COVID-19 prediction in South Africa: Understanding the unascertained cases -- the hidden part of the epidemiological iceberg

Author List: Xuelin Gu[1], Bhramar Mukherjee[1,2], Sonali Das[3], Jyotishka Datta[4]*

[1] Department of Biostatistics, University of Michigan, Ann Arbor, USA

[2]Department of Epidemiology, University of Michigan, Ann Arbor, USA

[3]Department of Business Management, University of Pretoria, Pretoria, South Africa

[4]Department of Mathematical Sciences, University of Arkansas, USA

*Corresponding author: Department of Mathematical Sciences, University of Arkansas, USA, E-mail: jd033@uark.edu.

Abstract

Background: Understanding the impact of non-pharmaceutical interventions remains a critical epidemiological problem in South Africa that reported the largest number of confirmed COVID-19 cases and deaths from the African continent. **Methods**: In this study, we applied two existing epidemiological models, an extension of the Susceptible-Infected-Removed model (eSIR) and SAPHIRE, to fit the daily ascertained infected (and removed) cases from March 15 to July 31 in South Africa. To combine the desirable features from the two models, we further extended the eSIR model to an eSEIRD model. **Results**: Using the eSEIRD model, the COVID-19 transmission dynamics in South Africa was characterized by the estimated basic reproduction number (R_0) at 2.10 (95%CI: [2.09,2.10]). The decrease of effective reproduction number with time implied the effectiveness of interventions. The low estimated ascertained rate was found to be 2.17% (95%CI: [2.15%, 2.19%]) in the eSEIRD model. The overall infection fatality ratio (IFR) was estimated as 0.04% (95%CI: [0.02%, 0.06%]) while the reported case fatality ratio was 4.40% (95% CI: [<0.01%, 11.81%]). As of December 31, 2020, the cumulative number of ascertained cases and total infected would reach roughly 801 thousand and 36.9 million, according to the long-term forecasting. **Conclusions**: The dynamics based on our models suggested a decline of COVID-19 infection and that the severeness of the epidemic might be largely mitigated through strict interventions. Besides providing insights on the COVID-19 dynamics in South Africa, we develop powerful forecasting tools that allow incorporating ascertained rate and IFR estimation and inquiring into the effect of intervention measures on COVID-19 spread.

Key words: COVID-19; South Africa; forecasting; unascertained cases; underreporting factors; infection fatality ratio

Key Messages:

- This study delineated the COVID-19 dynamics in South Africa from March 15 to July 31 and confirmed the effectiveness of the main non-pharmaceutical intervention—lockdown, and mandatory wearing of face-mask in public places using epidemiological models;
- COVID-19 spread in South Africa was found to be associated with both low ascertained rate and low infection fatality ratio;
- According to the long-term forecast, by December 31, 2020, the cumulative number of ascertained cases and total infected would reach roughly 801 thousand and 36.9 million respectively.

Word count: 4036

1 Introduction

The coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was first detected in early December 2019 in Wuhan, China. The first case was confirmed in South Africa on March 5, 2020. As of October 20, 2020, there are 1,262,476 confirmed cases (cumulative total) and 28,601 deaths confirmed in Africa[1]. South Africa remains the 'epicenter of the outbreak in the African continent'[1] with the largest number of confirmed cases (800,872) and deaths (21,803), contributing to 53% of the total confirmed cases and 89% of deaths, while accounting for only 5% of population in Africa as of December 4, 2020[2]. Although, there is no seroprevalence survey result published on the population in South Africa to our knowledge, an antibody survey on 3,000 blood donors in Kenya, a Sub-Saharan African country, estimated 1.6 million people with SARS-CoV-2 antibodies by the end of July 2020[3], implying the possibility of a large degree of underreporting/undetected cases in Africa, including South Africa. Thus, understanding the key epidemiological constructs for COVID-19 outbreak is paramount for containing the spread of COVID-19 in South Africa, as well as explaining the disparity between seroprevalence estimates and reported number of cases.

1.1 Interventions: With a universal goal to 'flatten the curve', a series of non-pharmaceutical interventions were implemented by the government in South Africa, that have been gradually lifted since early May 2020[4]. On March 27, 2020, South Africa adopted a three-week nationwide hard-lockdown (level 5) along with closure of its international borders, which was extended to April 30, 2020. Thereafter, to balance the positive health effects of strict interventions against their economic costs[5], South Africa began a gradual and phased recovery of economic activities with the lockdown restriction eased to level 4[4], allowing inter-provincial travel only for essential services. From June 1, national restrictions were lowered to level 3 allowing for inter-provincial travel and school opening (**Table 1**). Face-mask wearing was mandatory in public places at all times, with limitations on gatherings, and sale of alcohol and cigarettes were restricted[6]. Although these interventions implemented at an early stage had a higher potential for pandemic containment, previous studies[6–9] reported a consistently large value for the estimated basic reproduction number (R_0) ranging from 2.2 to 3.2 in South Africa by models trained with data in relatively early time windows. Using data observed under various intervention scenarios over a longer period of time, we carry out a thorough investigation to assess the current COVID-19 spread and the effect of these interventions, which will provide valuable insights into the transition dynamics of COVID-19 and intervention deployment in South Africa, and beyond.

1.2 Unascertained cases and deaths: Based on the clinical characteristics of COVID-19, a majority of patients are symptomatic (roughly 84% according to a recent study[10]), most of whom have mild symptoms[11] and tend to not seek testing and medical care. While private hospitals have reached maximum capacity, public and field hospitals beds have still some margin left with additional challenges due to scarcity of staff[12]. Several recent studies[13–15] reported that a nonnegligible proportion of unascertained cases contributed to the quick spreading of COVID-19[7]. It is suggested that only 1 in 4 mildly ill cases would be detected in South Africa[16]. The relatively lower testing rate in South Africa (**Table 1; Figure 1(b)**) coupled with a very high positive rate of testing especially in July and August[17], suggests inadequacy of testing, as well as the possibility of a large unobserved number of unascertained cases[18]. The WHO situation report dated October 20 reports an addition of 429 retrospective deaths over just 7 days from mortality audits in South Africa further questioning the reliability of COVID-

19 mortality data[19]. Thus, modeling both ascertained and unascertained cases and deaths can measure infection fatality ratios (IFRs, the proportion of deaths among all infected individuals[20]) of COVID-19, leading to a better understanding of the clinical severity of the disease.

