Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Regular Research Article

Why is the number of COVID-19 cases lower than expected in Sub-Saharan Africa? A cross-sectional analysis of the role of demographic and geographic factors

Pierre Nguimkeu,⇑ Sosson Tadadjeub

A Department of Economics, Andrew Young School of Policy Studies, Georgia State University, 55 Park Place, Atlanta, GA 30303, USA
b Faculty of Economics and Management, University of Dschang, P.O. Box 96, Dschang, Cameroon

Article info

Article history:
Available online 21 October 2020

Keywords:
COVID-19
Pandemic
Regression analysis
Africa

Abstract

Unlike initially predicted by WHO, the severity of the novel coronavirus pandemic has remained relatively low in Sub-Saharan Africa, more than two months after the first confirmed cases were identified. In this paper, we analyze the extent to which demographic and geographic factors associated to the disease explain this phenomenon. We use publicly available data from a cross-section of 182 countries worldwide, and we employ a regression analysis that accounts for possible misreporting of COVID-19 cases, as well as a Ramsey-type specification that preserves degree of freedom. We found that proportion of population aged 65+, population density, and urbanization are significantly positively associated with high numbers of active infected cases, while mean temperature around the first quarter (January-March) is negatively associated to this COVID-19 outcome. These factors are those for which Africa has a comparative advantage. In contrast, factors for which Africa has a relative disadvantage, such as income and quality of health care infrastructure, are found to be insignificant predictors of the spread of the pandemic. These results hold even when accounting for possible underreporting, as well as differences in the duration of the epidemic in each country, as measured by the time elapsed since the first confirmed case occurred. We conclude that differences in demographic and geographic characteristics help understand the relatively low progression of the pandemic in sub-Saharan Africa as well as the gap in the number of active cases between this region and the rest of the World. We also found, however, that this gap is insignificant beyond these factors, and is expected to narrow over time as the pandemic evolves. These results provide insights for relevant urban policies and kinds of development planning to consider in the fight against disease spreads of the coronavirus type.

1. Introduction

The coronavirus disease 2019 (COVID-19) caused by a novel SARS-CoV-2 that emerged in China last year, and has since spread to all regions of the world, is causing major ravages worldwide as reported by the World Health Organization (WHO). Since the burst of this disease, experts have been announcing a drastic surge and dramatic consequences of the epidemic in sub-Saharan Africa (SSA), a region already plagued by poverty and lack of health infrastructure. However, more than two months after the first confirmed case in the continent, the transmission and severity of the novel coronavirus pandemic has remained relatively low in sub-Saharan Africa, while other regions of the world (such as Europe and USA) have been more seriously hit. Sub-Saharan Africa is indeed the least affected region, with 28,848 infected cases and 1112 deaths recorded as of April 27, 2020 (Our world in data, 2020). Understanding why the severity of COVID-19 in sub-Saharan Africa remains comparatively low in spite of the weak health-care system, inadequate surveillance and laboratory capacity, scarcity of public health human resources, and limited financial means (Nkengasong & Mankoula, 2020) is therefore a question that merits attention.

Although there is no clear consensus about this seemingly puzzling situation, a number of hypotheses have been posited to try to explain the low numbers observed in sub-Saharan Africa. One argument that emerged was that, since very few African countries

⇑ Corresponding author.
E-mail address: nguimkeu@gsu.edu (P. Nguimkeu).

https://doi.org/10.1016/j.worlddev.2020.105251
0305-750X/© 2020 Elsevier Ltd. All rights reserved.
have sufficient and appropriate diagnostic capacities, the number of cases are largely underreported. However, as explained by Dr John Nkengasong, head of the African Center of Disease Control and Prevention, the fact that health facilities are still not overwhelmed by patients may rule out this hypothesis. Other hypotheses build from recent clinical studies (e.g. China, Italy) that have identified the presence of pre-existing non-communicable diseases such as cerebrovascular diseases (CVD), diabetes, hypertension, and cancer, as the main comorbidities associated with infected and death cases (Driggin et al., 2020; Yang et al., 2020). On the other hand, while it has been documented that dense communities, urban congestions or colder weather may favor the transmission of viruses of respiratory syndrome such as influenza, measles, tuberculosis, coronavirus (Alirol, Getaz, Stoll, Chappuis, & Loutan, 2011; Van de Poel, O’Donnell, & Van Doorslaer, 2012), recent clinical studies show that individuals aged 60 years or older are at higher risk of contamination and death (WHO, 2020; Zhou et al., 2020).

Against this backdrop, the present study aims to analyze the role of demographic and geographic (DG) factors in explaining the low severity of the epidemic in Sub-Saharan Africa (SSA) compared to other regions. We employ a regression analysis that estimates the number of active infected cases where these DG factors are used as explanatory variables. Based on the related literature, the demographic indicators considered are the median age and the proportion of population aged 65+, whereas geographical factors include population density, urbanization rate and mean first quarter temperature. Both these demographic and geographic factors are found to be significantly and positively associated with the number of active COVID-19 cases. Given that SSA countries exhibit relatively lower magnitudes in these factors compared to the rest of the world, they thus have a comparative advantage from these perspectives. In contrast, factors in which Sub-Saharan Africa has a considerable disadvantage such as income and quality of health infrastructure (measured by GDP per capita, and health expenditure), turn out to be insignificant predictors of the spread of the epidemic. Measures of epidemiological factors, especially the prevalence of diabetes, are also found to have no significant association with this COVID-19 outcome, possibly for endogenous behavioral reasons that we further discuss in Section 2.

