Understanding the correlation between population demographics and mortality causes through quantifying the population pyramid

Yasiru Ranasinghe[a,*], Harshana Weligampola[a], Gayanthi Ilangarathna[a], Roshan Godaliyadda[a], Vijitha Herath[a], Parakrama Ekanayake[a], Janaka Ekanayake[a], Samath Dharmaratne[b], Anuruddhika Rathnayake[c], Mallika Pinnawala[d], Sakunthala Yatigammana[d], Ganga Tilakaratna[e]

[a]*Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka, 20400*
[b]*Faculty of Medicine, University of Peradeniya, Peradeniya, Sri Lanka, 20400*
[c]*Ministry of Health, Colombo, Sri Lanka, 01000*
[d]*Faculty of Arts, University of Peradeniya, Peradeniya, Sri Lanka, 20400*
[e]*Institute of Policy Studies, Colombo, Sri Lanka, 00700*

Abstract

Background

While morbidity and mortality are of medical origin, many economic, sociological, anthropological, and environmental factors yield variation in morbidity and mortality statistics across nations and regions. We introduce a novel statistical measure, PoPStat, which enables the analysis of disease severity in the population pyramid, a ubiquitous measure in demographics that encapsulates many socioeconomic and anthropological factors. Specifically, we consider its application to the COVID-19 pandemic under $\text{PoPStat}_{\text{COVID-19}}$ statistic and explore the generalizability to other causes of morbidity and mortality.

Methods

We derived a scalar variable, $\text{PoPStat}_{\text{COVID-19}}$, to correlate with COVID-19 data as of April 08, 2022, with the population demographics conveyed by the population pyramid. The proposed variable was applied to other mortality causes to study the viability of correlating the information-rich population pyramid to these other types of disease severity.

Findings

The $\text{PoPStat}_{\text{COVID-19}}$ provided a mechanism to arrange the shape of the population pyramid in an orderly manner from progressive to regressive structure. The proposed mechanism displayed a better explanatory variable than other standard socioeconomic variables. The $\text{PoPStat}_{\text{COVID-19}}$ provides an excellent explanatory variable for natural mortality causes, especially communicable diseases and diseases dependent on living and healthcare standards.

Interpretation

A higher value in the correlation coefficient implicates that $\text{PoPStat}_{\text{COVID-19}}$ or the proposed mechanism is a better descriptor between the disease severity and demographic information. Countries with regressive pyramids typically have a better quality of life and a large elderly population vulnerable to COVID-19. Such countries generally have poor living conditions and inadequate healthcare facilities.

Funding

International Development Research Centre, Canada.

*Corresponding author

Email addresses: e14273@eng.pdn.ac.lk (Yasiru Ranasinghe), harshana.w@eng.pdn.ac.lk (Harshana Weligampola), ilangarathna.g@eng.pdn.ac.lk (Gayanthi Ilangarathna), roshangodd@ee.pdn.ac.lk (Roshan Godaliyadda), vijitha@eng.pdn.ac.lk (Vijitha Herath), mpb.ekanayake@ee.pdn.ac.lk (Parakrama Ekanayake), ekanayakej@eng.pdn.ac.lk (Janaka Ekanayake), samathd@pdn.ac.lk (Samath Dharmaratne), m29782@pgim.cmb.ac.lk (Anuruddhika Rathnayake), mallikap@pdn.ac.lk (Mallika Pinnawala), sakuyatigammana@arts.pdn.ac.lk (Sakunthala Yatigammana), ganga@ips.lk (Ganga Tilakaratna)

1. Introduction

Understanding the variation and attributions of mortality causes is vital in developing healthcare policies, and prevention and intervention programs. For that, several studies have researched and understood numerous variables that could be used to explain the variation in different mortality causes. For example, the role of cigarette smoking and the quality of the healthcare system[1,2] has been analyzed to explain variation in US mortality.[3] Though these factors are considered latent variables, their use is immense as it is difficult to measure risk factors directly because the observations are results of unobserved circumstances. Further, the influence of these latent variables is more prominent as they manifest in several causes of death. However, apart from the behavioural attributes that could be used as explanatory variables, it is also vital to shedding light on other demographic variables.

Research in context

Evidence before this study

There are few studies and reports about the impact of demographic and economic variables: median age, population density, and GDP per capita to several mortality causes. We searched PubMed, Scopus, and Google Scholar on July 1, 2022, with no date restrictions, for publications in English, using the keywords, "population pyramid" and "correlation" to identify correlation analysis performed using population demographics. We identified studies that utilize the structure of the population pyramids for a qualitative analysis of factors affecting different mortality causes.

Added value of this study

To our best knowledge, our work is the first analysis to correlate the population pyramid shape to the severity of mortality causes. The mechanism was first optimized for COVID-19 to maximize correlation with 163 countries. The proposed scalar variable recorded strong correlations with natural mortality causes, especially communicable diseases. Besides, the proposed mechanism displayed the association between socioeconomic standards and disease severity via the population pyramid shape. A strong positive correlation suggests that countries with more elderly populations are vulnerable to the disease due to higher life expectancy and despite the higher quality of life. In contrast, a strong negative correlation suggests that countries with a higher child population are vulnerable to the disease because of poor socioeconomic status and the healthcare system.

Implications of all the available evidence

The results depict the application of the proposed variable and the mechanism to understand the impact of the population pyramid shape on different mortality causes as the pyramid shape attributes to economic, health, and societal standards. According to the literature, the population pyramid has been instrumental in policy-making; policymakers could use the proposed variable or the mechanism in devising policies. Further, the proposed mechanism allows for identifying countries (or regions within a country as the mechanism can be easily adapted at the region or state level given data availability) with similar socioeconomic standards and benchmarks against their disease severity or vulnerability. Further, the proposed mechanism could be generalized, though it was initially proposed for COVID-19 in our previous work.