1.3 Epidemiological models: The Susceptible-Infectious-Removed or SIR model[21] is arguably the most commonly used epidemiological models for modeling the trajectory of an infectious disease. A recent extension of SIR, called extended-SIR or eSIR[22], was developed to incorporate user-specified non-pharmaceutical interventions and quarantine protocols into a Bayesian hierarchical Beta-Dirichlet state-space model, which was successfully applied to model COVID-19 dynamics in India[23]. One major advantage of this Bayesian hierarchical structure is that uncertainty associated with all parameters and functions of parameters can be calculated from posterior draws without relying on large-sample approximations[23]. Extending the simple compartment structure in eSIR model, the SAPHIRE model[24], delineated the full transmission COVID-19 dynamics in Wuhan, China with additional compartments by introducing unobserved categories[13]. In this article, we extended the eSIR approach to the eSEIRD model to combine the advantages of the two existing models, using a Bayesian hierarchical structure to introduce additional unobserved compartments and characterize uncertainty in critical epidemiological parameters including basic reproduction number, ascertained rate and IFR, with input data as observed counts for cases, recoveries and deaths. Furthermore, we applied these three models and compared the results of the eSEIRD model with the eSIR and SAPHIRE model, with the following primary objectives: (i) characterizing the COVID-19 dynamics from March 15 to July 31; (ii) evaluating the effectiveness of the main non-pharmaceutical intervention—lockdown, and mandatory wearing of face-mask in public places; (iii) capturing the uncertainty in estimating the ascertained rate and IFR; and (iv) forecasting the future of COVID-19 spread in South Africa.

2 Methods

2.1 Study Design and Data Source: COVID-19 data for South Africa were extracted from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[25] from the onset of the first 50 confirmed case (March 15) to November 29, 2020. We fitted the models using data up to July 31 and predicted the state of COVID-19 infection in South Africa in a short-term window, from August 1 to August 31, and a relatively long-term window up to December 31. To compare the model short-term prediction performance of different models, we used the symmetric mean absolute percentage error (SMAPE), given by: $\mathrm{SMAPE} = \frac{\mathbf{100\%}}{\boldsymbol{n}} \sum_{\boldsymbol{t}=\mathbf{1}}^{\boldsymbol{n}} \frac{|\boldsymbol{F_t}-\boldsymbol{A_t}|}{(|\boldsymbol{A_t}|+|\boldsymbol{F_t}|)/\mathbf{2}}$ where $\boldsymbol{A_t}$ is the observed value from August 1 to 31 and $\boldsymbol{F_t}$ is the forecast value in this time period. This design enabled us to select an optimal modeling strategy for South Africa data and check the robustness of prediction performance across different models.

2.2 Statistical Methodology: We considered two existing epidemiological methods, the eSIR and SAPHIRE model, and the extension of eSIR, *viz.*, the eSEIRD model, as described in Section 1.3. The infection transition schematic diagrams for the three models are shown in **Figure 2**.

Parameter settings: Table S.1.1 summarizes the list of notations and assumptions. We assumed a constant population size (N= 57,779,622) for all models and fixed a few transition parameters below in the SAPHIRE and eSEIRD model. First, we set an equal number of daily inbound and outbound travelers (n), in which $n = 4 \times 10^{-4}N$ from March 15 to 25 estimated by the number of international travelers to South Africa in 2018[26], otherwise $n = 0$ when border closed, *i.e.* after March 26. We fixed the transmissibility ratio between unascertained and ascertained cases at α=0.55 assuming lower transmissibility for unascertained cases[27], an incubation period of 5.2 days, and a pre-symptomatic infectious period of D_p=2.3 days[28,29], implying a latent period of D_e=2.9 days. The mean of total infectious period was D_i+D_p $= 5.2$ days[28], assuming constant infectiousness across the pre-symptomatic and symptomatic phases of ascertained cases[30], thus, the mean symptomatic infectious period was D_i=2.9 days. We set the period of ascertained cases from reporting to hospitalization D_q=7 days, the same as the median interval from symptom onset to admission reported[31,32]. The period from being admitted in hospital to discharge or death was assumed as $D_h = 8.6$ days[33].

Choice of Initial states: For the eSIR model, the prior mean for the initial infected/removed proportion was set at the observed infected/removed proportion on March 15, and that for the susceptible proportion was the total number of the population minus the infected and removed proportions[22].

For the SAPHIRE model, other than setting prior parameters for initial states, we set the number of initial latent cases $E(0)$ was the sum of those ascertained and unascertained cases with onset during March 15-17 as D_e=2.9 days[13] and the number of initial pre-symptomatic cases $P(0)$ was that from March 18-19 as D_p =2.3 days[13]. The number of ascertained symptomatic cases $I(0)$ was assumed as the number of observed infected cases on March 15 excluding $H(0), R(0)\ and\ D(0)$ (the initial numbers for hospitalized, recovered, and deaths). The initial ascertainment rate (r_0) was assumed as 0.10 as reported in literature[15,34], implying $A(0) = \frac{0.90}{0.10} I(0)$, and a sensitivity analysis with r_0=0.25 was conducted to address weak information for r_0 obtained in South Africa and variation of r_0 in different scenarios. $H(0)$ was assumed as 50% of the observed ascertained cases on March 9 (by assuming the period from reported to hospitalized was 7 days[31,32] at the early stage of the pandemic). In addition, we denoted $R(0)$ as the sum of observed recovered and death cases on March 15. The number of initial susceptible cases $S(0)$ was calculated as the total population (N) minus $E(0)$, $P(0)$, $I(0)$, $A(0)$ and $R(0)$.

In the eSEIRD model, we set the prior mean of initial ascertained, unascertained and hospitalized cases as $I(0)$, $A(0)$ and $H(0)$ discussed above. However, since the latent compartment incorporates the pre-symptomatic cases, the mean of the initial latent cases was set as the sum of those ascertained and unascertained cases with onset during March 15-19 as $D_e + D_p$ = 5.2 days[13].The prior mean of initial recoveries and deaths were fixed as the number of observed recovered and death cases on March 15, respectively. Therefore, the prior mean of initial susceptible compartment was set as the total population excluding the mean of other compartments.

