Socioeconomic and Healthcare Predictors of Life Expectancy Across Countries: Evidence from 2019, 2021, and 2022

Qilei Zhang

Abstract Life expectancy reflects the health status of a population and the level of social development in a given country. This study utilizes data from 217 countries and regions for the years 2019, 2021, and 2022 to analyze the relationships between life expectancy and various factors, including income, education, medical resources, sanitary conditions, healthcare expenditure, fertility rates, and urbanization. The year 2019 serves as a pre-COVID-19 baseline, while 2021 and 2022 allow for the observation of any shifts in these relationships during the pandemic. The study employs linear regression and random forest models for prediction, with results evaluated through cross-validation and out-of-sample residual analysis.

The findings indicate that basic sanitary facilities, GDP per capita, and fertility rates are strong predictors of life expectancy across all three years. Countries with higher coverage of sanitary facilities and higher income levels generally exhibit higher life expectancy, whereas those with higher fertility rates tend to have lower life expectancy. Other factors—such as the number of physicians, female secondary school enrollment rates, and the proportion of the urban population—show weaker predictive power or lack stability across the years examined, while healthcare expenditure is a moderate predictor. While both models largely agree on the primary determinants, the tuned random forest has somewhat better out-of-sample accuracy in all three years. In some countries, actual life expectancy is well below the model-predicted values, suggesting that income and medical resources alone cannot fully account for cross-national disparities in health.

These results suggest that, in efforts to increase life expectancy, governments should prioritize basic sanitary facilities and public health services, while also improving education, maternal and child health, and the efficiency with which medical resources are utilized. Policymakers should also pay close attention to countries where actual life expectancy consistently falls below predicted levels and further investigate factors such as the distribution of medical services, the burden of infectious diseases, social inequality, and public governance. It should be noted that this study identifies predictive associations and does not establish a causal relationship between these factors and life expectancy.

Keywords life expectancy; socioeconomic factors; basic sanitation; cross-national analysis

1. Introduction

Life expectancy at birth is a key indicator of population health and social development. Disparities in life expectancy across countries are closely linked to economic resources, education, healthcare services, and public health conditions; notably, income growth alone cannot fully account for improvements in health (Cutler et al., 2006). The COVID-19 pandemic has further altered this landscape: while some countries saw a recovery in life expectancy by 2021, others experienced a continued decline (Schöley et al., 2022). Consequently, understanding cross-national differences during the pandemic requires considering both long-term developmental conditions and the immediate shocks to mortality.

While existing research provides a basis for the aforementioned associations, the samples and methodologies employed across various studies differ, leaving an open question: when using identical metrics and validation procedures, does a random forest model consistently outperform linear regression both before and during the pandemic? To address this, the present study utilizes World Bank data for 217 countries and regions from 2019, 2021, and 2022—with 2019 serving as a pre-pandemic baseline—and incorporates seven indicators: GDP per capita, female education, physician density, access to basic sanitation, healthcare expenditure, fertility rate, and urbanization.

This study compares linear regression with a hyperparam-

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

eter-tuned random forest model within a unified framework of data preprocessing and out-of-sample validation, examining model performance, variable contributions, and prediction biases at the national level. Its contribution lies in providing comparable, cross-year predictive evidence; as pandemic-related indicators are not explicitly included, the results cannot be used to identify the causal effects of the pandemic or any individual factor.

The following sections present the relevant literature, data and methodology, research findings, and discussion, followed by conclusions, limitations, and policy implications, as well as information regarding the availability of data and code.

2. Literature Review

The relationship between economic development and life expectancy serves as an important starting point for cross-national health research. Preston (1975) compared per capita income with life expectancy across different time periods and found that life expectancy at a given income level varies over time, suggesting that economic growth alone cannot fully account for improvements in health. Cutler et al. (2006) further emphasized that scientific knowledge, medical technology, and their dissemination and application are critical factors in understanding the decline in mortality rates. Together, these studies indicate that while per capita GDP reflects economic resources, its relationship with life expectancy must be interpreted in the context of public health and technological conditions; it cannot be viewed simply as a direct causal effect of rising income.

The distribution of social resources and the status of women provide a more concrete explanation for this relationship. Marmot (2005) points out that health inequalities—both between and within nations—are closely linked to the social conditions in which people live; consequently, health issues cannot be resolved solely through the healthcare sector. In a cross-national study of developing countries, Williamson and Boehmer (1997) identified women's education, economic status, and reproductive autonomy as significant predictors of female life expectancy, while also noting that the choice of educational indicators can influence analytical outcomes. Meanwhile, Sart et al. (2024) utilized panel data from new European Union member states (covering 2010–2022) to incorporate both education and income inequality into an analysis of life expectancy. While these studies support the perspective that disparities in longevity should be understood through the lens of education and demographic development, the concepts used here—overall life expectancy and female secondary school enrollment—differ from those in the aforementioned research; therefore, the conclusions drawn in those studies cannot be directly equated to the findings of the present research.

