STEM Fellowship Inter-University Big Data Challenge 2022

Manuscript Title

Optional longer Manuscript Subtitle

Author One[1], Author Two[1], Author Three[1], and Author Four[2]

[1]One University
[2]B University

June 4, 2022

Abstract

Your abstract will get published whether or not your team wins the challenge, so make sure to do a good job! It should reflect the objectives, methodology and findings of the research project. When writing your abstract, choose the right (amount of) words to convey your information. Avoid long sentences and long-winded explanations as they could make the reader lose focus. The order of your writing is also important. Maintain a logical order to your writing so that the reader links each aspect of your work coherently. Introduction/Background: Explains what problem the study examined and why. You may provide some background to the project, and the motivation behind it. Materials and Methods: Describes the data sources used and methodologies adopted for the data analysis. Key Findings/Results: Outlines the discoveries/what was observed from the analysis. When describing your results, strive to focus on the main finding(s), and list no more than two or three points. Conclusion: Provide a general interpretation of the results, specifying what is new/innovative of your project, and give any important recommendation for future research.Once you've written these, delete the keywords, edit for flow, and you have your abstract.

The terms "spending on health" and "total health spending" in this report are used synonymously with "current health expenditure." Capital expenditure is not included.

Keywords pick, 3-5, good, keywords

1 Introduction

(——-Why do we predict suicide rates——) According to WHO SDG Progress Report in 2022, global suicide death rate decreased by 29% from 13.0 deaths per 100,000 population in 2000 to 9.2 deaths in 2019. Although the available data do not show an increase in suicide rates during the first months of the COVID-19 crisis, the pandemic has had a severe impact on the mental health and well-being of people around the world. In 2020, there was an 25% increase in prevalence of anxiety and depression worldwide[**?**]. As an indicator of the SDGs, suicide rate should be reduced by one third from non-communicable diseases through prevention and treatment and promote mental health and well-being by 2030[**?**].

2 Materials & Methods

2.1 Data source and data extraction

We collected open-source data on annual GDP current, GDP per capita, unemployment rate, suicide rate, mental health disorders, health expenditure (%GDP), and health expenditure per capita for 165 countries for different periods from multiple sources listed in Table 1. Since our data are not uniform in length and are not starting from the same year, we decided to extract and merge all collected data into a specific time frame beginning from 2000 to 2019 for further analysis. We then separated our dataset into two final datasets: one grouped by country and the other one grouped by income group. For the present study, we used the dataset grouped by income group for forecasting, and clustering purposes, while the detailed dataset grouped by country for analyzing trends and visualizing various attributes of suicide such as mental health disorders, GDP per capita, unemployment rate, etc.

2.2 Data processing

We used the `MinMaxScaler` method in the `scikit-learn.processing` library to scale some numeric features (i.e., GDP current, GDP per capita, Income, Health expenditure per capita, and Health expenditure (%GDP)) to a range of value from 0 to 1. We also converted other numeric features from percentage or rate into a ratio.

2.3 Time series clustering

We performed Dynamic Time Warping (DTW) Barycenter Averaging (DBA) with three clusters on our standardized dataset to learn a mapping that links our feature of interests to a time series. We used the `TimeSeriesKmeans` method with DTW distance metric in the `tslearn.clustering` library and changed the `np.average` with it. Given the cost matrix, $\Delta(x, y) := [\partial(x_i, y_i)]_{i,j}] \in R^{n*m}$, the inner product $< A, \Delta(x, y) >$ of that matrix with an alignment matrix $A \in \mathcal{A}_{n,m}$ gives the score of A. DTW considers the costs of all possible alignment matrices as follows:

$$DTW(x, y) = \min_{A \in \mathcal{A}_{n,m}} < A, \Delta(x, y) >$$

We then visualized the clusters using the `matplotlib` library. Each line in the plot represents a feature in our dataset.

2.4 Statistical Method

Time series forecasting model is one of the major analysis methods for the future prediction. In this report, in order to evaluate the possibility of achieving the indicator, we used Auto-Regressive Integrated Moving Average (ARIMA) model to predict the suicide rates and disease burden of mental health and substance use disorders based on the mean values from 2000 to 2019. ARIMA is one of the most frequently used stochastic time series models that is based on the idea that the information in the past observations of the time series can alone be used to predict future events. The model is generally denoted as ARIMA(p, d, q) where p, d, q refer to the order of autoregression(AR), the degree of difference, and the order of moving average(MA) respectively. In other words, p refers to the number of lags of the stationary time series and q refers to the number of lags of the forecast errors, and d denotes the minimum number of differencing needed to make the series stationary. An autoregression AR model of order p can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + e_t$$

A moving average MR model of order q uses past forecast errors in a regression

Source	Data
World Health Organization Database	Health expenditure (%GDP, and per capita)
	Suicide rate by country, and regions
Global Burden of Disease 2019 (GBD)	Death and mental substance use disorders
	DALY-Mental disorders
	Death rates by self harms
	Mental disorders prevalence (percent)
The World Bank Database	Unemployment (% total work force)
	GDP current, and GDP per capita
	Income

Table 1: Data sources

model as

$$y_t = c + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q}$$

Where c is a constant, e_t is a white noise, $\phi = (\phi_1, \phi_2, \ldots \phi_p)$ is the vector of model coefficients.