Prior distributions: In the eSIR model, the log-normal priors were used for the removed rate ν and the basic reproduction number R_0, in particular $\nu \sim LogN(-2.955, 0.910)$, with $E(\nu) = 0.082$ and $SD(\nu) = 0.1$ [22], and $R_0 (= \frac{\beta}{\nu}) \sim LogN(0.582, 0.223)$ with $E(R_0) = 3.2$ and $SD(R_0) = 1$[23]. Flat Gamma priors were used for the scale parameters of the Beta-Dirichlet distributions as follows: $\omega \sim Gamma(2, 0.0001)$, $\lambda^I \sim Gamma(2, 0.0001)$ and $\lambda^R \sim Gamma(2, 0.0001)$ [23]. In the eSEIRD model, apart from same prior for $R_0 \left(= \beta \left[a(1-r)D_i + \frac{r}{\frac{1}{D_i} + \frac{1}{D_q}} \right] \right)$, the ascertained rate $r \sim Beta(10, 90)$[35], the priors for IFR for non-hospitalized cases $\kappa_1 \sim Beta(0.03, 2.93)$ and for hospitalized cases $\kappa_2 \sim Beta(0.44, 1.76)$ with mean equal to 0.1% and 20%, respectively[33]. In addition, to account for the effect of time-varying contact rate on the transmission rate, we set a time-varying contact rate modifier $\pi(t)$ in the eSIR and eSEIRD model, $\pi(t)$= 1 before lockdown, $\pi(t)$= 0.75 during strict lockdown and $\pi(t)$=0.90[16,23,36] after September 20 when the interventions were largely eased. Note that the modifier $\pi(t)$ is a conjectural quantity and hence must be guided by empirical studies[23]. Using MCMC sampling method for the eSIR and eSEIRD model, we set the adaptation number to be 10^4, thinned by 10 draws to reduce autocorrelation, and set a burn-in period of 5×10^4 draws under 1×10^5 iterations for 4 parallel chains.

We fit the SAPHIRE model in four time periods: March 15-March 26, March 27- April 30, May 1- May 31 and June 1- July 31, separated by the change-points of the lockdown strictness level, and denote the ascertained rate and transmission rate in the time periods as r_1, r_2, r_3, r_4, β_1, β_2, β_3 and β_4. We used $r_1 \sim Beta(10{,}90)$ and reparameterized r_2, r_3 and r_4 by

$$logit(r_2) = logit(r_1) + \delta_1$$

$$logit(r_3) = logit(r_2) + \delta_2$$

$$logit(r_4) = logit(r_3) + \delta_3$$

where $logit(r) = log\left(\frac{r}{1-r}\right)$. We assumed δ_1, δ_2 and $\delta_3 \sim N(0{,}1)$, and a non-informative prior for all transmission rates $\beta_1, \beta_2, \beta_3$ and $\beta_4 \sim Unif(0{,}2)$, to reflect lack of information about these hyperparameters[13]. Therefore, β and r were assumed to follow different distributions for these four time periods. Finally, the effective reproduction number was given by $R_e = \beta\left[aD_p + a(1-r)D_i + \frac{r}{\frac{1}{D_i}+\frac{1}{D_q}}\right]$. Posterior samples were drawn using the delayed rejection adaptive metropolis algorithm implemented in the R package *BayesianTools* (version 0.1.7). We set a burn-in period of 10^5 iterations and continued to run 10^5 iterations with a sampling step size of 10 iterations.

Methodology implementation details were given in the Supplementary section **S.1**, with a comparison between the four models in the Supplementary **S.2**. All analyses were conducted in *R* (version 4.0.0), and source codes are available at https://github.com/umich-cphds/south_africa_modeling. Posterior mean and corresponding 95% credible interval (95% CI) were reported for the parameters of interests.

3 Results

3.1 Reproduction number and intervention evaluation: The estimated posterior mean of R_0 was similar in the eSIR (2.05 (95%CI: [1.81,2.31])) and eSEIRD (2.10 (95%CI: [2.09,2.10])) model and robust when $r_0 = 0.25$ (**Table 2**). To evaluate the time-varying effect of non-pharmaceutical interventions, we evaluated the effective reproduction number (R_e) in different lockdown periods using SAPHIRE model and it demonstrated that R_e decreased dramatically from 3.47 (95%CI: [3.32,3.61]) before lockdown to 1.39 (95%CI: [1.36,1.41]) after lockdown implementation though still significantly above 1, suggesting that the effective contact rate decreased 60% in the lockdown time period. When lockdown was eased to a relatively less strict level in the latest two time periods, the R_e increased slightly to 1.43 (95%CI: [1.42,1.45]) under lockdown level 4 and 1.58 (95%CI: [1.57,1.58]) under lockdown level 3 (**Table 2**; **Figure 3(e)**).

3.2 Short-term and long-term forecasts: We forecasted the total cumulative number of infections, including unascertained cases, in the SAPHIRE model up to August 31 depending on the time-period considered for estimating the trend. The estimated cumulative number of infections was: (a) 24.8 million if the trend of the strict lockdown (level 5) was assumed, (b) 28.2 million with if the trend of the lockdown level 4 was assumed, and (c) 35.2 million if the trend of lockdown level 3 was assumed. All the short-term forecasts in SAPHIRE model were robust under different r_0 settings which contradicts the intuition to some degree that different situation in the early stage may lead to different trajectory of pandemic **(Table 3)**. The eSEIRD model also output the predicted total cumulative number of cases which was 32.0 million under $r_0 = 0.10$, or 28.9 million under $r_0 = 0.25$, and the total deaths counts as 22 or 19 thousand when $r_0 = 0.10$ or 0.25, respectively, by August 31 **(Table 3).** Furthermore, we used the eSEIRD model to forecast the epidemic trajectory for a relatively longer time period, where we found that by December 31, the cumulative number of ascertained cases and total infected would reach roughly 801 thousand and 36.9 million (which is around 60% of the total population in South Africa), respectively. The number of total deaths was forecasted as 28 thousand at the same time.

3.3 Fitting and prediction performance: All the three main models fitted the COVID-19 data in South Africa with high accuracy as the estimated daily new cases were close to the observed numbers (**Figure 3(a)-(c)**). However, the SAPHIRE model performed best in terms of predicting cumulative infected cases with the smallest SMAPE (1.81% for 15 days and 2.96% for 31 days when $r_0 = 0.10$) while the eSEIRD model had the second smallest SMAPE (4.78% for 15 days and 6.02% for 31 days when $r_0 = 0.10$) **(Table 4).** Therefore, for selected important time points, the predicted number of cumulative ascertained infected cases for the SAPHIRE and eSEIRD model were closer to the observed numbers compared with the eSIR model(**Table 3**). The predictive accuracies for the three candidate methods substantiate their credibility in terms of capturing the transmission dynamics for the time-period considered in this study.

3.4 Unascertained cases and deaths: As demonstrated by SAPHIRE modeling results in **Figure 3(d)**, the large number of unascertained and pre-symptomatic cases contributed to the rapid spread of disease. The estimated ascertained rates were very low: 9.53% (95% CI: [8.70%, 10.40%]), 1.85% (95% CI: [1.74%, 1.98%]), 2.21% (95% CI: [2.16%, 2.26%]), and 1.84% (95% CI: [1.82%, 1.86%]) in the four time periods evaluated, respectively **(Table 2; Figure 3(f)).** Specifically, in the latest three time periods after lockdown, the ascertained rates estimates were almost consistent with time. Similarly, in the eSEIRD model, the estimated ascertained rate was also at a very low level as 2.17% (95%CI: [2.15%, 2.19%]) **(Table 2).** As of August 31, the overall under-reported factor for the infected cases is estimated as 46 and 54 in the eSEIRD and SAPHIRE model, respectively.

