Factors Influencing Worldwide Tuberculosis Incidence

Alyssa Grogorenz-de Oliveira - Student ID: 730079740

Academic Honesty Statement

I have familiarised myself with the academic misconduct and plagiarism guidelines in the Academic Honesty and Plagiarism module and the MTHM601 (Fundamentals of Data Science) ELE site's Assessment Information tab. This report constitutes my own work, and I have explicitly referenced and acknowledged those parts that draw on the literature, online sources, and the support of others. This includes acknowledgment of any use of Artificial Intelligence tools such as ChatGPT.

Introduction

Tuberculosis (TB), caused by the bacterium *Mycobacterium tuberculosis* [3], is present in every part of the globe and is one of the deadliest infectious diseases in the world [2], second only to COVID-19 [4]. Although the disease has accompanied humanity since antiquity [5], it was largely neglected during the later half of the 20th century, and has only relatively recently gained renewed attention [1].

The disease mostly affects the lungs and is an airborne infection only spread by those who exhibit symptoms. Treatments for tuberculosis exist [2], but their availability may be limited by a country's healthcare infrastructure. Additonally, the World Health Organisation (WHO) identifies those that live in close proximity to people with active tuberculosis to be at high risk of infection [6]. Population density may therefore be another factor that affects tuberculosis transmission, with countries with a higher population density potentially seeing a higher incident number.

As such, this project seeks to examine the relationship between a country's new and relapsed (incident) tuberculosis numbers and its gross domestic product (GDP) per capita, its health expenditure (HE) per capita, and its population density.

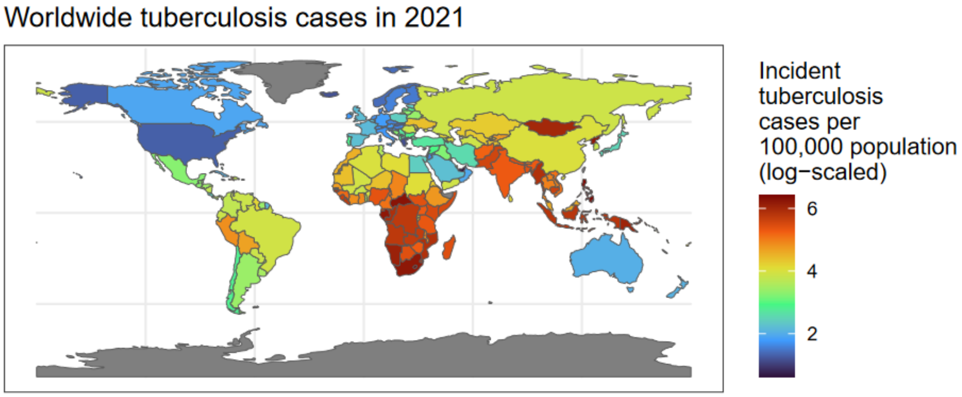
Worldwide tuberculosis cases in 2021
Incident tuberculosis cases per 100,000 population (log−scaled)
6
4
2

Figure 1: Global heatmap of incident tuberculosis cases per 100,000 population in 2021, scaled on a logarithmic scale.

Objectives & Methodology

To address this research question, relevant data was found and cleaned into one tidy data frame. Exploratory data analysis was performed and potential correlations between explanatory variables were examined. All variables were significantly correlated. Specifically, the variables GDP per capita and health expenditure per capita were strongly positively correlated (Pearson's correlation coefficient $r = 0.887$) and the variables GDP per capita and population density were moderately positively correlated ($r = 0.543$).

Imputation of missing values was considered using multiple imputations, however the results of the imputations were inadequate, so missing data was removed instead. This resulted in the data set containing 4,054 observations instead of 4,439, a loss of data of about 8.7 per cent.

During data exploration, it became apparent that the distribution of all four variables (GDP per capita, health expenditure per capita, population density, and incident TB numbers) was on a logarithmic scale, which is why the data was log-transformed to aid analysis. It also became apparent that observations were clustered based on WHO region, with incidence numbers being particularly high in Africa and low in Europe (Fig. 1). It was also reasonable to suspect explicit nesting of the random effects of region and country. This prompted the data to be analysed using a hierarchical linear regression model.