The best model is selected based on the minimum of AIC (Akaike Information Criterion), and BIC (Bayesian Information Criterion) values. P-value $<$ 0.05 is considered significantly. In this report parameters of the ARIMA models for forecasting the suicide rates and mental disorder disease burden from 2020 to 2030 were provided by `auto\_arima()` function from the Python package `pmdarima`. We analyzed the performance of two metrics (RMSE and MAE) to evaluate the forecast accuracy of our model. RMSE refers to Root Mean Square Error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2}$$

MAE refers to Mean Absolute Error.

$$MAE = \frac{1}{n}\sum_{1}^{n}|y_i - \hat{y}_i|$$

Where y_i is the actual value and $\hat{y}_i$ is the forecast value. Both the MAE and RMSE can range from 0 to ∞. The lower values have the better predictions.

3 Results

3.1 Global expenditure on health

Figure 1 shows the percentile distributions of global average spending on health in the past 20 years. We can see that there is a huge gap across countries. High health expenditure with more than 10 percentiles can be observed in North America, especially the United States (15.45%), in South Africa, and Europe; whereas some countries in North Africa and South Asia had the lowest proportion on the health spending, such as South Sudan (1.83%) and Equatorial Guinea (2.03%). The top and bottom five countries are listed in Table 2.

Figure 2 illustrates the relationship between health spending and the growth of GDP differed by income group. In 2019, health spending as a share of GDP ranged from 4.88% in lower-middle-income group to 8.30% in high income group, while health spending as a share of GDP was 5.56% in low-income group and 6.30% in upper-income group.

Table 2: The top five (left) and bottom five (right) countries for health expenditure a share of GDP (percentage).

Country	Health expenditure (%GDP)	Country	Health expenditure (%GDP)
United States	15.45	Bangladesh	2.49
Sierra Leone	11.88	Qatar	2.34
Zimbabwe	11.81	Congo	2.03
France	10.77	Equatorial Guinea	2.03
Germany	10.71	South Sudan	1.83

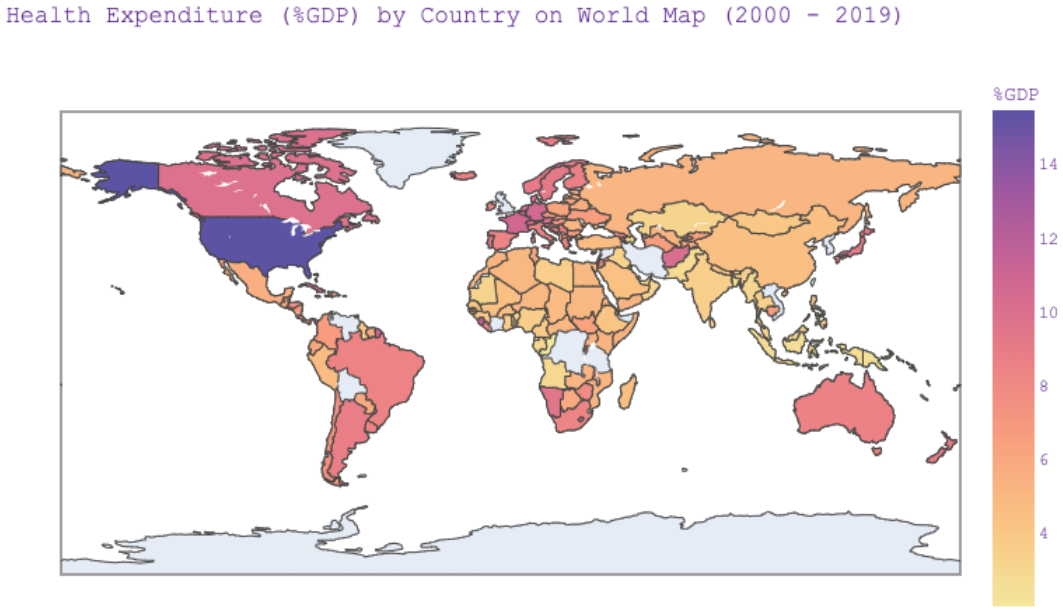

Figure 1: Health expenditure (%GDP) from 2000-2019

3.2 Global Suicide Rate from 2000-2019

Figure 3 highlights the suicide rate across different income groups in the world with two peaks; one small peak with a rising trend from 2004 to 2006 in the upper-middle-income group, and one major peak with a rising trend from 2006 to 2009 in the high and lower-middle-income groups. In contrast, low-income group has shown a steady decreasing trend from 2000 to 2019. For the next decade from 2010 to 2019, the suicide rate has gradually decreased among all income groups.