By the eSEIRD model, the overall IFR was estimated as 0.04% (95%CI: [0.02%, 0.06%]) while the observed overall case fatality ratio was estimated as 4.40% (95% CI: [<0.01%, 11.81%]) **(Figure 4).** Furthermore, the eSEIRD model provided Bayesian estimates for IFR and deaths among hospitalized and non-hospitalized cases. The estimated IFR for the hospitalized cases was 12.06% (95% CI: [11.76%, 12.35%]) which was much higher than that for non-hospitalized cases (less than 0.01%), and these estimates were robust to the choice of initial ascertained rate r_0. The under-reporting factor for deaths was estimated very close to 1, suggesting that most deaths occurred in hospitals**.**

4 Discussion

This modeling study investigates the spread process of COVID-19 in South Africa, 'the hardest hit country on the African continent'[19], considering the unascertained cases and population movement in different time periods at the same time, and evaluating the effect of the intervention strategy employed. Moreover, our study provides powerful methodological tools to estimate the IFR and predict deaths due to COVID-19 by making use of the reported deaths.

The SAPHIRE model characterizes the transmission dynamics of COVID-19 in South Africa as follows: it spread rapidly in South Africa before lockdown with a large effective reproduction number comparable to that in the early stage in Wuhan without interventions[13]. The lockdown intervention and mandatory face-mask wearing in public places employed in South Africa seemed to contain the spread of COVID-19 effectively as the R_e decreased dramatically and it increased slightly due to the relaxation of lockdown stringency afterwards. However, the R_e was consistently above 1 throughout the whole period analyzed, which implies the interventions failed to dampen the transmission fully, further substantiated by the basic reproduction number estimates in the eSIR and eSEIRD model as well. To stop the pandemic or prevent the resurgence, more strict intervention policies, such as lockdown, mandatory face-mask wearing, are suggested based on these results taking account their potential economic costs at the same time[5,37].

The estimated ascertained rate is very low in South Africa compared to that reported for many other countries[13,15,35], also implied by the low testing rate and high testing positive rate in South Africa[17]. As of September 21, the number of total tests conducted is 4.0 million, suggesting that about 7% population were tested[17]. Furthermore, the estimated ascertained rate is consistent with that in other multiple global epicenters under severe pandemic of COVID-19, such as France, the United States, Italy and Spain in March[34]. The large number of unascertained cases may contribute significantly to the rapid spread of COVID-19[27,38,39]. Therefore, even though the spread of COVID-19 is exhibiting an optimistic pattern of decline as indicated by the decay in daily ascertained cases starting at the end of July, with high probability, there is still a large number of active infectious cases as suggested by the low ascertained rate. Considering the unascertained infections, our findings suggest that there are roughly more than 40% of the total population in South Africa infected by July 31 and more than 60% by the end of year 2020. Our long-term forecasts for November 1 are much lower but closer to the observed numbers, compared to the long-term projection in NICD report in May[16], which also used a stochastic compartmental transmission model with a generalized SEIR structure accounting for disease severity and the treatment pathway, fitting early-stage data up to April 30[16]. For instance, as of November 1, the NICD report projected an estimated 3.4-3.7 million laboratory-confirmed cases, whereas the eSEIRD model prediction was 793 thousand, much closer to the observed count: 727 thousand confirmed cases. However, more surveillance testing and effective testing strategies under conditions of limited test availability, such as contact tracing of the contacts and confirmed cases, will be helpful to curtail the pandemic in South Africa[6].

Although highly transmissible and lowly ascertained, the COVID-19 IFR is estimated as 0.05% in South Africa, comparable to the estimates in other locations with similar low mortality rate based on serological data[40]. The low IFR may due to the entire South African population being relatively young such that decreases the fatal impact on general population to some extent[41]. Our estimates of the IFR of hospitalized cases are much higher than that for non-hospitalized cases, suggesting that the most severe cases may have been admitted to hospitals despite the relatively lack of the testing arrangements.

Comparison of the models: The eSIR and the SAPHIRE model have been successfully applied to the data in India and Wuhan, China, separately[22,31]. While SAPHIRE model still has a great robust prediction performance on COVID-19 cases, the eSIR model has relatively poor predictive capacity for capturing the change in the trend of the epidemic in time for neglecting some important clinical characteristics. The eSEIRD model has a comparable 15-day prediction performance to the SAPHIRE though relatively sensitive to the initial ascertained rate, which is more reasonable as the trajectory of pandemic would change with the number of initial infectious cases. Moreover, it is useful to measure the IFR of COVID-19 accurately accounting for the unascertained cases when evaluating the impact of pandemic.

Strengths and Limitations: Our research investigated and supported some important epidemiological and clinical characteristics of COVID-19 and estimated and projected the trend of the spread in South Africa accounting for some critical information obtained, such as the population movement and the prior distribution of the ascertained rate and IFR. It is worth noting that we provide useful statistical tools for predicting infections and deaths and accurately estimating for substantive parameters accounting for both the reported cases and deaths information at the same time.

However, there are some important limitations. First, the assumptions in the models were collected from previous reports from other countries because of the lack of such information for South Africa, especially the fixed values for hyper-parameters. Though the estimation of parameters and prediction of infections seem to be robust to these assumptions to some extent, the inference and prediction would be much more convincing when based on accurate information in South Africa using these statistical tools. Second, the ascertained rate was assumed to follow the same distribution across the whole time period in the eSEIRD model although it might be time-varying depending on the accumulating knowledge and deployment of clinical resources for COVID-19, given the spatial variation within South Africa regarding the population density and movement, as well as regarding location of COVID-19 hotspots and hospital resources. Further, the population density is highly heterogeneous in different regions in South Africa with higher concentration near high-density economic hub cities, such as Cape Town and Durban. COVID-19 cases are also diversely spread. For instant, Gauteng Province is a very small, highly dense province with roughly 30% of total cases in the nation, and 49% of confirmed cases cluster in KwaZulu-Natal, Eastern Cape and Western Cape Province. Without considering these heterogeneities and potential confounding factors in individual region, the conclusion on the national data might be biased. The burden of HIV and tuberculosis comorbidity, particularly among the less privileged socio-economic population, also adds to the complexity of analyzing the COVID-19 data from South Africa[42]. In addition, in this paper we implicitly assumed that the recovered cases would not be infected again but it is still inconclusive based on extant research for COVID-19[43]. It might lead to a resurgence if this assumption is not valid and the interventions are totally lifted. Thus, it may be needed to conduct some serological surveys on COVID-19 among the general population in South Africa to confirm the national, as well as provincial, seroprevalence and thus provide more powerful evidence to support the evolving benefits of nonpharmaceutical interventions decisions and of their uptake, furthermore, provide guidance to manage provincial level disparity.