Public health infrastructure represents another significant avenue for research. The World Health Organization's (2018) Guidelines on Sanitation and Health summarize the evidence linking sanitation interventions to health outcomes, underscoring the critical role of safe sanitation systems in reducing disease and improving health. This provides a rationale for including basic sanitation coverage in the present study, establishing it as a key indicator that bridges living conditions and population health. However, coverage rates capture only certain aspects of service conditions; they do not offer a comprehensive assessment of facility operation, safety management, or actual usage. Similarly, per capita healthcare expenditure and physician density—which reflect the scale of investment and the supply of personnel, respectively—cannot, on their own, serve as proxies for the quality of healthcare or the efficiency of resource allocation. Consequently, it is essential to distinguish between the resource conditions captured by these indicators and the actual outcomes of health services.

In recent years, research in machine learning has provided a complementary approach for examining complex relationships in life expectancy forecasting. Zhang et al. (2025) utilized panel data from 20 countries with high life expectancy (covering the period 2001–2023) and integrated traditional econometric methods with models such as random forests and gradient boosting, demonstrating the application of machine learning in this field. However, the scope of their sample—consisting of countries with high life expectancy—differs significantly from the sample in the present study, which encompasses countries and regions at varying levels of development. Furthermore, their research design differs from the year-by-year cross-sectional forecasting comparisons employed here. Consequently, existing applications of machine learning cannot serve as a substitute for independent validation within the current sample, nor can they establish, a priori, that random forests will necessarily outperform linear regression.

The COVID-19 pandemic has further complicated cross-year comparisons. Schöley et al. (2022) observed a divergence in life expectancy trends across the countries studied in 2021: while some nations experienced a recovery, others saw continued declines. This suggests that disparities in life expectancy during the pandemic may be shaped by both long-term structural conditions and the immediate impact of mortality shocks. Consequently, when interpreting outcomes for 2021 and 2022, it is essential to account for the pandemic context—using pre-pandemic years as a benchmark to assess whether observed patterns align with historical trends—rather than attributing annual variations solely to the pandemic itself.

In summary, the existing literature offers evidence regarding economic and social conditions, public health infrastructure, machine learning applications, and the mortality impact of the pandemic. However, given the variations

in samples, time periods, and analytical objectives across these studies, they do not directly address the specific question posed here: when using an identical set of seven indicators and a consistent validation protocol, does the Random Forest model consistently outperform linear regression in out-of-sample forecasting—both before and during the pandemic? Using 2019 as a reference point, this study analyzes cross-sectional data for 2021 and 2022, performing missing-value imputation and hyperparameter tuning for the Random Forest model within the training datasets. It further compares the models by examining variable contributions and country-level prediction errors. The contribution of this research lies in providing cross-year forecasting evidence within a unified analytical framework and clarifying the predictive power—as well as the limitations—of commonly used development indicators, rather than identifying the causal effects of the pandemic or any individual factor.

3. Data Source and Method

3.1 Data and variables

This study utilizes data from the World Bank's World Development Indicators (WDI) database for the years 2019, 2021, and 2022. After excluding aggregated data—such as those for regions or income groups—the dataset for each year comprises 217 countries and territories. Cross-sectional analyses are conducted separately for each year, rather than pooling the three years into a panel model. The year 2019 serves as a pre-pandemic reference, while 2021 and 2022 reflect conditions during the pandemic. Each country and territory is assigned an equal weight in the analysis; no weighting by population size is applied.

The dependent variable is life expectancy at birth, measured in years. Seven predictor variables are included, representing economic resources, female education, healthcare resources, public health infrastructure, and demographic structure. Specific indicators, their corresponding WDI codes, and original units of measurement are presented in Table 1.

Table 1. Variables and measurement

Variable	WDI series	Measurement / unit
Life expectancy	SP.DYN.LE00.IN	Life expectancy at birth, years
GDP per capita (PPP)	NY.GDP.PCAP.PP.KD	Constant PPP dollars per person (WDI series)
Female secondary enrollment	SE.SEC.ENRR.FE	Female gross secondary enrollment, percent; may exceed 100%
Physicians	SH.MED.PHYS.ZS	Physicians per 1,000 people
Basic sanitation	SH.STA.BASS.ZS	Population using at least basic sanitation services, percent
Health expenditure	SH.XPD.CHEX.PC.CD	Current health expenditure per capita, current US dollars
Fertility rate	SP.DYN.TFRT.IN	Births per woman
Urban population	SP.URB.TOTL.IN.ZS	Population living in urban areas, percent

Note: Gross enrollment rate for girls in secondary schools may exceed 100% because its numerator includes students who are not in the officially defined secondary school age range. GDP per capita is adjusted for purchasing power parity, while healthcare expenditure is measured in current US dollars; therefore, their price bases and measurement methods differ. Logarithmic transformation does not eliminate this difference.

These indicators are used to describe transnational development conditions and do not fully represent service quality, resource allocation, or institutional capacity. For example, the gross enrollment rate for girls in secondary schools reflects the current enrollment situation, rather than the cumulative education level of the adult population; and physician density and healthcare expenditure cannot directly measure the accessibility and efficiency of healthcare services. The study did not include direct indicators such as pandemic mortality rate, vaccination rate, or excess mortality; therefore, annual comparisons are not used to identify the independent impact of the pandemic.

