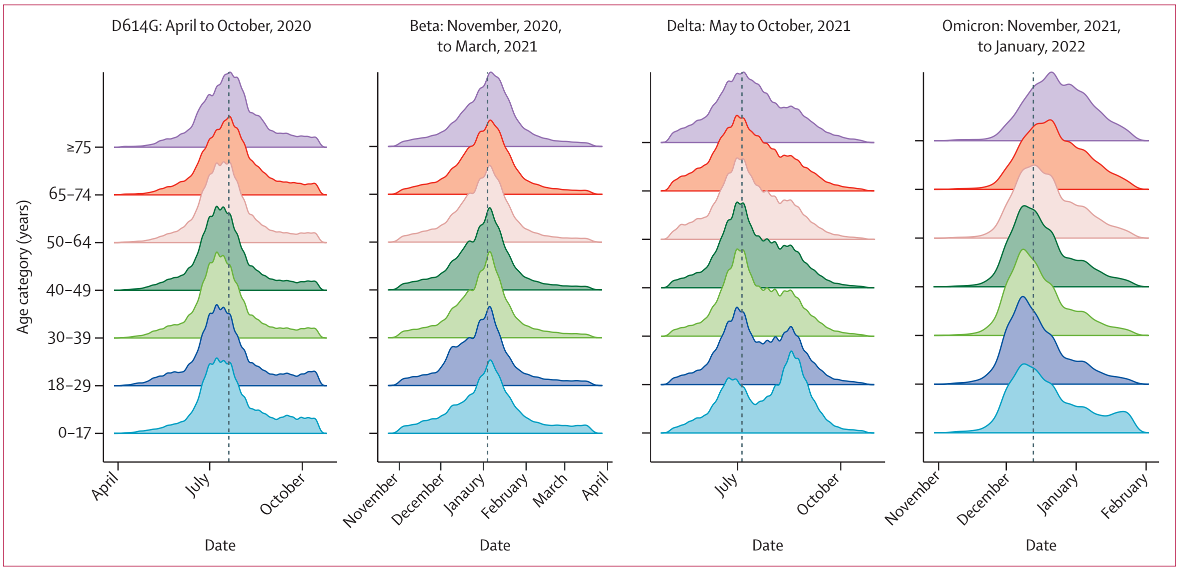

Figure 2: **Smoothed normalised density distribution of confirmed COVID-19 case counts over time, by age category and variant-dominated wave**
Vertical dashed lines indicate dates on which the daily reported COVID-19 case count, aggregated across age categories, peaked for each wave. During the omicron-dominated surge, infections appeared to first spread through younger age strata and gradually infiltrate older groups, albeit on a shorter timescale than previous variants of concern. For the delta-dominated wave, we noted a secondary peak dominant in the youngest age group (<18 years), corresponding to a return to school. In contrast, for the beta-dominated epidemic wave, case counts appeared to rise and fall almost uniformly across age divisions.

analyses comparing the morbidity and mortality of each variant. However, differences in the age structure of COVID-19 cases or hospitalisations over time, as well as differences in reporting, might confound these comparisons. Figure 2 was produced to further investigate the age structure of reported COVID-19 cases over time. Although difficult to discern from the epidemic curves in figure 1, we now note that the timing of infections differs substantially across age strata, within and across epidemic waves. Subsequent analyses were stratified by age to produce more informative comparisons.