The impact of the variations in demographics and cultures and how these groups responded to different diseases have been studied concerning variables such as median age, income levels, and population density and how these statistics are worthwhile for policy making.[4,5,6] Moreover, demographic parameters have been used to understand the progression of the disease across communities, and the attempt to compute correlation with demographic parameters has been a standard practice for both communicable[7,8,9,10] and non-communicable[11,12,13] diseases. Recently, these demographic variables were also used with COVID-19 data. For example, to understand the impact of the spatial distribution on COVID-19 progression in India, Bhadra et al.[14] correlated death data to population density information. Then, Wang et al.[15] studied the relationship between COVID-19 mortality cases and the age-based variables of countries. These scalar variables, especially median age, are derived from the country's population distribution. Further, the median is an average alike variable and therefore does not reveal any information about the spread or shape of the distribution, which is critical in identifying the population structure. Hence, properties characteristic of different population pyramid shapes: progressive and regressive, mainly, will not be encapsulated in the median age as it is indifferent to the demographic spread. Regardless, the vast usage of median age[16,17,18] proves the applicability of the information from the population distribution even in its most basic form is warranted.

Accordingly, to promote the application of a more informative representation of the demographic distributions, we proposed the $\text{PoPStat}_{\text{COVID-19}}$ metric as an explanatory variable for COVID-19 in our previous work. However, considering the mechanism presented to generate $\text{PoPStat}_{\text{COVID-19}}$, it is safe to claim that the variable is applicable or can be extended for other mortality causes of similar nature. Hence, this work discusses the application of the $\text{PoPStat}_{\text{COVID-19}}$ as an explanatory variable for different mortality causes, which allows the correlation of variances among the demographic distribution to the national fatality rate of other diseases. The variable was applied to recorded deaths per million available in all-cause mortality data in 2018 to remove the effect of COVID-19 following Gelfand et al.[19] to study the relationship to different mortality causes. The proposed work quantifies the age distribution (the shape of the population pyramid) to correlate with death data of mortality causes statistically.

Since the population pyramid is connected with different socioeconomic factors of the country, the proposed mechanism facilitates a comprehensive understanding between those factors and their impact on different mortality causes. Different correlation values with the proposed statistical measure for different causes provide various insights regarding mortality causes and the influence of socioeconomic standards and quality of living conditions. For example, a negative correlation suggests the disease severity is mainly due to poor living and socioeconomic standards. In contrast, a positive correlation predicts the fatality rate due to the high proportion of these countries' ageing population due to the high quality of life conditions and socioeconomic standards these countries have, which has resulted in a more significant proportion of vulnerable groups to particular types of diseases. As such, $\text{PoPStat}_{\text{COVID-19}}$ could be used in policy making[20,21] as it provides a succinct realization of the interplay between population demographics and disease severity. Also, as the mechanism was derived from statistical techniques, it could be used as a guideline to develop both global and national policy by considering relevant and appropriate disease and demographic data, then using the proposed mechanism to associate the influence of specific demographics on vulnerabilities for those diseases under consideration (mainly if a high level of correlation can be observed). Further, to signify the performance of the variable, we compared the correlation results of the proposed method to that of other commonly used variables: median age, population density, and GDP per capita.

2. Methods

2.1. Mortality data

For our analyses, we retrieved data available on "World Bank Open Data" for all-cause mortality rates from 2018, the most recent year before the COVID-19 outbreak. The data includes 181 countries and 32 causes of death for which mortality data were available, and records were downloaded on April 09, 2022, and have been made available online. Besides, we performed the specific-cause mortality analysis with the proposed $\text{PoPStat}_{\text{COVID-19}}$ as a controlled study to understand the suitability of the proposed $\text{PoPStat}_{\text{COVID-19}}$ to COVID-19 as well and, more importantly, the connection of the demographic profile to the progression of specific causes of death. This is a natural extension of the proposed work for COVID-19 to show what other causes of death have a high degree of correlation to the population pyramid or underlying factors influencing it.

2.2. *Applicability of the population pyramid*

The population pyramid breaks down the population according to gender: male and female, and age group. The pyramid shape is a result of the historical events[22] about policies, natural phenomena, and political events as these factors affect the birth and death rate of countries. For example, the Middle East population doubled from 25 million to 54 million between 2005 and 2015 as the number of migrant individuals and families increased due to economic opportunities.

Consequently, the population pyramid should display a correlation to socioeconomic factors: GDP per capita and quality of life index[23] apart from other derivative variables such as the median age of the demographic distribution. Further, the association of the population pyramid to the economy allows allocating funds[24,25] optimally to critical sectors identified as vulnerable zones based on their demographics now linked more effectively to different causes of death. This signifies the importance of the $\text{PoPStat}_{\text{COVID-19}}$ as it can correlate to various socioeconomic and healthcare attributes through the structure of the population pyramid because these socioeconomic and healthcare standards directly impact the death rate. While medical research is about the prevention and eradication of diseases, analysis with an essential tool such as a population pyramid is effective in understanding the factors that affect different rates for mortality causes other than medical.

2.3. *$\text{PoPStat}_{\text{COVID-19}}$ metric*

$\text{PoPStat}_{\text{COVID-19}}$ embeds the shape of a country's population pyramid for a reference distribution. This transformation from a multi-dimensional to a scalar representation encapsulates the demographic structure as opposed to median age to correlate with the fatality rate of different mortality causes. For the construction of $\text{PoPStat}_{\text{COVID-19}}$, we used the COVID-19 cases per million and were optimized to find the reference pyramid shape to generate the maximum correlation. Unlike other derivative variables of the population pyramid, the underlying mechanism of $\text{PoPStat}_{\text{COVID-19}}$ provides the option to switch between different demographics to ensure maximum correlation is attained so that the mortality rates can be explained reliably. Before calculating the $\text{PoPStat}_{\text{COVID-19}}$ value of a particular demographic, the population pyramid is converted to a probability distribution by combining the male and female histograms and then normalizing it by the total population to adhere to the definitions of a probability density function. Furthermore, the $\text{PoPStat}_{\text{COVID-19}}$ distance between the two distributions is computed using Kullback-Leibler (KL) divergence because the degree of inference between the two distributions is asymmetrical and nonreciprocating unless the two distributions are identical. i.e. KL-divergence measures the deformation between the two population pyramid shapes with one of the shapes as the origin.

In the construction of $\text{PoPStat}_{\text{COVID-19}}$, the optimization mechanism outputted Niger demography as the reference, which yields the maximum correlation and with that, the $\text{PoPStat}_{\text{COVID-19}}$ can be written as,

$$D_{KL}(P \,\|\, Q) = \sum_{i \in I} P_{(i)} \log\left(\frac{P_{(i)}}{Q_{(i)}}\right) \tag{1}$$

where, i is the class index of each age interval of the population distributions P and Q and $P_{(i)}$ and $Q_{(i)}$ are the individual probabilities of the corresponding i^{th} age interval. $D_{KL}(P \,\|\, Q)$ is called the relative entropy between the two random vectors concerning Q, and here P represents the population distribution of any country in the sample while Q represents that of Niger.