Candidate models were chosen based on their Akaike Information Criterion (AIC), conditional R^2 (cR^2) value, and adjusted interclass correlation coefficient (ICC). The residuals of each candidate model were examined for homoscedasticity and normality, and the final model chosen following k-fold validation and based on the principle of parsimony.

Data

The data used for this report spanned from the year 2000 to 2021 to allow for a big data set. Time series data for GDP per capita (in current USD) for 217 countries and for the population density (people per square km of land area) for 218 countries was acquired from the World Bank Group's DataBank. Data for health expenditure per capita (in current USD) for 192 countries and data for new and relapsed cases of tuberculosis per 100,000 population per year for 194 countries was obtained from the WHO's data repository.

Each of these four data sets was transformed into a tidy format with columns that contained the year, country, and the variable of interest. The incidence data set also included confidence intervals for the number of incident cases, which were removed from the column.

Since the WHO and the World Bank Group use different country name variations, the country names from the population density data set from the World Bank were used. Country codes (unique three letter codes representing one country) were also extracted from the population density data set and added to the incidence and health expenditure data sets to allow the data to be combined.

Following the classification used by the WHO [7], another variable called "Region" was added to the data set, grouping the countries into either Africa, the Americas, South-East Asia, Europe, the Mediterranean, or the Western Pacific.

All four variables were log-transformed before further analysis. Because there were some observations with an incident tuberculosis number of zero, the incidence numbers were increased by one to avoid taking the logarithm of zero. The completed data set contained the variables year, country, and region, as well as log-transformed incidence, GDP, health expenditure, and population density.

Some data points were missing, since the data sets came from two different sources and spanned a relatively long time period of more than twenty years. Specifically, there was no GDP data available for the Cook Islands, Niue, or the Democratic People's Republic of Korea, and the GDP was only partially available for seven other countries. No population density data was available for the Cook Islands or Niue, with three other countries missing one datum for population density each. There was no health expenditure data available for Somalia and the Democratic People's Republic of Korea, very limited data available for Venezuela, and there were ten other countries that had missing values for health expenditure for less than half of the time period considered.

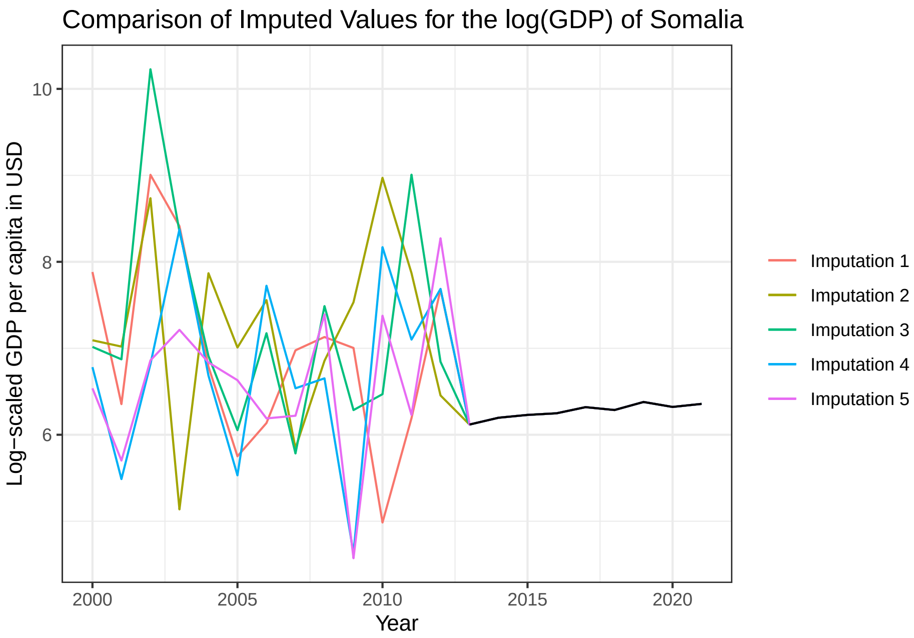

Figure 2: Graph showing annual log(GDP) per capita in USD for Somalia. The black line is not imputed, the other five lines are 5 different imputations of the country's log(GDP).

