Using Google Health Trends to investigate COVID-19 incidence in Africa

Alexander Fulk[1], Daniel Romero-Alvarez[1,2,3]*, Qays Abu-Saymeh[1], Jarron M. Saint Onge[4], A. Townsend Peterson[1,2], Folashade B. Agusto[1]

[1]Department of Ecology & Evolutionary Biology, University of Kansas, Lawrence, Kansas, United States

[2]Biodiversity Institute, University of Kansas, Lawrence, Kansas, United States

[2]OneHealth Research Group-Facultad de Ciencias de la Salud, Universidad de las Américas, Quito, Ecuador

[3]University of Kansas Medical Center, Kansas City, Kansas, United States

[4]Department of Sociology, University of Kansas, Lawrence, Kansas, United States

*Corresponding author: Daniel Romero-Alvarez. Biodiversity Institute and Department of Ecology & Evolutionary Biology, 1345 Jayhawk Blvd. Lawrence, Kansas, 66045. U.S.

daromero88@gmail.com || da.romero@ku.edu

ABSTRACT:

Objectives: In the present manuscript we used Internet-derived data to assess whether Google Health Trends (GHT) search counts are able to track COVID-19 incidences in Africa.

Methods: We collected COVID-19 case and death incidence for 54 African countries from February 2020 to January 2021. We used GHT to characterize COVID-19 incidence up to 24 January 2021, collecting the number of searches of four terms: 'coronavirus', 'coronavirus symptoms', 'COVID19', and 'pandemic'. The terms were related to weekly COVID-19 case incidences for the study period via multiple linear regressions. We also collected 72 predictors assessing Internet accessibility, demographics, economics, etc., to explain the potential mechanisms linking the relationship between GHT searches and COVID-19.

Results: Important increases for COVID-19 death incidence were observed for South Africa and Tunisia at the end of the study period. Our study demonstrated a lack of correlation between GHT and COVID-19 incidence for most African countries. The predictors analyzed were unhelpful in explaining the pattern of GHT statistics and their relationship to COVID-19, complicating interpretability of GHT.

Conclusions: According to our results, GHT-based surveillance for an ongoing epidemic might be useful only in specific situations. Future studies might assess the algorithm in different epidemic contexts.

KEYWORDS: COVID-19, Google Health Trends, Digital Epidemiology

INTRODUCTION

Coronavirus disease 2019 (COVID-19) is a respiratory disease caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) discovered in China in 2019. People infected experience a range of symptoms including headache, fever, difficulty breathing, and loss of taste and smell, or they may be completely asymptomatic (CDC, 2020a). Since its discovery, SARS-CoV-2 has spread around the globe, with over 100 million confirmed cases as of 20 February 2021 according to the John Hopkins University (CRC, 2020; Dong et al., 2020). The elderly (>65 years old), as well as those with pre-existing conditions, have the highest risk of mortality if infected (Williamson et al., 2020). COVID-19 spreads via respiratory particles, which allows it to infect others mainly through contaminated aerosols and droplets suspended in the air in closed spaces (CDC, 2020b). Asymptomatic carriers account for a significant amount of secondary transmissions with some reports showing ~80% of infections may occur without symptoms constituting the source of the majority of secondary COVID-19 cases (CDC, 2020a; Li et al., 2020; Zhang et al., 2020).

After the Ebola outbreak in 2015, African leaders were aware that swift and decisive action was needed to avert the spread of COVID-19 and prevent healthcare system collapse. This awareness led to wide adoption of mitigation and control efforts that avoided an overwhelming first epidemic wave with a partially structured continental response (Loembé et al., 2020). Regardless, testing in Africa has been limited; about 75% of COVID-19 diagnoses came from tests conducted in only 10 countries (CDC, 2020b; Loembé et al., 2020).

Given difficulties to obtain accurate and timely data on case counts and other epidemiological metrics for COVID-19 worldwide (Fairchild et al., 2018; Wu et al., 2020), the

current pandemic represents an opportunity to use digital epidemiology tools to fill gaps in information. Digital epidemiology is an area of epidemiology that uses digital data to gain insight into disease dynamics (Salathé, 2018; Vlajinac, 2008). The digital data used for this method of surveillance vary widely and may or may have been not intended for epidemiological purposes, and thus it can come from unexpected sources such as restaurant receipts, tweets, Facebook posts, or Google search queries (Copeland et al., 2013; Nsoesie et al., 2014; Pollet et al., 2017a).

Google developed two specific algorithms to address infectious diseases, Google Flu Trends (GFT) in 2009 and Google Dengue Trends (GDT) in 2011 (Copeland et al., 2013), which, after inquiries into their usefulness, were shut down in 2015 (Lazer et al., 2014). Currently, Google maintains two portals by which to harvest search query data: Google Trends (GT) and Google Health Trends (GHT). GT inquiries yield a ranked score from 0 to 100 based on the highest frequency of searches for a term in a particular time period. GHT provides search counts from a relative proportion of a random sample of the overall Google search dataset for any particular term in a selected time interval (Romero-Alvarez et al., 2020). Both of these portals have limitations, such as possibly excluding certain groups (e.g., the elderly, rural residents, low income populations), a lack of detail on who is searching certain terms, and the underlying motivations of the searches (Arora et al., 2019).

Digital tools have been used in many instances to predict disease incidence (Althouse et al., 2011; Xu et al., 2020; Zhang et al., 2017) including COVID-19. Kurian et al., 2020, evaluated the applicability of Google Trends (GT) in predicting COVID-19 cases in the United States (U.S.) in a state-by-state analysis. They found that certain keywords had a strong correlation with COVID-19 cases and concluded that GT may be a useful tool for predicting

COVID-19 outbreaks. Brodeur et al., 2021, used GT to see how lockdowns affected well-being in the U.S. Once lockdowns were implemented, well-being likely decreased, as searches for certain terms such as 'stress', 'suicide', and 'worry' increased over the lockdown period. Ahmad et al., 2020 used gastrointestinal-related symptom search terms to determine whether GT could predict COVID-19 incidence, and found correlations between the search terms and increases of COVID-19 cases in multiple regions across the U.S. with a four-week lag.