2.4. *Role of the funding body*

The funding body of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

3. Results

In Fig. 1, the transformation of the population pyramid shape from progressive to regressive as PoPStat$_{\text{COVID-19}}$ is increased is illustrated. Hence, the proposed mechanism quantifies the shape of the population pyramid and characteristics such as economic standards, living conditions, and healthcare quality affiliated with the population pyramid. It, as shown in the Fig. 2, traces an optical path on how the demographic changes of countries result in a gradual increase (positive correlation) or gradual decrease (negative correlation) in susceptibility or vulnerability to certain diseases. This enables policymakers to infer the socio-economical or political factors that must have influenced those particular demographic changes.

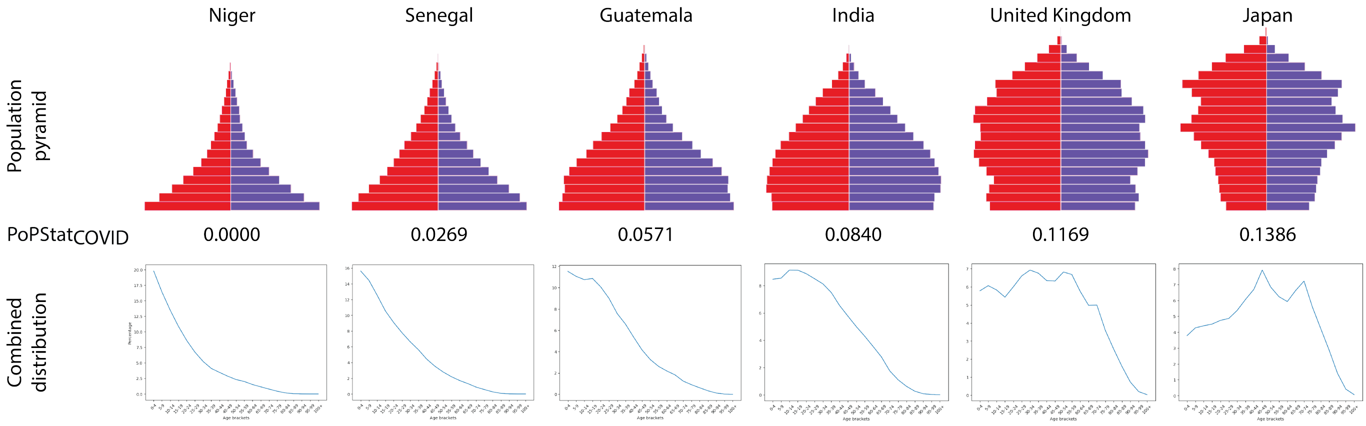

Fig. 1. Transformation of the population pyramid shape with PoPStat$_{\text{COVID-19}}$

Table 1. Correlation coefficients between mortality data and the proposed PoPStat$_{\text{COVID-19}}$ for communicable diseases

Cause of death	PoPStat$_{\text{COVID-19}}$	Median age	GDP per capita	Population density
Hepatitis	-0.71(p<0.0001 [-0.78,-0.63])	-0.62(p<0.0001 [-0.71,-0.52])	-0.07(p=0.3811 [-0.22,0.08])	-0.50(p<0.0001 [-0.61,-0.38])
HIV/AIDS	-0.57(p<0.0001 [-0.66,-0.46])	-0.39(p<0.0001 [-0.51,-0.26])	-0.08(p=0.3363 [-0.22,0.08])	-0.27(p=0.0004 [-0.41,-0.12])
Malaria	-0.73(p<0.0001 [-0.82,-0.6])	-0.54(p<0.0001 [-0.68,-0.37])	-0.08(p=0.5028 [-0.29,0.15])	-0.33(p=0.0036 [-0.51,-0.11])
Meningitis	-0.89(p<0.0001 [-0.91,-0.85])	-0.67(p<0.0001 [-0.75,-0.58])	-0.08(p=0.288 [-0.23,0.07])	-0.47(p<0.0001 [-0.58,-0.34])
Tuberculosis	-0.75(p<0.0001 [-0.81,-0.67])	-0.60(p<0.0001 [-0.69,-0.49])	-0.08(p=0.3372 [-0.22,0.08])	-0.46(p<0.0001 [-0.57,-0.33])
Diarrheal diseases	-0.79(p<0.0001 [-0.84,-0.73])	-0.66(p<0.0001 [-0.73,-0.56])	-0.08(p=0.3201 [-0.23,0.08])	-0.46(p<0.0001 [-0.57,-0.33])
Maternal disorders	-0.88(p<0.0001 [-0.91,-0.84])	-0.71(p<0.0001 [-0.78,-0.63])	-0.09(p=0.2406 [-0.24,0.06])	-0.53(p<0.0001 [-0.63,-0.4])