Multiple imputations of the missing data were attempted, but the resulting imputations did not match the trends seen in the data. As an example, the imputed values for log(GDP) for Somalia between 2000 and 2012 did not follow the expected distribution seen for the years 2013 to 2021 (Fig. 2), although this was not just the case for Somalia or the imputations for log(GDP). This may be because the data was Missing Not At Random (MNAR). As noted previously, data was only missing for specific countries, which means that the missingness itself most likely depended on a different, unmeasured variable (such as political leaning or leadership, perhaps) that made it less likely that countries reported metrics such as GDP, health expenditure, or population density for a given year. The multiple imputations method can only be applied to data that is Missing At Random (MAR), however, and was therefore not a suitable method for dealing with missing data in this data set.

As noted previously, the removal of missing observations resulted in a loss of data of about 8.7 per cent, which was deemed acceptable considering the size of the remaining data set. The final data set included data from 188 countries.

Results

Identifying candidate models

The first step in identifying suitable models was the consideration of only one explanatory variable. Linear mixed effect models, accounting both for regional differences and nested country-level differences, were fit to the log-transformed data set and the ICC, AIC, and cR^2 values recorded (Tab. 1). The adjusted ICC captures how much of the variance in the data is explained by the random effect, the AIC is an estimator of the relative quality of a model which trades off goodness of fit with simplicity (and therefore discourages overfitting), and the cR^2 of a linear mixed model describes how much of the overall variance is described by the model.

A model was deemed suitable when all its fixed effect parameters were significant (at a significance level of $P \leq 0.05$), its AIC value was minimal, and the ICC and cR^2 were maximal.

Table 1: Overview of linear mixed models with one log-transformed predictor variable, including whether model parameters were significant as well as the adjusted ICC, AIC, and conditional R-squared values for each model. The predictors were GDP per capita in USD (GDP), health expenditure per capita in USD (HE), and population density (Density). The models were either a random intercept (Type intercept) or additionally a random slope (Type slope) model. Random effects were either only the WHO region (Region), or additionally a nested random effect of country (Region:Country).

Predictor (Fixed Effect)	Type	Random Effect	Significant?	ICC	AIC	cR^2
GDP	intercept	Region	yes	0.303	10,875	0.570
		Region:Country	yes	0.955	2,055	0.959
	slope	Region	yes	0.461	10,465	0.643
		Region:Country	yes	0.975	1,220	0.977
HE	intercept	Region	yes	0.283	10,923	0.569
		Region:Country	yes	0.955	2,033	0.959
	slope	Region	yes	0.485	10,508	0.654
		Region:Country	yes	0.977	1,082	0.979
Density	intercept	Region	yes	0.504	12,273	0.529
		Region:Country	yes	0.973	2,295	0.979
	slope	Region	yes	0.529	12,093	0.550
		Region:Country	no	0.999	1,180	0.999

The inclusion of the nested random effect of country within a region improved model performance, as evidenced by the decrease in AIC and increase in ICC and cR^2 for models that included nested effects compared to their counterparts that only included a regional effect. Analysis of all subsequent models was therefore only performed on explicit nested random effect models.

Additionally, the inclusion of a random slope and random intercept improved model performance slightly when compared to models that only included a random intercept.

Candidate models after this first step were m1, which was a random slopes model with GDP as its predictor variable (ICC = 0.975, AIC = 1,220, cR^2 = 0.977), and m2, which was a random slopes model with health expenditure as its predictor (ICC = 0.977, AIC = 1,082, cR^2 = 0.979). Health expenditure seemed to be a slightly better predictor of incident tuberculosis numbers.