Funding

This work was supported by grants from the National Science Foundation [grant numbers DMS- 1712933 (to B.M.) and DMS-2015460 (to J.D.)] and from National Institute of Health [grant number 1 R01 HG008773-01 (to B.M.)].

Acknowledgements

The first and second author (X.G. and B.M.) would also like to thank the Center for Precision Health Data Sciences at the University of Michigan School of Public Health, The University of Michigan Rogel Cancer Center and the Michigan Institute of Data Science for internal funding that supported this research.

Reference

1. Coronavirus Disease (COVID-19) Situation Reports. Accessed August 14, 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

2. Johns Hopkins University. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE). https://github.com/CSSEGISandData/COVID-19

3. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Kenyan blood donors | medRxiv. Accessed November 29, 2020. https://www.medrxiv.org/content/10.1101/2020.07.27.20162693v1

4. South Africa Department of Health. COVID-19 Online Resource and News Portal. https://sacoronavirus.co.za

5. Arndt C, Davies R, Gabriel S, et al. Covid-19 lockdowns, income distribution, and food security: An analysis for South Africa. *Global Food Security*. 2020;26:100410. doi:10.1016/j.gfs.2020.100410

6. Garba SM, Lubuma JM-S, Tsanou B. Modeling the transmission dynamics of the COVID-19 Pandemic in South Africa. *Mathematical Biosciences*. 2020;328:108441. doi:10.1016/j.mbs.2020.108441

7. Nyabadza F, Chukwu W, Chirove F, fatmawati fatmawati, Gatyeni P. Application of Optimal Control to Long Term Dynamics of COVID-19 Disease in South Africa. *medRxiv*. Published online January 1, 2020:2020.08.10.20172049. doi:10.1101/2020.08.10.20172049

8. Mukandavire Z, Nyabadza F, Malunguza NJ, Cuadros DF, Shiri T, Musuka G. Quantifying early COVID-19 outbreak transmission in South Africa and exploring vaccine efficacy scenarios. *PLOS ONE*. 2020;15(7):e0236003. doi:10.1371/journal.pone.0236003

9. Zhao Z, Li X, Liu F, Zhu G, Ma C, Wang L. Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya. *Science of The Total Environment*. 2020;729:138959. doi:10.1016/j.scitotenv.2020.138959

10. He J, Guo Y, Mao R, Zhang J. Proportion of asymptomatic coronavirus disease 2019: A systematic review and meta-analysis. *J Med Virol*. Published online July 21, 2020:10.1002/jmv.26326. doi:10.1002/jmv.26326

11. WHO, WHO. *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)*. Geneva; 2020.

12. New booze, curfew rules for Covid hotspots: Govt to take decision today. BusinessInsider. Accessed December 1, 2020. https://www.businessinsider.co.za/new-booze-curfew-rules-for-covid-hot-spots-govt-to-take-decision-today-2020-12

13. Hao X, Cheng S, Wu D, Wu T, Lin X, Wang C. Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature*. Published online July 16, 2020:1-5. doi:10.1038/s41586-020-2554-8

14. Rahmandad H, Lim TY, Sterman J. Estimating COVID-19 under-reporting across 86 nations: implications for projections and control. *medRxiv*. Published online 2020.

15. Bhattacharyya R, Bhaduri R, Kundu R, Salvatore M, Mukherjee B. Reconciling epidemiological models with misclassified case-counts for SARS-CoV-2 with seroprevalence surveys: A case study in Delhi, India. *medRxiv*. Published online 2020.

16. Consortium SAC-19 M. *Estimating Cases for COVID-19 in South Africa Long-Term National Projections*. May; 2020.

17. Our World in Data. Coronavirus (COVID-19) Testing. https://ourworldindata.org/coronavirus-testing

18. WHO. COVID-19 virtual press conference on 30 March 2020. https://www.who.int/docs/default-source/coronaviruse/transcripts/who-audio-emergencies-coronavirus-press-conference-full-30mar2020.pdf?sfvrsn=6b68bc4a_2

19. World Health Organization. *Coronavirus Disease (COVID-19): Weekly Epidemiological Update*.; 2020. https://www.who.int/publications/m/item/weekly-epidemiological-update---20-october-2020

20. Organization WH. *Estimating Mortality from COVID-19: Scientific Brief, 4 August 2020*. World Health Organization; 2020.

21. Kermack WO, McKendrick AG, Walker GT. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character*. 1927;115(772):700-721. doi:10.1098/rspa.1927.0118

22. Song PX, Wang L, Zhou Y, et al. An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China. *MedRxiv*. Published online 2020.

23. Ray D, Salvatore M, Bhattacharyya R, et al. Predictions, Role of Interventions and Effects of a Historic National Lockdown in India's Response to the the COVID-19 Pandemic: Data Science Call to Arms. *Harvard Data Science Review*. Published online June 9, 2020. doi:10.1162/99608f92.60e08ed5

24. Wang C, Liu L, Hao X, et al. Evolving epidemiology and impact of non-pharmaceutical interventions on the outbreak of Coronavirus disease 2019 in Wuhan, China. *MedRxiv*. Published online 2020.

25. CSSEGISandData. *CSSEGISandData/COVID-19*.; 2020. Accessed August 16, 2020. https://github.com/CSSEGISandData/COVID-19

26. The World Bank. International tourism, number of arrivals - South Africa. https://data.worldbank.org/indicator/ST.INT.ARVL?locations=ZA

27. Li R, Pei S, Chen B, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*. 2020;368(6490):489-493. doi:10.1126/science.abb3221

28. He X, Lau EHY, Wu P, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*. 2020;26(5):672-675. doi:10.1038/s41591-020-0869-5

29. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N Engl J Med*. 2020;382(13):1199-1207. doi:10.1056/NEJMoa2001316

30. Ferretti L, Wymant C, Kendall M, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*. 2020;368(6491):eabb6936. doi:10.1126/science.abb6936

31. Garg S. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—COVID-NET, 14 States, March 1–30, 2020. *MMWR Morbidity and mortality weekly report*. 2020;69.

32. Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *Jama*. 2020;323(11):1061-1069.

33. U.S. CDC. Provisional Death Counts for Coronavirus Disease 2019 (COVID-19). U.S. CDC. https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm

34. Lau H, Khosrawipour T, Kocbach P, Ichii H, Bania J, Khosrawipour V. Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters. *Pulmonology*. Published online 2020.