For each of the variant-dominated waves, we conducted a time series analysis to measure the

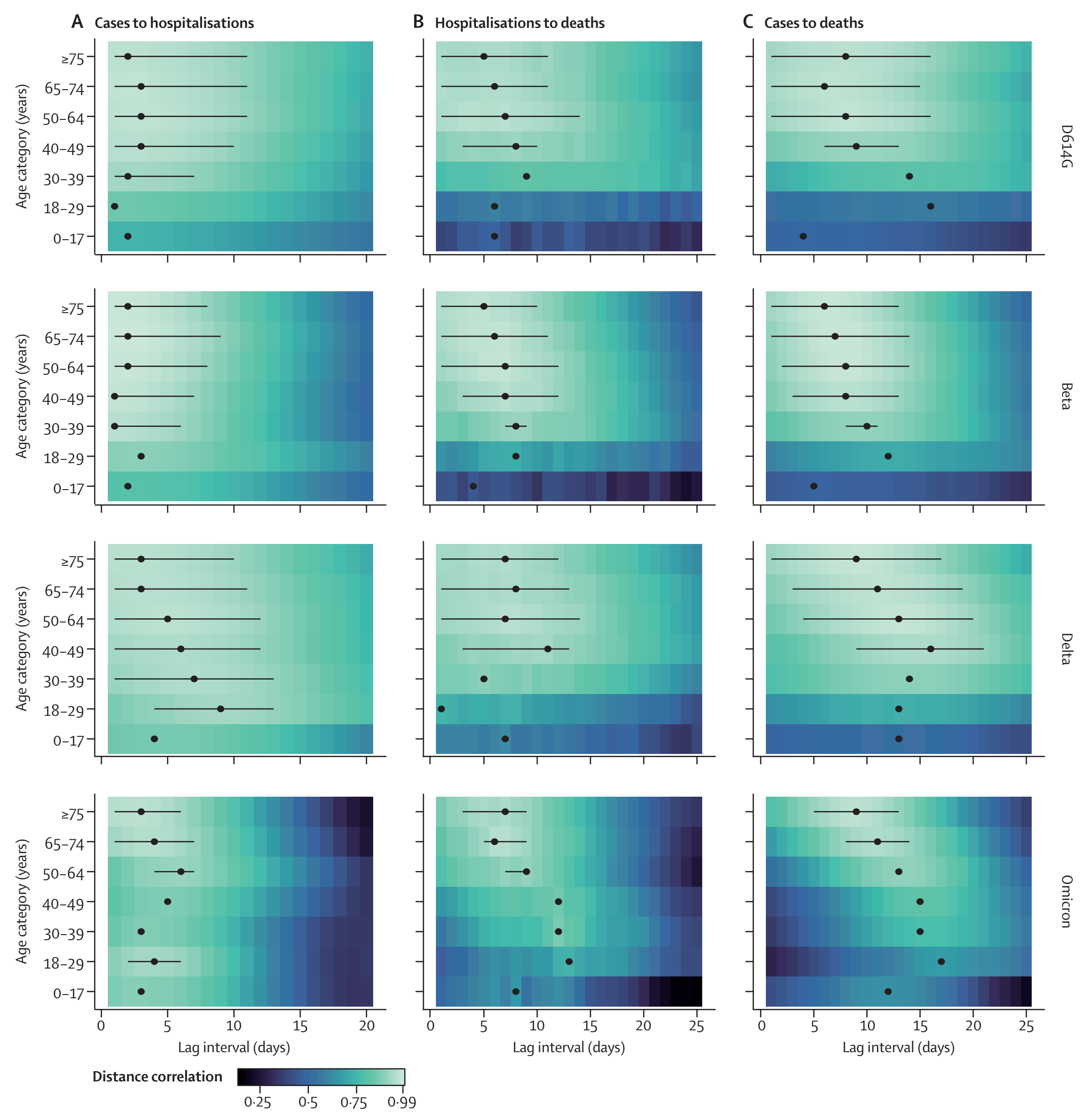

Figure 3: **Distance correlation coefficients by lag interval, stratified by age category and variant-dominated wave**
Heatmaps of distance correlations between (A) 7-day average COVID-19 confirmed cases and daily hospital admissions for lag intervals of 1–20 days, by variant and age category; (B) daily COVID-19 hospital admissions and daily within-hospital deaths for lag intervals of 1–25 days, by variant and age category; (C) 7-day average COVID-19 confirmed cases and daily within-hospital deaths for lag intervals of 1–25 days, by variant and age category. Distance correlation is bound between 0 and 1, and is 0 if and only if the leading and lagging indicators are independent. Optimal lag intervals, defined by the maximum value of the distance correlation coefficient, are indicated by black dots for each age category and variant. Apparent differences in lag intervals between variants and age categories are on the magnitude of days, not weeks. The black lines span lag intervals for which the distance correlation coefficient is ≥0·90. Plots are faceted vertically by the COVID-19 indicators being compared, and horizontally by variant-dominated waves. Each row represents an age category and each column a lag interval in days.