3.3 Global prevalence of mental disorders from 2000-2019

Figure 5 (left) shows that high-income countries have the highest percentage (11.95%) of mental and substance use disorders, whereas low-income countries have the lowest percentage (0.24%) of mental and substance use. Similarly, figure 5 (right) shows that high-income countries have the highest percentage (7.28%) of sharing mental disorders burden, whereas low-income countries have the lowest percentage (2.27%) of sharing mental disorders burden.

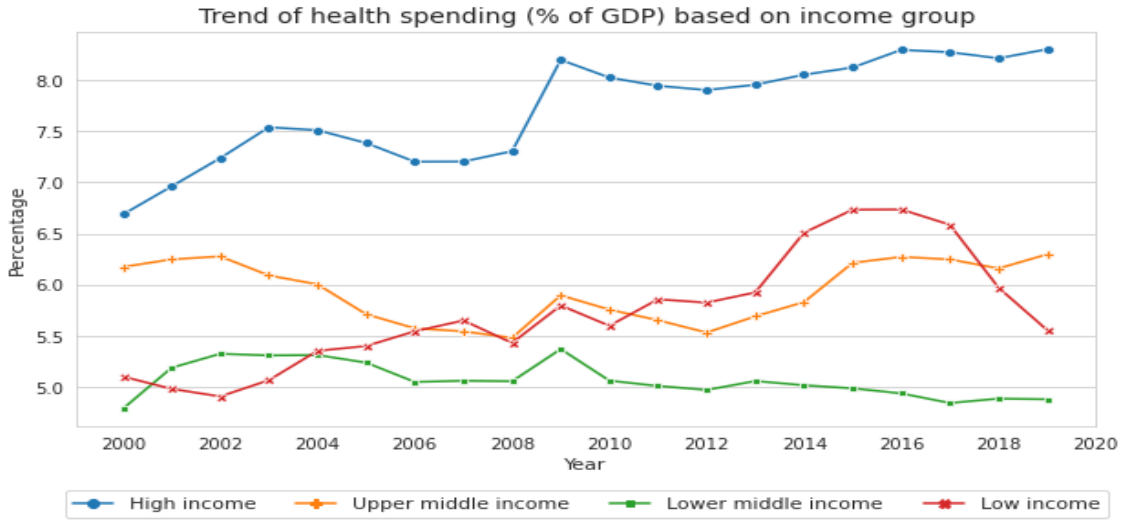

Figure 2: Trend of health expenditure as a share of GDP in different income groups from 2000-2019

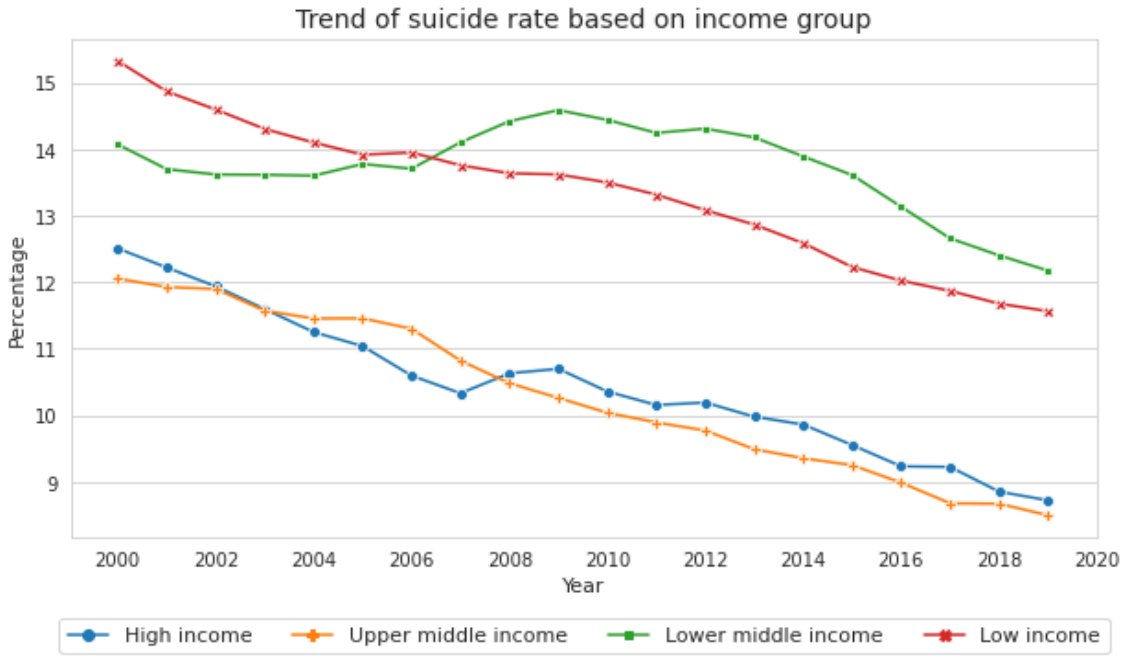

Figure 3: Trend of suicide rate of both sexes with standardized age in different income categories from 2000-2019

