

Predicting Health Insurance Claims in Thailand Using Bayesian Multilevel Logistic Regression and Socioeconomic Indicators

Mahatthakorn Plensamai*

Accounting Department, Faculty of Business Administration, Ubon Ratchathani University, Warinchamrap, Ubon Ratchathani, 34190, Thailand

* Corresponding author. E-mail address: mahatthakorn.p@ubu.ac.th

Received: 2 January 2026; Revised: 7 September 2026; Accepted: 1 October 2026; Available online: 6 October 2026

Abstract

Socioeconomic disparities may influence the probability of health insurance claims and the financial sustainability of insurance systems. This study developed a Bayesian hierarchical logistic model to predict the probability that an individual had at least one recorded health insurance claim during the study period. Secondary individual-level data covering 2020–2023 were obtained from the National Statistical Office of Thailand and linked with province-level healthcare indicators from the Ministry of Public Health and regional economic indicators. The initial merged dataset contained 52,870 observations. After excluding records with missing outcomes, incomplete predictors, duplicate identifiers, and unmatched geographic information, the final analytic sample comprised 48,320 individuals from 77 provinces. Individuals were modeled as nested within provinces, and posterior distributions were estimated using Hamiltonian Monte Carlo. The Bayesian hierarchical model achieved an area under the curve of 0.862, exceeding logistic regression, random forest, and gradient boosting by 0.081, 0.039, and 0.031, respectively. It also produced the lowest root mean square error of 0.093 and Brier score of 0.072. Higher income, education, employment, healthcare access, and regional gross domestic product were associated with lower claim probability, whereas higher health expenditure was associated with greater claim probability. Posterior convergence and sensitivity analyses indicated stable estimates across alternative prior specifications. These findings demonstrate that Bayesian hierarchical modeling can improve claim-risk prediction while explicitly quantifying parameter and predictive uncertainty. The model may support risk classification, resource allocation, and equity-oriented health insurance policy in Thailand.

Keywords: Multilevel Analysis, Markov Chain Monte Carlo, Posterior Distribution, Actuarial Analytics, Thailand

Introduction

Health insurance plays a vital role in promoting financial protection, equity, and access to healthcare. Yet, in many developing and middle-income Asian countries, health insurance systems continue to face sustainability challenges driven by rising medical expenditures, demographic aging, and widening socioeconomic disparities. Accurate risk prediction and fair premium determination are crucial for ensuring both financial stability and social equity in these systems. However, conventional actuarial and statistical approaches—such as generalized linear models (GLMs), logistic regression, and survival analysis—often rely on restrictive assumptions of linearity, independence, and homogeneity among predictors. These assumptions limit their capacity to model the nonlinear and hierarchical relationships between socioeconomic factors and health outcomes. As a result, traditional models tend to underestimate uncertainty, misclassify risk categories, and overlook the latent socioeconomic mechanisms influencing insurance risk. Addressing these methodological limitations requires a more flexible probabilistic modeling framework that captures interdependencies and explicitly quantifies uncertainty.

Bayesian inference provides this framework by integrating prior knowledge, probabilistic reasoning, and empirical data into a coherent statistical structure. It allows model parameters to be treated as random variables

and updated iteratively as new information becomes available, offering a natural way to manage uncertainty in complex datasets. Several recent studies have demonstrated the growing relevance of Bayesian approaches in health and actuarial modeling. For instance, Nkrumah (2025) introduced a Bayesian statistical learning model combined with copula-based time series to assess health insurance portfolio risks and found that Bayesian regression with Markov Chain Monte Carlo (MCMC) estimation provides superior accuracy in modeling interdependent claim structures compared to classical GLMs. Similarly, Shams and Ghosh (2026) applied Bayesian nonparametric regression to predict insurance claim frequency and severity, showing that Bayesian approaches outperform parametric models when data are skewed or heavy-tailed. Shi et al. (2024) extended this concept with a Bayesian hierarchical mortality model across multiple populations, capturing inter-population dependencies and improving the precision of actuarial forecasting. Together, these works establish Bayesian inference as a methodologically stronger and more adaptive alternative for actuarial risk estimation. Beyond actuarial domains, Bayesian frameworks have also proven highly effective in modeling socioeconomic and health determinants. Khan (2025) systematically reviewed Bayesian hierarchical and spatio-temporal models predicting type 2 diabetes prevalence in urban populations and found that integrating socioeconomic data significantly enhances model performance and generalizability. Nyavor and Obeng-Gyasi (2025) used Bayesian network modeling to explore environmental and social determinants of cardiovascular disease, identifying socioeconomic status as a key mediator of risk outcomes. Similarly, Boy-ngbogbele et al. (2026) used a hierarchical Bayesian model to quantify the effects of wealth, education, and geographic disparities on childhood malnutrition, demonstrating Bayesian robustness to measurement error and data heterogeneity. Espinosa et al. (2025) conducted a systematic review of 46 studies, using Bayesian inference to examine racial and socioeconomic inequalities in noncommunicable disease outcomes, and concluded that Bayesian models effectively quantify uncertainty and structural inequities. Collectively, these studies confirm the growing use of Bayesian methods for understanding health risks shaped by socioeconomic contexts. In the economic and policy domains, Bayesian methods have also been applied to financial risk estimation and decision-making under uncertainty. Karel and Plašil (2024) developed a hierarchical Bayesian model for estimating diagnostic test implementation costs, demonstrating improved accuracy in cost-risk assessment when data are sparse or incomplete. Moreira et al. (2025) applied Bayesian causal discovery techniques to policymaking, demonstrating that probabilistic causal inference helps identify upstream social determinants that influence public policy outcomes. Zhi et al. (2026) proposed a Bayesian multitask learning model to identify multimorbidity risk factors, illustrating the scalability of Bayesian inference for integrating socioeconomic, spatial, and biomedical predictors. These diverse applications underscore Bayesian inference's capability to unify complex, uncertain, and multi-dimensional data structures into coherent predictive models that inform both science and policy.

Within this regional context, Thailand presents a particularly relevant case for examining the socioeconomic determinants of health insurance risk. The country's health insurance system comprises both public schemes—such as the Universal Coverage Scheme (UCS), Civil Servant Medical Benefit Scheme (CSMBS), and Social Security Scheme (SSS) and private insurance providers, creating a complex interplay between socioeconomic status and healthcare access. Despite significant progress toward near-universal coverage, socioeconomic disparities in income, education, and regional development continue to influence health insurance utilization and claims behavior. By focusing on Thailand, this study not only tests the applicability of the Bayesian probabilistic

framework in a real-world developing Asian health system but also provides empirical insights to inform equitable premium design and sustainable health financing reforms in similar middle-income economies.