strength of the association between COVID-19 case volume and hospital admissions, for a range of lag intervals (figure 3A). The optimal lag time, defined by the maximum distance correlation coefficient, represents the number of days the epidemic curve of COVID-19 cases should be shifted forward to optimise its alignment with hospital admissions. This analysis was repeated for the correlation between (1) hospital admissions and within-hospital deaths (figure 3B), and (2) case volume and within-hospital deaths (figure 3C). Optimal lag intervals determined for each VOC and age category (appendix p 3) were all derived from statistically significant dCor values ($p<0{\cdot}0001$) and were shorter on average for some variants (D614G and beta) than others. An inverse association between the lag interval and age was observed across variants, with the decrease in lag as age increases most evident for cases-to-deaths. Hospital admissions generally lagged cases by less than 1 week, whereas deaths lagged cases by 1–2 weeks.

CHRs, CFRs, and within-hospital CFRs for each age and variant category are shown in the appendix (pp 4–5). Figure 4 visualises the dependency of severity metrics on the selected lag interval, with optimal lag intervals obtained from our results in figure 3 marked on the graph to indicate our best severity estimates. Of note, the period by which hospitalisations and deaths are lagged from confirmed case data has a substantial impact on severity estimates, particularly for the omicron-dominated wave in which cases rose and fell rapidly.

Although biases in case reporting probably inflate severity estimates, assuming that these biases remain relatively constant over time, we can compare these morbidity and mortality measures across variants (figure 5). As expected, reductions in CHR, CFR, and within-hospital CFR were observed nearly consistently for omicron compared with previous variants, with reductions in morbidity and mortality appearing most pronounced in the oldest age groups. For example, compared with the most recent delta VOC, the CFR for omicron appeared to stay relatively constant in youth (≤17 years), to decrease monotonically by 24% to 73% in age strata encompassing those aged between 18 and 64 years, and to stabilise around a 60–70% reduction in those 65 years and older. CHR and within-hospital CFR