Neonatal disorders	-0.91(p<0.0001 [-0.94,-0.88])	-0.80(p<0.0001 [-0.85,-0.74])	-0.09(p=0.2385 [-0.24,0.06])	-0.59(p<0.0001 [-0.68,-0.48])
Neoplasms	0.82(p<0.0001 [0.77,0.86])	0.89(p<0.0001 [0.86,0.92])	-0.01(p=0.9055 [-0.16,0.14])	0.56(p<0.0001 [0.45,0.66])
Nutritional deficiencies	-0.75(p<0.0001 [-0.81,-0.67])	-0.61(p<0.0001 [-0.7,-0.5])	-0.08(p=0.3216 [-0.23,0.08])	-0.43(p<0.0001 [-0.55,-0.3])
Protein-energy malnutrition	-0.75(p<0.0001 [-0.81,-0.67])	-0.61(p<0.0001 [-0.7,-0.5])	-0.08(p=0.3253 [-0.23,0.08])	-0.43(p<0.0001 [-0.55,-0.3])
Alzheimer disease and other dementias	0.85(p<0.0001 [0.81,0.89])	0.87(p<0.0001 [0.83,0.91])	0.00(p=0.9765 [-0.15,0.15])	0.61(p<0.0001 [0.5,0.7])
Cardiovascular diseases	0.71(p<0.0001 [0.63,0.78])	0.67(p<0.0001 [0.58,0.75])	-0.06(p=0.4079 [-0.21,0.09])	0.19(p=0.0125 [0.04,0.34])
Chronic kidney disease	0.31(p<0.0001 [0.17,0.44])	0.26(p=0.0008 [0.11,0.39])	0.00(p=0.9534 [-0.15,0.16])	0.05(p=0.4939 [-0.1,0.21])
Chronic respiratory diseases	0.46(p<0.0001 [0.33,0.57])	0.42(p<0.0001 [0.29,0.54])	-0.05(p=0.5604 [-0.2,0.11])	0.21(p=0.0058 [0.06,0.36])
Cirrhosis and other chronic liver diseases	0.19(p=0.0124 [0.04,0.33])	0.24(p=0.0018 [0.09,0.38])	-0.12(p=0.1399 [-0.26,0.04])	-0.11(p=0.1523 [-0.26,0.04])
Diabetes mellitus	0.11(p=0.1686 [-0.05,0.25])	0.08(p=0.2803 [-0.07,0.23])	-0.01(p=0.8675 [-0.16,0.14])	-0.10(p=0.2025 [-0.25,0.05])
Digestive diseases	0.29(p=0.0002 [0.14,0.42])	0.44(p<0.0001 [0.31,0.55])	-0.13(p=0.0951 [-0.28,0.02])	0.04(p=0.5818 [-0.11,0.2])
Lower respiratory infections	-0.45(p<0.0001 [-0.56,-0.32])	-0.42(p<0.0001 [-0.54,-0.29])	0.03(p=0.688 [-0.12,0.18])	-0.30(p=0.0001 [-0.43,-0.15])
Parkinson disease	0.85(p<0.0001 [0.8,0.89])	0.90(p<0.0001 [0.86,0.92])	0.00(p=0.9804 [-0.15,0.15])	0.61(p<0.0001 [0.5,0.69])
Conflict and terrorism	-0.25(p=0.0505 [-0.47,0])	-0.07(p=0.5726 [-0.32,0.18])	-0.08(p=0.5385 [-0.33,0.18])	0.13(p=0.3236 [-0.13,0.37])
Drowning	-0.41(p<0.0001 [-0.53,-0.28])	-0.30(p<0.0001 [-0.43,-0.16])	-0.10(p=0.1794 [-0.25,0.05])	-0.40(p<0.0001 [-0.52,-0.27])
Drug use disorders	0.48(p<0.0001 [0.35,0.58])	0.35(p<0.0001 [0.21,0.48])	-0.07(p=0.3479 [-0.22,0.08])	0.40(p<0.0001 [0.26,0.52])

Environmental heat and cold exposure	0.05(p=0.5119 [-0.1,0.2])	0.22(p=0.0034 [0.08,0.36])	-0.09(p=0.2296 [-0.24,0.06])	0.01(p=0.9452 [-0.15,0.16])
Exposure to forces of nature	-0.09(p=0.4802 [-0.33,0.16])	-0.14(p=0.2612 [-0.38,0.11])	-0.07(p=0.6121 [-0.31,0.19])	-0.15(p=0.2506 [-0.39,0.11])
Fire, heat, and hot substances	-0.31(p<0.0001 [-0.44,-0.17])	-0.18(p=0.0211 [-0.32,-0.03])	-0.17(p=0.0309 [-0.31,-0.02])	-0.28(p=0.0004 [-0.41,-0.13])
Interpersonal violence	-0.28(p=0.0003 [-0.41,-0.13])	-0.16(p=0.036 [-0.31,-0.01])	-0.07(p=0.3922 [-0.22,0.09])	-0.22(p=0.004 [-0.36,-0.07])
Road injuries	-0.44(p<0.0001 [-0.55,-0.31])	-0.43(p<0.0001 [-0.55,-0.3])	-0.17(p=0.0274 [-0.32,-0.02])	-0.42(p<0.0001 [-0.54,-0.28])
Self-harm	0.43(p<0.0001 [0.3,0.55])	0.46(p<0.0001 [0.33,0.57])	-0.06(p=0.4541 [-0.21,0.09])	0.20(p=0.0117 [0.04,0.34])
Terrorism (deaths)	-0.46(p<0.0001 [-0.63,-0.25])	-0.29(p=0.0174 [-0.5,-0.05])	-0.11(p=0.4045 [-0.34,0.14])	-0.12(p=0.3469 [-0.35,0.13])
Poisonings	-0.43(p<0.0001 [-0.54,-0.29])	-0.17(p=0.0295 [-0.31,-0.02])	-0.15(p=0.0583 [-0.29,0.01])	-0.30(p=0.0001 [-0.43,-0.15])
Alcohol use disorders	0.42(p<0.0001 [0.29,0.54])	0.42(p<0.0001 [0.28,0.54])	-0.08(p=0.3185 [-0.23,0.08])	0.17(p=0.0326 [0.01,0.31])

Since the correlation optimization was performed for a communicable disease, we considered the correlation results with other communicable diseases to ascertain if any patterns in the correlations with the proposed distance measure existed and whether they are explainable. The variation between the communicable diseases used in the all-cause mortality database and the proposed $\text{PoPStat}_{\text{COVID-19}}$ are provided in Fig. 2 and the correlation with the other variables for different communicable diseases are given in Table 1. As depicted in Fig. 2, the linear variation of the deaths per million with the proposed $\text{PoPStat}_{\text{COVID-19}}$ suggests a strong correlation between the proposed $\text{PoPStat}_{\text{COVID-19}}$ and the deaths from communicable diseases.