Global patterns of mental disorders prevalence are presented in figure 4. Table 3 (left) shows that Australia (Oceania) (19.07%), New Zealand (Oceania) (18.97%), and Portugal (Southern Europe) (18.71%) are top 3 countries that have the highest percentage of mental disorders prevalence. On the other hand, Mali (West Africa) (10.44%), Mauritania (Africa) (10.52%), and Poland (Europe) (10.55%) are the top 3 countries with the lowest percentage of mental disorders prevalence.

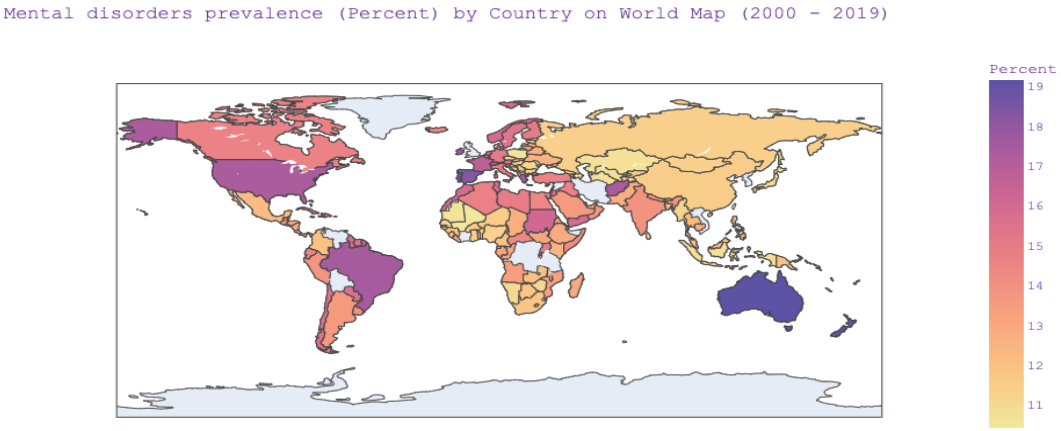

Figure 4: Mental disorder prevalence distribution from 2000-2019

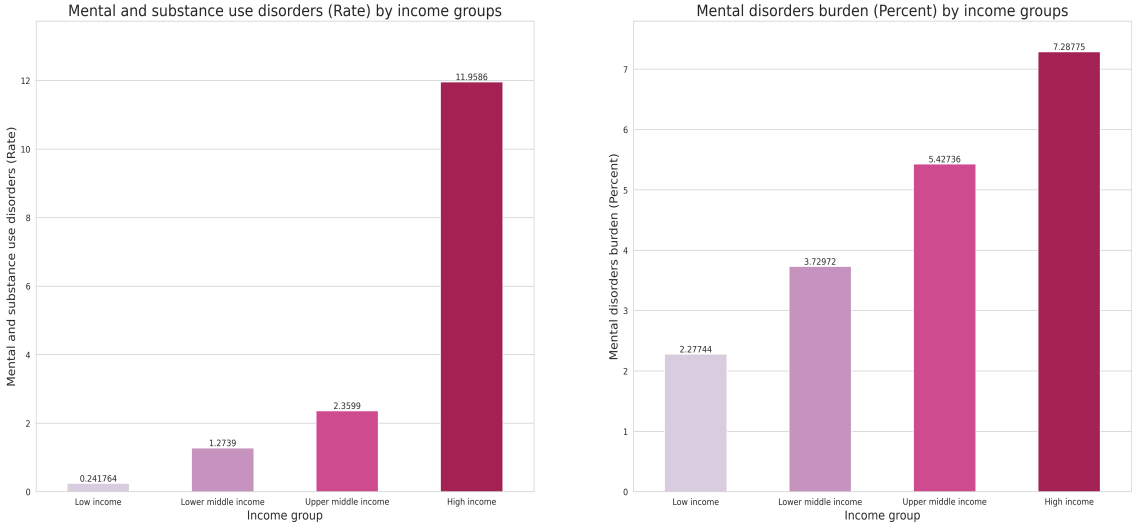

Figure 5: Left: Mental and substance use disorders (rate). Right: Mental disorders burden (percent) by income groups