Here, we explored whether GHT search query data correlate with COVID-19 incidence at the country-level in Africa as a potential complementary source for more customary forms of COVID-19 surveillance. We collected case and death data for 54 African countries, and used four COVID-19-related search terms (see below) for each African country. We then assessed whether Internet accessibility, demographic, economic, and health variables, could explain GHT usefulness. Lastly, we calculated a standardized volatility index to illuminate whether variability in the signal of case incidence led to less accurate predictions by GHT.

METHODS

COVID-19 Incidence data

Daily COVID-19 new cases and death counts were obtained for all 54 African countries from 2 February 2020 to 28 January 2021. Country-level case data were obtained via the Johns Hopkins COVID-19 global time series on the pandemic (CSSE, 2020), and were constrained to lab-confirmed cases only. We explored the progression of average daily COVID-19 case and death incidence in Africa using four three-month time periods: (a) 2 February to 30 April, (b) 1 May to 31 July, (c) 1 August to 31 October, and (d) 1 November 2020 to 28 January 2021. We then converted daily new cases into weekly new cases for each of the countries to match the

GHT data up to 24 January 2021, for a total of 51 observations. We calculated weekly incidence rates by dividing the number of cases per week by the total population per country in millions (Vandenbroucke and Pearce, 2012). Country-level population data were collected from the forecasted midyear 2020 estimates from the U.S. Census Bureau (US Census Bureau, 2020).

Google Health Trends data

We downloaded four English terms from the GHT application programming interface (API): ‘coronavirus’, ‘coronavirus symptoms’, ‘COVID19’, and ‘pandemic’. Although the four terms are conceptually related, they have the potential to capture a broad spectrum of information related to the disease (Asseo et al., 2020; Romero-Alvarez et al., 2020). We matched the relative search proportions of these words—which is the raw output provided by GHT (Romero-Alvarez et al., 2020)—with the weekly COVID-19 case incidence for the selected time period.

Statistical analysis

We used a multiple linear regression model fitted with the four GHT search terms as predictors of COVID-19 incidence at the country level for each of the 54 African countries being evaluated. The primary outcome measure was the adjusted R^2 statistic. If one or more of the four terms chosen did not retrieve search counts from GHT, it was removed from the analysis for that country. A minimum of two terms were included for each region.

Next, we used the adjusted R^2 statistics collected from the 54 African countries as our dependent variable and explored whether different categories of predictors might explain the pattern obtained. Predictors for the African countries included Internet access, demographic,

economic, and health indicators (Table 1); data were gathered from the World Bank (The World Bank, 2020). We explored logarithmic transformations of each of these predictors to determine if normalization of the indicators led to stronger correlations. Finally, we included a predictor based on a standardized volatility index calculated using the standardized normalized case incidence data of each country as follows:

$$Volatility = \frac{1}{n-1}\sum_{i=2}^{n} |Y_i - Y_{i-1}|$$

in which *n* is the total number of observations and *Y* is the normalized case incidence per country. The average of the absolute difference (i.e., volatility) summarizes the case signal reflecting if it is relatively constant or fluctuates broadly from week to week (Romero-Alvarez et al., 2020). Overall, we explored a total of 72 potential explanatory variables (Table 1 and Supplementary Table 1).

Table 1. Predictors explored in the present study. Different categories were selected based on their potential to explain patterns of the GHT and COVID-19 regression models. GDP = gross domestic product; HIV = human immunodeficiency virus. Raw values of the variables can be found in Supplementary Table 1.

Category	Indicator
Internet access	1. Percentage of population with access to electricity. 2. Fixed total number of broadband subscriptions in a country. 3. Fixed broadband subscriptions per 100 people. 4. Fixed total number of telephone subscriptions in a country. 5. Fixed telephone subscriptions per 100 people. 6. Percentage of individuals using the Internet. 7. Total number of mobile cellular subscriptions in a country. 8. Mobile cellular subscriptions per 100 people. 9. Secure Internet servers per 1 million people.
Demographics	10. Percentage of people 15 years and above that are literate. 11. Percentage of people using at least basic drinking water services. 12. Percentage of people using at least basic sanitation services. 13. Percentage of people using safely managed drinking water services. 14. Percentage of people using safely managed sanitation services. 15. Percentage of people with basic handwashing facilities. 16. Total population. 17. Population density as people per square km of land area. 18. Total urban population. 19. Percentage of urban population.
Economics	20. Percentage of GDP for current health expenditure.

	21. GDP based on current U.S.$ value. 22. GDP per capita based on current U.S.$ value.
Health	23. Average weekly cases over the studied period. 24. Community health workers per 1000 people. 25. Cumulative total deaths over the study period. 26. Hospital beds per 1000 people. 27. Total life expectancy (years) at birth. 28. Nurses and midwives per 1000 people. 29. Physicians per 1000 people. 30. Percentage of population 15-49 years with HIV. 31. Prevalence of moderate or severe food insecurity in the population. 32. Prevalence of severe food insecurity in the population. 33. Percentage of people at risk of catastrophic expenditure for surgical care. 34. Percentage of people at risk of impoverishing expenditure for surgical care. 35. Smoking prevalence for people above 15 years.
Miscellaneous	36. Volatility score for a country calculated using weekly incidence.