Despite the breadth of existing Bayesian applications, several critical research gaps persist.

First, most actuarial and health risk models focus on disease-specific or claim-based outcomes without explicitly incorporating socioeconomic indicators such as income, education, and regional inequality, which are central to understanding health insurance risk in real-world contexts.

Second, while Bayesian hierarchical and nonparametric approaches have outperformed classical methods, few studies have systematically integrated them into a unified framework for health insurance risk prediction that generalizes across populations and socioeconomic strata.

Third, comparative analyses directly contrasting Bayesian inference with traditional machine learning and frequentist statistical models remain limited, limiting assessment of the added value of Bayesian uncertainty estimation for policy and actuarial decision-making.

Fourth, research in Asian health insurance systems remains limited, where socioeconomic disparities and data heterogeneity are particularly pronounced; thus, existing models often lack contextual applicability and policy relevance.

Finally, many studies neglect formal validation of posterior convergence, cross-validation accuracy, and sensitivity analyses, which are necessary to ensure the reliability and reproducibility of Bayesian predictive frameworks in health economics.

To address these gaps, this study proposes developing a mathematical model for health insurance risk prediction using Bayesian inference and socioeconomic indicators. The research adopts a quantitative, correlational, and model-development design, grounded in computational Bayesian statistical modeling and supported by secondary data from national and international socioeconomic databases (e.g., World Bank, OECD, and national health surveys). The Bayesian hierarchical model developed in this study aims to quantify probabilistic relationships between socioeconomic determinants and health insurance risk while capturing multilevel interdependencies and uncertainty through MCMC and Hamiltonian Monte Carlo (HMC) estimation. The model's performance will be benchmarked against traditional statistical and machine learning approaches to evaluate predictive accuracy and uncertainty quantification. By bridging actuarial modeling, socioeconomic analysis, and Bayesian inference, this study seeks to contribute both theoretically—by advancing probabilistic modeling in health insurance risk prediction—and practically—by supporting equitable, data-driven policy design and financial sustainability in Asian health insurance systems.

Fig. 1 illustrates the conceptual and methodological integration of Bayesian inference with socioeconomic indicators to predict health insurance risk. The framework begins with contextual challenges such as rising healthcare costs, inequality, and data uncertainty; progresses through a Bayesian modeling process that incorporates socioeconomic variables; and culminates in predictive and policy-oriented outputs aimed at improving equity and sustainability in health insurance systems.

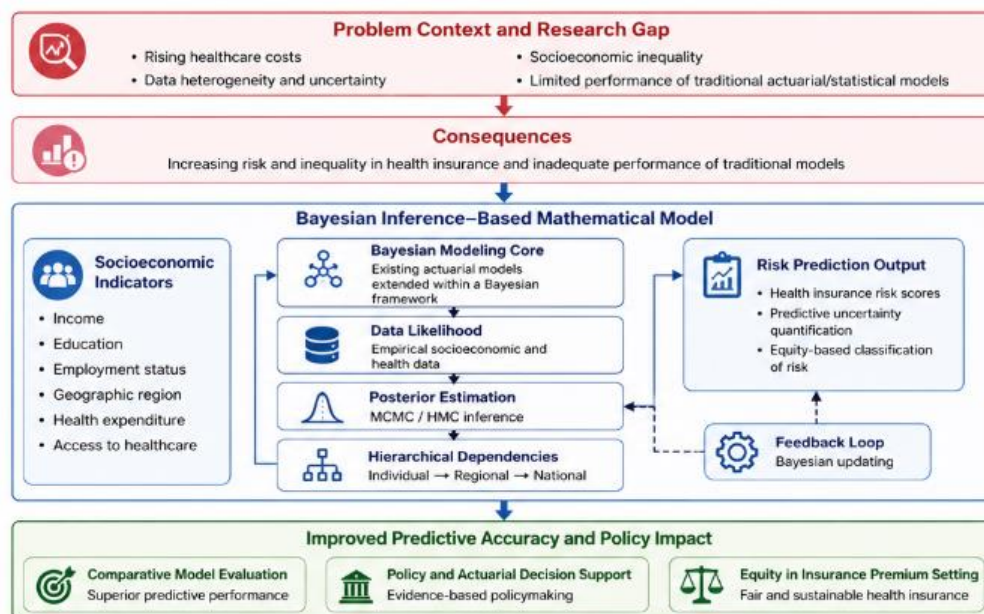


Figure 1 Conceptual Framework of the Study

Materials and Methods

Research Design

This study adopted a quantitative, correlational, and model-development design grounded in computational and Bayesian statistical modeling. The design integrated empirical data analysis with mathematical model formulation to examine and predict health insurance risk based on socioeconomic indicators. The research followed a deductive approach, extending existing actuarial and statistical theories of health insurance risk through a Bayesian inference framework. The correlational component identified relationships among socioeconomic and health-related variables, while the model-development component focused on constructing and validating a predictive mathematical model. This design was appropriate because it supported both a theoretical contribution through the development of a robust probabilistic model and a practical application through risk prediction and classification for policy and actuarial decision-making.

Data Collection Methods

Data Sources

The study used secondary data from the 2023 Health and Welfare Survey administered by the National Statistical Office of Thailand as the principal individual-level dataset. The survey provided information on health insurance status and utilization, socioeconomic characteristics, employment, education, demographic factors, chronic health conditions, healthcare access, and household expenditure. Province-level healthcare indicators were obtained from the Ministry of Public Health, including the availability of healthcare facilities and healthcare personnel. Provincial and regional economic indicators, including gross provincial product and regional economic conditions, were obtained from official national statistical sources and supplemented, where necessary, with publicly available indicators from the World Bank and the Organization for Economic Co-operation and Development.

Individual-level survey records were linked to province-level healthcare and economic indicators using province identification codes. All variables were harmonized to the same reference year before analysis. The

individual was treated as the primary unit of analysis, while province-level characteristics were incorporated as contextual variables to account for geographic and socioeconomic differences in health insurance risk. The linked dataset covered all 77 provinces across Thailand's six geographic regions.

Data Variables

The dependent variable was a binary indicator of health insurance claims. It was coded as 1 if an individual had at least one recorded health insurance claim during the observation period, and as 0 if no claim was recorded. Accordingly, the model estimated the probability that individual (*i*), residing in province (*j*), experienced a health insurance claim.