3.4 DBA K-means Clustering Shows Distinct Clusters between Suicide Rate, Mental Disorder Prevalence, Unemployment Rate, and Health Expenditure across Different Income Groups

Figure 6 illustrates that there are three distinct clusters for high-income and lower-middle-income groups, while there are only two distinct clusters for upper-middle-income and low-income groups. Table 4 shows that all income groups have suicide rate, share of deaths from suicide, and mental and substance use disorder deaths as shared features in a cluster. Besides these shared features, suicide rate is also clustered with unemployment rate and health expenditure (%GDP) in the upper-middle-income group, whereas health expenditure per capita is clustered with suicide rate in lower-middle-income and low-income groups.

Table 3: The top five (left) and bottom five (right) countries for mental disorder prevalence (percentage).

Country	Mental disorder prevalence (%)	Country	Mental disorder prevalence (%)
Australia	19.07	Mali	10.44
New Zealand	18.97	Mauritania	10.52
Portugal	18.71	Poland	10.55
Spain	17.98	Azerbaijan	10.61
Afghanistan	17.49	Kazakhstan	10.74

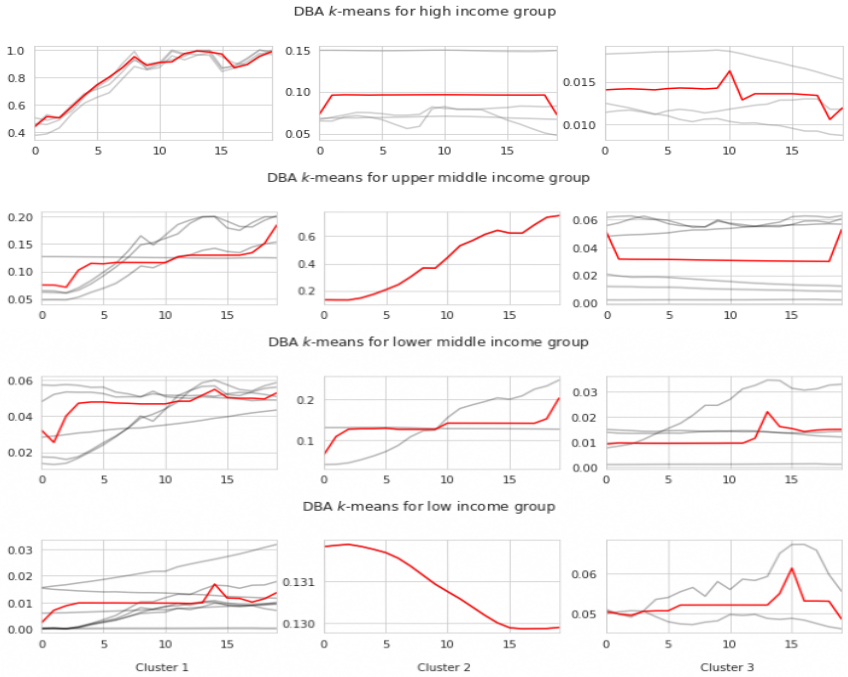

Figure 6: DBA K-means clustering in different income groups

High-income	
Series	**Cluster**
gdp_current	Cluster 1
gdp_per_capita	Cluster 1
health expenditure (per capita)	Cluster 1
Income	Cluster 1
Mental disorders prevalence (Percent)	Cluster 2
health expenditure (% GDP)	Cluster 2
unemployment rate (percent)	Cluster 2
Mental disorders burden (Percent)	Cluster 2
Share of deaths from suicide (Percent)	Cluster 3
Mental and substance use disorder deaths (Rate)	Cluster 3
Suicide rate	Cluster 3
Lower-middle-income	
Series	**Cluster**
gdp_per_capita	Cluster 1
health expenditure (% GDP)	Cluster 1
unemployment rate (percent)	Cluster 1
Income	Cluster 1
Mental disorders burden (Percent)	Cluster 1
Mental disorders prevalence (Percent)	Cluster 2
gdp_current	Cluster 2
Share of deaths from suicide (Percent)	Cluster 3
Mental and substance use disorder deaths (Rate)	Cluster 3
health expenditure (per capita)	Cluster 3
Suicide rate	Cluster 3