Variables were analyzed individually using a pair-wise univariate linear regression and collectively in a stepwise regression in which predictors were iteratively added and removed to obtain a subset of predictors that provided the best model outcome as measured by the Akaike Information Criterion (AIC). Countries with missing variable information were removed from the analysis of that particular variable (38/72; 53% of variables had at least one country removed, Supplementary Table 1) and only variables with information for every country were used in the stepwise regression. All the analyses were performed in R (R Core Team, 2017). Data and scripts to replicate the results of this study are available in a GitHub repository accompanying this publication (https://github.com/alxjfulk/GHT-and-COVID19-code).

RESULTS

Examining the distribution of first cases among the 54 African countries, we observed that dates of first reported COVID-19 cases were centered around March 2020. Egypt (EGY) reported the first case of COVID-19 on the continent on 14 February 2020, 15 days after the World Health Organization (WHO) declared the COVID-19 epidemic an emergency of international concern (WHO, 2020). Comoros (COM) and Lesotho (LSO) were the last countries to report COVID-19 introductions, on 30 April and 13 May 2020, respectively (Figure 1).

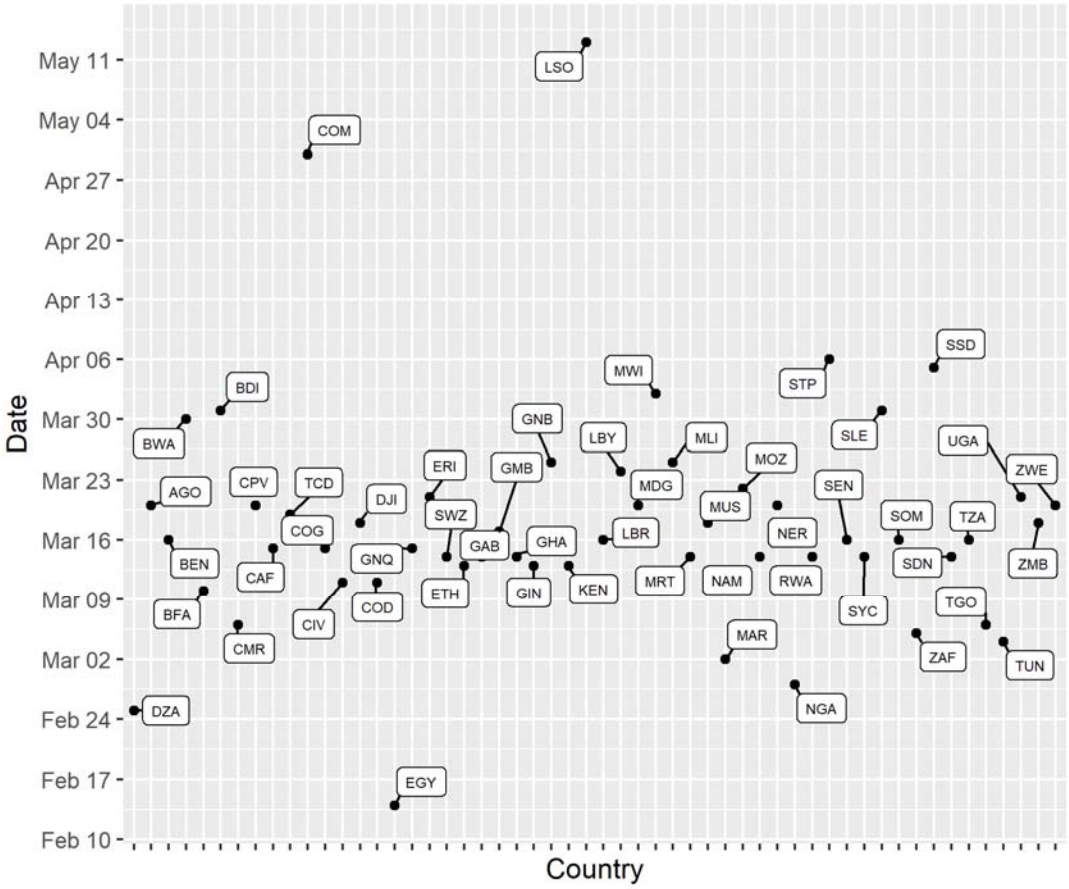

Figure 1. Distribution of the day of the first COVID-19 reported case in 54 African countries. The plot depicts the dates of the first reports of COVID-19 cases in the 54 studied African countries as reported by the Johns Hopkins global time series on the pandemic (CRC, 2020; Dong et al, 2020). The countries in this

distribution are designated by their three-letter Alpha-3 codes: DZA: Algeria, AGO: Angola, BEN: Benin, BWA: Botswana, BFA: Burkina Faso, BDI: Burundi, CPV: Cabo Verde, CMR: Cameroon, CAF: Central African Republic, TCD: Chad, COM: Comoros, COD: Democratic Republic of the Congo, COG: Congo, CIV: Côte d'Ivoire, DJI: Djibouti, EGY: Egypt, GNQ: Equatorial Guinea, ERI: Eritrea, SWZ: Eswatini, ETH: Ethiopia, GAB: Gabon, GMB: Gambia, GHA: Ghana, GIN: Guinea, GNB: Guinea-Bissau, KEN: Kenya, LSO: Lesotho, LBR: Liberia, LBY: Libya, MDG: Madagascar, MWI: Malawi, MLI: Mali, MRT: Mauritania, MUS: Mauritius, MAR: Morocco, MOZ: Mozambique, NAM: Namibia, NER: Niger, NGA: Nigeria, RWA: Rwanda, STP: Sao Tome and Principe, SEN: Senegal, SYC: Seychelles, SLE: Sierra Leone, SOM: Somalia, ZAF: South Africa, SSD: South Sudan, SDN: Sudan, TZA: United Republic of Tanzania, TGO: Togo, TUN: Tunisia, UGA: Uganda, ZMB: Zambia, ZWE: Zimbabwe.