The only source of COVID-19 data available to us for this exercise is a publicly available one whose quality is, unfortunately, very uncertain. Our econometric specification attempts to solve this uncertainty by explicitly accounting for possible underreporting in the official number of confirmed cases used, as well as for the lag in the disease introduction in each country, as measured by the time elapsed since the first confirmed case was detected. The latter also allows to capture the learning effect, as countries that experience the epidemic relatively later are likely to learn from successful coping strategies adopted by those that experienced it relatively earlier. To test whether and by how much the estimated effects of the DG factors differ between Sub-Saharan Africa and the rest of the world, we employ a Ramsey-type device that preserves degrees of freedom. This consists in assessing whether an interaction between a SSA dummy and these factors help explain the outcome variable. Subject to the above-mentioned data quality caveats, our results provide conclusive evidence that the relatively low progression of the epidemic in Sub-Saharan Africa and the gap observed in the number of active cases compared to the rest of the world can be partly explained by the differences in demographic and geographic factors. However, this gap narrows down with the duration of the epidemic and is not significant beyond these factors. These results call for strategies to implement mitigation efforts and containment measures that pertain to SSA situation, and provide insights on policies and program interventions that could be considered to prevent the spreads of disease of coronavirus type. This paper is organized as follows. Section 2 discusses the background and descriptive statistics. Section 3 presents the estimation approach and the results of the regression analysis. Section 3 concludes.

2. Data and descriptive statistics

The data consists of a cross-section of 182 countries affected by the Coronavirus pandemic. We collated the most recent data available from various sources including Worldometer Coronavirus (2020), World Development Indicators (WDI), Global Health Observatory (GHO), World Bank Climate Change Knowledge Portal, World Population Prospects (2019) and the Institute for Health Metrics and Evaluation (IHME). Data on COVID-19 spread includes the total number of confirmed cases, the total number of deaths and the total number of active cases (which is the total number of confirmed cases net of recoveries and deaths).3

Fig. 1 compares the trends in the average number of active cases in Sub-Saharan Africa versus the average of the rest of the world for the first 60 days since the pandemic has erupted in the given regions. Overall, this trend has been consistently lower in sub-Saharan Africa compared to the rest of the world, even when adjusting for the lag in the timing of disease occurrence in both regions.

Table 1 presents descriptive statistics of the variables of interests in our whole sample. We denote by Duration, the variable that accounts for differences in the timing of disease eruption, i.e. the number of days elapsed since the first case of COVID-19 to the observed date.

The first panel summarizes the characteristics of the COVID-19, including the total number of cases, total number of deaths, the total number of active cases, and the duration of the epidemic from our data source. The second panel summarizes the epidemiological factors including the prevalence of CVD and the prevalence of diabetes (i.e. the proportion of people aged 20–79 that have type 1 or 2 diabetes). These data are based on the Worldometer Coronavirus database (2020), whose quality is very uncertain. These uncertainties are especially more serious when comparing vastly different countries such as those of Africa with those of Western countries. Not only recorded cases but also deaths attributed to covid-19 are very unequally measured across the globe including within Europe and the USA.

Data on active cases is thus doubly uncertain since they are obtained by using both recorded cases and deaths. Data on excess mortality seem the most satisfactory in general, but they are still very uncertain in the case of Africa, making them similarly problematic for this study.

Data for these measures come from the Worldometer Coronavirus database (2020), whose quality is very uncertain. These uncertainties are especially more serious when comparing vastly different countries such as those of Africa with those of Western countries. Not only recorded cases but also deaths attributed to covid-19 are very unequally measured across the globe including within Europe and the USA. Data on active cases is thus doubly uncertain since they are obtained by using both recorded cases and deaths. Data on excess mortality seem the most satisfactory in general, but they are still very uncertain in the case of Africa, making them similarly problematic for this study.

This is the average number of cases per country. This means, e.g., we are comparing the first 60 days of the occurrence of the disease in sub-Saharan Africa with the first 60 days of the occurrence of the disease in the rest of the world to account for lags in the spread. The starting time is the day when the average of the given region exceeds 1.

---

3 For a report, see https://newschs.com/covid-19-africa-steps-up-testing-le-point/
type 2 diabetes). The choice of these variables are based on WHO (2020) report emphasizing that those with such pre-existing medical conditions (i.e. cerebrovascular disease, diabetes, chronic respiratory disease, and cancer) are at higher risk. While COVID-19 infects people of all ages, Zhou et al. (2020) recently found that older people (e.g. 60 + years) are at higher risk of the disease. We therefore consider the proportion of the population aged 65 + as a fraction of the total population (denoted Pop 65 + ), and the median age of the population as our main demographic indicators, whose statistics are summarized in the third panel of the table. As for geographic factors which are given in the fourth panel of the table, we consider three indicators: urbanization (i.e. the proportion of people living in urban areas), population density (i.e. number of people per squared kilometer, km2), mean temperature (in Celsius degrees) of the first quarter of the year (January-March).

Table 2 presents summary statistics of our variables of interest by region. It shows that compared to all other regions, sub-Saharan Africa has the lowest prevalence of CVD and diabetes, the lowest proportion of population aged 65 +, the youngest population i.e. lowest median age, one of the lowest urbanization rate (apart from South Asia), one of the lowest population density (apart from North America), and the highest average temperature around the first quarter of the year. In particular, the Pop 65 + and Median age in Sub-Sahara is 3.3% and 20 years, compared to 16.5% and 40 years in North America. This is a huge difference in terms of age-related vulnerability. There is also a huge difference in the average temperature across regions. While SSA experiences an average of 26 °C during the months of January through March, other regions experience much lower temperature, especially Europe and Central Asia with an average of 1.89 °C, as well as North America with a negative average of –8.92 °C (see Table 2).