Upper-middle-income	
Series	**Cluster**
Mental disorders prevalence (Percent)	Cluster 1
gdp_per_capita	Cluster 1
health expenditure (per capita)	Cluster 1
Income	Cluster 1
gdp_current	Cluster 2
Share of deaths from suicide (Percent)	Cluster 3
Mental and substance use disorder deaths (Rate)	Cluster 3
health expenditure (% GDP)	Cluster 3
unemployment rate (percent)	Cluster 3
Mental disorders burden (Percent)	Cluster 3
Suicide rate	Cluster 3
Low-income	
Series	**Cluster**
Share of deaths from suicide (Percent)	Cluster 1
Mental and substance use disorder deaths (Rate)	Cluster 1
gdp_current	Cluster 1
gdp_per_capita	Cluster 1
health expenditure (per capita)	Cluster 1
Income	Cluster 1
Mental disorders burden (Percent)	Cluster 1
Suicide rate	Cluster 1
Mental disorders prevalence (Percent)	Cluster 2
health expenditure (% GDP)	Cluster 3
unemployment rate (percent)	Cluster 3

Table 4: DBA K-means clustering distribution in different income groups

3.5 Correlation between Mental health and Global Economics

Figure 7 shows correlation coefficients among mental health indicators and global economics by using Pearson's correlation. The correlation between mental disorder prevalence and suicide rate is -0.13. The correlation between global spending on health and Mental and substance use disorders death rate is 0.33. In addition, the correlation between suicide rate and suicide as a share of total death rate is 0.45. The correlation between mental disorder prevalence and mental disorder as a share of total burden of disease is 0.58.

3.6 Forecasts for Suicide Rates and Mental Disorder Burdens up to Year 2030 by Income Group

The forecast trends up to 2030 with its 95% forecasted interval of the suicide rates for four income groups is shown in Figure 8. The suicide rate was expected to experienced a similar trend in all income groups, with a slight rise in the year 2020 and then falling steadily in the following 10 years. For the high-income group (top left), the suicide rate was supposed to decrease from 8.72 per 100,000 in 2019 to 6.16 per 100,000 in 2030. There was a deduction from 8.50 per 100,000 in 2019 to 6.90 per 100,000 in 2030 in the upper-income countries (top right). In the meanwhile, the suicide rate was predicted to reach 10.53, 8.29 per 100,000 in 2030 in lower-middle-income (bottom right) and low-income countries (bottom right) respectively.

Figure 9 highlights time trend for mental disorders on the burden of disease in four income groups from 2000 to 2019 and the predictions up to 2030. For the past two decades, the lowest mental disorders as a share of total disease were observed in 2019, i.e., 6.72%, 5.68%, 4.34%, and 3.19% in high-income group (top left), upper-income group (top right), lower-middle-income (bottom right) and low-income countries (bottom right) respectively. It was supposed to be observed a rising trend up to 2020 (6.82%) followed by a major fall from 2021 (6.74%) onwards up to 2030 (5.75%) in high-income groups (top left), while in the other income groups, it may be seen a decrease trend up to 2020 with an upward trend in the next 10 years.

In the report we used the best fit ARIMA model with the lowest AIC and BIC values. Table 5 shows the accuracy values of the ARIMA model for predictions.

4 Discussion

Suicide continues to be the 10^{th} leading cause of death across eastern Europe, central Europe, high-income Asia Pacific, Australasia, and high-income North America [1]. Recently, several studies have shown that the global change in economic activities has deeply affected people's mental health and may have influenced the suicide rate [2]. Although the decrease in suicide rates has been substantial from 2000 to 2019 (Figure 3), we need to consider the complex interplay of factors such as levels of economic development, unemployment, and economic events and; the distribution of risk factors like exposure to mental and substance use disorders.

4.1 The economics of mental health

Table 4 suggests a relationship between health expenditure (% GDP) and suicide rate in the upper-middle, lower-middle, and low-income groups. Recent studies have suggested that although at low