Countries with the highest COVID-19 case incidences for the first time period include Djibouti ($1.33x10^{-5}$), Equatorial Guinea ($4.23x10^{-6}$), and Central African Republic ($3.44x10^{-6}$) (Fig. 2). During the second period, South Africa ($9.39x10^{-5}$), Equatorial Guinea ($5.86x10^{-5}$), and Djibouti ($4.71x10^{-5}$) were most affected (Fig. 2). For the third and fourth periods, countries across the continent reported increased COVID-19 incidences, with Libya (third period = $9.07x10^{-5}$; fourth period = $9.08x10^{-5}$), Morocco (third period = $5.95x10^{-5}$; fourth period = $7.9x10^{-5}$), Cameroon (third period = $1.18x10^{-4}$; fourth period = $9.61x10^{-5}$), Tunisia (third period = $5.4x10^{-5}$; fourth period = $1.39x10^{-4}$), and South Africa (third period = $4.47x10^{-5}$; fourth period = $1.42x10^{-4}$) ranking top among the studied countries (Fig. 2). Tanzania had an incidence of 0 for the third and fourth time periods, which will be discussed below.

COVID-19 death incidence was recorded for all the African countries in the second time period except for Eritrea, Seychelles, Mauritius, and Burundi, although the latter two reported $8.15x10^{-8}$ and $9.47x10^{-10}$ death incidences during the first period, respectively. Further, South Africa ($1.52x10^{-6}$) and Equatorial Guinea ($1.07x10^{-6}$) were the countries reporting the highest death incidence between 1 May and 31 July 2020. For the third period, the highest death

incidences were reported in South Africa (2.17×10^{-6}) and Cameroon (1.34×10^{-6}), while for the fourth period, highest incidences were in Tunisia (4.98×10^{-6}), South Africa (4.74×10^{-6}), and Eswatini (4.28×10^{-6}).

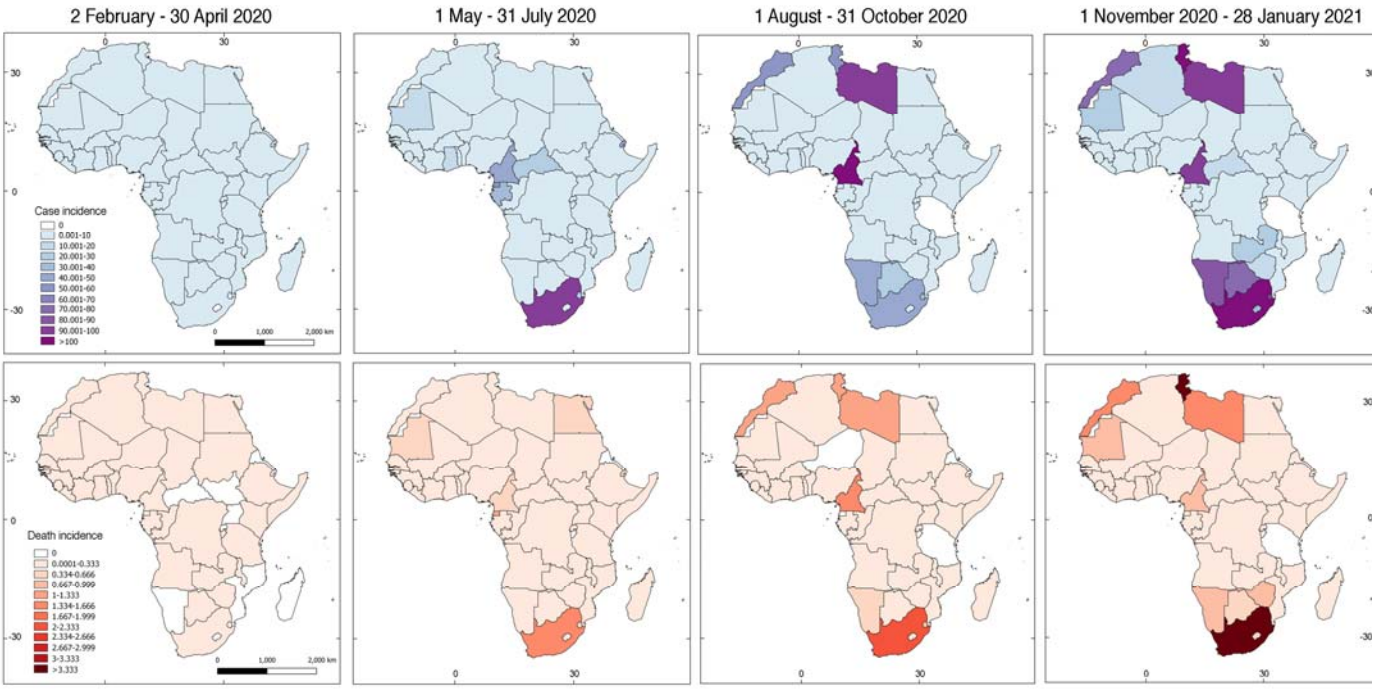

Figure 2. Average case and death incidences of COVID-19 over four three-month time periods in Africa. Eight plots showing the average case incidences (upper panels in purple) and average death incidences (bottom panels in orange) over four three-months time periods from 2 February 2020 to 28 January 2021. Scale is the same for all case/death incidence maps and is depicted in the left panels.

Several countries had no information for one or two of the chosen terms (6/54; 11.1%); only 'coronavirus' and 'COVID19' always recovered search query counts (Supplementary Table 2). The adjusted R^2 collected to depict the relationship between GHT search queries and COVID-19 weekly incidence was generally low, never going above 0.6 for any of the countries (Fig 3). The largest adjusted R^2 results were obtained from Mauritius (0.54), Tanzania (0.45), and Cabo

Verde (0.3; Fig. 4). The countries with the lowest adjusted R^2 results included Equatorial Guinea (-0.076), Niger (-0.071), and Seychelles (-0.046; Figs. 3, 4, and Supplementary Table 2).