3.2 Preprocessing and modeling

This study retains the original observations and, prior to any imputation, calculates the number and proportion of missing values for each variable by year. Descriptive statistics—including valid sample size, mean, standard deviation, minimum, and maximum—are also computed based on the original, non-missing observations. Because the degree of data completeness varies across variables, the sample sizes used for these descriptive statistics may

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

differ.

For both models, GDP per capita and healthcare expenditure per capita are transformed using the natural logarithm. This compresses the scale of these monetary variables and allows for relationships with life expectancy that deviate from the linear patterns associated with the original, untransformed values. Since natural logarithms are undefined for zero or negative values, such entries are treated as missing during the transformation process and are subsequently addressed through imputation. Life expectancy and the remaining predictor variables are not log-transformed.

Missing values in the predictor variables are imputed using the median calculated from the training data. Log-transformation, imputation, and model fitting are integrated into a unified processing pipeline. For linear regression, the data are standardized following imputation, whereas no such standardization is applied to the random forest models. During both hold-out testing and cross-validation, the medians used for imputation—as well as the means and standard deviations required for standardization—are estimated exclusively from the respective training subsets and then applied to the validation or test subsets, thereby preventing information leakage. Missing values for life expectancy are not imputed; there are 217 target observations for each of the three years under consideration.

3.3 Model Estimation and Parameter Selection

This study compares multiple linear regression with random forest regression. Linear regression is used to summarize conditional associations after controlling for other predictor variables. Because the inputs are standardized, each coefficient shows the change in the expected life expectancy, in years, for a one standard deviation increase in that predictor, holding the others constant. For GDP and health expenditure, this applies to the log values after imputation. The coefficients describe associations and should not be interpreted as causal effects.

The random forest model consists of 400 decision trees, designed to capture non-linear relationships and interaction patterns between the predictor variables and life expectancy. The study employed a random search to select eight parameter combinations from the candidate range listed in Table 2, and then used three-fold internal cross-validation to select the combination with the lowest mean root mean square error (RMSE). After selecting the parameters, the model is refitted on all the training data provided for this search. A random seed of 42 was used for data splitting, model training, and the random search process.

Table 2. Random forest hyperparameter search space

Hyperparameter	Parameter name	Candidate values
Maximum tree depth	max_depth	None (unrestricted), 5, 8, 12, 20
Minimum samples per leaf	min_samples_leaf	1, 2, 4, 6
Proportion of features considered at each split	max_features	0.6, 0.8, 1.0

Notes: The number of trees is fixed at 400. Randomized search evaluates eight parameter combinations using three-fold inner cross-validation, selecting the combination with the lowest mean validation RMSE. The selected model is then refitted on the corresponding training sample. Tuning is conducted separately within the holdout training set and each outer training fold of the five-fold cross-validation. The random seed is set to 42.

3.4 Model Evaluation and Interpretation of Results

The study employs two complementary evaluation approaches. First, the training set and the test set were divided into 75% and 25% for each year, respectively, resulting in 162 training observations and 55 test observations. Both models use the same data partitioning; the random forest only tunes parameters within the training set, and the test set does not participate in parameter selection.

Secondly, a shuffled five-fold cross-validation is performed on the complete samples of each year, with both models using the same external partition. For the Random Forest, this is implemented as a nested cross-validation, where a three-fold internal tuning process is carried out within the training portion of each fold. The external validation portion is excluded from both the estimation of preprocessing parameters and the model selection process. The reporting metrics include the coefficient of determination (R^2), RMSE, and mean absolute error (MAE), and summarize the mean and standard deviation of five external folds. Higher R^2 values and lower RMSE and MAE values indicate superior predictive performance. As the dependent variable is not log-transformed, the units for RMSE and MAE remain in years. The standard deviation between folds is used to describe the degree to which the results vary with the division, and is not considered as a significance test for confidence intervals or model differences.

Variable interpretation integrates standardized linear coefficients with Random Forest feature importance. For the Random Forest, both impurity-based importance and permutation-based importance (calculated on the test set) are assessed. The latter is derived by randomly permuting each variable 30 times and recording the average decline

in the test-set R^2. These two importance metrics differ in both meaning and scale; permutation importance reflects the model's predictive reliance on a given variable but does not indicate the direction of the effect, nor does it imply a causal contribution.

At the national level, prediction deviations are quantified using out-of-fold residuals—calculated as the actual life expectancy minus the predicted value for a given observation within an external validation fold. A negative residual indicates that the actual life expectancy is lower than the predicted value, while a positive residual indicates the opposite. These deviations are used to identify cases warranting further investigation; they do not, in isolation, establish that factors such as the pandemic, disease burden, or institutional conditions are the underlying causes.