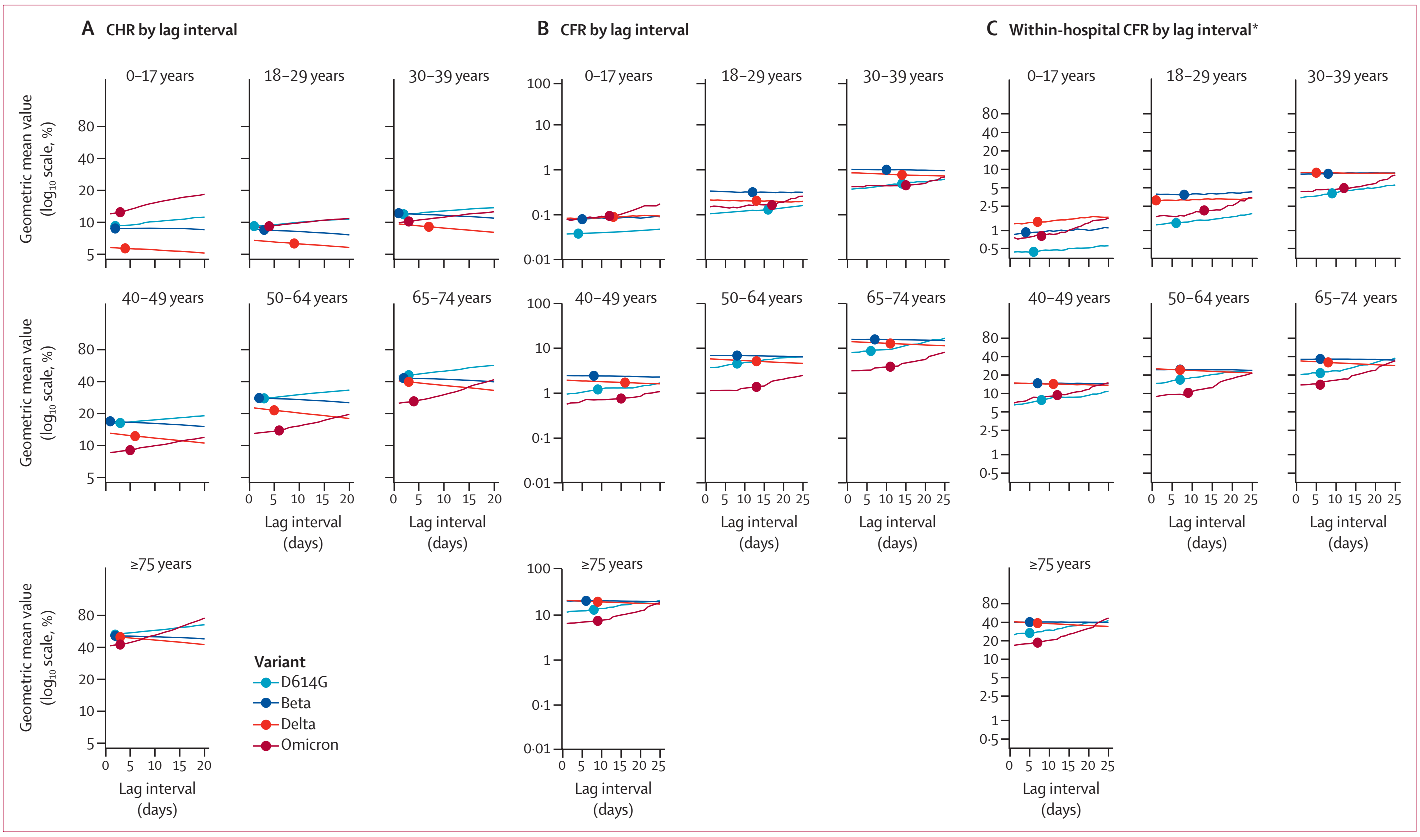

Figure 4: **Age-stratified CHRs, CFRs, and within-hospital CFRs by variant**

Age-specific CHR and CFR values were calculated using a range of lag intervals (1–20 days and 1–25 days, respectively) for each variant-dominated wave. Optimal lag intervals, as defined by the maximum value of the distance correlation coefficient, and their corresponding CHR and CFR values are denoted with a point for each age category and variant. Geometric mean values of CHR and CFR are plotted for each potential lag interval after performing the calculations across the entirety of each epidemic wave (t_0 to t_1). CHR=case-hospitalisation ratio. CFR=case-fatality ratio. *Within-hospital CFR refers to the estimated CFR among patients with COVID-19 admitted to hospital.