None of the 72 indicators were able to predict the pattern of adjusted R^2 statistics obtained for the 54 African countries. All linear analyses yielded R^2 values of 0.25 or less (Supplementary Table 3). Stepwise regression analysis showed that a model including percentage of GDP for current health expenditure, life expectancy (years) at birth, total population, GDP per capita based on current U.S.$ value, percentage of people using at least basic drinking water services, and secure Internet servers per 1 million people yielded the lowest AIC score with untransformed data with an adjusted R^2 value of 0.39. A model including average weekly cases over the studied period, percentage of GDP for current health expenditure, percentage of individuals using the Internet, life expectancy (years) at birth, total population, GDP per capita based on current U.S.$ value, and percentage of people using at least basic drinking water services yielded the lowest AIC scored with logarithmically transformed data with an adjusted R^2 value of 0.35. Both of these multiple linear regression models yielded adjusted R^2 values larger than any of the univariate analyses (Supplementary Table 3); however, they were still in a lower performance threshold (James et al., 2013; Gluskin et al., 2014; Romero-Alvarez et al., 2020).

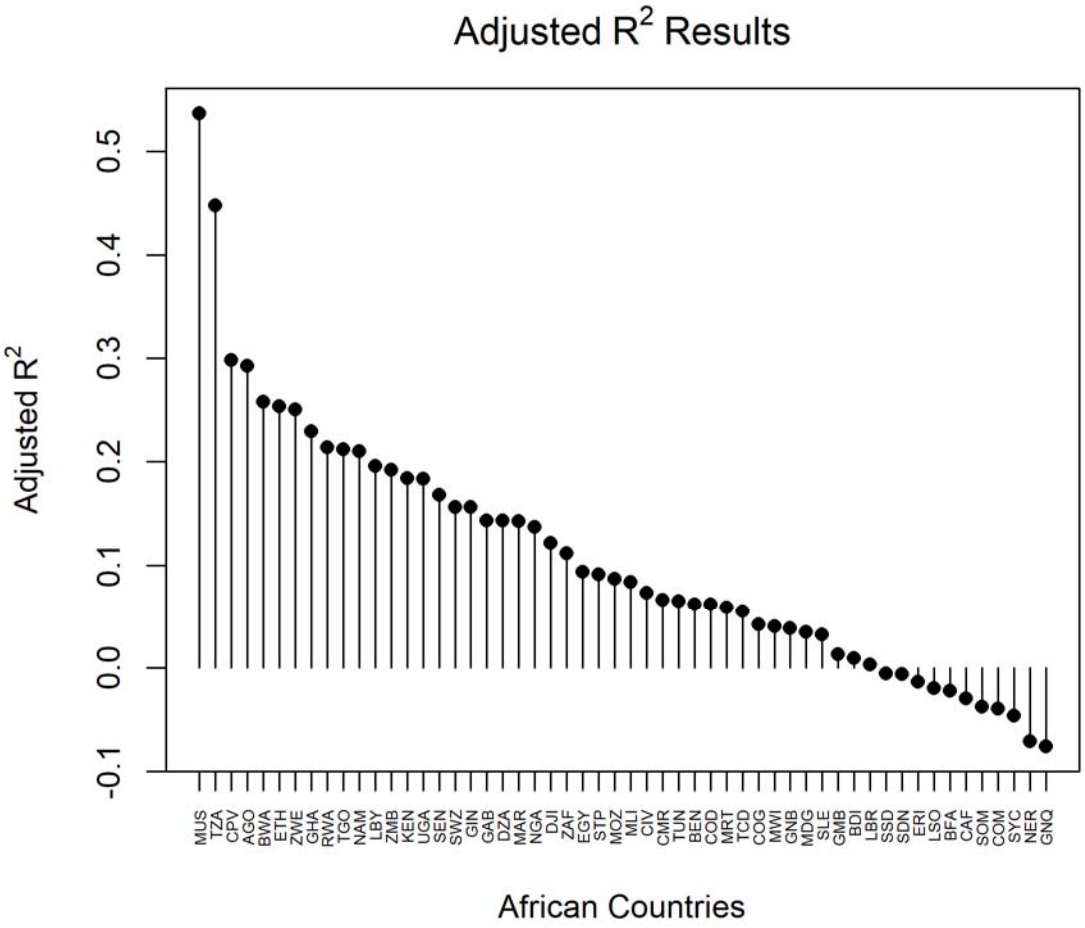

Figure 3: Results of multiple linear regression analysis between COVID-19 incidence and Google Health Trends search terms. The adjusted R^2 of each analysis is depicted here to represent visually countries from the highest to lowest performance. The countries in this figure are designated by their three-letter Alpha-3 code as in Figure 1.