Following related literature, we use GDP per capita as a measure of income, and health expenditure per capita as a measure of healthcare infrastructure. Both can be understood as Economic indicators. For these factors, Sub-Saharan Africa is very disadvantaged, and has the lowest scores along with South Asia. For example, GDP per capita in Sub-Saharan Africa is about 20 times lower than in North America.

The observed differences in these summary statistics suggests, as it will become clearer in the regression analysis below, that there should be important differential effects of the corresponding factors between SSA and the other regions.

3. Regression analysis

This section presents the regression model and the estimation approach, and discusses the associated results. We consider the log total number of active infected cases of COVID-19 as our response variable, which is the total number of cases net of total deaths and recovered.

Table 1
Summary statistics of the main factors of interest.

| Variable                        | Observations | Mean       | Std. Dev. | Min    | Max    |
|---------------------------------|--------------|------------|-----------|--------|--------|
| Total cases (in thousands)      | 182          | 9.3020     | 43.123    | 0.002  | 501.65 |
| Total death (in thousands)      | 145          | 0.7074     | 0.289     | 0.001  | 18.849 |
| Active cases (in thousands)     | 182          | 6.6787     | 36.257    | 0.001  | 455.70 |
| Duration (in days)              | 182          | 35.720     | 19.230    | 3.000  | 143.00 |
| CVD cases (in thousands)        | 170          | 7.013      | 3.373     | 3.218  | 16.470 |
| Diabetes prevalence (in %)      | 181          | 7.920      | 4.140     | 1.000  | 22.110 |
| Population > 65 ages (in %)     | 172          | 30.60      | 9.170     | 15.10  | 48.200 |
| Median Age (in years)           | 172          | 172.00     | 132.60    | 15.10  | 27.570 |
| Population density (per km²)    | 177          | 265.6      | 858.2     | 2.040  | 795.29 |
| Urban population (in %)         | 181          | 60.66      | 22.73     | 13.03  | 100.00 |
| Temperature (°C)                | 168          | 16.57      | 11.740    | 11.740 | 81.83  |
| GDP per capita (in $1000)       | 165          | 15.071     | 20.904    | 210.8  | 110.74 |
| Health expenditure (% GDP)      | 170          | 4.020      | 2.230     | 0.780  | 10.750 |

Source: Authors calculations.

Table 2
Means of variables of interest by region (as of April 10, 2020).

| Variable                        | SSA          | MENA         | EAP          | ECA          | LAC          | NA           | SA           |
|---------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Total Cases*                    | 0.149        | 5.499        | 5.887        | 17.4         | 1.556        | 261.9        | 1.682        |
| Total Death*                    | 0.052        | 0.279        | 0.361        | 1.482        | 76.16        | 9.628        | 0.722        |
| Active Cases*                   | 123.65       | 2.967        | 1.244        | 11.98        | 1.338        | 235.6        | 1.433        |
| Duration                        | 21.63        | 40.61        | 55.47        | 41.28        | 27.05        | 75           | 49.25        |
| Diabetes prev. (%)              | 5.29         | 11.1         | 9.42         | 6.47         | 9.89         | 9.2          | 10.76        |
| CVD Cases*                      | 4.525        | 6.093        | 5.999        | 11.49        | 5.831        | 10.23        | 4.811        |
| Pop. 65+ (%)                    | 3.30         | 5.64         | 9.41         | 18.12        | 8.93         | 16.51        | 5.517        |
| Median Age                      | 20.16        | 30.3         | 33.17        | 40.47        | 30.85        | 39.85        | 27.01        |
| Urban Pop. (%)                  | 45.32        | 81.10        | 62.18        | 68.40        | 64.62        | 81.83        | 31.46        |
| Pop. Density (km²)              | 117.3        | 323.1        | 837.7        | 180.9        | 150.8        | 19.92        | 538.4        |
| Temp. (°C)                      | 26.05        | 17.36        | 19.07        | 1.89         | 24.33        | –8.92        | 14.55        |
| GDP per capita                  | 2.713        | 18.53        | 18.51        | 28.51        | 8.45         | 52.99        | 2.622        |
| Health Expend. (% GDP)          | 3.793        | 3.702        | 3.342        | 5.661        | 4.068        | 7.845        | 2.965        |

Note: SSA = Sub-Saharan Africa; MENA = Middle East and North Africa; EAP = East Asia and Pacific; ECA = Europe and Central Asia; LAC = Latin America and Caribbean; NA = North America; SA = South Asia. The list of countries that are part of these regional stratifications can be found in https://data.worldbank.org/country. In thousands of the corresponding unit.

Source: Authors calculations.
3.1. Estimation

We use a linear regression model framework to analyze the relationship between disease outcome and demographic and geographic factors while controlling for epidemiologic, economic and health system infrastructure indicators:

\[
Y_i' = \beta_0 + \gamma_i A_i + \beta_1 t_i + \gamma_1 A_t i + \sum_{k=2}^{K} X_k \beta_k + \epsilon_i',
\]

where \(Y_i'\) denotes the true COVID-19 outcome variable in country \(i\) (e.g., \(\log\) Active Cases); \(A_i\) is a binary indicator (dummy variable) for sub-Saharan Africa, which equals 1 if country \(i\) is a sub-Sahara African country, and equals 0 otherwise; \(t_i\) is the duration of the epidemic in country \(i\) (i.e., the number of days elapsed since the first confirmed case was reported in country\(i\)); \(X_i = [X_{12}, \ldots, X_{16}]\) is a vector of explanatory variables including epidemiological, demographic, environmental, economic and health infrastructure factors in country \(i\); \(K\) is the total number of explanatory variables (excluding the dummy variable \(A_i\)). The error term, \(\epsilon_i'\), is assumed to be mean zero and captures all other factors driving the outcome that are not accounted for by our model specification.