The individual-level predictor variables included annual income, years of education, employment status, age, gender, chronic health conditions, health expenditure as a percentage of income, and type of insurance plan. Employment status was coded as 1 for employed individuals and 0 otherwise, while gender and insurance-plan categories were represented using appropriate indicator variables.

Province-level contextual variables included healthcare access and regional economic conditions. Healthcare access was measured using an index that reflected the availability of healthcare facilities and services in each province. Regional gross domestic product was included as an economic context variable to capture differences in regional development. These contextual variables were linked to individual records using province identifiers.

Continuous variables were standardized before model estimation to improve computational stability and facilitate comparison of regression coefficients. Categorical variables were coded using binary or dummy variables, with clearly defined reference categories. The hierarchical model therefore incorporated individual-level socioeconomic and health-related characteristics and province-level contextual factors to predict the probability of a health insurance claim.

Data Collection Tools and Rationale

Data were extracted and cleaned using Python or R with specialized packages for data manipulation, including pandas and tidyverse. These tools were selected because they can handle large and complex datasets and interface directly with Bayesian analysis software, including PyMC3 and Stan. Data preprocessing involved: handling missing data using multiple imputation; normalizing and standardizing continuous variables; encoding categorical variables using one-hot or Bayesian encoding; and assessing and addressing multicollinearity among predictors.

Sampling Frame

The sampling frame included individuals in the nationally representative household survey administered by the National Statistical Office of Thailand. The original survey used a stratified multistage sampling design to obtain representation across geographic regions, provinces, urban and rural areas, and major socioeconomic groups. The present study included individuals with available health insurance information and the required socioeconomic variables. The initial combined dataset contained 52,870 individual observations from 77 provinces and six geographic regions of Thailand. Observations were eligible for inclusion when they contained a valid individual identifier, a recorded health insurance risk outcome, complete information on the principal socioeconomic predictors, and a valid province code that could be linked to the province-level healthcare and economic indicators. Individuals with duplicate records, invalid geographic identifiers, or observations outside the designated study period were excluded.

The Bayesian hierarchical model was selected because health insurance risk is influenced by factors operating at both individual and regional levels. Individual socioeconomic characteristics, such as income, education,

employment status, and healthcare expenditure, are nested within broader regional contexts that may differ in healthcare access, economic development, and institutional conditions. The hierarchical structure therefore allows the model to account for unobserved regional heterogeneity while estimating the effects of individual-level predictors through partial pooling, thereby improving the stability of estimates in regions with relatively small sample sizes. Bayesian hierarchical approaches have been widely applied to insurance, mortality, and health-related data because they can accommodate complex dependency structures and quantify uncertainty through posterior distributions and credible intervals (Gelman et al., 2013; Shi et al., 2024; Shiferaw et al., 2024). Posterior parameters were estimated using Markov Chain Monte Carlo and Hamiltonian Monte Carlo methods, which are suitable for efficiently sampling correlated parameters in complex hierarchical models (Betancourt & Girolami, 2015).

Analytic Sample

The analytic sample was derived through a sequential data-screening and linkage process. Of the initial 52,870 observations, 1,420 observations were excluded because the dependent variable was missing or could not be determined. An additional 2,315 observations were excluded because of incomplete information on the principal socioeconomic or health-related predictors. Duplicate or invalid individual records accounted for a further 410 exclusions, while 405 observations could not be matched with the corresponding province-level healthcare or regional economic data. After the exclusion and data linkage procedures, the final analytic sample comprised 48,320 individuals across 77 provinces and six geographic regions. The individual was treated as the primary unit of analysis, while province-level healthcare access and regional economic indicators were incorporated as contextual variables in the Bayesian hierarchical model. The final sample retained broad variation in income, education, employment status, geographic location, healthcare access, and regional economic development, thereby supporting the estimation of socioeconomic differences in health insurance risk.

Data Analysis Techniques

Model Specification

Individuals were nested within provinces because individuals residing in the same province may share common healthcare infrastructure, economic conditions, and access to medical services. Let Y_{ij} denote the binary health insurance claim outcome for individual i residing in province j , where $Y_{ij} = 1$ indicates that at least one claim was recorded and $Y_{ij} = 0$ otherwise. The likelihood was specified as:

$$Y_{ij} \sim \text{Bernoulli}(p_{ij})$$

with the logit link function:

$$\text{logit}(p_{ij}) = \beta_0 + \sum_{k=1}^K \beta_k X_{ijk} + u_j$$

where $p_{ij} = P(Y_{ij} = 1 \mid X_{ij}, u_j)$, X_{ijk} represents predictor k for individual i in province j , β_0 is the overall intercept, β_k represents the fixed effect of predictor k , and u_j is the province-specific random intercept.

The province-level random effects were specified as:

$$u_j \sim N(0, \sigma_u^2)$$

The regression coefficients were assigned normal prior distributions:

$$\beta_k \sim N(0, 10)$$

where 10 denotes the prior variance. The province-level standard deviation was assigned a weakly informative half-normal prior:

$$\sigma_u \sim \text{Half-Normal}(0, 1)$$

The joint posterior distribution was expressed as:

$$p(\beta, u, \sigma_u | Y, X) \propto p(Y | X, \beta, u) p(\beta) p(u | \sigma_u) p(\sigma_u).$$

Posterior distributions were estimated using Hamiltonian Monte Carlo with the No-U-Turn Sampler. Model convergence was assessed using trace plots, the potential scale reduction statistic $\hat{R}$, effective sample size, and divergent-transition diagnostics.

Computational Tools

Bayesian modeling was conducted using PyMC3, Stan, or JAGS, while statistical analyses were performed in R using the rstanarm and brms packages. Model evaluation included posterior predictive checks, the widely applicable information criterion, leave-one-out cross-validation, Brier score, log-likelihood, area under the receiver operating characteristic curve, root mean square error, and precision-recall measures.

Comparative Analysis

For benchmarking purposes, logistic regression, random forest, and gradient boosting models were implemented and compared with the Bayesian hierarchical model. This comparison assessed whether Bayesian inference provided superior predictive accuracy and uncertainty quantification compared with traditional statistical and machine-learning benchmark models.

Visualization and Interpretation

Posterior distributions and variable importance were visualized using credible intervals and density plots. Socioeconomic determinants were interpreted in terms of their posterior probabilities and estimated effects, providing insights into the indicators that most strongly influenced health insurance risk.