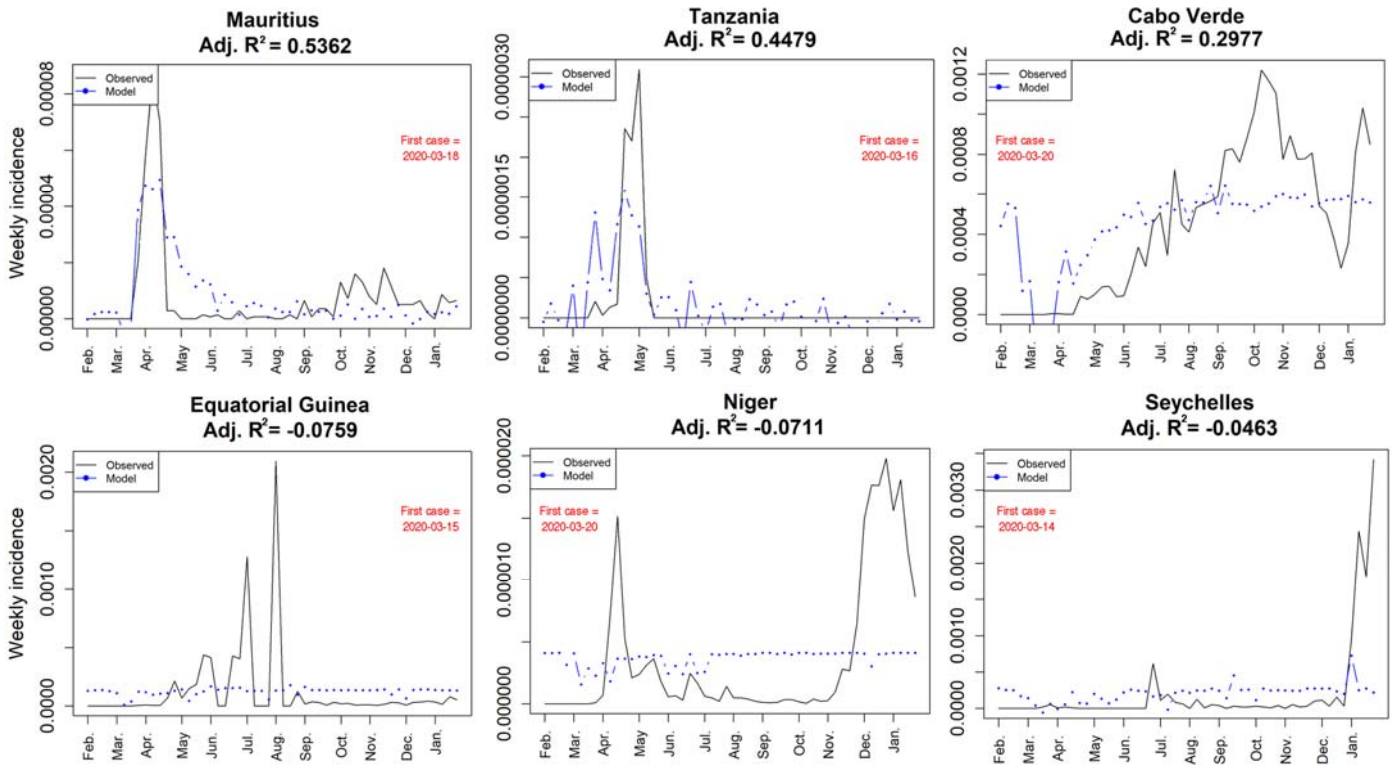

Figure 4. Best and worst Google Health Trends (GHT) performing countries. When analyzing whether GHT correlated with case incidence (black line) via a multiple linear regression analysis (blue line), the three best performing countries were Mauritius, Tanzania, and Cabo Verde, respectively (upper panels). The three worst performing countries were Equatorial Guinea, Niger, and Seychelles, respectively (bottom panels).

DISCUSSION

Despite successful demonstrations of the GHT algorithm to aid infectious disease surveillance for influenza, dengue, and other diseases (Arora et al., 2019; Gluskin et al., 2014; Messina et al., 2019), our study demonstrates that, in the context of the COVID-19 epidemic, GHT appeared to be difficult to implement. None of the predictors, either individually or in combination, allowed us to clearly describe the pattern of adjusted R^2 statistics collected from the correlation analysis between GHT terms and COVID-19 incidence, complicating the

interpretability of GHT usefulness even in the countries with the larger adjusted R^2 values (Figs. 3 and 4).

Mauritius and Tanzania seemed to have a similar COVID-19 incidence signal type. Cases in each country spiked early and were followed by a rapid decrease in the reporting of the disease (Fig. 4, upper panels). This may indicate that GHT might be best suited for earlier epidemic stages where a rapid growth of infection coincides with interest in the topic and Internet search volume for disease-specific terms is likely to be high. On the other hand, the three countries with the worst GHT prediction (Fig. 4, lower panels), showed growth of case detectability later. Thus, even for the best performing countries, we remain wary about endorsing GHT usefulness before further research is able to explain this relationship.

As in the rest of the world, incidence of both COVID-19 cases and COVID-19 related death, increased across Africa steadily during the study period. However, beginning the third study period, Tanzania showed zero COVID-19 cases, a pattern that continued up to the fourth period (Fig. 2). Upon closer examination, the country stopped reporting coronavirus cases and deaths in April of 2020. It appears that the higher adjusted R^2 value for Tanzania is actually reflecting the lack of data compared to other countries (Makoni, 2021). While COVID-19 numbers are concerning on the continent, Africa has been observed to have a lower disease burden in comparison to other regions of the world (BBC News, 2020a,b; European CDC, 2020; Maeda and Nkengasong, 2021; but see Mwananyanda et al., 2021). As of August 2020, Africa had reported approximately 900,000 cases and 17,000 deaths in nearly seven months since COVID-19 was declared a pandemic. In comparison, the U.S. at that point in time had surpassed five million cases and Brazil confirmed nearly 100,000 deaths (CRC, 2020; Dong et al, 2020), even though each has only a fraction of the population of Africa (Maeda and Nkengasong, 2021).