4. Results

4.1 Descriptive Statistics and Data Integrity

Significant cross-national variations are evident across all three years (Table 3). The unweighted average life expectancy across the countries and regions fell from 73.03 years in 2019 to 71.80 years in 2021, before rebounding to 73.11 years in 2022. While this trend offers a descriptive comparison of the periods before, during, and after the pandemic, it cannot be attributed solely to the impact of the pandemic, nor does it signify a complete recovery. In 2021, the observed values ranged from 40.28 to 85.53 years; in 2022, the lowest value—18.82 years, recorded for the Central African Republic—is the value published in the WDI series, although it lies far outside the country's usual range.

The standard deviations for GDP per capita and healthcare expenditure are substantial, indicating marked disparities in resources between countries. Missing data are primarily concentrated in the variables for physician density (with 37.79%–43.32% of values missing over the three years) and female secondary school enrollment (31.80%–35.48% missing); conversely, there are no missing values for life expectancy, fertility rate, or the proportion of the urban population.

Table 3. Descriptive statistics before transformation and imputation

Variable	2019, Mean (SD)	2021, Mean (SD)	2022, Mean (SD)
Life expectancy (years)	73.03 (7.97)	71.80 (7.71)	73.11 (7.94)
GDP per capita (constant 2021 international $)	26,352.95 (26,295.48)	26,003.83 (26,141.11)	26,809.48 (26,569.07)
Health expenditure per capita (current US$)	1,190.49 (1,958.20)	1,407.31 (2,323.67)	1,367.14 (2,221.82)
Physicians per 1,000 people	2.24 (1.84)	2.39 (1.82)	2.18 (1.86)
Basic sanitation coverage (%)	78.33 (26.82)	79.17 (26.44)	79.84 (26.00)
Female secondary enrollment (gross %)	92.75 (25.21)	91.94 (25.39)	94.55 (23.17)
Total fertility rate (births per woman)	2.54 (1.27)	2.46 (1.23)	2.41 (1.22)
Urban population (%)	61.56 (23.58)	62.09 (23.43)	62.35 (23.33)

Notes: Statistics use available raw observations and equal country weights. Sample sizes for 2019/2021/2022 are 217/217/217 for life expectancy, fertility, and urban population; 199/199/199 for GDP; 193/192/191 for health expenditure; 135/123/125 for physicians; 203/201/199 for sanitation; and 141/148/140 for female enrollment. Extreme life-expectancy observations are retained as recorded in the supplied dataset; the Central African Republic's 2022 value of 18.82 is published in WDI but lies far outside its usual range.

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

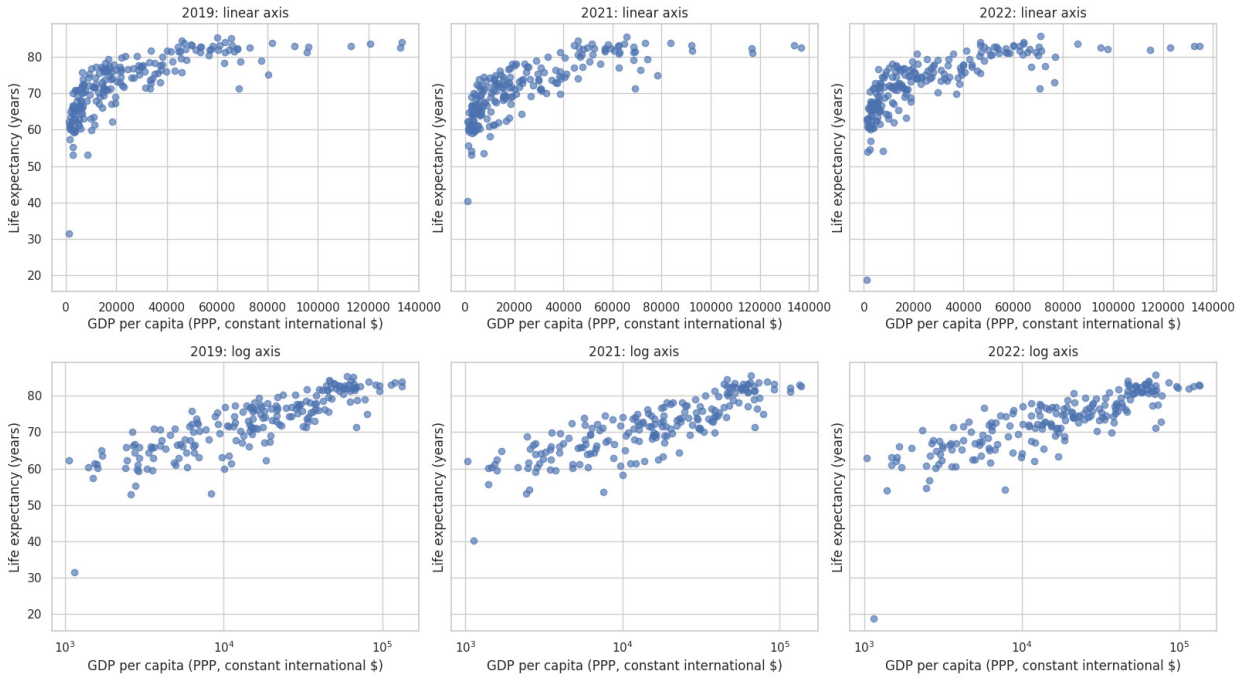


Figure 1: Life Expectancy and GDP per Capita, 2019, 2021, and 2022

Notes: Each point represents a country or territory with available observations. The upper panels use a linear GDP axis; the lower panels use a logarithmic GDP axis. Life expectancy remains measured in years. Extreme recorded observations are retained; the Central African Republic's 2022 value of 18.82 is published in WDI but lies far outside its usual range.