Unfortunately, the true COVID-19 outcome variable \(Y_i'\) is unobserved due to uncertainties in data quality and we can only observe a possibly misreported surrogate \(Y_i\) defined as

\[
Y_i = Y_i' - u_i
\]

where \(u_i\) is the measurement error, or the amount of underreporting. If we assume that, for the reasons already evoked, data from sub-Saharan Africa are possibly more underreported than those of other regions, then we can write

\[
u_i = \tau_0 + \kappa_0 A_i + \zeta_i, E(\tau_i|A_i) = 0
\]

where \(\kappa_0 \geq 0\) represents the average amount of excess underreported log number of active cases in SSA compared to other regions (where average amount of underreporting is \(\tau_0 \geq 0\)), and \(\zeta_i\) is the residual measurement error that is assumed to be uncorrelated with \(A_i, t_i\) and \(X_i\). We therefore have \(Y_i = Y_i' - \tau_0 - \kappa_0 A_i - \zeta_i\), and the relationship between the reported cases and the factors of interest to be estimated can be summarized as:

\[
Y_i = \beta_0 + \gamma_i A_i + \beta_1 t_i + \gamma_1 A_t i + \sum_{k=2}^{K} X_k \beta_k + \epsilon_i,
\]

where \(Y_i\) is the reported number of active cases in country \(i\) itself. The coefficient \(\beta_0 = \beta_0' - \tau_0\) is the intercept and \(\gamma_i = \gamma_0' - \kappa_0\) is the average difference in outcome between sub-Saharan Africa and the rest of the world, both exacerbated by the amounts of misreporting \(\tau_0\) and \(\kappa_0\), respectively, conditional on the vector of factors \(X_i\). The coefficient \(\beta_1\) is the average ceteris paribus effect of the length of the pandemic on the outcome, whereas \(\gamma_1\) captures how this duration effect varies between Sub-Saharan Africa and the rest of the world. These coefficients also capture the “learning” effect, given that countries that experienced the epidemic relatively later may have learned and adopted the most successful strategies adopted by those that faced it earlier. The vector \(\beta = [\beta_2, \ldots, \beta_k]\) is the vector of parameters associated with the factors \(X = [X_{12}, \ldots, X_{16}]\). It should be noted that, for \(k = 2, \ldots, K\), the coefficient \(\beta_k\) captures the average ceteris paribus effect of the specific factor \(X_k\) on the outcome. Notice that the new error term, \(\epsilon_i\), has an increased variance as it includes both the measurement error component and the original error expected in the regression. Overall, if this model is correctly specified, the outcome of the regression will be an unbiased estimate of \([\beta_1, \beta_2, \ldots, \beta_k]\), but with reduced precision in these estimates, lower \(t\)-statistics and a reduced \(R^2\).

Since a focus in this study is on the conditional comparison of the outcome of sub-Saharan Africa with the rest of the world, a useful modification of Model (4) is to interact the sub-Saharan Africa dummy variable, \(A_i\), with all the explanatory variables of the model, \(X_i = [X_{12}, \ldots, X_{16}]\) and add these interaction terms, \(A_i X_{12}, \ldots, A_i X_{16}\), as additional regressors to the model to assess heterogeneity in the effects of the initial factors. However, this would significantly reduce the degree of freedom of the model (which is already low), and thus lead to low power in the significance testing of the coefficients. To overcome this issue, we adopt a device similar in spirit to the Ramsey RESET test, which allows us to assess whether the interaction terms \(A_i X_{12}, \ldots, A_i X_{16}\) significantly explain \(Y_i\), while conserving in degrees of freedom. This consists in augmenting the interaction between \(A_i\) and the OLS fitted values as an additional regressor of the model to see whether it significantly explains the response variable. Specifically, recall that the fitted values from Eq. (4) is defined by

\[
\bar{Y}_i = \hat{\beta}_0 + \bar{\gamma}_i A_i + \bar{\beta}_1 t_i + \bar{\gamma}_1 A_t i + \sum_{k=2}^{K} X_k \hat{\beta}_k
\]

where the components with “\(\hat{\}\)”, represent the OLS estimates of their underlying quantities. These fitted values are therefore just linear functions of the independent variables. If we interact the fitted values with the dummy variable, \(A_i\), we get a particular function of the desired interactions terms, \(A_i, A_t, A_i X_{12}, \ldots, A_i X_{16}\). This suggests estimating a model of the form

\[
Y_i = \delta_0 + \theta_0 A_i + \delta_1 t_i + \theta_1 A_t i + \sum_{k=2}^{K} X_k \delta_k + \theta_2 A_i \bar{Y}_i + \eta_i
\]

where \(\bar{Y}_i\) stands for the fitted values obtained above. Notice that model (5) is equivalent to a “reduced-form” regression model of the form

\[
Y_i = \psi_0 + \varphi_0 A_i + \psi_1 t_i + \varphi_1 A_t i + \sum_{k=2}^{K} X_k \psi_k + \sum_{k=2}^{K} A_i X_k \varphi_k + \eta_i
\]

where the relationship between the two last sets of coefficients is given by

\[
\begin{align*}
\psi_0 &= \delta_0; \\
\varphi_0 &= \theta_0 + \theta_2 \left( \hat{\beta}_0 + \bar{\gamma}_0 \right); \\
\psi_1 &= \delta_1; \\
\varphi_1 &= \theta_1 + \theta_2 \left( \hat{\beta}_1 + \bar{\gamma}_1 \right); \\
\psi_k &= \delta_k; \\
\varphi_k &= \theta_2 \hat{\beta}_k, \\
&\text{for } k \geq 2
\end{align*}
\]