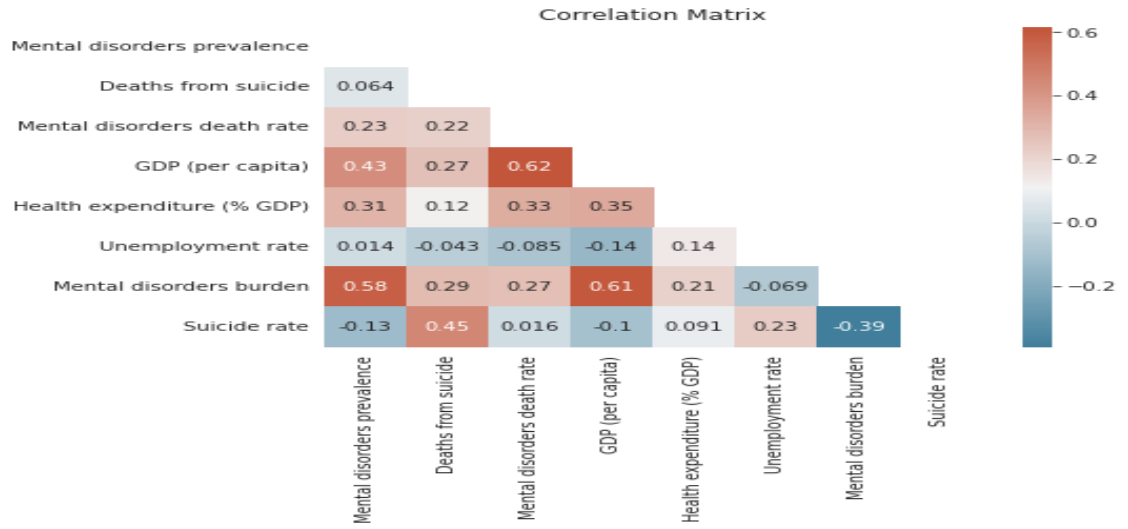

Figure 7: Mental disorder prevalence distribution from 2000-2019

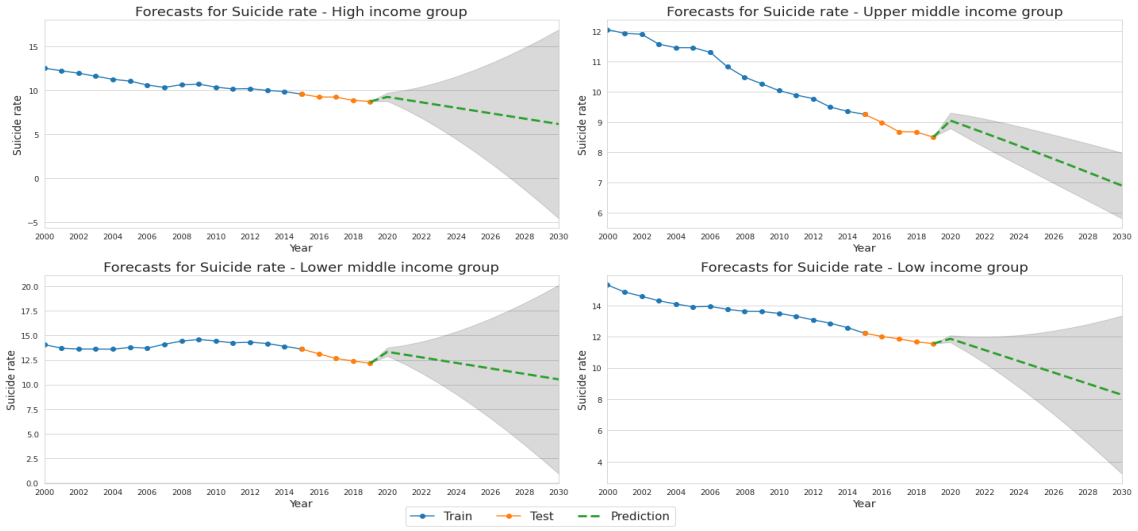

Figure 8: Trend for suicide rate from 2000 to 2019 and forecast up to 2030 by income groups. Solid line denote suicide rates from 2000 to 2019 estimated by Global Burden of Disease 2019. The fitted data is separated with the training data (blue dots) and test data (orange dots). Green dashed lines denote suicide rates by income group from 2019 to 2030 predicted by ARIMA model based on the mean rates in the last 20 years. Grey areas denote 95% confidence intervals of the predicted suicide rates.

GDP levels, increases in GDP are associated with increases in suicide rates, once a given threshold of economic development is reached, further increases in GDP do not correlate with further increases in suicide rate [2]. Our findings are consistent with this suggestion since GDP per capita is only observed in the same cluster as suicide rate (Table 4) in the low-income group.

Moreover, recent research has shown that people who suffer from severe and common mental disorders are more likely to be unemployed than people with no conditions [3]. Table 4 also shows that unemployment rate is associated with mental disorders prevalence in all income groups. However, more fundamental intervention research is required to understand better the effects of mental disorders and unem-