Furthermore, in Fig. 3, the effect of the population pyramid has been illustrated for neonatal disorders (r=-0.91, p<0.0001, CI[-0.94,-0.88])) and depicts which factors could be ascribed to have had a higher impact for the disease. For example, with the quadrant placement, the neonatal disease had a higher impact where socioeconomic conditions are poor. In Fig. 3, the axes' origin were centered around the normalized means of the deaths and $\text{PoPStat}_{\text{COVID-19}}$ value. The negative correlation recorded with $\text{PoPStat}_{\text{COVID-19}}$ for neonatal deaths suggests that countries with progressive population pyramids were more affected by the cause than countries with regressive population pyramids. Since the population pyramid's structure is related to countries' socioeconomic standards, the proposed $\text{PoPStat}_{\text{COVID-19}}$ has provided a tool of inference between socioeconomic standards and disease severity. For further confirmation of the link between our proposed $\text{PoPStat}_{\text{COVID-19}}$ and how it links to socioeconomic conditions, we examined causes

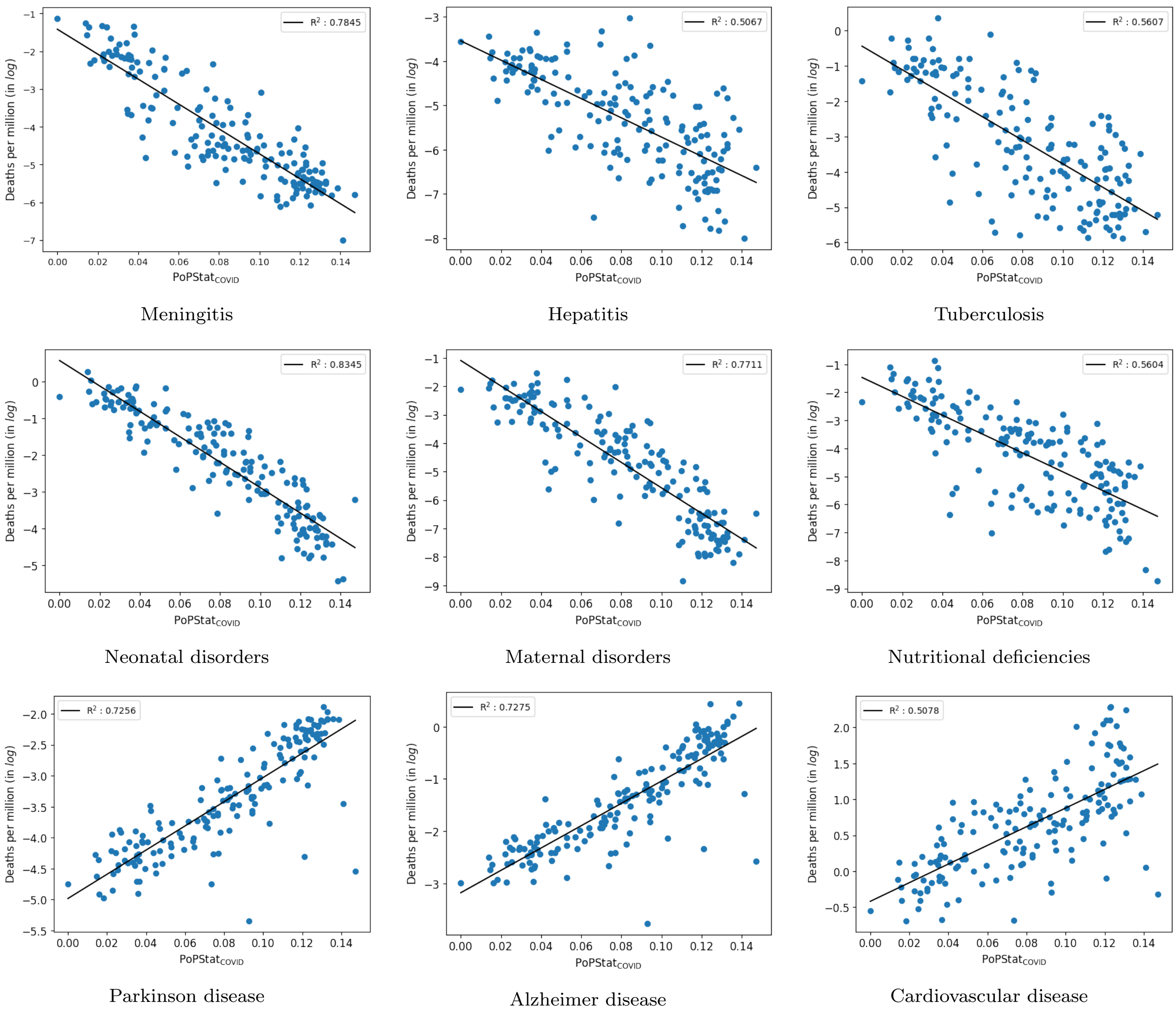

Fig. 2. Relationship between the proposed $\mathrm{PoPStat}_{\mathrm{COVID\text{-}19}}$ and communicable diseases

of death for numerous non-communicable diseases and, interestingly, those with a link to conditions of living and socioeconomic standards such as diarrheal (r=-0.79, p<0.0001, CI[-0.84,-0.73]) and maternal disorders (r=-0.88, p <0.0001, CI[-0.91,-0.84]) showed a high magnitude of correlation.

Accordingly, we considered the viability of the proposed $\mathrm{PoPStat}_{\mathrm{COVID\text{-}19}}$ for diseases under this category, and we observed a strong relationship between the proposed $\mathrm{PoPStat}_{\mathrm{COVID\text{-}19}}$ and the diseases related to the country's socioeconomic level through a negative correlation. As depicted in Fig. 2, a linear variation was recorded with the diseases under this bracket. The numerical results for the diseases related to socioeconomic levels are tabulated in Table 1 and the proposed $\mathrm{PoPStat}_{\mathrm{COVID\text{-}19}}$ has been a better descriptor for these causes of death than the other commonly used statistical variables. The high negative correlation demonstrates the higher impact of these diseases on progressive population pyramid countries where healthcare and quality of life are subpar according to the quadrant placement (mainly on the two labeled quadrants as would be for a negative correlation) in Fig. 3 for neonatal disorders.