4.2 Model prediction performance

After hyperparameter tuning, the Random Forest model achieved higher R^2 values and lower RMSE and MAE values across all three years in the hold-out test (Table 4). The performance gap between the two models was most pronounced in 2021: the Random Forest yielded an R^2 of 0.785 compared to 0.682 for linear regression, with corresponding RMSE values of 3.669 years and 4.461 years. Five-fold cross-validation yielded the same relative ranking. The average R^2 for the Random Forest was higher by approximately 0.027, 0.072, and 0.032 across the three years, respectively, while the average RMSE was consistently lower. Consequently, given the current dataset and validation setup, the Random Forest demonstrates a consistent predictive advantage. However, as both models exhibited significant variability across folds in 2022—and because no formal significance testing was conducted to compare the models—the observed superiority in average metrics cannot be characterized as statistically significant.

Table 4. Holdout and five-fold cross-validation performance

Year	Model	Holdout R^2	Holdout RMSE	Holdout MAE	CV R^2, Mean (SD)	CV RMSE, Mean (SD)
2019	Linear regression	0.773	3.740	2.822	0.757 (0.050)	3.879 (0.784)
2019	Tuned random forest	0.805	3.472	2.549	0.784 (0.074)	3.640 (0.991)
2021	Linear regression	0.682	4.461	3.576	0.718 (0.067)	4.021 (0.620)
2021	Tuned random forest	0.785	3.669	2.944	0.790 (0.055)	3.467 (0.606)
2022	Linear regression	0.742	3.804	3.068	0.709 (0.080)	4.226 (1.412)
2022	Tuned random forest	0.794	3.404	2.686	0.741 (0.113)	3.974 (1.641)

Notes: Each annual holdout contains 162 training and 55 test observations. Both models use identical outer folds. Random forest tuning uses three-fold inner validation within each training subset. GDP and health expenditure are natural log transformed in both models. RMSE and MAE are measured in years; SD denotes variation across five outer folds.

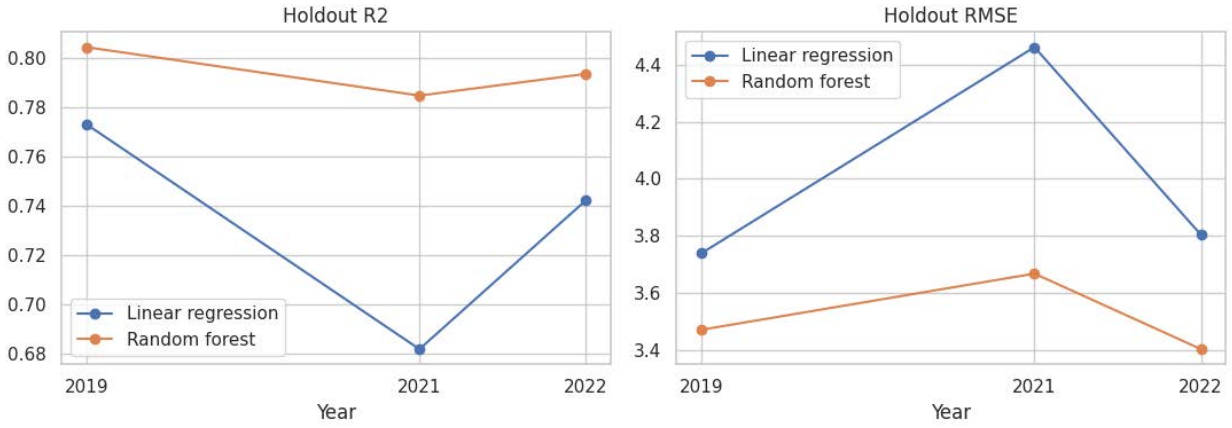


Figure 2: Holdout predictive performance by model and year.

Notes: The panels compare test-set R^2 and RMSE for linear regression and the tuned random forest. Higher R^2 and lower RMSE indicate better performance. RMSE is measured in years. Each annual test set contains 55 observations. Lines connect the three analyzed years for visual comparison.

4.3 Linear Association and Random Forest Importance

In the linear regression analysis, basic sanitation coverage and the logarithm of GDP per capita show a positive association across all three years, while fertility rates show a negative association; the absolute values of their standardized coefficients are relatively large (Table 5). The coefficients for healthcare expenditure are also positive in all three years. The coefficients for female school enrollment are close to zero, and the direction of the association for physician density varies by year. As this analysis does not estimate confidence intervals for the coefficients, these results should not be used to determine statistical significance.