The same process was repeated for models including two and three variables. Out of all the models tested, only those with significant fixed effects (at a significance level of $P \leq 0.05$) were presented in Tab. 2.

In addition to m1 and m2 from earlier considerations, models m6, m9, m10, and m11 were final candidate models due to their smaller AIC and larger ICC and cR^2 compared to the other models.

Table 2: Overview of linear mixed models with multiple log-transformed predictor variable, including the adjusted ICC, AIC, and conditional R-squared values for each model. The predictors were GDP per capita in USD (GDP), health expenditure per capita in USD (HE), and population density (Density). The models either included a random intercept only, or also included a random slope on one of the fixed effects. The random effects were the WHO region (Region) and a nested random effect of country.

Model	Fixed Effects	Random Effect type	ICC	AIC	cR^2
m3	GDP, HE	intercept only	0.955	2,023	0.959
m4	GDP, HE	slope (HE)	0.977	1,084	0.978
m5	GDP, Density	intercept only	0.960	1,961	0.967
m6	HE, Density, HE:Density	slope (HE)	0.981	909	0.985
m7	HE, Density	intercept only	0.959	1,950	0.966
m8	GDP, HE, Density	intercept only	0.980	1,012	0.984
m9	GDP, HE, Density, GDP:HE	slope (GDP)	0.981	948	0.985
m10	GDP, HE, Density, GDP:HE	slope (HE)	0.981	908	0.986
m11	GDP, HE, Density, GDP:HE, HE:Density	slope (GDP)	0.981	944	0.985

Assessing Model Fit

All candidate models were assessed based on the normality of their residuals and the homoscedasticity of their lowest-level residuals.

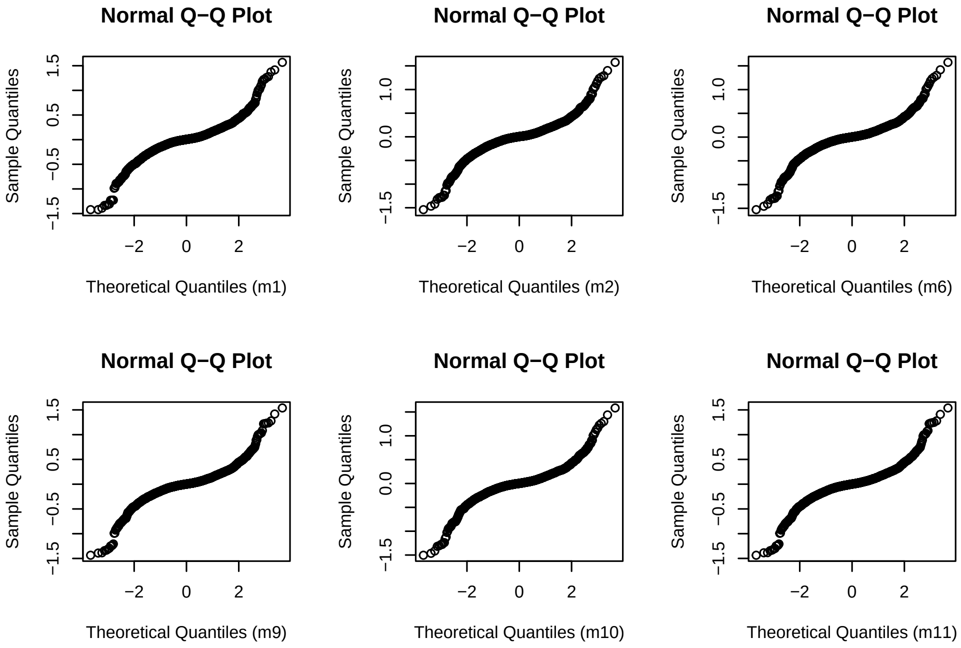

Figure 3: Quantile-Quantile Plots for all six candidate models.

The quantile-quantile plots (Fig. 3) did not differ strongly between the models. There were hints that the residuals might not be normally distributed. This may be due to the log-transformation of the data.