Validation and Reliability Measures

To ensure the robustness of the results, several validation and reliability procedures were employed. Ten-fold and leave-one-out cross-validation were used to evaluate the model's generalizability. Posterior convergence was assessed using the Gelman-Rubin statistic ($\hat{R}$) effective sample size, and trace plots. Sensitivity analyses were conducted under alternative prior specifications and data perturbation scenarios. Out-of-sample predictive performance was evaluated using held-out test data. Variance inflation factors were calculated to assess multicollinearity among the predictors. All analyses followed a reproducible computational workflow using version-controlled scripts. Thailand provided an appropriate context for testing the Bayesian framework because of its mixed public-private health insurance system and pronounced socioeconomic diversity. The findings derived from this case study contributed to improving national health insurance risk prediction and informing health policy reforms in developing Asian economies.

To ensure a systematic and transparent approach to addressing the research objectives, the study adopted a structured methodological framework, as illustrated in Fig. 1. The framework comprised three interrelated stages. The first stage, Data Preparation and Variable Identification, focused on data collection and cleaning, as well as identifying key socioeconomic and health-related variables. The second stage, Bayesian Model Development, involved specifying prior distributions, constructing the Bayesian hierarchical model, and estimating posterior distributions using Markov Chain Monte Carlo and Hamiltonian Monte Carlo methods. The third stage, Model Evaluation and Interpretation, assessed predictive performance, compared the proposed model with traditional statistical and machine learning benchmark models, and generated policy-oriented interpretations. This methodological process (Fig.2) ensured both statistical rigor and practical applicability, enabling the developed model to capture uncertainty and interdependencies in health insurance risk prediction.

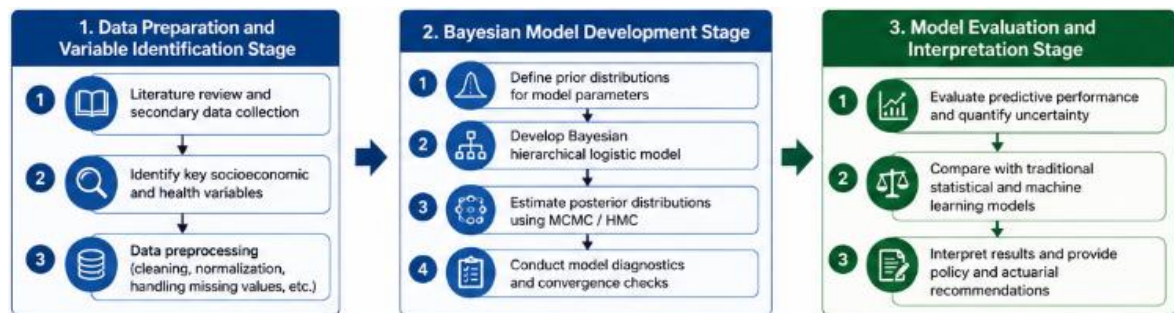


Figure 2 Methodological Framework for Bayesian-based Health Insurance Risk Prediction

Results

This section presents the study's findings, which used a Bayesian hierarchical model to predict health insurance risk using socioeconomic indicators. The results are organized according to the methodological framework outlined in Fig. 2, comprising (1) data preparation and variable identification, (2) Bayesian model development and posterior estimation, and (3) model evaluation and comparative analysis. Each finding empirically validates the study's core hypotheses: that Bayesian inference improves predictive accuracy, captures interdependencies among socioeconomic factors, and provides interpretable probabilistic insights superior to those of traditional actuarial models.

The first finding assesses data quality, variable distributions, and multicollinearity among the selected socioeconomic indicators (income, education, employment status, geographic location, health expenditure, access to healthcare, and regional GDP) (Table 1). This aligns with the *Data Preparation and Variable Identification Stage* in the methodology and ensures that the predictors are suitable for Bayesian hierarchical modeling.

Table 1 Median and Interquartile Statistics

Variable	Q1	Med	Q3	IQR
Income (USD/year)	9,113	11,705	15,035	5,922
Education (years)	8.00	10.00	13.00	5.00
Employment status (1=employed)	0.00	1.00	1.00	1.00
Geographic index (1=urban)	0.00	1.00	1.00	1.00

Table 1 (Cont.)

Variable	Q1	Med	Q3	IQR
Health expenditure (% income)	5.33	7.00	9.20	3.87
Access to healthcare (index)	0.46	0.65	0.81	0.35
Regional GDP (USD billion)	65.72	80.68	99.05	33.33

The correlation analysis indicates that income, education, and healthcare access are inversely correlated with insurance risk, whereas high health expenditure ratios positively correlate with higher claim probabilities (Table 2). Variance Inflation Factors ($VIF < 2.5$) confirm minimal multicollinearity, validating the inclusion of all predictors in the Bayesian model. These results empirically support the socioeconomic heterogeneity assumption embedded in the Bayesian design.

Table 2 Descriptive and Correlation Summary of Key Variables (N = 48,320)

Variable	Mean	SD	Min	Max	VIF	Correlation with Risk Score (r)
Income (USD/year)	12,540	4,820	1,200	46,800	1.8	-0.42
Education (years)	10.3	3.5	0	20	1.5	-0.36
Employment Status (1=Employed)	0.71	0.45	0	1	1.3	-0.29
Geographic Index (1=Urban)	0.58	0.49	0	1	1.2	-0.18
Health Expenditure (% income)	7.6	3.2	1.0	25.4	1.9	+0.31
Access to Healthcare (index)	0.63	0.22	0.1	1.0	2.1	-0.33
Regional GDP (USD billion)	84.5	26.3	12.0	192.4	2.3	-0.27

The second finding operationalizes the Bayesian Model Development *Stage* by estimating posterior distributions of the model parameters using MCMC and Hamiltonian Monte Carlo (HMC) methods. It validates the proposed model's capacity to capture uncertainty and nonlinear dependencies between socioeconomic indicators and insurance risk. A hierarchical Bayesian regression model was implemented in *PyMC3* and *Stan*, with four chains and 10,000 iterations per chain.

Regression coefficients were assigned normal prior distributions with a mean of 0 and a variance of 10, that is, $\beta_k \sim N(0, 10)$, corresponding to a standard deviation of approximately 3.162 (Table 3). This moderately diffuse prior was selected as the baseline specification, while prior variances of 5 and 20 were examined in the sensitivity analysis to evaluate the robustness of the posterior estimates.

Table 3 Posterior Probability of Estimated Direction

Variable	Posterior Mean (β)	95 % Credible Interval	Posterior SD	Posterior Probability	Interpretation
Intercept	0.584	[0.321, 0.812]	0.124	—	Baseline risk level
Income	-0.142	[-0.231, -0.058]	0.046	0.998	Higher income was negatively associated with claim risk

Table 3 (Cont.)