35. Rahmandad H, Lim TY, Sterman J. Estimating COVID-19 under-reporting across 86 nations: implications for projections and control. *medRxiv*. Published online August 3, 2020:2020.06.24.20139451. doi:10.1101/2020.06.24.20139451

36. University of Oxford. CORONAVIRUS GOVERNMENT RESPONSE TRACKER. https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker

37. Stiegler N, Bouchard J-P. South Africa: Challenges and successes of the COVID-19 lockdown. *Ann Med Psychol (Paris)*. 2020;178(7):695-698. doi:10.1016/j.amp.2020.05.006

38. Zou L, Ruan F, Huang M, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *New England Journal of Medicine*. 2020;382(12):1177-1179.

39. Cereda D, Tirani M, Rovida F, et al. The early phase of the COVID-19 outbreak in Lombardy, Italy. *arXiv preprint arXiv:200309320*. Published online 2020.

40. Ioannidis J. The infection fatality rate of COVID-19 inferred from seroprevalence data. *medRxiv*. Published online 2020.

41. Statista. South Africa: Average age of the population from 1950 to 2050. https://www.statista.com/statistics/578976/average-age-of-the-population-in-south-africa/

42. Boulle A, Davies M-A, Hussey H, et al. Risk factors for COVID-19 death in a population cohort study from the Western Cape Province, South Africa. *Clinical Infectious Diseases*. 2020;(ciaa1198). doi:10.1093/cid/ciaa1198

43. Gousseff M, Penot P, Gallay L, et al. Clinical recurrences of COVID-19 symptoms after recovery: Viral relapse, reinfection or inflammatory rebound? *Journal of Infection*. Published online June 30, 2020. doi:10.1016/j.jinf.2020.06.073

Figures and tables

Figure 1. (a) Total cases by country in the African continent on September 21; (b) The 7-day average testing positive rate of COVID-19 in South Africa during March 5 – September 21.

Figure 2. Schematic diagram of the three models (a) eSIR; (b) SAPHIRE; (c) eSEIRD.

Figure 3. (a)-(c) Daily new number of ascertained infections cases estimated by the models compared with observed data: (a) eSIR, (b) SAPHIRE, and (c) eSEIRD; (d) Current pre-symptomatic/unascertained/ascertained infectious in the SAPHIRE model; (e)-(f) Estimated effective reproduction number (R_e) and ascertained rate (r) in the SAPHIRE model in four time periods. (Assume initial ascertained rate (r_0) equal to 0.10.)

Figure 4. Case fatality ratio (CFR) and estimated infection fatality ratio (IFR) in the eSEIRD model.

Table 1. Timeline of COVID-19 preventions and interventions in South Africa.

Table 2. The posterior mean and credible intervals of the basic/effective reproduction number (R_0/R_e) and ascertained rate (r) obtained from different models and settings.

Table 3. Comparison of the models regarding the cumulative ascertained infected and death with the observed (in thousands).

Table 4. Symmetric mean absolute percentage error (SMAPE) of short-term forecasting.

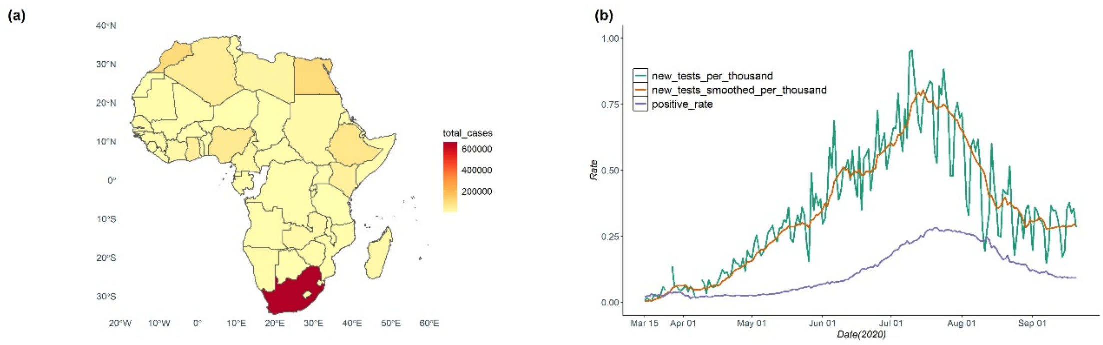

Figure 1. (a) Total cases by country in the African continent on September 21; (b) The 7-day average testing positive rate of COVID-19 in South Africa during March 5 – September 21.