In addition, we observed similar linear variations as in Fig. 2 for the proposed $\mathrm{PoPStat}_{\mathrm{COVID\text{-}19}}$ with natural causes of death. It is conclusive that the proposed $\mathrm{PoPStat}_{\mathrm{COVID\text{-}19}}$ correlates well with most natural causes of death considering the causes of deaths tabulated in Table 1. This is because, as many of these natural causes of death are prominent during certain ages, there is a strong connection between age and the natural death causes. Hence, a statistic quantifying the age distribution will perform well in representing the relationship between age and these

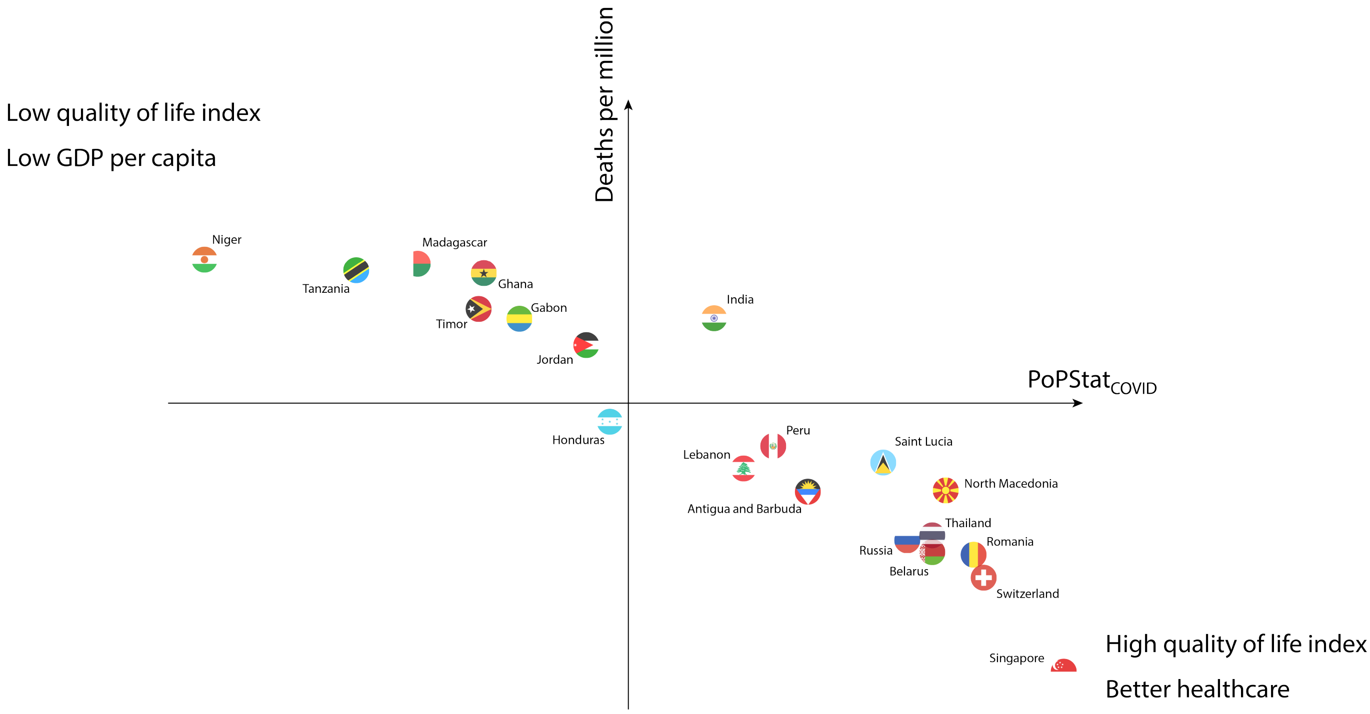

Fig. 3. Quadrant placement for selected countries according to the variation with the PoPStat$_{\text{COVID-19}}$ for Neonatal disorders.

disease cause severity. Then, on the contrary, the proposed PoPStat$_{\text{COVID-19}}$ did not have strong correlations with the causes of death on the opposing extreme of the natural causes of death, which are premeditated actions such as road injuries, violence, and poisoning. According to Table 1, all these causes of death have recorded weak correlations with the proposed PoPStat$_{\text{COVID-19}}$ as expected because the proposed PoPStat$_{\text{COVID-19}}$ was proven to be a good descriptor for natural causes of death. Causes of these types of deaths have a comparatively less correlation and, at times, are more random or independent of age distributions and conditions of living compared to the aforementioned natural causes of death. The illustrations for the variations are provided in the appendix for the causes of death.

4. Discussion

The application of such demographic descriptor as proposed in this work, with a superior ability to understand the impact of various mortality causes on demographics, is worth investing in considering the association between the population pyramid and various socioeconomic factors of a country; accordingly, in this work, we established the application of PoPStat$_{\text{COVID-19}}$ to correlate the population pyramid information to mortality records under different disease categories. It is found that diseases which are of public health interest: tuberculosis (r=-0.75, p<0.0001, CI[-0.81,-0.67]), meningitis (r=-0.89, p<0.0001, CI[-0.91,-0.85]), and malaria (r=-0.73, p<0.0001, CI[-0.82,-0.6]), are correlated with the population pyramid, while cuases like poisoning (r=-0.43, p<0.0001, CI[-0.54,-0.29]), road injuries (r=-0.44, p<0.0001, CI[-0.55,-0.31]), drowning (r=-0.41, p<0.0001, CI[-0.53,-0.28]) had no such correlation. Those diseases which are correlated are precisely the ones which are generally accepted to be linked with the socioeconomic standing of a country or a region.

Therefore, this work provides yet another quantitative justification that the socioeconomic condition is linked to the impact of diseases such as neonatal disorder, maternal disorder, protein malnutrition, and nutritional deficiencies but now with a quantifiable mechanism that is linked through statistical analysis to population demographics. Since PoPStat$_{\text{COVID-19}}$ provides an explanatory variable that links demographic distribution changes to disease vulnerability, it could be employed in developing and implementing public health policies and schemes by interpreting those demographic shape changes under a socio-economical lens, thereby linking it to public policy. For example, the negative correlation $\text{PoPStat}_{\text{COVID-19}}$ with communicable diseases, which indicates the vulnerable groups in terms of quality of life indices or socioeconomic standing, can be used for macro level decision making in-terms of resource focus by WHO or similar national bodies.