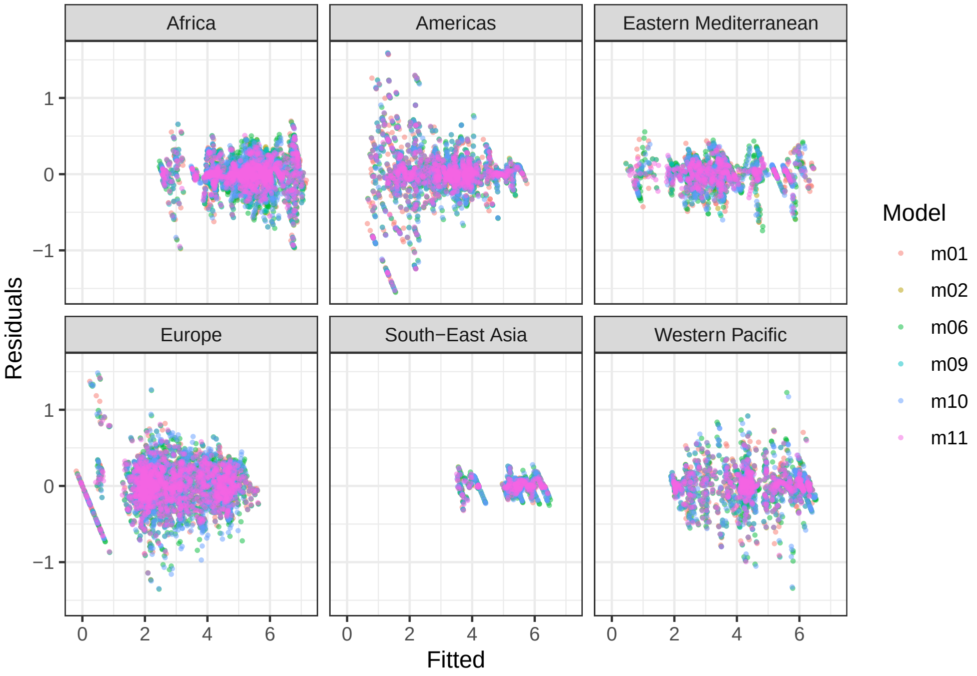

Figure 4: Lowest level residuals for all six candidate models, facetted by region.

Overall, the lowest level residuals of all six models were homoscedastic (Fig. 4). There are some hints of heteroscedasticity for residuals in Europe and the Americas, which may be due to the residuals stemming from data points describing the same country.

K-fold Cross Validation

In order to assess model fit, and to prevent overfitting of the model, k-fold cross validation was applied to each of the candidate models. In this case, $k = 10$ partitions were created. In turn, each of the partitions was removed from the data, the model fit on the remaining data, and then the remaining partition was used to validate the model.

Training and validation errors were recorded for each iteration for each model. The training error reflects how well the training data set fits the model, with an error of zero indicating a perfect fit. The validation error indicates how well the model fits new data that was not included in the building of the model. The validation error can therefore be expected to be bigger than the training error, but both numbers should be similar; a big difference between training and validation error may indicate that the model was overfitted because the model does not predict the validation data set well enough. Additionally, the best model will have a small validation error, indicating a good overall fit.

The mean training and validation errors were similar for all considered models, indicating that none of them were overfitted (Tab. 3). Model fit for models m6, m10, and m11 seemed to be slightly better, considering their comparatively low validation error.

Table 3: Mean training and validation errors for all six models.

Model	Training Error	Validation Error
m1	0.056	0.061
m2	0.055	0.060
m6	0.053	0.058
m9	0.053	0.059
m10	0.053	0.058
m11	0.053	0.058

The Final Model

All considered models had a cR² value of at least 0.95, meaning that they all explained at least 95 per cent of the variation in annual incident tuberculosis numbers. Models m6, m9, m10, and m11 had the lowest AIC, highest cR², and similarly high ICC, which made all of them equally suitable to predict incident TB numbers. It should be noted however, that m6 only included two predictor variables (health expenditure and population density), whereas m9, m10, and m11 included three predictor variables. The addition of GDP as a predictor therefore did not drastically increase the precision of the linear mixed effects model, most likely due to the strong correlation between GDP and health expenditure as well as moderate correlation between GDP and population density noted earlier.