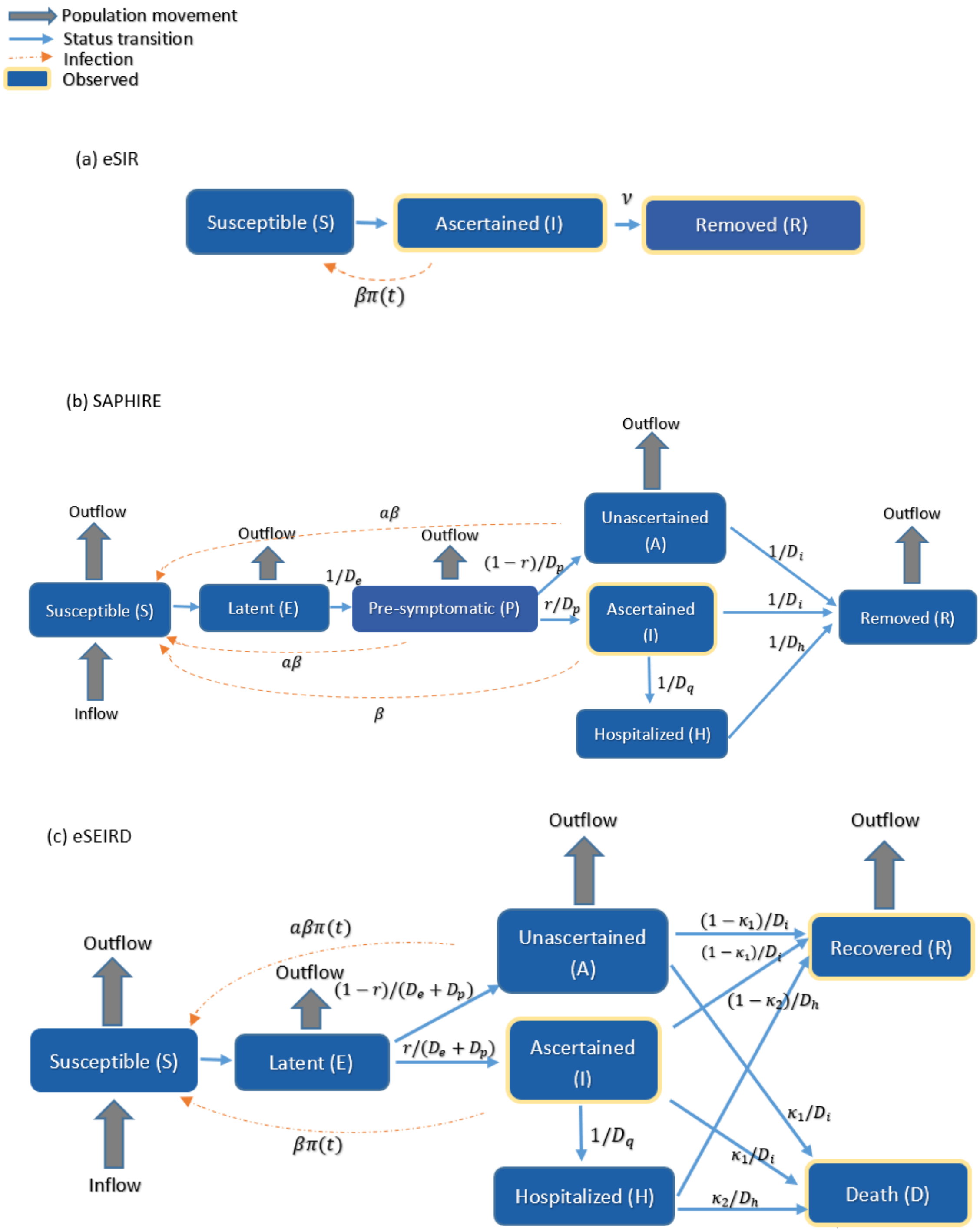

Figure 2. Schematic diagram of the three models (a) eSIR; (b) SAPHIRE; (c) eSEIRD.

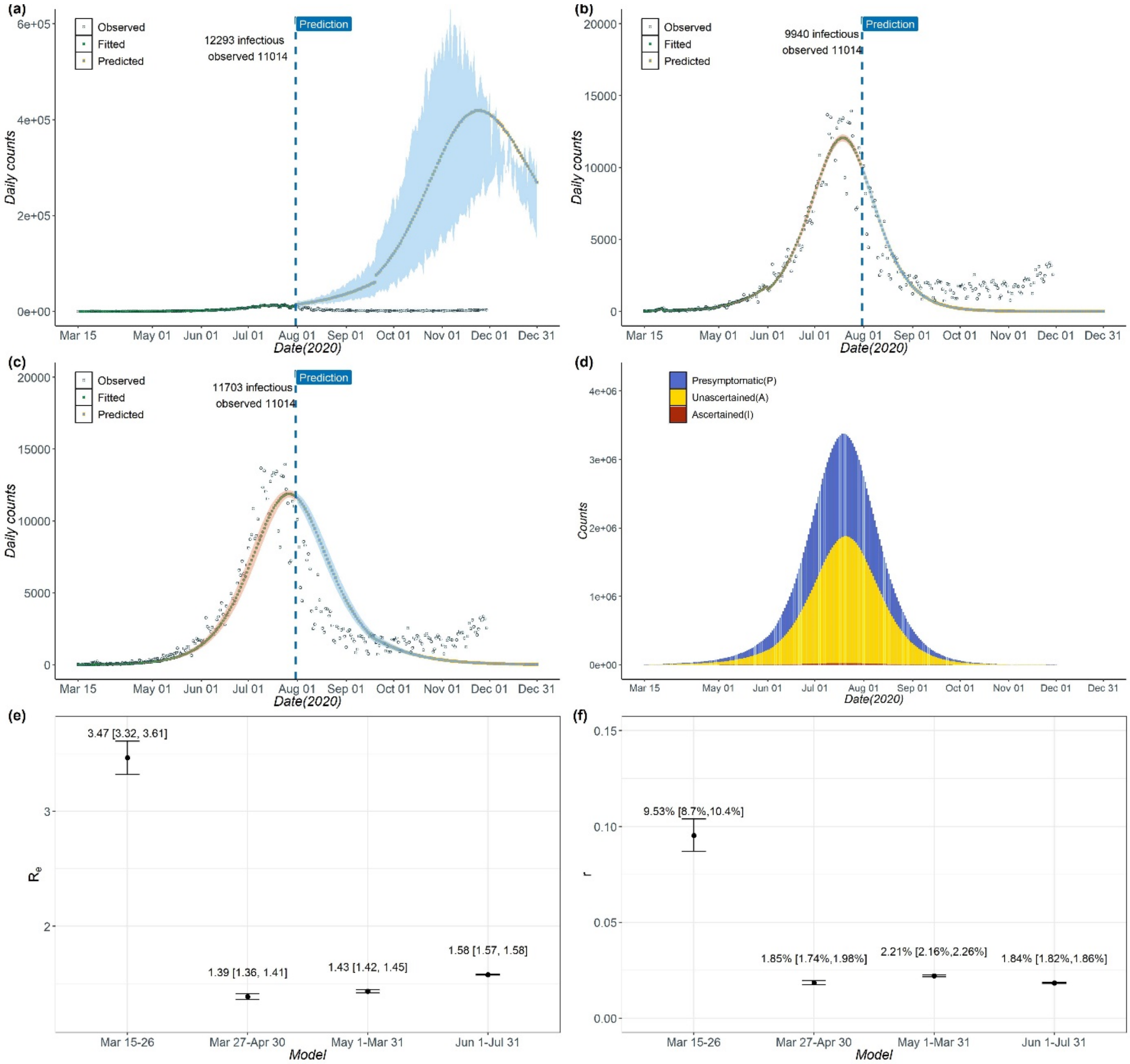

Figure 3. (a)-(c) Daily new number of ascertained infections cases estimated by the models compared with observed data: (a) eSIR, (b) SAPHIRE, and (c) eSEIRD; (d) Current pre-symptomatic/unascertained/ascertained infectious in the SAPHIRE model; (e)-(f) Estimated effective reproduction number (R_e) and ascertained rate (r) in the SAPHIRE model in four time periods. (Assume initial ascertained rate (r_0) equal to 0.10.)