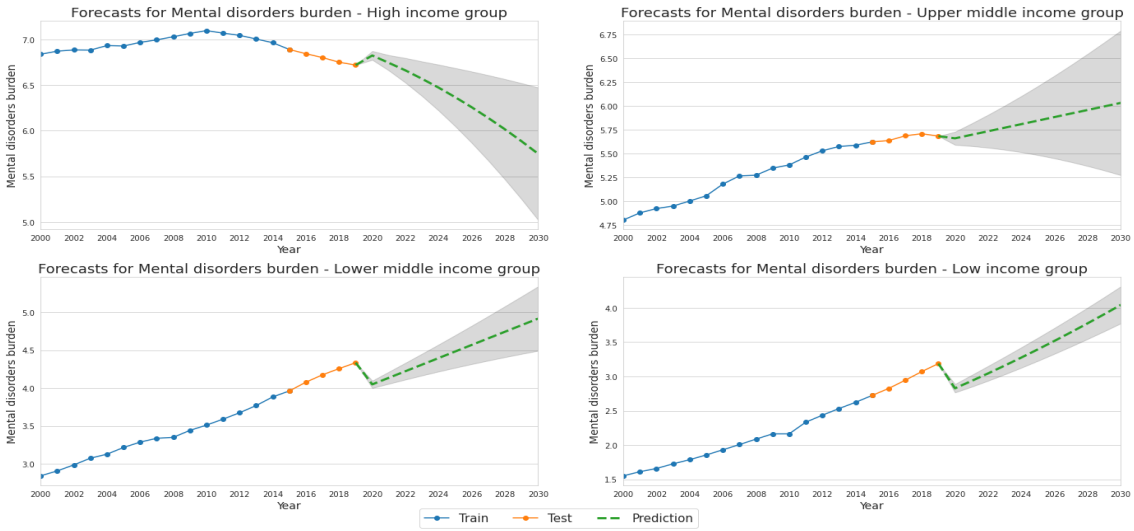

Figure 9: Mental health disorders as a share of total disease burden from 2000 to 2019 and predictions up to 2030 by income groups. Solid line denote mental disorders on the total disease burden from 2000 to 2019 estimated by Global Burden of Disease 2019. The fitted data is separated with the training data (blue dots) and test data (orange dots). Green dashed lines denote mental disorders on the total disease burden by income group from 2019 to 2030 projected by ARIMA model based on the mean rates in the past 20 years. Grey areas denote 95% confidence intervals of the projected mental disorders as a share of total disease burden.

ployment in different income groups.

4.2 Mental burdens

The low income group has a very small number of responding countries in comparison to the other income groups, which could be attributed to reporting difficulties due to the COVID-19 pandemic.

As we can see from Figure 1, most of the countries with high health expenditure as a percentage of GDP were in Europe, North America, Australia, and East Asia, while most countries with low health expenditure as a percentage of GDP were in Africa and South Asia. As is shown in the Figure 2, countries in high income group have a higher health expenditure proportion. It may be paradoxical that low income group spend a higher percentage of GDP on health than low middle income group. The main reason is external aid for health plays in a large role in many low income countries. For example, in 2019, out-of-pocket spending and external aid accounted for roughly three-quarters of health spending in low income countries. In 2019, the share was 1.7 times larger in high income group (8.30%) than lower middle income group (4.88%). Health expenditure as a share of GDP has climbed in all income groups from 2000 to 2019; however, there has been a declining trend in low and lower middle income groups in the recent three years.

The burden of mental health disorders is sometimes underestimated since health impacts are commonly assessed in terms of total numbers of suicide. Measuring health impact simply on the basis of suicide rate ignores the influence mental health illness have on an individual's well-being[4]. To further validate the relationships among the mental health and economics, a correlation matrix was created (Figure 7). As expected, mental dis-

Table 5: RMSE and MAE for Best-fit ARIMA Models for Predictions.

Income Group	Suicide Rate		Mental Disorder Burden	
	RMSE	MAE	RMSE	MAE
High-income	0.2731	0.2326	0.0926	0.0793
Upper-middle-income	0.0965	0.0838	0.0480	0.0373
Lower-middle-income	0.3260	0.3172	0.0326	0.0322
Low-income	0.5055	0.4521	0.0204	0.0172

order as a share of total burden of disease with mental disorder prevalence's correlation coefficient was greater than 0.5, which means that they were strongly positively correlated with each other. Global spending on health and mental and substance use disorders death rate also had a positive correlation. Unemployment rate was weak positively related to suicide rate. However, we did not see that there was a direct relationship between the prevalence of mental disorders and suicide rates.

The forecasts show all income groups may not achieve a one-third reduction

Conclusions

What are the long-term implications of your findings? Wrap up your discussion succinctly while pointing out the significance of your work as well as it what it means for the fields you examined as much as possible. Lastly, suggest ideas for future studies that could build on your work, and justify why they might be useful. Otherwise, you're all done!

Acknowledgements

Anyone to thank/credit for helping your team along the way? This is the place to do it!

References

[1] Mohsen Naghavi. Global, regional, and national burden of suicide mortality 1990 to 2016: systematic analysis for the global burden of disease study 2016. *BMJ Journals*, 364(194):1–11, 2019.

[2] Blasco-Fontecilla H, Perez-Rodriguez MM, and et al Garcia-Nieto R. Worldwide impact of economic cycles on suicide trends over 3 decades: differences according to level of development. a mixed effect model study. *BMJ Journals*, 2(e000785):1–7, 2012.

[3] Evelien P.M Brouwers. Social stigma is an underestimated contributin factor to unemployment in people with mental illness or mental health issues: position paper and future directions. *BMC Pyschology*, 8(36):1–7, 2020.

[4] Martin Prince, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R Phillips, and Atif Rahman. No health without mental health. *The Lancet*, 370(9590):859–877, 2007.