Following the principle of parsimony, the simpler model m6 was chosen as the final model.

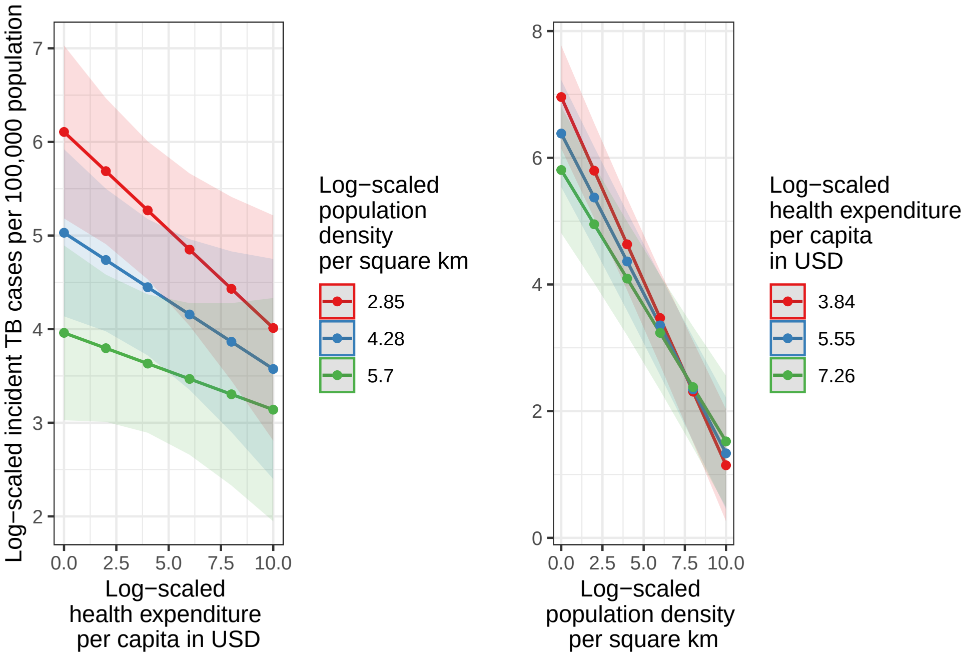

Figure 5: Predicted values of incident tuberculosis based on health expenditure and population density.

With an increase in public health expenditure, the incident tuberculosis numbers are expected to decrease. The higher the population density, the less stark this decrease is. Alternatively, the higher the population density, the more incident tuberculosis numbers are expected to decrease. From this point of view, the effect of health expenditure is more pronounced at lower population densities (Fig. 5).

The overall relationship between incident tuberculosis numbers (Inc) per 100,000 population and its predictors health expenditure (HE) per capita in USD and population density (PD) in people per square km of land area can be described as follows:

$$\log(Inc) = 8.25 - 0.33 \cdot \log(HE) - 0.75 \cdot \log(PD) + 0.04 \cdot \log(HE + PD)$$

An increase in health expenditure per capita in USD is correlated with a significant reduction in incident TB cases, a fact that makes intuitive sense. However, an increase in population density is even more strongly correlated with a decrease in incident tuberculosis numbers. Because the disease is an airborne infection, it might be expected that a higher population density may be correlated with higher tuberculosis incidence. The reason this does not seem to be the case is because average population density may not be a good indicator for how closely people of a given country actually live together. Instead, it is more likely a indicator of development, with more rural (and therefore less densely populated) countries seeing higher incident TB numbers.