		Percent change from D614G			Percent change from beta			Percent change from delta		
		CHR	CFR	Within-hospital CFR*	CHR	CFR	Within-hospital CFR*	CHR	CFR	Within-hospital CFR*
Beta	Overall†	–0·9%	67·1%	61·0%						
	0–17 years	–5·2%	100·0%	108·9%						
	18–29 years	–7·7%	**146·2%**	**189·5%**						
	30–39 years	2·4%	104·1%	108·6%						
	40–49 years	3·7%	100·8%	87·8%						
	50–64 years	0·7%	50·2%	46·5%						
	65–74 years	–6·2%	80·9%	69·9%						
	≥75 years	–2·4%	57·3%	50·4%						
Delta	Overall†	–21·8%	33·3%	53·6%	–21·1%	–20·2%	–4·6%			
	0–17 years	–38·4%	**125·0%**	**208·9%**	–35·0%	12·5%	47·9%			
	18–29 years	–31·3%	61·5%	134·6%	–25·6%	–34·4%	–19·0%			
	30–39 years	–23·7%	57·1%	117·2%	–25·5%	–23·0%	4·1%			
	40–49 years	–24·8%	42·1%	82·8%	–27·5%	–29·2%	–2·7%			
	50–64 years	–22·8%	12·5%	44·5%	–23·4%	–25·1%	–1·3%			
	65–74 years	–13·4%	46·8%	49·4%	–7·7%	–18·9%	–12·0%			
	≥75 years	–5·0%	50·2%	44·8%	–2·6%	–4·5%	–3·7%			
Omicron	Overall†	–30·9%	–53·6%	–26·9%	–30·3%	–72·2%	–54·6%	–11·6%	–65·2%	–52·4%
	0–17 years	35·0%	**125·0%**	82·2%	42·4%	12·5%	–12·8%	**119·2%**	0·0%	–41·0%
	18–29 years	–0·3%	23·1%	60·2%	8·0%	–50·0%	–44·7%	45·1%	–23·8%	–31·7%
	30–39 years	–14·3%	–8·2%	21·9%	–16·3%	–55·0%	–41·6%	12·3%	–41·6%	–43·9%
	40–49 years	–44·5%	–37·2%	20·2%	–46·5%	–68·7%	–36·0%	–26·2%	–55·8%	–34·2%
	50–64 years	–50·2%	–69·7%	–38·9%	–50·5%	–79·9%	–58·3%	–35·5%	–73·1%	–57·8%
	65–74 years	–43·3%	–55·2%	–35·5%	–39·6%	–75·2%	–62·0%	–34·5%	–69·5%	–56·8%
	≥75 years	–19·3%	–43·5%	–30·5%	–17·3%	–64·1%	–53·8%	–15·1%	–62·3%	–52·0%

Figure 5: **Percent change in estimated CHRs, CFRs, and within-hospital CFRs across variant-dominated waves**
Severity metrics are calculated separately for each age category and variant using the lag interval defined by the optimal distance correlation coefficient. Green indicates an improvement (or decrease) in the severity estimate relative to the variant of comparison, whereas red indicates a worsening (or increase) in the severity estimate. The colour intensity corresponds to the magnitude of the percent change. CHR=case-hospitalisation ratio. CFR=case-fatality ratio. *Within-hospital CFR refers to the estimated CFR among patients with COVID-19 admitted to hospital. †Overall refers to age-standardised severity estimates, an aggregate measure weighted to the distribution of cases in each age category throughout the pandemic in South Africa (March 2, 2020, to Jan 28, 2022) to enable cross-variant comparisons.

reductions for omicron relative to previous variants followed similar patterns across age strata.

Finally, sensitivity analysis revealed that estimates obtained using our proposed methods began to stabilise earlier on in the progression of each variant-dominated wave compared with those obtained from conventional methods that do not account for the lag interval (appendix pp 6–7). These estimates not only stabilised but also converged to the conventional CFR measure calculated post-wave using all available data from t_0 to t_1. The sensitivity analysis supports the utility of this method for rapid, real-time assessment of a novel VOC's characteristics.

Discussion

The COVID-19 pandemic has highlighted how a limited understanding of emerging variant's characteristics can challenge public health responses. Using surveillance data of the SARS-CoV-2 epidemic waves in South Africa, we characterise age-specific estimates for the lag time between fluctuations in COVID-19 cases, hospital admissions, and within-hospital deaths. We then use the optimal lag intervals for each age strata to produce variant-specific, population-level severity estimates. Despite only stratifying by age, relative results by variant are consistent with those recently published, estimating the same metrics with adjustment for additional individual-level covariates.[18–20]

Our work highlights that applying parameters from previous strains of SARS-CoV-2 or not accounting for differences in age structures across VOC-dominated waves might lead to biases in projecting future hospitalisation and death toll trajectories. For VOCs with rapid growth rates such as omicron—which lead to a rapid rise and fall in cases, hospitalisations, and deaths—characterising accurate lag intervals is all the more important; overestimating these lag intervals can greatly distort disease severity estimates. The methods we outline here prove feasible and accurate using limited real-time data relative to those that do not take the lag interval into account.