The sign and significance of the coefficient \(\theta_2\) can therefore be used to assess how the effects of \(X_i = [X_{12}, \ldots, X_{16}]\) on the outcome variable \(Y_i\) differ between Sub-Saharan Africa and the rest of the world. Once Eqs. (4) and (5) have been estimated, the coefficients of the reduced form model given by Eq. (6) can be inferred using the formulas given by Eq. (7), and their standard errors can be
approximated using the delta method. Notice that the coefficients of the main baseline factors $t_i$ and $X_i$ do not change across specifications, but only the coefficients of the terms interactions with $A_i$ are likely to change across specifications. Our ultimate targets are those from the reduced form relationship given by Eq. (6), although in this case, we are really only interested in computing $\varphi_0$ and $\varphi_1$, which can be readily obtained from the estimation of Eq. (5) using the formulas in Eq. (7).

In addition, given that $Y_i$ is an estimate of the expected value of $Y_i$ given $A_i, t_i, \text{ and } X_i$ using Eq. (5) to estimate the outcome variable is also useful to correct for potential heteroscedasticity, in case the error variance in Eq. (4) is thought to change with $A_i, A_i t_i, \text{ and } A_i X_i$. In summary the estimation method proceeds as follows.

- **Step 1**: Estimate the model given by Eq. (4) by OLS, and obtain the fitted values $\hat{Y}_i$. Compute the interaction terms $A_i \hat{Y}_i$.
- **Step 2**: Run the regression model given by Eq. (5) by OLS. The estimates of $\delta_i$ and $\delta_k$ for $k = 2, \cdots, K$, provide the ceteris paribus effects of $t_i$ and $X_i = [X_{i2}, \cdots, X_{ik}]$ on the COVID-19 outcome, respectively. The estimate of $\theta_2$ represents the SSA average differential effect of $X_i = [X_{i2}, \cdots, X_{ik}]$ on the disease outcome. These estimates can then be used to compute $\varphi_0$ and $\varphi_1$, the reduced-form coefficients of $A_i$ and $A_i t_i$, representing respectively the average difference in outcome and the differential effect of $t_i$ on the disease outcome between SSA and the rest of the world, conditional on all other factors.

Given the small size of the sample, classic standard errors may not be correctly estimated from the usual asymptotic variance–covariance matrix. We therefore use the bootstrap method as a complementary estimation approach for these standard errors. To account for possible heteroscedasticity in the error term given in Eqs. (4) and (5), we also compute robust standard errors.10

An obvious limitation of this econometric model is its inability to account for endogenous behavioral response to the disease outcome, which could be substantial. For instance, it is possible that people at higher risk are responding to the disease in a way that could mitigate the initially perceived impact. This is issue is especially true for epidemiologic factors. Addressing it would require ancillary information about the underlying behavior and possibly a different methodological approach. The empirical results for these factors should therefore be taken with extra caution. In contrast, demographic and geographic factors are less responsive to outbreaks so that the effect from these DG factors should be the most meaningful estimates in the empirical assessment.

3.2. Results

The regression results of the baseline estimates with log total number of active cases as the main dependent variable are presented in Table 3. The estimation results for this outcome are summarized in columns (1) through (3). The first column presents the baseline estimates without the SSA regional dummy. The second column presents the estimates that include the SSA dummy variable (corresponding to Eq. (4)), while the third column presents these results where this dummy variable is interacted with the fitted values of the regressions in Column (2) which corresponds to Eq. (5).

Preliminary analysis has shown strong pairwise correlation between CVD and diabetes prevalence, as well as median age and Pop 65+, respectively. These strong correlations have also been noticed in earlier work such as Halter (2014). Hence, to avoid multi-collinearity, only one of these factors was included in the main regressions. In particular, Table 3 presents the results using the variables that have less missing values among highly correlated alternatives. This means that we focused on results with diabetes prevalence and Pop 65 + in our main discussion. Those with the alternative measures, i.e., CVD prevalence and median age, were less conclusive because of poor fit and statistical power (some due, e.g., to high rates of missing values), but are available from the authors.11 Several functional forms were also considered (e.g., which explanatory variables to include in log form) and the reported results are based on those that gave the highest fit in terms of adjusted R-squared and/or pseudo-log likelihood value.

The epidemiological predictors we use to assess the spread of the pandemic are the duration of the pandemic and the prevalence of diabetes in the country. While recent clinical studies suggest that cerebrovascular diseases and diabetes are some of the most distinctive comorbidities among patients under intensive care of COVID-19 (e.g., Fang, Karakilçak, & Roth, 2020; Yang et al., 2020), our current cross-country data does not provide enough evidence to support this hypothesis. Our regression estimation results show no statistical significance in the association between diabetes prevalence and the number of active COVID-19 cases worldwide. One possible explanation for this puzzling result, as already mentioned in the methodological discussion, is the presence of endogenous behavioral responses among diabetes and cerebrovascular diseases patients. Being more at risk, these individuals are likely to be more careful in adopting safety and social distancing measures than others, which could eventually cancel their higher mortality or contamination risk and lead to no significance in an empirical assessment. Our econometric model does not deal with this important issue. On the other hand, as one would expect, the duration of the pandemic appears to be a strong predictor of the disease spread. We estimate that any additional day in the duration of the epidemic is associated with a 4% increase in the number of active cases, and this effect is significant at 1% (see Columns (1)–(3)). Although the first cases of coronavirus occurred in Sub-Saharan Africa relatively later than other parts of the world (e.g., China, Europe, USA), the interaction between the SSA dummy and the duration of the epidemic shows that the longer the epidemic would last, the more severe the consequence would be for Sub-Saharan Africa countries compared to the rest of the world, everything else equal. In particular, our results show that any additional day of the epidemic is associated with a 12.27% increase in the number of active cases in SSA, and this effect is significant at 1%.12 This duration effect is much higher in SSA compared to the rest of the world.