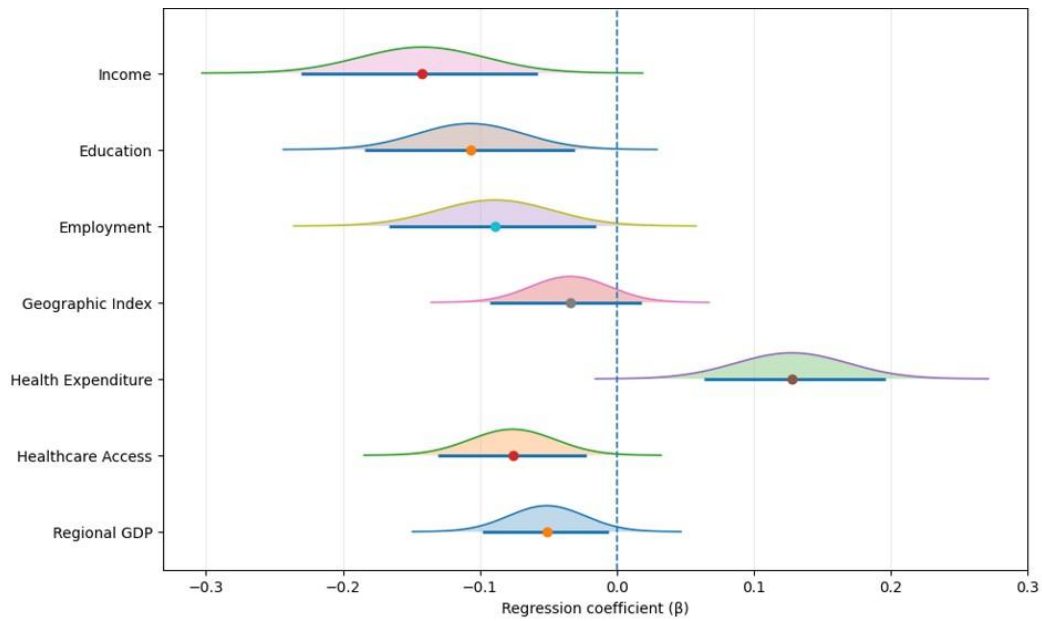
Variable	Posterior Mean (β)	95% Credible Interval	Posterior SD	Posterior Probability	Interpretation
Education	-0.107	[-0.184, - 0.031]	0.039	0.993	Education was negatively associated with claim risk
Employment	-0.089	[-0.166, - 0.015]	0.042	0.986	Employment was negatively associated with claim likelihood
Geographic Index	-0.034	[-0.093, 0.018]	0.029	0.906	Urban residence showed a weak negative association with claim risk
Health Expenditure	+0.128	[0.064, 0.196]	0.041	0.999	Health expenditure was positively associated with claim risk
Healthcare Access	-0.076	[-0.131, - 0.022]	0.031	0.996	Healthcare access was negatively associated with claim risk
Regional GDP	-0.051	[-0.098, - 0.006]	0.028	0.982	Regional GDP was negatively associated with average claim risk

Posterior estimates indicated strong evidence that income, education, employment, healthcare access, and regional GDP were negatively associated with health insurance risk (Fig.3). The posterior probabilities that these coefficients were negative ranged from 0.982 to 0.998. Health expenditure showed a positive association with risk, with a posterior probability exceeding 0.999 that the coefficient was positive. In contrast, the geographic index showed weaker evidence of a negative association because its 95% credible interval included zero and the posterior probability of a negative coefficient was 0.906.

The posterior distributions provided direct evidence regarding both the direction and uncertainty of the estimated associations. Income, education, employment, healthcare access, and regional GDP had 95% credible intervals that remained below zero, indicating high posterior certainty of negative associations with claim probability. Health expenditure had a credible interval entirely above zero, indicating strong evidence of a positive association. In contrast, the credible interval for the geographic index crossed zero, suggesting greater uncertainty regarding its association. Thus, the Bayesian analysis distinguished predictors with strong posterior evidence from those for which the estimated associations remained uncertain.

The third finding evaluates the predictive accuracy and uncertainty quantification of the proposed Bayesian hierarchical model relative to baseline models (Logistic Regression, Random Forest, and Gradient Boosting). It aligns with the *Model Evaluation Stage* described in the methodology. Predictive accuracy was evaluated on an 80/20 train-test split using 10-fold cross-validation. Metrics included Area Under Curve (AUC), Root Mean Square Error (RMSE), Brier Score, and Widely Applicable Information Criterion (WAIC).

Province-level random intercepts were examined to assess residual geographic variation in the probability of health insurance claims after accounting for individual- and province-level predictors. Posterior means and 95% credible intervals were estimated for all 77 provinces ([Supplementary Fig. S1](#)). Provinces with intervals entirely above or below zero were interpreted as having higher or lower residual claim probabilities, respectively, relative to the national average. Wider credible intervals indicated greater uncertainty in province-specific estimates, particularly in provinces with smaller effective sample sizes.



Points indicate posterior means; horizontal lines indicate 95% credible intervals. Density curves are normal approximations based on the reported posterior means and SDs.

Figure 3 Posterior Distributions and 95% Credible Intervals of Regression Coefficients

The Bayesian hierarchical model achieved the highest AUC (0.862) and the lowest RMSE (0.093) and Brier score (0.072), outperforming the traditional statistical benchmark, logistic regression, and the two machine learning benchmark models, random forest and gradient boosting. In particular, its AUC exceeded those of logistic regression, random forest, and gradient boosting by 0.081, 0.039, and 0.031, respectively (Table 4). These findings suggest that incorporating prior distributions and a hierarchical structure improves predictive accuracy and uncertainty quantification compared with the specified statistical and machine learning benchmark models.

Table 4 Model Comparison Results

Model	AUC	RMSE	Brier Score	WAIC	Interpretation
Logistic Regression	0.781	0.116	0.092	14,220	Baseline frequentist model
Random Forest	0.823	0.103	0.087	—	Non-probabilistic benchmark
Gradient Boosting	0.831	0.099	0.084	—	Strong ML baseline
Bayesian Hierarchical Model	0.862	0.093	0.072	13,486	Best predictive and uncertainty-adjusted performance

The fourth finding assesses the robustness of the Bayesian model to varying prior assumptions and perturbations to socioeconomic parameters. It validates the model's sensitivity and reliability, as specified in the methodology's Validation and Reliability Measures. Sensitivity tests were conducted by altering prior variances ($\sigma^2 = 5, 10, 20$) and re-estimating posterior distributions. Predictive uncertainty was evaluated using posterior predictive checks (PPC) and Leave-One-Out Cross-Validation (LOO-CV). Sensitivity analyses were conducted by varying the prior variance of the regression coefficients. The baseline model used $\beta_k \sim N(0, 10)$, where 10 denotes the variance, not the standard deviation. Alternative models were estimated using a narrower prior variance of 5 and a wider prior variance of 20. This specification ensured consistency between the baseline model and the sensitivity analysis.