Our proposed methodology faces several key limitations that must be acknowledged. First, as we compare results across SARS-CoV-2 epidemic waves and age groups, any change over time in case ascertainment, testing, or reporting policy might bias results. Relying on confirmed COVID-19 cases, which are chronically under-reported, almost certainly inflates disease severity metrics; here, we assume cases are under-reported at the same rate over time, enabling comparison across epidemic waves. However, the extent to which this assumption is true for a specific setting must be evaluated, and an upward bias in CHRs and CFRs (the extent of which is also context-specific, dependent on the comprehensiveness of case surveillance) must be acknowledged.[21,22] Death counts can also be affected by under-reporting and are estimated to be around three times higher in South Africa than reported, according to excess mortality estimates.[23] Our proposed methods are best suited for comparison of severity measures to those of previous VOCs in the same setting; absolute severity estimates should not be overinterpreted and must be accompanied by a consideration of the data limitations in the setting of interest. In South Africa, national reporting of both positive PCR and antigen tests for SARS-CoV-2 makes for a more robust denominator of confirmed cases. However, differences in testing recommendations over time, as well as testing strategies (targeted *vs* mass testing) which vary at the provincial level, might introduce biases.[24] South Africa's national surveillance data rely on reporting from both the public and private sector, which might also introduce variability in terms of patient populations, testing and admission policies, and in-hospital management.

We additionally acknowledge the uncaptured, differential effect that vaccination roll-out might have had on our findings. Vaccination did not become widely available until mid-2021, after the D614G and beta waves had swept through the country, and is still not available for those younger than 12 years.[25] It is reasonable that immunity derived from both vaccination and previous infection contributed to the reduced disease severity of omicron we observed for most age strata, and the exception we see in youth (≤17 years) might reflect this cohort's limited protection from vaccination.

It is unknown what proportion of hospitalised patients were admitted primarily due to COVID-19 versus what proportion were incidental hospitalisations, or primarily hospitalised for other conditions but subsequently identified as cases in screening. In late 2020, subsequent to the D614G wave, South Africa shifted to routine testing of all admitted patients, probably increasing the proportion of incidental hospitalisations. The risk of nosocomial transmission is not trivial either; we cannot quantify the extent of hospital-acquired infections in our data, which might also increase over time due to increased transmissibility of later VOCs such as omicron. We acknowledge that a high volume of incidental hospitalisations might lead to a downward bias in lag interval estimates for the time from cases to hospitalisations and an upward bias in CHR point estimates, and that this proportion might differ with successive waves and by age group.

Finally, as individual identifiers for linking COVID-19 cases, hospitalisations, and deaths were not available, we aggregated three different datasets containing these age-stratified daily counts. One must avoid extrapolating results to individual-level questions (eg, how long is the average time interval from COVID-19 diagnosis to death?), as we can characterise only population-level dynamics from these data. No information on the estimated time of infection or symptom onset was available, which also constrained our ability to study the individual-level progression of disease. Recently published CHR and within-hospital CFR estimates for each of the four variants in South Africa based on individual-level data (and adjusted for age, sex, race, comorbidities, public or private sector, and province) vary by up to 8·3% from our age-standardised estimates; however, relative reductions in these severity metrics across VOCs are highly similar.[20]

It is unpredictable where the next SARS-CoV-2 VOC will first be detected; the virus is agnostic to which countries are best equipped to quickly respond and collect data to inform disease severity estimates. Data infrastructure capable of collecting longitudinal individual-level data is key for pandemic preparedness, and our proposed methods do not replace the need for gold standard epidemiological case investigation; rather, we propose alternative methods that can be used to supplement existing approaches, particularly when time or data are limited. Here, using NICD's robust national surveillance data, we propose a unique set of methods that can be applied in real-time in a variety of settings to crudely estimate severity measures relative to previous VOCs using population-level surveillance data. Our proposed methods provide a starting point for parsing out information from epidemic curves and estimating age-specific indicators of severity to inform public health responses.