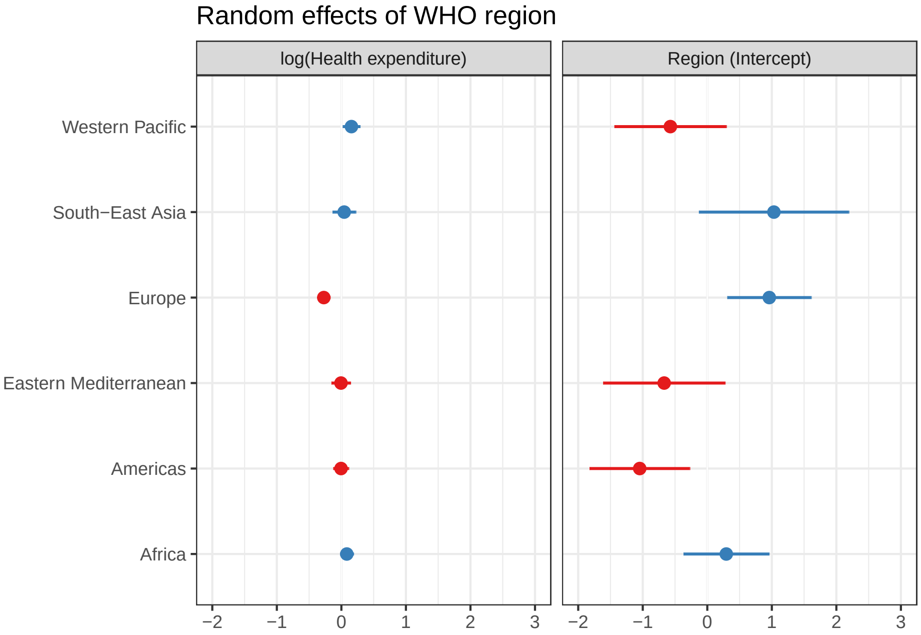

Figure 6: Random effects of WHO region on slope and intercept for the random slopes model m6.

The random effect of region on the slope of the variable health expenditure was scattered around zero for most regions, with Europe being a notable exception (Fig. 6). An increase in health expenditure was correlated with a reduction in incident TB numbers in Europe. The random effect of the region on the model intercept was more pronounced, although the standard error spaned zero for most of the estimates, indicating that they may not significantly differ from zero themselves. Only Europe and the Americas had an intercept that is significantly different from zero.

The nested effect of country on the random slope of the model was, similarly to the regional effect, scattered around zero (Fig. 7). There was more variation in the effect of country on the random intercept, indicating that a country-level effect on incident tuberculosis numbers may be more important than a regional effect.

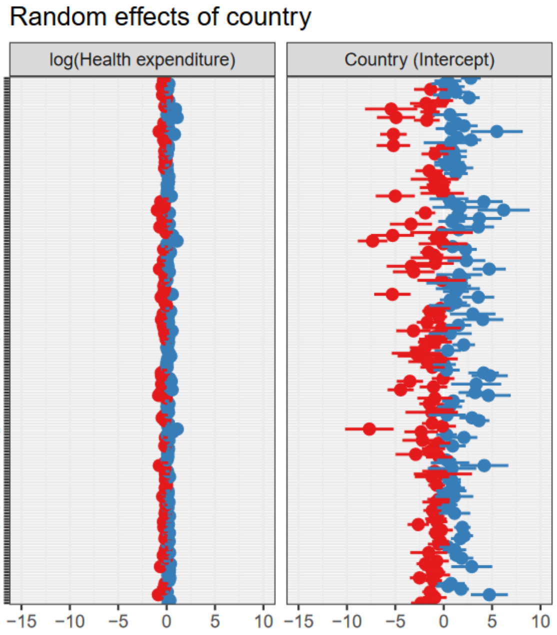

Figure 7: Random effects of country on slope and intercept for the random slopes model m6.

Limitations

As noted previously, population density was not a good proxy for assessing how closely a country's population lives together. Instead, a metric such as average number of individuals per household could be used as a more accurate predictor variable if this analysis were to be expanded upon.

Additionally, the model may not apply to every country. Missing data was more common for countries such as the People's Democratic Republic of Korea or Niue, meaning that conclusions drawn from the model may not be applicable to these countries.