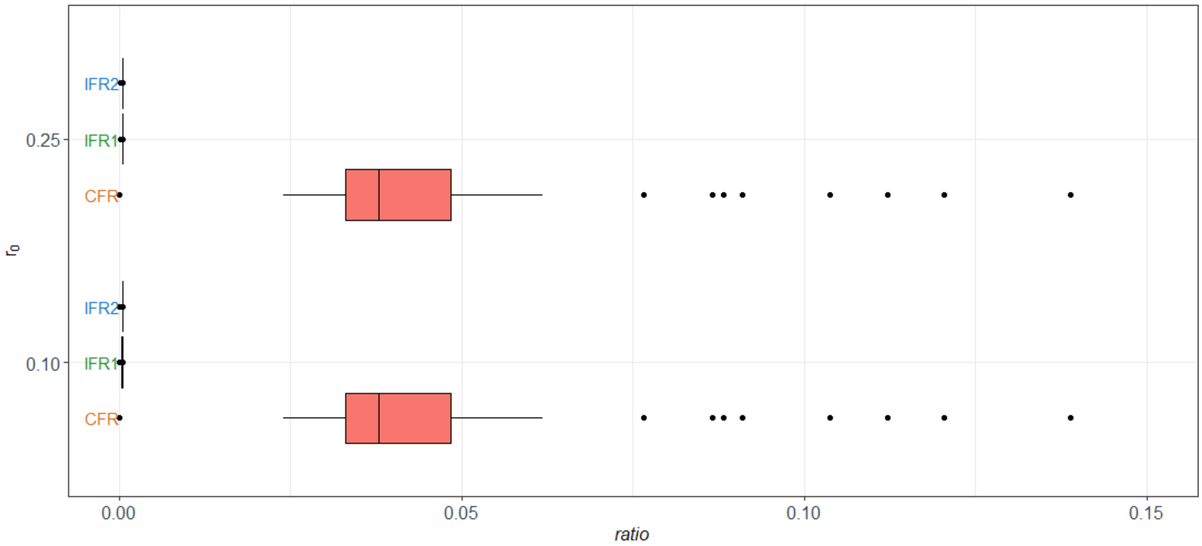

Figure 4. Case fatality ratio (CFR) and estimated infection fatality ratio (IFR) in the eSEIRD model. CFR $= \frac{\text{Number of reported deaths}}{\text{Number of reported deaths and recovered}}$; IFR1 = $\frac{\text{Number of reported deaths}}{\text{Number of of reported and unreported cases}}$; and IFR2 = $\frac{\text{Number of reported and unreported deaths}}{\text{Number of of reported and unreported cases}}$[1].

[1] World Health Organization, "Estimating Mortality from COVID-19: Scientific Brief, 4 August 2020" (World Health Organization, 2020), 19.

Table 1. Timeline of COVID-19 preventions and interventions in South Africa.

Date (2020)	Confirmed[2]	Death	Testing rate[3]	Interventions and update[4]
5 March	1	0	-	
10 March	7	0	-	Screening at ports of entry has intensified and escalated.
15 March	51	0	-	Self-quarantine for COVID-19 is recommended. Visas to visitors from high-risk countries (Italy, Iran, South Korea, Spain. Germany, US, UK) are cancelled and previously granted visas are hereby revoked. Gatherings of more than 100 are prohibited. Mass celebrations are canceled.
16 March	62	0	-	Of the 53 land ports, 35 are shut down.
18 March	116	0	-	A travel ban on foreign nationals from high-risk countries such as Germany, US, UK and China.
27 March	1170	1	-	A national lockdown is implemented. Alert level 5 is in effect from midnight 26 March to 30 April.
1 May	5951	116	0.004	A less strict lockdown is in place. Alert level 4 is in effect from 1 to 31 May. Borders will remain closed to international travel, no travel will be allowed between provinces, except for the transport of goods and exceptional circumstances.
1 June	34,357	705	0.013	From 1 June 2020 alert level 3 will be in effect. Restrictions on many activities, including at workplaces and socially, to address a high risk of transmission.
8 June	50,879	1080	0.016	More than half of the cases since the outbreak started have occurred in the last 2 weeks
17 July	337,594	4804	0.041	Recommendations that isolation period for those confirmed to drop from 14 to 10 days
21 July	434,200	6655	0.044	South Africa has the 5th most confirmed Covid-19 infections in the world. An acceleration of cases has increased by 30% in the last week.

[2] Johns Hopkins University, "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE)," n.d., https://github.com/CSSEGISandData/COVID-19.
[3] Our World in Data, "Data on COVID-19 (Coronavirus)," n.d., github.com/owid/covid-19-data/tree/master/public/data/testing.
[4] South Africa Department of Health, "COVID-19 Online Resource and News Portal," n.d., https://sacoronavirus.co.za.

Table 2. The posterior mean and credible intervals of the basic/effective reproduction number (R_0/R_e) and ascertained rate (r) obtained from different models and settings.

Model	r_0		R_0/R_e*		r (%)	
			Mean	95%CI	Mean	95%CI
eSIR	**0.10**		2.05	[1.81,2.31]	-	-
	0.25		2.04	[1.82,2.31]	-	-
SAPHIRE	**0.10**	Mar 15-26	3.47	[3.32,3.61]	9.53	[8.70,10.40]
		Mar 27-Apr 30	1.39	[1.36,1.41]	1.85	[1.74,1.98]
		May 1-31	1.43	[1.42,1.45]	2.21	[2.16,2.26]
		Jun 1-July 31	1.58	[1.57,1.58]	1.84	[1.82,1.86]
	0.25	Mar 15-26	4.65	[4.48,4.83]	14.12	[12.80,15.50]
		Mar 27-Apr 30	1.41	[1.39,1.43]	1.95	[1.83,2.08]
		May 1-31	1.44	[1.43,1.45]	2.22	[2.17,2.27]
		Jun 1-July 31	1.58	[1.57,1.58]	1.84	[1.82,1.86]
eSEIRD	**0.10**		2.10	[2.09,2.10]	2.17	[2.15,2.19]
	0.25		2.10	[2.09,2.10]	2.17	[2.15,2.20]

*R_0 in eSIR and eSEIRD model; R_e in SAPHIRE model.

Table 3. Comparison of the models regarding the cumulative ascertained infected and death with the observed (in thousands). Bold-faced entries indicate column winners regarding the closeness to the observed.

		Infected			Death		
Model	r_0	Estimation	Prediction		Estimation	Prediction	
		Jul 31	Aug 15	Aug 31	Jul 31	Aug 15	Aug 31
eSIR	**0.10**	496	769	1,230	-	-	-
	0.25	496	768	1,228	-	-	-
SAPHIRE	**0.10**	**493**	603	**653**	-	-	-
	0.25	**493**	**603**	653	-	-	-
eSEIRD	**0.10**	439	**594**	698	11	17	22
	0.25	308	483	**633**	7	13	19
Observed	-	493	584	627	8	12	14

Table 4. Symmetric mean absolute percentage error (SMAPE) of short-term forecasting. Bold-faced entries indicate column winners regarding prediction performance.

Model	r_0	Cumulative ascertained cases		Cumulative ascertained deaths	
		Testing SMAPE		Testing SMAPE	
		Aug 1 -Aug 15	Aug 1 – Aug 31	Aug 1 -Aug 15	Aug 1 – Aug 31
eSIR	**0.10**	13.57%	30.96%	-	-
	0.25	13.47%	30.85%		
SAPHIRE	**0.10**	**1.81%**	**2.96%**	-	-
	0.25	**1.80%**	**2.95%**		
eSEIRD	**0.10**	4.78%	6.02%	36.28%	38.97%
	0.25	31.38%	18.96%	4.12%	12.46%