Contributors

NG, RW, AMS, CDI, TML, IM-M, DM, WJ, and CC were involved in data collection and curation. WH, MS, and DS can be credited with study conceptualisation, design, and supervision. ERe, BS, SG, and ERu conducted the formal data analysis and wrote the original draft. All authors participated in manuscript review and editing. NG, RW, AMS, CDI, TML, IM-M, DM, WJ, CC, ERe, BS, and SG accessed and verified the data. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

WH reports his position as a member of Biobot Analytics' scientific advisory board and has received stock options in Biobot Analytics, as well as payment for expert witness testimony on the expected course of the COVID-19 pandemic. DS reports previous employment at Pfizer before the initiation of this analysis and stock in Pfizer. DS also reports compensation for occasional 1 h blinded consultancies for several consulting companies over the past 36 months. MS reports receipt of institutional research funds from the Johnson and Johnson Foundation and from Pfizer; neither funder had a role in the design or content of the current manuscript. CC reports grant funding from the Wellcome Trust, South African Medical Research Council, US Centers for Disease Control and Prevention, and Sanofi Pasteur to institute COVID-19 research in the past 36 months, as well as a role on the scientific advisory committee for "BCHW: Burden of COVID-19 among health care workers, assessing infection, risk factors, vaccine effectiveness, working experiences and one-health implications: a mixed methodology, multisite international study". RW reports shareholding stock in the following health or pharmaceutical companies in South Africa, none related to this work: Adcock Ingram Holdings, Dischem Pharmacies, Discovery, Netcare, and Aspen Pharmacare Holdings. All interests listed are outside the current work. All other authors declare no competing interests.

Data sharing

Aggregate data of reported COVID-19 cases, hospital admissions, and attributable deaths by province, sector, ward of admission, and date are available to the public through NICD's DATCOV reports. Further third-party requests for research purposes require a data sharing agreement. Data inquiries can be directed to datcov19@nicd.ac.za. The R code used to implement the analysis is available online.

For the **code** see https://github.com/emreichert13/covid19lag

Acknowledgments

MS was partially funded by the National Institute of General Medical Sciences of the US National Institutes of Health under award number R01GM130668. The contents of this work are the sole responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health. The national COVID-19 case surveillance and DATCOV are national surveillance systems, funded by the NICD and the South African National Government. No additional funding was obtained towards the completion of this analysis and the development of this manuscript. We thank the South African National Department of Health for implementation support, the NICD for support and oversight, and the DATCOV team. We also thank the health professionals in all public and private sector hospitals and hospital groups submitting data to DATCOV.

References

1 European Centre for Disease Prevention and Control. Implications of the emergence and spread of the SARS-CoV-2 B.1.1.529 variant of concern (Omicron) for the EU/EEA. Stockholm: European Centre for Disease Control and Prevention, 2021.

2 Rambaut A, Holmes EC, O'Toole Á, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020; **5:** 1403–07.

3 Viana R, Moyo S, Amoako DG, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* 2022; **603:** 679–86.

4 Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet* 2021; **398:** 2126–28.

5 Faes C, Abrams S, Van Beckhoven D, Meyfroidt G, Vlieghe E, Hens N. Time between symptom onset, hospitalisation and recovery or death: statistical analysis of Belgian COVID-19 patients. *Int J Environ Res Public Health* 2020; **17:** E7560.

6 Pellis L, Scarabel F, Stage HB, et al. Challenges in control of COVID-19: short doubling time and long delay to effect of interventions. *Philos Trans R Soc Lond B Biol Sci* 2021; **376:** 20200264.