Permutation importance analysis using a random forest model indicates that basic sanitation ranks first in all three years, followed primarily by GDP, fertility rates, and healthcare expenditure, though the specific rankings vary. Physician density and female school enrollment contribute minimally to the model's predictive power for 2019 and 2021. Low or slightly negative importance values suggest that the current model derives limited additional predictive information from these indicators; this does not imply that the corresponding public services lack health-related value.

Table 5. Standardized linear coefficients and random forest permutation importance

Predictor	β, 2019	β, 2021	β, 2022	PI, 2019	PI, 2021	PI, 2022
Log GDP per capita	2.101	2.387	2.138	0.105	0.148	0.125
Log health expenditure per capita	0.968	1.584	0.567	0.059	0.086	0.057
Physician density	−0.190	−0.367	0.147	0.000	0.001	0.034
Basic sanitation coverage	2.295	1.792	2.017	0.287	0.283	0.195
Female secondary enrollment	−0.122	−0.030	−0.039	−0.002	−0.005	0.001
Total fertility rate	−2.067	−1.448	−2.384	0.110	0.074	0.083
Urban population	0.541	0.413	0.265	0.011	0.018	0.022

Notes: β denotes coefficients from the holdout-training linear model, expressed in years per one training-sample SD of the processed predictor. PI denotes the mean decrease in test-set R^2 across 30 permutations. Coefficients and PI measure different quantities and should not be compared numerically. A displayed PI of 0.000 reflects rounding.

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

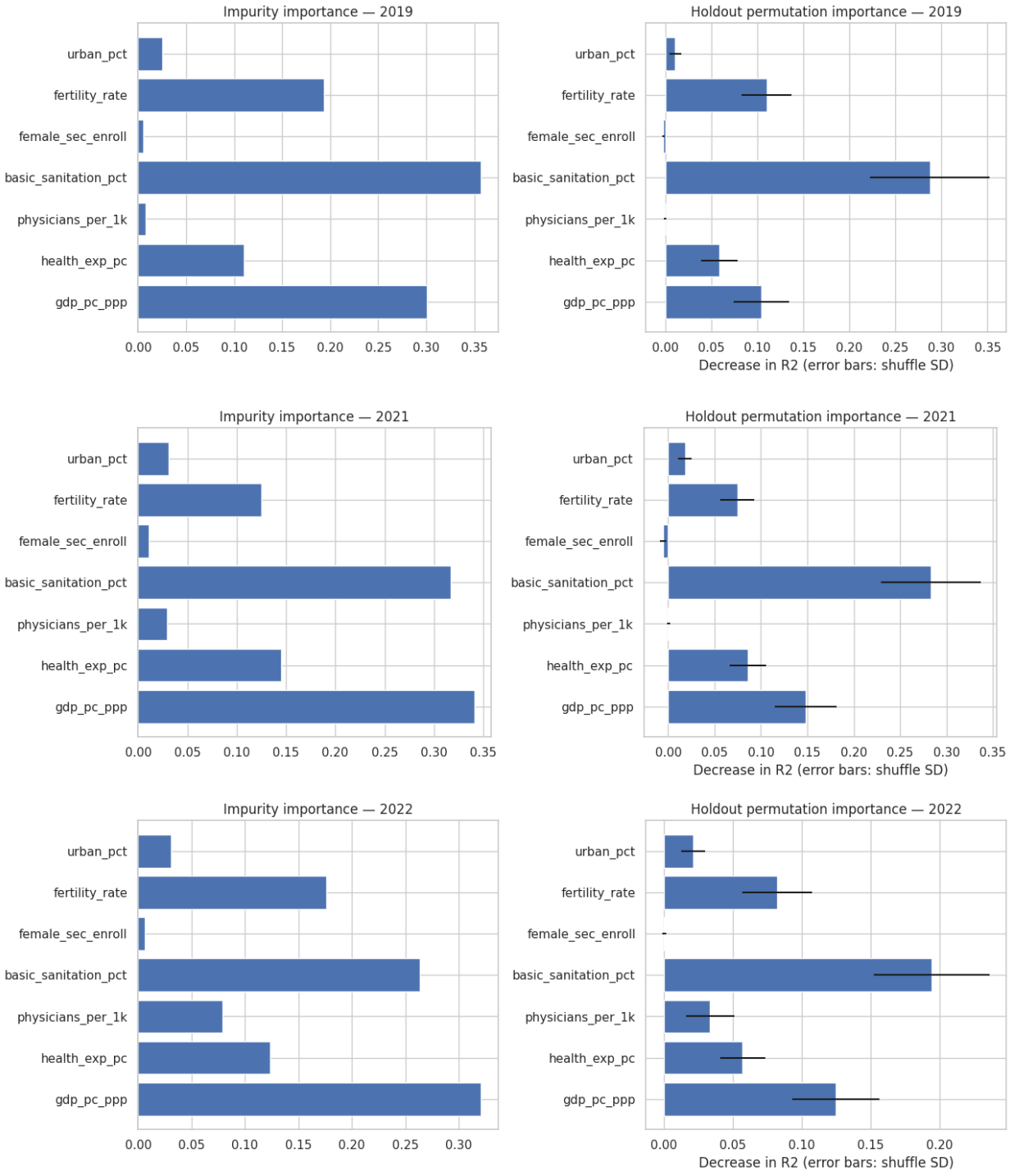


Figure 3: Holdout permutation importance of the tuned random forest in 2019, 2021, and 2022.