The optimization proposed in constructing $\text{PoPStat}_{\text{COVID-19}}$ offers flexibility in selecting a reference demography for different diseases, maximizing the correlation. However, this reference demography is essential as $\text{PoPStat}_{\text{COVID-19}}$ measures the Kullbeck-Leibler divergence (KL-divergence) between probability distributions. In this work, we set the reference as Niger by optimizing for COVID-19 data as a standard. The scalar transformation of the population pyramid shape distills age statistics and other influential socioeconomic status factors to reflect the connection to diseases. Niger has a monotonically decreasing population with age. Hence with Niger as the reference, $\text{PoPStat}_{\text{COVID-19}}$ can measure the divergence of a given demographic structure from being a progressive pyramid shape (how far is the country from the demographics of Niger). Further, since Niger has the lowest median age, $\text{PoPStat}_{\text{COVID-19}}$ can reproduce any result derived from using median age as the explanatory variable but with much more insights. As depicted in Fig. 1, the pyramid shape alters from progressive to regressive with increasing values for $\text{PoPStat}_{\text{COVID-19}}$.

However, the proposed $\text{PoPStat}_{\text{COVID-19}}$ neither portrays the age-specific correlation to the disease's high-risk age groups nor is it an indication of the most vulnerable age of the disease. Instead, it illustrates how specific population structures are affected by the disease and thereby the effect of different socioeconomic factors generally associated with the pyramid shape on the disease severity. For example, in the United States, the age-specific risk is highest in persons older than 65; more than 60% of Tuberculosis cases occur in persons aged 25-64 years[7] which cannot be extracted from this measure. However, what is conveyed under the proposed $\text{PoPStat}_{\text{COVID-19}}$, a negative correlation would imply that the population normalized death count for tuberculosis is higher in countries with progressive population structures than regressive population structures. It latches on to the overall shape of the population pyramid of a given region or country and ranks how vulnerable the entire demographic is to a given disease.

As observed in Table 1 and Table 1, strong correlations were observed for non-communicable diseases as well. However, these are prevalent in countries with lower economic status, as validated by the correlation results derived for specific mortality causes with GDP per capita. Further, according to research[26, 27, 28] and the WHO reports, most of the deaths from these causes are reported from low- or middle-income countries. As aforementioned, most countries with progressive population pyramids belong to low-income countries. Hence, using the population demographics has reproduced the impact of socioeconomic status. As mentioned above, when age becomes a non-factor as opposed to communicable diseases or natural causes of death where age is a factor, the high negative correlation kicks in with the socioeconomic level symbolization.

The relative nature inherited in the proposed $\text{PoPStat}_{\text{COVID-19}}$ allows it to be suitably transformed from disease to disease. For example, a different reference population pyramid could be selected for causes that had a weak correlation with $\text{PoPStat}_{\text{COVID-19}}$to improve the correlation with the population pyramid. Because a weak correlation for the current statistic only suggests that the relationship between the disease severity and divergence from the currently selected reference pyramid shape is not strong. So to examine if the disease is uncorrelated from the population pyramid, this specific reference computation is a must. If, for all possible references, the statistic and cause of death remain uncorrelated, only then is that cause of death truly uncorrelated from the population demographics.

Furthermore, we used the country-wise death data and population statistics in this work. However, the correlation optimization could be reduced to data corresponding lower administrative levels; hence the proposed method is a tool that could help national policymakers. In addition, the mechanism in its construction was not biased towards any specific age group. Therefore, the proposed $\text{PoPStat}_{\text{COVID-19}}$ does not find the relative impact among various age groups on a given mortality cause. Accordingly, further optimization could be introduced to the population pyramid to find an optimal weight vector that generates the relative impact from each age group to a mortality cause during the reference demography optimization.

4.1. Contributors

YR conceptualized and designed the mechanism with support from HW, GI, RG, VH, and PE. YR, curated the data and carried the analysis with support from JE, SD, and AR. YR, RG, MP, SY, and GT visualised the results for the manuscript. YR compiled the manuscript with help and revisions from RG, PE, VH, MP, SY, and GT. All authors commented on the manuscript and approved the final version. All authors had full access to all the data in the study and accept responsibility to submit for publication.

5. Declaration of interests

We declare no competing interests.

References

[1] Richard Peto, Jillian Boreham, Alan D Lopez, Michael Thun, and Clark Heath, "Mortality from tobacco in developed countries: indirect estimation from national vital statistics," *The Lancet*, vol. 339, no. 8804, pp. 1268–1278, 1992.

[2] National Research Council, Committee on Population, et al., "International differences in mortality at older ages: Dimensions and sources," 2011.

[3] Ellen Nolte and Martin McKee, *Does health care save lives? Avoidable mortality revisited*, The Nuffield Trust, 2004.

[4] Gerard Joseph Abou Jaoude, Ines Garcia Baena, Peter Nguhiu, Andrew Siroka, Tom Palmer, Lara Goscé, Kasim Allel, Edina Sinanovic, Jolene Skordis, and Hassan Haghparast-Bidgoli, "National tuberculosis spending efficiency and its associated factors in 121 low-income and middle-income countries, 2010–19: a data envelopment and stochastic frontier analysis," *The Lancet Global Health*, vol. 10, no. 5, pp. e649–e660, 2022.

[5] Neff Walker, Jennifer Bryce, and Robert E Black, "Interpreting health statistics for policymaking: the story behind the headlines," *The Lancet*, vol. 369, no. 9565, pp. 956–963, 2007.

[6] Carla AbouZahr, Sam Adjei, and Churnrurtai Kanchanachitra, "From data to policy: good practices and cautionary tales," *The Lancet*, vol. 369, no. 9566, pp. 1039–1046, 2007.

[7] Matthew J Saunders, Tom Wingfield, Marco A Tovar, Matthew R Baldwin, Sumona Datta, Karine Zevallos, Rosario Montoya, Teresa R Valencia, Jon S Friedland, Larry H Moulton, et al., "A score to predict and stratify risk of tuberculosis in adult contacts of tuberculosis index cases: a prospective derivation and external validation cohort study," *The Lancet Infectious Diseases*, vol. 17, no. 11, pp. 1190–1199, 2017.

[8] Shevanthi Nayagam, Mark Thursz, Elisa Sicuri, Lesong Conteh, Stefan Wiktor, Daniel Low-Beer, and Timothy B Hallett, "Requirements for global elimination of hepatitis b: a modelling study," *The Lancet Infectious Diseases*, vol. 16, no. 12, pp. 1399–1408, 2016.

[9] Neeraj Dhingra, Prabhat Jha, Vinod P Sharma, Alan A Cohen, Raju M Jotkar, Peter S Rodriguez, Diego G Bassani, Wilson Suraweera, Ramanan Laxminarayan, Richard Peto, et al., "Adult and child malaria mortality in india: a nationally representative mortality survey," *The Lancet*, vol. 376, no. 9754, pp. 1768–1774, 2010.