7 Vekaria B, Overton C, Wiśniowski A, et al. Hospital length of stay for COVID-19 patients: data-driven methods for forward planning. *BMC Infect Dis* 2021; **21:** 700.

8 Marschner IC. Estimating age-specific COVID-19 fatality risk and time to death by comparing population diagnosis and death patterns: Australian data. *BMC Med Res Methodol* 2021; **21:** 126.

9 Jin R. The lag between daily reported Covid-19 cases and deaths and its relationship to age. *J Public Health Res* 2021; **10:** 2049.

10 Testa CC, Krieger N, Chen JT, Hanage WP. Visualizing the lagged connection between COVID-19 cases and deaths in the United States: an animation using per capita state-level data (January 22, 2020 – July 8, 2020). HCPDS working paper volume 19, number 4. July 10, 2020. https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1266/2020/07/HCPDS-WP_19_4_testa-et-al_Visualizing-Lagged-Connection-Between-COVID-19-Cases-and-Deaths-in-US_final_07_10_with-cover.pdf (accessed Feb 8, 2022).

11 Institute for Health Metrics and Evaluation. COVID-19 projections. 2022. https://covid19.healthdata.org/global?view=cumulative-deaths&tab=trend (accessed March 30, 2022).

12 Ward T, Johnsen A. Understanding an evolving pandemic: an analysis of the clinical time delay distributions of COVID-19 in the United Kingdom. *PLoS One* 2021; **16:** e0257978.

13 Gamio L, Jones LW, Walker AS. How to think about covid data right now. The New York Times. Jan 7, 2022. https://www.nytimes.com/interactive/2022/01/07/us/covid-data-explained.html (accessed Feb 8, 2022).

14 Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *Ann Stat* 2007; **35:** 2769–94.

15 Székely GJ, Rizzo ML. The distance correlation t-test of independence in high dimension. *J Multivariate Anal* 2013; **117:** 193–213.

16 Rizzo M, Szekely G. Energy: E-Statistics: multivariate inference via the energy of data. 2022. https://CRAN.R-project.org/package=energy (accessed March 31, 2022).

17 Hodcroft EB. CoVariants: SARS-CoV-2 mutations and variants of interest. 2021. https://covariants.org/per-country (accessed Feb 8, 2022).

18 Ferguson N, Ghani A, Hinsley W, Volz E. Report 50: Hospitalisation risk for Omicron cases in England. Imperial College London. December, 2021. http://spiral.imperial.ac.uk/handle/10044/1/93035 (accessed Feb 10, 2022).

19 Wolter N, Jassat W, Walaza S, et al. Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study. *Lancet* 2022; **399:** 437–46.

20 Jassat W, Abdool Karim SS, Mudara C, et al. Clinical severity of COVID-19 in patients admitted to hospital during the omicron wave in South Africa: a retrospective observational study. *Lancet Glob Health* 2022; **10:** e961–69.

21 South African Government. Regulations and guidelines—coronavirus COVID-19. 2022. https://www.gov.za/covid-19/resources/regulations-and-guidelines-coronavirus-covid-19 (accessed Feb 25, 2022).

22 South African Government. Education–coronavirus COVID-19. https://www.gov.za/covid-19/individuals-and-households/education-coronavirus-covid-19 (accessed Feb 25, 2022).

23 South African Medical Research Council. Report on weekly deaths in South Africa. 2022. https://www.samrc.ac.za/reports/report-weekly-deaths-south-africa?bc=254 (accessed May 18, 2022).

24 National Institute for Communicable Diseases. Weekly testing summary. 2022. https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/surveillance-reports/weekly-testing-summary/ (accessed March 6, 2022).

25 National Institute for Communicable Diseases. COVID-19 vaccine rollout strategy FAQ. 2022. https://www.nicd.ac.za/covid-19-vaccine-rollout-strategy-faq/ (accessed Feb 25, 2022).