Furthermore, the incident tuberculosis numbers were only reported cases of the disease - especially in countries where a tuberculosis infection is heavily stigmatised and those affected shun testing and treatment [8], the true number of incident cases may be higher, reducing the accuracy of this model.

Finally, the WHO recognises a weakened immune system to increase the risk of a TB infection. Specifically, people living with HIV are 16 times more likely to become infected with tuberculosis than those without HIV [2]. As such, countries with a high percentage of HIV-positive people most likely also experience a higher tuberculosis incidence. The observed correlation in this analysis may be masked by an effect of increased health expenditure on reduction in HIV numbers. If the number of HIV infections were controlled for in a future analysis, then a country's health expenditure per capita may in fact not be significantly correlated with a reduction in tuberculosis incidence after all.

Conclusions

This project's aim was to examine the relationship between a country's incident tuberculosis numbers and its potential explanatory variables GDP, health expenditure, and/or population density. This analysis was based on data sets provided by the WHO and the World Bank Group for 188 countries. Due to apparent regional clustering of incidence, a linear mixed effects model was chosen to explore this relationship.

Random slopes models had higher conditional R^2 values and could therefore explain more of the observed variance when compared to their random intercept-only counterparts. Additionally, the inclusion of a country-level random effect, nested within a region-level effect, increased model performance, prompting the inclusion of a nested random effect.

Following the principle of parsimony, the effect of GDP (correlated with both health expenditure and population density) was excluded from the final model, since the inclusion of this predictor variable only marginally improved model fit.

The final model included the fixed effects health expenditure, population density, and an interaction between these two variables. This model explained about 98.5 per cent of the variation in incident tuberculosis cases. A reduction in incidence was correlated with an increase in health expenditure (especially in Europe) and in population density, although the latter was found to have a stronger effect. This may be due to population density being a proxy for urban development, with more rural countries having a higher tuberculosis load, as can be expected. Based on this model, an increase in health expenditure has a stronger effect on the reduction of tuberculosis numbers when population densities are lower.

In summary, this analysis provides evidence that a country's overall level of urban development and ability to allocate funds in its healthcare system may be connected with a lower number of new or relapsed tuberculosis numbers. To combat this deadly disease, a viable strategy may be a collective, global effort to provide additional funding of the healthcare systems of less densely-populated countries, since the effect of increased funding may be even more pronounced here.

References

1. Sudre P, Ten Dam G, Kochi A. Tuberculosis: a global overview of the situation today. Bulletin of the World Health Organization. 1992;70(2):149.

2. Fact sheet: Tuberculosis [Internet]. World Health Organisation (WHO); 2023 Nov 7 [cited 2024 Jan 14]. Available from: https://www.who.int/news-room/fact-sheets/detail/tuberculosis

3. Fact sheet: Basic TB Facts [Internet]. Centers for Disease Control and Prevention (CDC); 2016 Mar 20 [cited 2024 Jan 14]. Available from: https://www.cdc.gov/tb/topic/basics/default.htm

4. Tuberculosis remains one of the deadliest infectious diseases worldwide, warns new report [Internet]. European Centre for Disease Prevention and Control (ECDC); 2022 [cited 2024 Jan 14]. Available from: https://www.ecdc.europa.eu/en/news-events/tuberculosis-remains-one-deadliest-infectious-diseases-worldwide-warns-new-report

5. Lawn SD, Zumla AI. Tuberculosis. The Lancet. 2011 Jul;378(9785):57–72.

6. World Health Organization, editor. Implementing the WHO Stop TB Strategy: a handbook for national TB control programmes. Geneva: World Health Organization; 2008. 184 p.

7. Regional Offices [Internet]. World Health Organization (WHO); [cited 2024 Jan 14]. Available from: https://www.who.int/about/who-we-are/regional-offices

8. Stigma and myths [Internet]. TB Alert; [cited 2024 Jan 14]. Available from: https://www.tbalert.org/about-tb/global-tb-challenges/stigma-myths/