Our most meaningful findings are those related to demographic and geographic factors. The results show that the proportion of population aged 65 and above, an important demographic indicator capturing population ageing, is positively associated to the number of active cases, and this correlation is significant across all specifications. Specifically, everything else equal, a 1 percent increase in the fraction of this population is associated with about 0.09 percent increase in the number of active cases on average. This result is in line with recent evidence suggesting that relatively older adults are at a higher risk of COVID-19 (e.g., Zhou et al., 2020; WHO, 2020). Since Sub-Saharan Africa has an extremely young population (e.g., half of the population is aged below 20, and only 3.3% are above 65), they are likely at a relatively lower risk on this dimension.

---

10 We only reported the results with bootstrapped standard errors. Those with the robust standard errors are similar.

11 Some of these variables, such as CVD prevalence and health worker density have up to 50% of missing values, especially for SSA countries which is our main focus, thus yielding less reliable results.

12 This is the reduced-form effect that can be calculated from Equations (4) which combines estimates from Columns (2) and (3).
All the geographic factors considered are also significantly associated with the severity of the disease. The coefficient for log population density is estimated to range between 0.23 and 0.27 across specifications, implying that a 1% increase in the density of the population is associated with a 0.23%-0.27% increase in the number of active cases. Indeed, higher population density would tend to increase the likelihood of inter-community contagion, even under active cases. Given that population density and urbanization are becoming important hubs for the transmission of infectious diseases, particularly true for diseases transmitted via respiratory and fecal-oral routes (such as influenza, measles, tuberculosis, severe acute respiratory syndrome, etc.), given the increase in the amount of shared airspace. These authors also showed that cities are acute respiratory syndrome, etc.), given the increase in the amount of shared airspace. These authors also showed that cities are

All the geographic factors considered are also significantly associated with the severity of the disease. The coefficient for log population density is estimated to range between 0.23 and 0.27 across specifications, implying that a 1% increase in the density of the population is associated with a 0.23%-0.27% increase in the number of active cases. Indeed, higher population density would tend to increase the likelihood of inter-community contagion, even under social distancing measures. Alirol et al. (2011) explain that this is increase the likelihood of inter-community contagion, even under of active cases. Indeed, higher population density would tend to population is associated with a 0.23%-0.27% increase in the number of active cases. Indeed, higher population density would tend to increase the likelihood of inter-community contagion, even under social distancing measures. Alirol et al. (2011) explain that this is

| Dependent variable: Log Active Cases | (1) | (2) | (3) |
|-------------------------------------|-----|-----|-----|
| Constant                            | 0.5878 | 1.1490 | 0.1817 |
| (1.2388)                            | (1.3775) | (1.5711) | |
| SSA                                 | – | –1.8027** | –0.2496 |
| (0.8838)                            | (1.0689) | | |
| **Duration**                        | 0.0404*** | 0.0371*** | 0.0352*** |
| (0.0129)                            | (0.0141) | (0.0132) | |
| **SSA × Duration**                  | – | 0.0633** | 0.2275*** |
| (0.0289)                            | (0.0785) | | |
| Log Diabetes prevalence            | 0.0014 | –0.0138 | 0.0077 |
| (0.0440)                            | (0.0544) | (0.0589) | |
| Log Population aged ≥65            | 0.0939*** | 0.0854* | 0.0899*** |
| (0.0442)                            | (0.0454) | (0.0463) | |
| Log Population density             | 0.2217** | 0.2304** | 0.2638** |
| (0.1099)                            | (0.1092) | (0.1116) | |
| Urban population                    | 0.0328*** | 0.0313*** | 0.0371*** |
| (0.0102)                            | (0.0098) | (0.0102) | |
| Temperature                         | –0.0381** | –0.0372** | –0.0375** |
| (0.0191)                            | (0.0190) | (0.0192) | |
| Log GDP per capita                  | 0.0890 | 0.0597 | 0.1072 |
| (0.1950)                            | (0.2016) | (0.2218) | |
| Health expenditure                  | –0.1344 | –0.0135 | –0.1224 |
| (0.3745)                            | (0.3662) | (0.3779) | |
| **SSA × Predicted**                | – | – | –1.3870** |
|                                  | – | – | (0.6194) |
| Observations                        | 154 | 154 | 154 |
| R-squared                           | 0.6006 | 0.6084 | 0.6257 |
| Adjusted R-squared                  | 0.5785 | 0.5810 | 0.5967 |
| *Waldχ²                             | 195.09 | 235.70 | 277.30 |

Notes: Significance code are *, **, and ***. 10%, 5% and 1% statistical significance, respectively. Bootstrap standard errors (using 500 replications) are reported in parenthesis. * Wald χ² reports the test statistic of global significance based on bootstrap standard errors. All the values reported in columns (1), (2) and (3) indicate global significance at 1%.