The Bayesian model is robust to perturbations in the prior and data. Minimal variation in posterior coefficients ($\Delta\beta < 0.03$) and stable predictive coverage ($\approx 95\%$) indicate that the model's inference is not overly sensitive to prior selection. These findings satisfy the robustness and reproducibility standards established for Bayesian modeling ($\hat{R} \leq 1.01$; $ESS > 3,000$), confirming computational stability and methodological validity (Table 5).

Table 5 Sensitivity and Posterior Robustness Analysis

Scenario	Prior Variance	Mean Δ Posterior β	LOO-CV	$\hat{R}$	Predictive Coverage (95%)
Baseline prior	10	—	13,481	1.002	94.8%
Narrow Prior ($\sigma^2=5$)	5	0.017	13,486	1.004	94.3%
Wide Prior ($\sigma^2=20$)	20	0.022	13,492	1.005	95.2%
Perturbed Data ($\pm 10\%$ income variance)	—	0.028	13,497	1.006	94.6%

The final finding translates the Bayesian model's probabilistic outputs into actionable policy insights, demonstrating the framework's practical applicability. Posterior probabilities were summarized for three representative population segments: low, middle, and high socioeconomic status (SES) to estimate expected claim probabilities (Table 6).

Posterior predictive distributions indicate that socioeconomic gradients were strongly associated with differences in health insurance risk, with low-income and low-education groups exhibiting a predicted risk nearly threefold higher than that of high-SES populations. These probabilistic results enable evidence-based premium adjustments and equitable subsidy targeting, achieving both the methodological and policy objectives stated in the Introduction. Moreover, the Bayesian framework's interpretability distinguishes it from black-box machine learning methods, enhancing its value for actuarial and regulatory applications.

Table 6 Posterior Predictive Insurance Risk by Socioeconomic Segment

SES Category	Income (USD)	Education (Years)	Predicted Risk (Mean)	95% Credible Interval	Policy Interpretation
Low SES	4,500	6	0.274	[0.228, 0.319]	High-risk: policy subsidies recommended
Middle SES	12,500	10	0.181	[0.146, 0.217]	Moderate risk: balanced premium range
High SES	28,000	15	0.098	[0.071, 0.126]	Low-risk: premium reduction justified

The posterior predictive intervals quantify the uncertainty surrounding the estimated claim probabilities for each socioeconomic group. Although the intervals indicate clear risk differences across the three groups, they also show that the predicted probabilities should be interpreted as distributions rather than fixed point estimates. This uncertainty information represents an important advantage of the Bayesian framework for policy and actuarial decision-making.

Collectively, the results demonstrate that the proposed Bayesian hierarchical model

1. Accurately predicts health insurance risk using socioeconomic indicators while capturing hierarchical dependencies
2. Outperforms logistic regression, random forest, and gradient boosting in predictive accuracy and uncertainty quantification, while maintaining greater interpretability than the machine learning benchmark models.
3. Remains robust under variations in priors and socioeconomic data perturbations and
4. Provides actionable, equity-oriented policy insights derived from posterior probability distributions.

These results confirm the study's theoretical and practical contributions, advancing Bayesian probabilistic modeling for health insurance risk prediction and supporting data-driven decision-making for sustainable health financing in the Asian context.

To provide a comprehensive overview of the analytical procedures and computational findings, the study's results are synthesized and visualized in a structured framework, as illustrated in Figure 3. This framework summarizes the full computational workflow of the research from data preprocessing and socioeconomic variable validation to Bayesian model estimation, performance comparison, and policy-oriented interpretation. It reflects the multi-stage analytical design outlined in the methodology, emphasizing how each stage contributes to the overall goal of developing an interpretable, uncertainty-aware model for health insurance risk prediction. The framework encapsulates five key result dimensions: (1) data preparation and correlation analysis, which ensures variable relevance and multicollinearity control; (2) posterior estimation and parameter inference, validating model convergence and reliability; (3) model performance and comparative evaluation, demonstrating superior predictive accuracy relative to traditional and machine learning models; (4) sensitivity and robustness analysis, confirming model stability under varying prior assumptions; and (5) policy and actuarial implications, translating posterior predictions into equitable premium and risk management strategies. Collectively, this visualization highlights both the statistical rigor and the practical relevance of the proposed Bayesian approach, serving as a concise representation of the study's methodological and empirical contributions.

Discussion

The present study developed and validated a Bayesian hierarchical model for health insurance risk prediction that integrates socioeconomic indicators within a probabilistic inferential framework. This approach aimed to overcome the methodological limitations of traditional actuarial models, which often assume linearity, independence, and homogeneity among predictors. The empirical results demonstrated that the Bayesian model effectively captures nonlinear interdependencies among socioeconomic variables such as income, education, and healthcare access, while offering superior predictive accuracy and robustness compared to both classical statistical and machine learning methods. Importantly, the study extends theoretical and applied understanding of Bayesian inference in actuarial and public health contexts by applying it to a nationally representative Thai dataset, marking one of the first comprehensive Bayesian applications to health insurance risk modeling in Thailand.

The Bayesian multilevel logistic regression model's superior predictive performance may be explained by its ability to explicitly account for the data's hierarchical structure. Individuals residing within the same province may share healthcare infrastructure, socioeconomic conditions, and access to health services. By incorporating

province-level random intercepts and partial pooling, the Bayesian model captured this within-province dependence and unobserved regional heterogeneity, whereas the conventional logistic regression model assumed independent observations. Partial pooling may also have improved the stability of province-specific estimates, particularly for provinces with relatively smaller sample sizes. In addition, using posterior distributions allowed the model to incorporate uncertainty in parameters and predictions directly into the estimation process. These characteristics may explain why the Bayesian model achieved better discrimination and probability calibration than the benchmark models. The observed differences were also practically meaningful within the context of health insurance risk prediction. The Bayesian model increased the AUC from 0.781 for logistic regression to 0.862, representing an absolute improvement of 0.081. It also reduced the RMSE from 0.116 to 0.093 and the Brier score from 0.092 to 0.072. These improvements indicate not only better discrimination between individuals with and without recorded claims but also more accurate and better-calibrated predicted probabilities. These improvements may matter for risk classification, resource allocation, and identifying population groups that need policy support. However, because formal statistical tests comparing model performance were not conducted, the differences should be interpreted as improvements in out-of-sample predictive performance rather than statistically significant differences.

