Design and Methodology:

This study evaluates how imputation can improve analytical validity, statistical power, and model performance when assessing global healthcare expenditure. Data were collected from the World Bank Open Datahub for 211 economies from 2000-2025 (N=5,425). Seven indicators were analyzed: GDP per capita, unemployment rate, secondary school enrollment, female labor force participation, official development assistance per capita, agriculture value added, and healthcare expenditure per capita as the outcome variable. On average, 27% of observations from the original dataset were not missing completely at random (MCAR) based on the Little's MCAR test, confirming imputation appropriateness.

Original Data and Results:

Five imputation techniques were compared using scaled data and a complete-case reference model: mean substitution, regression imputation, multiple imputation by chained equations (MICE), k-nearest neighbors (KNN) imputation, and random forest imputation. Model performance was assessed using root-mean-square error (RMSE) and coefficient of determination (R^2). Random-forest (RMSE = 0.037) and MICE (0.041) achieved the lowest errors, while regression imputation provided the best model fit (R^2= 0.57), improving explanatory power compared to the complete-case baseline model (R^2 = 0.31).

Conclusion:

Global health econometric assessments led by the World Bank have explored multiple imputation and machine-learning-based methods to address systematic data gaps that hinder cross-country analyses of global health expenditure. Missing data reduce database accuracy and distort how relationships among economic and health indicators are interpreted, leading to biased projections and ill-informed policy decisions. Results demonstrate that advanced, data-driven imputation can support reliability and strengthen the evidence base for global health financing and development decision-making, particularly in data-scarce, low-income countries.