Average temperature around the first quarter of the year (January–March) is also another relevant geographic factor, not only because recent research has suggested that temperature and climatological factors could influence the spread of this novel coronavirus in general (de Ángel Solá, Wang, Vázquez, & Lázaro, 2020; Liu et al., 2020), but also because this particular quarter of the year corresponds to when the novel coronavirus has been initially spreading. We found this indicator to be negatively associated with the pandemic spread in our estimations, showing that a 1°C decrease in average temperature around this quarter is associated with about 3.8% increase in the number of active cases of COVID-19. This result means that countries with relatively higher temperature at this time of the year, such as sub-Sahara African countries, would tend to have a relatively lower number of cases, everything else equal.

An important aspect of this regression analysis is that it allows to formally assess how the effects of the demographic and geographic factors discussed above differ between SSA and the rest of the world. The key parameter for this assessment is the coefficient of the interaction term between SSA and the predicted values of the regression in Column (2), presented in Column (3). This parameter is estimated at −1.39 and is significant at 5%. It implies that any factor that is positively associated with the spread of COVID-19 worldwide would have, on average, a 1.39% lower effect on the number of active cases in sub-Saharan Africa compared to the rest of the World, for any percent shift. This is the case for all demographic and geographic factors considered. However, when we control for all these factors, the number of active cases in sub-Saharan Africa is not significantly different from those of other parts of the world. This is evidenced from the non significance of the SSA dummy in both the reduced form estimate and in the specification given in Column (3). This implies that any assumption about SSA being relatively safer through unobserved heterogeneity such as some pre-existing immunity (e.g. Guerrini & Oshadiya, 2020) should be taken with great caution. This may also mean that any number of cases by which SSA is exceeding the rest of the world is being offset on average by the amount of underreporting, conditional on the DG factors.14

13 Using the formulas in (4) and the delta method, the reduced form coefficient for the SSA dummy is inferred from estimates in Columns (2) and (3) is computed at

\[ \hat{q}_2 = -0.25 - 1.39 \times (-1.8 + 1.15) - 0.661, \]

with a standard error of 0.957.

14 We cannot, however, develop this argument further, given that these numbers are not identified.
The last set of indicators whose role are examined in this analysis are those related to income and the quality of the health system. A large body of the health economics literature shows that these factors contribute to improving health outcomes (Cutler, Deaton, & Lleras-Muney, 2006). However, our estimation results show that GDP per capita, our measure of aggregate income, is insignificant in all our specifications (see Columns (1)–(3)). This means that, although this factor provides the opportunity to improve material conditions, subsidize effective containment measures such as social distancing, and improve related public goods as shown by many research (e.g. Marmot, 2002, Condliffe & Link, 2008), it is not associated with a significantly lower level of COVID-19 spread, when demographic and geographical factors are controlled for. Thus, given these DG factors, the fact that SSA is relatively poor compared to other regions of the world does not seem to be a crucial issue in addressing the pandemic in the continent, unlike initially thought. Another important issue that has spurred concerns about this pandemic for the African continent is the fragile health systems in most Sub-Saharan African countries, and the fear that new or re-emerging disease outbreaks such as the current COVID-19 pandemic could potentially paralyze health systems at the expense of primary healthcare requirements (Velavan & Meyer, 2020). However, using public health expenditure as a recognized indicator for the quality of healthcare infrastructure (see Ssozi & Amlani, 2015; Gallet & Doucouliagos, 2017; Obrizan & Wehby, 2018), we found no statistically significant association with the active spread of COVID-19, or its containment thereof. This suggests that the relative fragility of health infrastructure in SSA countries and their relatively weak capacity to diagnose and handle outbreaks compared to other regions does not constitute a significant catalyst of the COVID-19 spread.

The results discussed above are subject to the caveats of data quality already raised in Section 2, which may remain pending even after attempting to mitigate it with our econometric strategy. The results on epidemiologic indicators (i.e. prevalence of diabetes or CVD) can not bear a causal interpretation given the possible underlying endogenous behavioral adjustment discussed in the estimation section and unaccounted by our model. Our most credible estimates are the effects of demographic and geographic factors which are largely exogenous and are found to be significant and robust across all our specifications. They provide compelling evidence that may help understand why the number of infected cases of COVID-19 has been growing slower in sub-Saharan Africa and has remained relatively low compared to other regions of the world. These findings are however credible to the extent that the measurement errors in the dependent variable are uncorrelated with these factors as is usually assumed in the econometric literature (see Bound, Brown, & Mathiowetz, 2002, Hausman, 2001). Otherwise, a more sophisticated model of misreporting is needed and may require other methodological investigations that are beyond the scope of this work.

4. Concluding remarks

The goal of this paper was to assess the role of demographic and geographic factors in explaining the spread of COVID-19, with the aim of understanding why the epidemic is progressing relatively slower in sub-Saharan Africa. We employ a Ramsey-type device that preserves degrees of freedom in a regression analysis framework that accounts for possible misreporting to estimate the number of active COVID-19 cases as a function of these factors. We found that the proportion of population aged 65+, population density and urban population rate are positively associated with the number of active cases, whereas average temperature around the first quarter of the year (January-March) is negatively associated with this epidemic outcome. Because sub-Sahara African countries exhibit both lower rates of the former factors and higher levels of the latter, they are less affected than other countries by these drivers. As a consequence, these factors are found to have lower marginal effects on the number of active cases in sub-Saharan Africa compared to the rest of the world. These results help understand the relatively low progression of the pandemic in sub-Saharan Africa, compared to the rest of the world. However, this advantage that sub-Saharan Africa seems to have regarding the spread of COVID-19 disappears once we take away demographic and geographic characteristics. This suggests that any assumption that sub-Saharan African countries could be benefiting from pre-existing immunity conditions beyond the above-mentioned factors should be taken with caution.