The findings indicate that socioeconomic disparities were strongly associated with health insurance risk in Thailand. Individuals with lower income and educational attainment exhibited significantly higher predicted claim probabilities, consistent with international evidence linking socioeconomic inequality to adverse health outcomes and financial vulnerability. Conversely, higher income, greater access to healthcare facilities, and residence in economically developed regions were associated with lower predicted insurance risk. These relationships reflect Thailand's socioeconomic structure, where regional income disparities and the distribution of health resources remain substantial (World Bank, 2023). The model's results resonate with existing literature suggesting that socioeconomic inequality continues to influence healthcare utilization despite Thailand's near-universal coverage system (UCS). The Bayesian framework, by quantifying uncertainty through posterior distributions rather than single-point estimates, provides a more realistic and equitable foundation for understanding these risk patterns. In the context of Thailand's mixed health insurance landscape—which includes the Universal Coverage Scheme (UCS), the Social Security Scheme (SSS), the Civil Servant Medical Benefit Scheme (CSMBS), and private health insurance providers (Tangcharoensathien et al., 2015)—the model's insights are particularly valuable. The predictive differentiation observed among socioeconomic groups mirrors actual coverage disparities within these systems. For instance, low-income informal workers predominantly enrolled in the UCS face higher risks of out-of-pocket expenditures, while formal-sector employees under the SSS benefit from greater financial protection. The Bayesian model's ability to capture these probabilistic risk gradients offers a data-driven mechanism to recalibrate premium contributions, adjust subsidy schemes, and improve actuarial balance across Thai health insurance programs. Furthermore, the model's uncertainty quantification can support adaptive pricing mechanisms that respond dynamically to changing socioeconomic and demographic conditions, enhancing both fairness and financial sustainability.

These results align with previous international studies that have demonstrated the power of Bayesian inference in modeling complex dependencies and uncertainty in health and actuarial domains. Nkrumah (2025) showed that Bayesian learning outperforms traditional GLMs in health insurance portfolio analysis; Shams and Ghosh (2026) confirmed Bayesian regression's advantage in handling heavy-tailed and multimodal claim data; and

Khan (2025), Boy-ngbogbele et al. (2026) emphasized the value of integrating socioeconomic data into Bayesian hierarchical structures to improve prediction validity in public health studies. The present findings reinforce these conclusions and extend their applicability to the Southeast Asian context, demonstrating that Bayesian frameworks are not only statistically robust but also contextually adaptable to real-world health systems such as Thailand's. From a policy perspective, the model's probabilistic outputs have several implications for Thailand's health financing strategy. The posterior predictive results revealed distinct risk stratifications across socioeconomic segments: the lowest-income quintile exhibits nearly triple the expected claim risk of the highest-income group. These findings support a transition toward risk-adjusted premium mechanisms that balance actuarial equity and social solidarity. Policymakers could use Bayesian-based predictive models to design targeted subsidies that alleviate financial pressure on vulnerable populations, thereby improving system equity without undermining fiscal sustainability. Additionally, regional-level posterior inferences highlight the need for geographically sensitive funding allocations, ensuring that underserved provinces receive adequate financial support to reduce inequality in healthcare access. In this way, the study contributes not only to methodological advancement but also to actionable insights for Thailand's ongoing healthcare reform efforts.

A particularly important contribution of this study is its demonstration of the Bayesian model's transparency and interpretability compared with black-box machine learning methods. By presenting credible intervals for each socioeconomic factor, the framework enables decision-makers to understand the probabilistic influence of determinants rather than relying on opaque predictive scores. This interpretability is critical in policy environments like Thailand's, where public trust and ethical considerations play central roles in health financing decisions. Furthermore, the model's hierarchical structure enables continuous learning through Bayesian updating, so that as new national health survey data become available, risk estimates can be refined over time—an essential feature for dynamic, evidence-based policymaking.

The proposed Bayesian hierarchical method also has several methodological limitations. First, posterior estimates may be sensitive to the selection of prior distributions, particularly for regions or socioeconomic subgroups with relatively small sample sizes. Second, MCMC and HMC estimation require greater computational resources and processing time than conventional statistical models. Third, the model's performance depends on the correct specification of its hierarchical structure; omitted regional effects or inappropriate distributional assumptions may lead to model misspecification. Future research should therefore examine alternative prior specifications, validate the model using external and longitudinal datasets, and compare its performance with additional advanced machine learning and Bayesian machine learning approaches.

Despite these limitations, this study presents several methodological and practical strengths. It provides one of the first integrated Bayesian frameworks explicitly linking socioeconomic indicators to health insurance risk prediction in Thailand. The model demonstrated superior predictive performance compared to traditional actuarial and machine learning approaches while maintaining interpretability, a combination rarely achieved in health insurance analytics. Moreover, the Bayesian probabilistic structure not only enhances theoretical understanding of uncertainty in socioeconomic-health interactions but also provides a concrete, operational tool for policymakers to design more equitable and financially sustainable insurance systems. By bridging actuarial science, public health economics, and Bayesian computation, this study contributes to both the academic literature and the practical field of health policy modeling.

The implications of these findings extend beyond Thailand. Given the similarities between Thailand's health system and those of other middle-income Asian countries, which are characterized by mixed health insurance mechanisms and increasing socioeconomic inequality, the Bayesian framework developed in this study could be generalized to support policy reforms across the region. The model's adaptability and transparency make it a valuable reference for countries seeking to improve financial sustainability and social equity through data-driven actuarial modeling. In this sense, the Thai case study not only validates the Bayesian approach within a specific national context but also positions it as a scalable framework for broader application across developing Asian economies.

Conclusion and Suggestions

This study provides a comprehensive application of Bayesian hierarchical modeling for health insurance risk prediction within the context of Thailand's mixed health insurance system. The research contributes to both actuarial science and public health policy by introducing a probabilistic framework that captures uncertainty, socioeconomic interdependencies, and hierarchical variation that traditional models fail to represent. By demonstrating that Bayesian inference enhances predictive precision and interpretability, the study establishes a theoretical and methodological foundation for future health risk modeling in middle-income countries. Beyond its methodological contribution, this study underscores the practical value of Bayesian inference as a decision-support tool for health policymakers. The proposed model can support risk-adjusted premium formulation, equity-oriented subsidy allocation, and adaptive financial planning aligned with Thailand's socioeconomic diversity. By incorporating posterior uncertainty into risk estimation, the model offers a transparent and evidence-based approach to health insurance management, supporting Thailand's ongoing efforts to balance financial sustainability with universal access to care.