Although digital epidemiology represents the next frontier of infectious disease surveillance (Pollett et al., 2017a; Salathé, 2018), the present modeling effort demonstrates that search queries from GHT might be difficult to correlate with incidence of disease in the context of emerging epidemics. In contrast with diseases such as influenza or dengue that are consistently studied in a seasonal pattern or are endemic to multiple regions (Gluskin et al., 2014; Pollett et al., 2017b; Romero-Alvarez et al., 2020), COVID-19 represented an unprecedented case study that might render Google-based information mining ineffective for several reasons: (a) partial or incomplete COVID-19 case reporting (Loembé et al., 2020; Schneider, 2020), (b) media-induced search behavior (Southwell et al., 2016), or even (c) information fatigue (Asseo et al., 2020). Thus, we encourage caution regarding the interpretation of COVID-19 modeling experiments based on Google search engines. For example, Ahmad et al., 2020, found a correlation between gastrointestinal search terms obtained through GT and COVID-19 cases and suggested that Internet searches may be useful in predicting COVID-19 cases using a four-week lag in the U.S. This correlation might be an artifact since none of the gastrointestinal terms are specific to COVID-19, and the only COVID-19 specific term—'ageusia'—increased during the time that the pandemic was declared (i.e., 11 March) and decreased while cases started to increase (Figure 1 in Ahmad et al., 2020). The U.S. showed an increase in case numbers driven by increasing test capacity, thus, these case numbers inaccurately were reflecting disease incidence (Stokes et al., 2020). Thus, although our findings are framed on the GHT algorithm, we are cautious about interpreting our own results, and that of others, while characterizing COVID-19 via Google search engines. Similar to our findings, Asseo et al., 2020, found correlations between GT search queries related with smell and taste at the beginning of the pandemic in Italy and the U.S., which then faded in consecutive

epidemiological weeks. More importantly, Asseo et al., 2020, also showed how correlation patterns fall apart when analyzing Google search queries and COVID-19 incidence in nonconsecutive weeks.

We enumerate some limitations of the present modeling exercise. Because of the timeframe of the study and the availability of GHT data as weekly counts, we had to convert daily cases to weekly cases, limiting the data for the analysis to only 51 observations, decreasing the statistical power of our approach. Moreover, the four terms related with COVID-19 that were selected might not be as popular in the region as expected. English is not the primary language used in many African countries, although language has shown to be a permeable barrier (Althouse et al., 2011; Romero-Alvarez et al., 2020). Finally, we lacked complete data for some of the predictors (e.g., prevalence of severe food insecurity in the population; Supplementary Table 3) which impedes the interpretability of several of the indicators used; however, even the ones available for all the countries and that have proved useful in other research studies (e.g., total population, signal volatility, disease incidence, etc; Gluskin et al., 2014; Romero-Alvarez et al., 2020), were inconclusive here.

CONCLUSIONS

According to our results, surveillance for an ongoing epidemic via GHT might be useful in very specific situations. Google instruments to recover population search counts—GT and GHT—are powerful digital epidemiology tools that can lead to greater insight into disease dynamics, and should be studied and implemented depending on the particular context of an outbreak (Kurian et al., 2020; Nuti et al., 2014; Pelat et al., 2009; Romero-Alvarez et al., 2020; Samaras et al., 2020; Valdivia and Monge-Corella, 2010). Future directions to examine GHT on

COVID-19 research include expansion of our analysis to a larger dataset both in time and space. Moreover, other refinements can be implemented combining other forms of digital data (e.g., Twitter, Wikipedia) to determine if the addition of more information improves the predictive power of the model.

ACKNOWLEDGMENTS AND FUNDING

DRA thanks the W.O.G. for support with writing. Our research is supported by the National Science Foundation with the grant number DMS 2028297.

SUPPLEMENTARY MATERIAL:

Supplementary Table 1. Raw data for the predictors used to explore patterns of Google Health Trends search queries and COVID-19 incidence in 54 African countries.

Supplementary Table 2. Date of the first COVID-19 case reported in Africa. Results of multiple linear regression analysis performed between COVID-19 incidence and Google Health Trends search queries from four selected terms.

Supplementary Table 3. Univariate and multivariate linear regression analysis to explore associations between adjusted R^2 of Google Health Trends search queries and COVID-19 incidence in 54 African countries.

Supplementary Figures. Plots depicting the multiple linear regression models between COVID-19 case counts and Google Health Trends search queries for the 54 African countries studied in the present manuscript.

REFERENCES

1. Ahmad I, Flanagan R, Staller K. Increased Internet search interest for GI symptoms may predict COVID-19 cases in US hotspots. Clin Gastroenterol Hepatol 2020;18:2833-34.
2. Althouse BM, Ng YY, Cummings DAT. Prediction of dengue incidence using search query surveillance. PLoS Negl Trop Dis 2011;5:e1258.
3. Arora VS, McKee M, Stuckler D. Google Trends: opportunities and limitations in health and health policy research. Health Policy 2019;123:338-41.
4. Asseo K, Fierro F, Slavutsky Y, Frasnelli J, Niv MY. Tracking COVID-19 using taste and smell loss Google searches is not a reliable strategy. Sci Rep 2020;10:20527.
5. BBC News. Coronavirus in Africa: five reasons why Covid-19 has been less deadly than elsewhere. 2020a. https://www.bbc.com/news/world-africa-54418613 (accessed 22 March 2021).
6. BBC News. Coronavirus: What's happening to the numbers in Africa? 2020b. https://www.bbc.com/news/world-africa-53181555?intlink_from_url=https://www.bbc.com/news/world&link_location=live-reporting-story (accessed 7 December 2020).
7. Brodeur A, Clark A, Fleche S, Powdthavee N. COVID-19, lockdowns and well-being: evidence from Google Trends. J Public Econ 2021:193;104346.
8. CDC. Coronavirus Disease 2019 (COVID-19). 2020a. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html (accessed 3 November 2020).
9. CDC. Coronavirus disease (COVID-19): how is it transmitted? 2020b. https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted (accessed 7 December 2020).

10. Center for Systems Science and Engineering (CSSE). COVID-19 data repository. Johns Hopkins University. 2020. https://github.com/CSSEGISandData/COVID-19 (accessed 7 December 2020).
11. Copeland P, Romano R, Zhang T, Hecht G, Zigmond D, Stefansen C. Google disease trends: an update. Int. Soc. Neglected Trop. Dis. 2013. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/41763.pdf (accessed 23 March 2021).
12. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20:533-34.
13. European Centers Disease Prevention and Control. COVID-19 situation update worldwide, as of 14 December 2020. 2020. https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases (accessed 15 December 2020).
14. Fairchild G, Tasseff B, Khalsa H, Generous N, Daughton AR, Velappan N, et al. Epidemiological data challenges: planning for a more robust future through data standards. Front Public Health 2018;6:336.
15. Gluskin R, Johansson M, Santillana M, Brownstein J. Evaluation of Internet-based dengue query data: Google Dengue Trends 2014. PLoS Negl Trop Dis 2011;8:e2713.
16. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with applications in R. 1sd ed. New York: Springer; 2013.
17. Johns Hopkins Coronavirus Resource Center (CRC). COVID-19 Dashboard. 2020. https://coronavirus.jhu.edu/ (accessed 7 December 2020).