While the number of active cases increases with the duration of the epidemic, our results show that the perverse effect of time is exacerbated in sub-Sahara African countries compared to the rest of the world, in spite of the former having a learning advantage.
This means that the comparative advantage that SSA seems to have now could narrow and possibly reverse in the future as the pandemic evolves amid no medical solutions. This therefore calls for awareness and strategies to implement mitigation efforts and containment measures that pertain to the SSA situation. Our results provide insights for policies that could be implemented to overcome disease spreads of the coronavirus type. In particular, given that geographic factors such as urbanization and dense populations appear to have the largest and most significant impacts in our analysis, successful policies and programs to address the spread and severity of such diseases should leverage on geographical eco-system. This includes sensible planning of the expansion of cities as well as the integration of health and social distancing concerns into urban policies.

The most evident limitation of our analysis is the quality of the publicly available data that we used and the associated misreporting in the outcome variable. Econometric approaches to deal with these issues such as the one we employed may not fully mitigate it or fully identify some relevant components of the relationship, especially if the measurement errors are correlated with explanatory factors. Another important limitation is the inability of the model to measure the endogenous behavioral responses of some of the key explanatory variables. Addressing these concern would require better quality data as well as ancillary information about these behavioral responses. These considerations are left as possible avenues of future research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alirol, E., Getaz, L., Stoll, B., Chappuis, F., & Loutan, L. (2011). Urbanisation and infectious diseases in a globalised world. The Lancet Infectious Diseases, 11(2), 131–141.

Bound, J., Brown, C., & Mathiowetz, N. (2002). Measurement error in survey data. In J. Heckman & E. Leamer (Eds.), Handbook of econometrics (Volume 5, pp. 3705–3843). New York: Springer-Verlag.

Condiffe, S., & Link, C. R. (2008). The relationship between economic status and child health: Evidence from the United States. American Economic Review, 98(4), 1605–1618.

Cutler, D., Deaton, A., & Lleras-Muney, A. (2006). The determinants of mortality. Journal of Economic Perspectives, 20(3), 97–120.

de Ângel Solá, D. E., Wang, L., Vázquez, M., & Lázaro, P. A. M. (2020). Weathering the pandemic: How the Caribbean Basin can use viral and environmental patterns to predict, prepare and respond to COVID-19. Journal of Medical Virology.

Driggin, E., Madhavan, M. V., Bikdeli, B., Chuah, T., Laracy, J., Bondi-Zoccai, G., et al. (2020). Cardiovascular considerations for patients, health care workers, and health systems during the coronavirus disease 2019 pandemic. Journal of the American College of Cardiology.

Fang, L., Karakulakis, G., & Roth, M. (2020). Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? The Lancet Respiratory Medicine.

Gallet, C. A., & Dosouliages, H. (2017). The impact of healthcare spending on health outcomes: A meta-regression analysis. Social Science & Medicine, 179, 9–17.

Greif, M. J., Dodoo, F. N. A., & Jayaraman, A. (2011). Urbanisation, poverty and sexual behaviour: The tale of five African cities. Urban Studies, 48(5), 947–957.

Guerrini, L., & Oshadiya, M. (2020). Potential link between anti malaria prophylaxis and the prevention of COVID-19 infection.

Halter, J. B. et al. (2014). Diabetes and cardiovascular disease in older adults: Current status and future directions. Diabetes, 63, 2578–2589.

Hausman, J. (2001). Mis-measured variables in econometric analysis: Problems from the right and problems from the left. Journal of Economic Perspectives, 15(4), 57–67.

Liu, J., Zhou, J., Yao, J., Zhang, X., Li, L., Xu, X., et al. (2020). Impact of meteorological factors on the COVID-19 transmission: A multi-city study in China. Science of The Total Environment, 138513.

Marmot, M. (2002). The influence of income on health: Views of an epidemiologist. Health Affairs, 21(2), 31–46.

Nkengasong, J. N., & Mankoula, W. (2020). Looming threat of COVID-19 infection in Africa: Act collectively, and fast. The Lancet, 395(10227), 841–842.

Obrizan, M., & Wehby, G. L. (2018). Health expenditures and global inequalities in longevity. World Development, 191, 28–36.

Our world in data. (2020). Data retrieved from Our world in data https://ourworldindata.org/coronavirus-source-data/ (Accessed April 10, and 27 April 2020).

Sozzi, J., & Amlani, S. (2015). The effectiveness of health expenditure on the proximate and ultimate goals of healthcare in Sub-Saharan Africa. World Development, 76, 165–179.

Stillwaggon, E. (2002). HIV/AIDS in Africa: Fertile terrain. Journal of Development Studies, 38(6).

Van de Poel, E., O’Donnell, O., & Van Doorslaer, E. (2012). Is there a health penalty of China’s rapid urbanization?. Health Economics, 21(4), 367–385.

Velavan, T. P., & Meyer, C. G. (2020). The Covid-19 epidemic. Tropical Medicine and International Health. https://doi.org/10.1111/tmi.13383.

World Health Organization. (2020). Coronavirus disease 2019 (COVID-19): situation report, 72.

World Population Prospects. (2019). United Nations, Department of Economic and Social Affairs. Available at https://population.un.org/wpp/Download/Standard/Population/.

Worldometer Coronavirus. (2020). COVID-19 Coronavirus Pandemic. Available at https://www.worldometers.info/coronavirus/ (Accessed 10 April 2020).

Yang, X., Yu, Y., Xu, J., Shu, H., Liu, H., Wu, Y., et al. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: A single-centered, retrospective, observational study. The Lancet Respiratory Medicine.