Notes: Panels are arranged from top to bottom by year. Bars show the mean decrease in test-set R^2 across 30 permutations; error bars indicate ±1 standard deviation across permutations, not confidence intervals. Higher values indicate greater predictive contribution. Horizontal scales differ across panels. GDP and health expenditure are natural log transformed within the model pipeline.

4.4 Prediction Biases at the National Level

Country-level residuals are reported for the linear model to provide an interpretable baseline for identifying systematic prediction errors. The country-level residual examples below use the linear model's out-of-fold predictions; the random forest is compared with it in Section 4.2 on aggregate predictive performance. These linear-model predictions reveal that, for Nigeria, actual life expectancy in 2019, 2021, and 2022 fell below the predicted values by 11.09, 11.10, and 10.12 years, respectively—a persistent negative deviation. For South Africa, the corresponding

deviations were −7.45, −10.36, and −7.51 years, with the largest gap occurring in 2021.

The negative deviation for the Central African Republic is even more extreme, reaching −38.80 years in 2022—corresponding to an observed life expectancy of 18.82 years. The 18.82-year observation is published in the WDI series but lies far outside the country's usual range, so this residual should be interpreted cautiously. The persistent deviations observed in other countries suggest that the seven indicators currently employed do not fully capture all cross-national differences; the specific reasons for this, including omitted factors and data limitations, will be analyzed in the discussion section.

5. Discussion

This study finds that basic sanitation facilities, GDP per capita, and fertility rates are the most consistent predictors of life expectancy. These three factors ranked highly in 2019, 2021, and 2022. This suggests that a nation's life expectancy depends not only on medical services but also on living conditions, economic resources, and demographic structure.

The role of basic sanitation facilities is the most stable. Access to safe drinking water, toilets, and sewage treatment systems can reduce the incidence of infectious diseases and other preventable causes of death (WHO, 2018). Therefore, the government needs to continue to improve basic health facilities, especially in rural and low-income areas. Policies also need to focus on whether facilities can be used normally, rather than just on coverage.

Countries with higher incomes generally possess more resources to allocate toward housing, nutrition, education, and healthcare (Cutler et al., 2006). However, the health improvements derived from income growth are not guaranteed; such gains are typically more pronounced in low-income nations (Preston, 1975). Therefore, economic growth does not automatically lead to higher life expectancy. Governments must also improve resource allocation and ensure that public funds are effectively directed toward primary healthcare and public health services (Marmot, 2005).

There is a strong negative correlation between fertility rates and life expectancy. Higher fertility rates can place increased strain on both families and public health systems (Williamson & Boehmer, 1997). Nevertheless, this finding does not imply that high fertility rates directly cause a decline in life expectancy; rather, fertility rates may also reflect factors such as women's education, household income, and the availability of maternal and child healthcare. Relevant policies should improve women's education, family planning, and maternal and child health services, rather than simply aiming to reduce the number of births.

Physician density, female secondary school enrollment rates, and the proportion of the urban population show weaker or less consistent predictive contributions across years, while healthcare expenditure is a moderate predictor: its linear coefficient is positive in all three years, and it ranks around fourth in permutation importance. This does not mean these factors are unimportant. For instance, medical personnel and healthcare funding may be concentrated in a few urban centers, failing to translate into effective primary-level services. Healthcare expenditure may have different implications depending on how effectively resources reach patients. Tripathi and Maiti (2023) also find that well-managed urbanization is associated with better health outcomes. Thus, beyond merely increasing investment, governments must also focus on the equitable allocation of resources and the quality of services provided.

The COVID-19 pandemic likely changed outcomes in 2021 and 2022. Existing research indicates that life expectancy declined in many countries during the pandemic, though the extent of this decline varied (Aburto et al., 2022; Schöley et al., 2022). Data from 2019 provide a pre-pandemic baseline, which helps distinguish between long-standing disparities and changes that emerged during the pandemic.

National residuals can help identify countries where actual life expectancy is substantially higher or lower than predicted values. Countries with large negative residuals may face challenges related to infectious diseases, social inequality, or the allocation of healthcare resources. Conversely, countries with large positive residuals may have achieved better health outcomes despite limited resources. However, residuals do not serve as direct proof of the success or failure of any specific policy; they merely highlight countries that warrant further investigation.

The predictive performance of linear regression and random forest models is generally similar, with the tuned random forest achieving somewhat better out-of-sample accuracy in all three years. More importantly, both models assign high importance to factors such as sanitation facilities, income, and fertility rates, suggesting that the primary findings are not entirely dependent on any single model. This study analyzes predictive relationships and cannot establish causality. Given that the three years examined are close to one another, the results reflect only short-term consistency.