18. Kurian SJ, Bhatti A ur R, Alvi MA, Ting HH, Storlie C, Wilson PM, et al. Correlations between COVID-19 cases and Google Trends data in the United States: a state-by-state analysis. Mayo Clin Proc 2020;95:2370-81.
19. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science 2014;343:1203-5.
20. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science 2020;368:489-93.
21. Loembé MM, Tshangela A, Salyer SJ, Varma JK, Ouma AEO, Nkengasong JN. COVID-19 in Africa: the spread and response. Nat Med 2020;26:999–1003.
22. Maeda JM, Nkengasong JN. The puzzle of the COVID-19 pandemic in Africa. Science 2021;371:27-8.
23. Makoni M. Tanzania refuses COVID-19 vaccines. The Lancet 2021;397:566.
24. Messina JP, Brady OJ, Golding N, Kraemer MUG, Wint GRW, Ray SE, et al. The current and future global distribution and population at risk of dengue. Nat Microbiol 2019;4:1508–15.
25. Mwananyanda L, Gill CJ, MacLeod W, Kwenda G, Pieciak R, Mupila Z, et al. Covid-19 deaths in Africa: prospective systematic postmortem surveillance study. BMJ 2021;372:n334.
26. Nsoesie EO, Buckeridge DL, Brownstein JS. Guess who’s not coming to dinner? Evaluating online restaurant reservations for disease surveillance. J Med Internet Res 2014;16:e22.

27. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of Google Trends in health care research: a systematic review. PLoS ONE 2014;9 e109583.

28. Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron A-J. More diseases tracked by using Google Trends. Emerg Infect Dis 2009;15:1327-8.

29. Pollett S, Althouse BM, Forshey B, Rutherford GW, Jarman RG. Internet-based biosurveillance methods for vector-borne diseases: Are they novel public health tools or just novelties? PLoS Negl Trop Dis 2017a;11:e0005871.

30. Pollett S, Boscardin WJ, Azziz-Baumgartner E, Tinoco YO, Soto G, Romero C, et al. Evaluating Google Flu Trends in Latin America: important lessons for the next phase of digital disease detection. Clin Infect Dis 2017b;64:34–41.

31. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. 2017. https://www.R-project.org/ (accessed 7 December 2020).

32. Romero-Alvarez D, Parikh N, Osthus D, Martinez K, Generous N, del Valle S, et al. Google Health Trends performance reflecting dengue incidence for the Brazilian states. BMC Infect Dis 2020;20:252.

33. Salathé M. Digital epidemiology: what is it, and where is it going? Life Sci Soc Policy 2018;14:1-5.

34. Samaras L, García-Barriocanal E, Sicilia MA. Comparing social media and Google to detect and predict severe epidemics. Sci Rep 2020;10:4747.

35. Schneider EC. Failing the test—the tragic data gap undermining the US pandemic response. N Engl J Med 2020;383:299–302.

36. Southwell BG, Dolina S, Jimenez-Magdaleno K, Squiers LB, Kelly BJ. Zika virus–related news coverage and online behavior, United States, Guatemala, and Brazil. Emerg Infect Dis 2016;22:1320–1.

37. Stokes EK, Zambrano LD, Anderson KN, Marder EP, Raz KM, Felix SEB, et al. Coronavirus disease 2019 case surveillance—United States, January 22–May 30, 2020. Morb Mortal Wkly Rep 2020;69:759–65.

38. The World Bank. Indicators. 2020. https://data.worldbank.org/indicator (accessed 7 December 2020).

39. United States Census Bureau. International Data Base (IDB). 2020. https://www.census.gov/data-tools/demo/idb/ -/country?YR_ANIM=2020&FIPS_SINGLE=AG (accessed 15 December 2020).

40. Valdivia A, Monge-Corella S. Diseases tracked by using Google Trends, Spain. Emerg Infect Dis 2010;16:168.

41. Vandenbroucke J, Pearce N. Incidence rates in dynamic populations. Int J Epidemiol 2012;41:1472-79.

42. Vlajinac H. Epidemiology, aims and scope. In: Kirch W, editor. Encyclopedia of public health. Dordrecht: Springer Netherlands; 2008, p. 350-2.

43. WHO. COVID-19 Public Health Emergency of International Concern (PHEIC) Global research and innovation forum. 2020. https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum (accessed 5 January 2021).

44. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature 2020;584:430-6.

45. Wu SL, Mertens AN, Crider YS, Nguyen A, Pokpongkiat NN, Djajadi S, et al. Substantial underestimation of SARS-CoV-2 infection in the United States. Nat Commun 2020;11:4507.

46. Xu C, Bai Y, Yang H, Cao X, Jia P, Wang Y. Detecting lung cancer trends by leveraging real-world and Internet-based data: infodemiology study. J Med Internet Res. 2020;22:e1684.

47. Zhang Q, Perra N, Perrotta D, Tizzoni M, Paolotti D, Vespignani A. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. Proc. 26th Int. Conf. World Wide Web 2017; 311-19.

48. Zhang R, Li Y, Zhang AL, Wang Y, Molina MJ. Identifying airborne transmission as the dominant route for the spread of COVID-19. Proc Natl Acad Sci 2020;117:14857-63.