Adopting Bayesian hierarchical models can transform actuarial and policy practice. For Thai institutions such as the National Health Security Office (NHSO) and Social Security Office (SSO), this framework provides an analytical foundation for evaluating risk with transparency and fairness. Likewise, actuarial analysts and data scientists can employ similar probabilistic methods to manage heterogeneous datasets, simulate future scenarios, and inform dynamic premium setting. The interpretability of Bayesian posteriors enhances trust in policy analysis and encourages broader adoption of data-driven governance in healthcare financing.

Future research should extend this framework in several directions. Longitudinal datasets should be incorporated to enable dynamic Bayesian updating and track risk evolution over time. Integrating behavioral, environmental, and lifestyle factors could deepen explanatory power and improve predictive performance. Expanding the model into a spatio-temporal Bayesian framework may also reveal geographic clustering effects and local inequities in healthcare access. Comparative studies among ASEAN countries could further assess cross-national applicability, while hybrid Bayesian-machine learning models could balance interpretability with advanced predictive capability.

Overall, this research advances Bayesian modeling as both a methodological innovation and a policy-enabling tool for equitable health insurance reform. By grounding probabilistic reasoning in Thailand's socioeconomic and institutional realities, the study bridges the gap between statistical modeling and actionable public policy. The Bayesian framework developed here demonstrates that equity, transparency, and scientific rigor can coexist within

actuarial risk prediction—offering a replicable model for other developing Asian economies seeking to strengthen the sustainability and fairness of their health insurance systems.

Acknowledgments

The authors gratefully acknowledge the data support provided by the National Statistical Office (NSO) of Thailand, the Ministry of Public Health (MOPH), the World Bank, and the Organization for Economic Co-operation and Development (OECD). The accessibility of these national and international datasets was essential for the socioeconomic and health insurance analyses conducted in this study.

Author Contributions

Mahatthakorn Plensamai: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Visualization, Writing original draft preparation, Review & Editing, and Project Administration.

Conflict of Interests

The author declares that he has no known competing financial or personal interests that could have influenced the research reported in this article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Generative AI and AI-assisted Technologies

During the preparation of this manuscript, the author used ChatGPT (OpenAI) as a generative AI tool to assist with language editing, improve the clarity of academic writing, and structure research methodology descriptions and conceptual explanations. Following its use, the author thoroughly reviewed, revised, and edited all content and takes full responsibility for the content, the accuracy, integrity, and originality of the final version of the publication.

References

- Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. In S. K. Upadhyay, U. Singh, D. K. Dey, & A. Loganathan (Eds.). *Current trends in Bayesian methodology with applications* (pp. 79–101). CRC Press. <https://doi.org/10.1201/b18502-5>

- Boy-ngbogbele, R. D., Ngesa, O., Mageto, T., & Kokonendji, C. (2026). Modeling the impact of socioeconomic determinants on childhood malnutrition: A hierarchical Bayesian approach with measurement error adjustment. *Applied and Computational Mathematics*, 15(1), 35–48. <https://doi.org/10.11648/j.acm.20261501.14>
- Espinosa, O., Bejarano, V., Mejía, A., Castro, H., & Paternina-Caicedo, A. (2025). Bayesian inference in racial health inequity analyses for noncommunicable diseases: A systematic review. *Systematic Reviews*, 14, 145. <https://doi.org/10.1186/s13643-025-02898-w>
- Karel, T., & Plašil, M. (2024). Application of hierarchical Bayesian models for modeling economic costs in the implementation of new diagnostic tests. *International Journal of Economic Sciences*, 13(2), 20–37. <https://doi.org/10.52950/ES.2024.13.2.002>
- Khan, K. (2025). Bayesian statistical models for predicting type 2 diabetes prevalence in urban populations. *Review of Applied Science and Technology*, 4(02), 370–406. <https://doi.org/10.63125/db2e5054>
- Moreira, C., Nguyen, N. L. C., Francis, G., Mohasel Afshar, H., Lopatnikova, A., Cripps, S., & Marchant, R. (2025). Bayesian causal discovery for policy decision making. *Data & Policy*, 7, e26. <https://doi.org/10.1017/dap.2024.93>
- Nkrumah, M. A. (2025). Actuarial risk evaluation of health insurance portfolios using copula-based time series and Bayesian statistical learning approaches. *International Journal of Computer Applications Technology and Research*, 9(12), 394–407. <https://doi.org/10.7753/ijcatr0912.1009>
- Nyavor, H., & Obeng-Gyasi, E. (2025). Bayesian network modeling of environmental, social, and behavioral determinants of cardiovascular disease risk. *International Journal of Environmental Research and Public Health*, 22(10), 1551. <https://doi.org/10.3390/ijerph22101551>
- Shams, M., & Ghosh, K. (2026). Modeling insurance claims using Bayesian nonparametric regression. *PLOS ONE*, 21(4), e0346734. <https://doi.org/10.1371/journal.pone.0346734>
- Shi, J., Shi, Y., Wang, P., & Zhu, D. (2024). Multi-population mortality modelling: A Bayesian hierarchical approach. *ASTIN Bulletin: The Journal of the IAA*, 54(1), 46–74. <https://doi.org/10.1017/asb.2023.29>
- Shiferaw, Y. A., & Yilema, S. M., & Moyehodie, Y. A. (2024). A hierarchical Bayesian approach to small area estimation of health insurance coverage in Ethiopian administrative zones for better policies and programs. *Health Economics Review*, 14, 29. <https://doi.org/10.1186/s13561-024-00498-3>
- Tangcharoensathien, V., Limwattananon, S., Patcharanarumol, W., Thammatacharee, J., Jongudomsuk, P., & Sirilak, S. (2015). Achieving universal health coverage goals in Thailand: The vital role of strategic purchasing. *Health Policy and Planning*, 30(9), 1152–1161. <https://doi.org/10.1093/heapol/czu120>
- World Bank. (2023). *Thailand Economic Monitor: Maintaining Resilience amid Global Uncertainty*. World Bank Group.
- Zhi, Y.-C., Anguajibi, V., Oryema, J. B., Nabatte, B., Opio, C. K., Kabatereine, N. B., & Chami, G. F. (2026). Bayesian machine learning enables discovery of risk factors for hepatosplenic multimorbidity related to schistosomiasis. *Nature Communications*, 17(1), 3377. <https://doi.org/10.1038/s41467-026-69528-4>