6. Conclusion, Limitations, and Policy Recommendations

This study compares the performance of cross-national life expectancy forecasting before the pandemic (2019) and during the pandemic (2021–2022). Using a standardized preprocessing and validation pipeline, a tuned random forest model consistently yielded superior out-of-sample average predictive performance across all three years. Basic

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

sanitation facilities emerged as the most important predictor based on permutation importance. GDP per capita and fertility rates also provided strong predictive information, while healthcare expenditure was a moderate predictor. The findings indicate that these development indicators can capture some of the cross-national disparities in life expectancy; however, significant country-level prediction errors persist.

The three annual cross-sections are insufficient for identifying causal relationships or the independent impact of the pandemic, nor do they constitute independent, repeated validations. National-level averages mask internal inequalities, and the models fail to account for factors such as the burden of disease, governance, and the quality of healthcare. Several predictor variables contain significant amounts of missing data, and the use of median imputation fails to capture the uncertainty associated with these missing values; furthermore, the Central African Republic's published WDI life expectancy value of 18.82 in 2022 lies far outside its usual range. Finally, this analysis does not assess the statistical significance of differences between the models, and the conclusions are specific only to the sample and model specifications employed.

The direct implication of this study is that insufficient coverage of sanitation facilities and persistent negative prediction bias should serve as indicators for further national-level assessments, rather than as a basis for directly prioritizing policy investments solely according to variable importance. In line with independent evidence from the World Health Organization (2018), there is a clear public health rationale for improving safe sanitation facilities; however, specific interventions must account for local needs, implementation costs, and service quality. For countries exhibiting significant prediction bias, the initial step should be to verify the data, followed by an investigation that considers the local burden of disease and service accessibility.

Data and code accessibility

Data comes from the World Bank's World Development Indicators database. The analysis code and saved output are available in the following Colab notebook:

https://colab.research.google.com/drive/1u_T4UuVFUNv7CAhvl-EhVOaXTXzYAgoa?usp=sharing

The notebook records the data extraction, missing value processing, logarithmic transformation, model tuning and verification process, and exports the annual raw data and results table. The random seed is set to 42; since the WDI data may be updated, the results obtained after re-downloading may be different from the data version used in this article.

References

[1] Aburto, J. M., Schöley, J., Kashnitsky, I., Zhang, L., Rahal, C., Missov, T. I., Mills, M. C., Dowd, J. B., & Kashyap, R. (2022). Quantifying impacts of the COVID-19 pandemic through life-expectancy losses: A population-level study of 29 countries. *International Journal of Epidemiology, 51*(1), 63–74. https://doi.org/10.1093/ije/dyab207

[2] Cutler, D., Deaton, A., & Lleras-Muney, A. (2006). The determinants of mortality. *Journal of Economic Perspectives, 20*(3), 97–120. https://doi.org/10.1257/jep.20.3.97

[3] Marmot, M. (2005). Social determinants of health inequalities. *The Lancet, 365*(9464), 1099–1104. https://doi.org/10.1016/S0140-6736(05)71146-6

[4] Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population Studies, 29*(2), 231–248. https://doi.org/10.1080/00324728.1975.10410201

[5] Sart, G., Bayar, Y., & Danilina, M. (2024). Impact of education and income inequalities on life expectancy: Insights from the new EU members. *Frontiers in Public Health, 12*, Article 1397585. https://doi.org/10.3389/fpubh.2024.1397585

[6] Schöley, J., Aburto, J. M., Kashnitsky, I., Kniffka, M. S., Zhang, L., Jaadla, H., Dowd, J. B., & Kashyap, R. (2022). Life expectancy changes since COVID-19. *Nature Human Behaviour, 6*(12), 1649–1659. https://doi.org/10.1038/s41562-022-01450-3

[7] Tripathi, S., & Maiti, M. (2023). Does urbanization improve health outcomes: A cross country level analysis. *Asia-Pacific Journal of Regional Science, 7*(1), 277–316. https://doi.org/10.1007/s41685-022-00268-1

[8] Williamson, J. B., & Boehmer, U. (1997). Female life expectancy, gender stratification, health status, and level of economic development: A cross-national study of less developed countries. *Social Science & Medicine, 45*(2), 305–317. https://doi.org/10.1016/S0277-9536(96)00346-2

[9] World Bank. (n.d.). *World Development Indicators* [Data set]. Retrieved October 24, 2026, from https://databank.worldbank.org/source/world-development-indicators

[10] World Health Organization. (2018). *Guidelines on sanitation and health*. https://www.who.int/publications/i/item/9789241514705

[11] Zhang, H., Qiao, H., Li, X., Uddin, I., Li, R., & Zhang, X. (2025). Determinants of life expectancy in high longevity countries: Evidence from machine learning. *Frontiers in Public Health, 13*, Article 1687871. https://doi.org/10.3389/fpubh.2025.1687871