[10] JoelI Ward, MiltonK W Lum, HaroldS Margolis, DavidW Fraser, ThomasR Bender, and Porter Anderson, "Haemophilus influenzae disease in alaskan eskimos: characteristics of a population with an unusual incidence of invasive disease," *The Lancet*, vol. 317, no. 8233, pp. 1281–1285, 1981.

[11] Michael Drozd, Mar Pujades-Rodriguez, Patrick J Lillie, Sam Straw, Ann W Morgan, Mark T Kearney, Klaus K Witte, and Richard M Cubbon, "Non-communicable disease, sociodemographic factors, and risk of death from infection: a uk biobank observational cohort study," *The Lancet Infectious Diseases*, 2021.

[12] Rachel Gitau, Alan Cameron, Nicholas M Fisk, and Vivette Glover, "Fetal exposure to maternal cortisol," *The Lancet*, vol. 352, no. 9129, pp. 707–708, 1998.

[13] Houra Haghpanahan, Jim Lewsey, Daniel F Mackay, Emma McIntosh, Jill Pell, Andy Jones, Niamh Fitzgerald, and Mark Robinson, "An evaluation of the effects of lowering blood alcohol concentration limits for drivers on the rates of road traffic accidents and alcohol consumption: a natural experiment," *The Lancet*, vol. 393, no. 10169, pp. 321–329, 2019.

[14] Arunava Bhadra, Arindam Mukherjee, and Kabita Sarkar, "Impact of population density on covid-19 infected and mortality rate in india," *Modeling Earth Systems and Environment*, vol. 7, no. 1, pp. 623–629, 2021.

[15] Xue-Qiang Wang, Ge Song, Zheng Yang, Ren-Jie Chen, Yi-Li Zheng, Hao-Yu Hu, Xuan Su, and Pei-Jie Chen, "Association between ageing population, median age, life expectancy and mortality in coronavirus disease (covid-19)," *Aging (Albany NY)*, vol. 12, no. 24, pp. 24570, 2020.

[16] Dina N Greene, Michael L Jackson, David R Hillyard, Julio C Delgado, and Robert L Schmidt, "Decreasing median age of covid-19 cases in the united states—changing epidemiology or changing surveillance?," *PLoS One*, vol. 15, no. 10, pp. e0240783, 2020.

[17] Tegan K Boehmer, Jourdan DeVies, Elise Caruso, Katharina L van Santen, Shichao Tang, Carla L Black, Kathleen P Hartnett, Aaron Kite-Powell, Stephanie Dietz, Matthew Lozier, et al., "Changing age distribution of the covid-19 pandemic—united states, may–august 2020," *Morbidity and Mortality Weekly Report*, vol. 69, no. 39, pp. 1404, 2020.

[18] Kushan Tharuka Lulbadda, Dhanushka Kobbekaduwa, and Malika Lakmali Guruge, "The impact of temperature, population size and median age on covid-19 (sars-cov-2) outbreak," *Clinical Epidemiology and Global Health*, vol. 9, pp. 231–236, 2021.

[19] Michele J Gelfand, Joshua Conrad Jackson, Xinyue Pan, Dana Nau, Dylan Pieper, Emmy Denison, Munqith Dagher, Paul AM Van Lange, Chi-Yue Chiu, and Mo Wang, "The relationship between cultural tightness–looseness and covid-19 cases and deaths: a global analysis," *The Lancet Planetary Health*, vol. 5, no. 3, pp. e135–e144, 2021.

[20] Mariachiara Di Cesare, Young-Ho Khang, Perviz Asaria, Tony Blakely, Melanie J Cowan, Farshad Farzadfar, Ramiro Guerrero, Nayu Ikeda, Catherine Kyobutungi, Kelias P Msyamboza, et al., "Inequalities in non-communicable diseases and effective responses," *The Lancet*, vol. 381, no. 9866, pp. 585–597, 2013.

[21] J Ties Boerma and Sally K Stansfield, "Health statistics now: are we making the right investments?," *The Lancet*, vol. 369, no. 9563, pp. 779–786, 2007.

[22] Yi Zeng and Therese Hesketh, "The effects of china's universal two-child policy," *The Lancet*, vol. 388, no. 10054, pp. 1930–1938, 2016.

[23] Ibrahim Abubakar, Sarah L Dalglish, Blake Angell, Olutobi Sanuade, Seye Abimbola, Aishatu Lawal Adamu, Ifedayo MO Adetifa, Tim Colbourn, Afolabi Olaniyi Ogunlesi, Obinna Onwujekwe, et al., "The lancet nigeria commission: Investing in health and the future of the nation," *The Lancet*, vol. 399, no. 10330, pp. 1155–1200, 2022.

[24] Ansley Johnson Coale and Edgar M Hoover, "Population growth and economic development," in *Population Growth and Economic Development*. Princeton University Press, 2015.

[25] PHYSICAL EDUCATION IN KENYA, "Economy, education and physical education in kenya in the light of statistical indicators," *Journal of the Institute for African Studies №*, vol. 3, no. 56, pp. 18, 2021.

[26] Joseph Akuze, Simon Cousens, Joy E Lawn, Peter Waiswa, Vladimir Sergeevich Gordeev, Fred Arnold, Trevor Croft, Angela Baschieri, and Hannah Blencowe, "Four decades of measuring stillbirths and neonatal deaths in demographic and health surveys: historical review," *Population health metrics*, vol. 19, no. 1, pp. 1–14, 2021.

[27] Daniel GP Leventhal, Inácio Crochemore-Silva, Luis P Vidaletti, Nancy Armenta-Paulino, Aluísio JD Barros, and Cesar G Victora, "Delivery channels and socioeconomic inequalities in coverage of reproductive, maternal, newborn, and child health interventions: analysis of 36 cross-sectional surveys in low-income and middle-income countries," *The Lancet Global Health*, vol. 9, no. 8, pp. e1101–e1109, 2021.

[28] Emily Rose Smith, "Global nutrient availability: a call for accountability and action," *The Lancet Planetary Health*, vol. 2, no. 9, pp. e380–e381, 2018.