# Interpretable Machine Learning to Predict Catastrophic Health Expenditure Risk in China: Evidence from Nationally Representative Survey Data

Danyang Wei, Min Hu*
Fudan University, Shanghai, China

MSR138

## OBJECTIVE

- The incidence of catastrophic health expenditure (CHE) among Chinese households remains high by global standards, revealing gaps in financial risk protection.
- Prior studies have largely relied on conventional statistical approaches to examine correlates of CHE.
- This study aimed to develop a machine learning–based prediction model for household CHE risk in China and to identify key predictive factors, moving beyond traditional regression to capture non-linear interactions and improve prediction accuracy.

## METHODS

- **Data:** 2022 China Family Panel Studies; 8,000 households after data cleaning.
- **Outcome:** CHE, defined as medical spending ≥40% of non-food expenditure; 25% and 10% thresholds were used for robustness checks.
- **Predictors:** selected using the Andersen behavioral model and grouped as predisposing, enabling, and need factors.
- **Model development and evaluation：** Decision tree, random forest, and XGBoost classifiers were developed. Data were split into training and testing sets at an 80:20 ratio. Five-fold cross-validation and class-imbalance handling were applied. Model performance was assessed using AUROC, accuracy, precision, recall, and F1 score.
- **Interpretation method:** SHAP was applied to the best-performing model to identify key predictors of CHE risk.

## RESULTS

**CHE incidence：**

- 9.69% at the 40% threshold
- 17.39% the 25% threshold
- 38.01% the 10% threshold

**Subgroup analyses（Tab 1）：**

- Age: ≥75y (32.7%) vs Young (3.7%) , p<0.001
- Income: Lowest (20.4%) vs Highest (3.5%) , p<0.001
- Hospitalization: Yes (30.5%) vs No (7.2%) , p<0.001

Table 1. Catastrophic Health Expenditure by Household and Individual Characteristics

| Characteristic | N | CHE cases | CHE rate, % (95% CI) | p value | Characteristic | N | CHE cases | CHE rate, % (95% CI) | p value |
|---|---|---|---|---|---|---|---|---|---|
| Age | | | | < 0.001 | Per capita household income/year | | | | < 0.001 |
| 16–44 | 3426 | 127 | 3.71 (3.10–4.39) | | 0–6,500 | 754 | 154 | 20.42 (17.60–23.48) | |
| 45–59 | 2748 | 240 | 8.73 (7.70–9.85) | | 6,501–20,000 | 2637 | 290 | 11.00 (9.83–12.25) | |
| 60–74 | 1514 | 306 | 20.21 (18.21–22.32) | | 20,001–30,000 | 1372 | 126 | 9.18 (7.71–10.84) | |
| ≥75 | 312 | 102 | 32.69 (27.51–38.20) | | 30,001–50,000 | 1585 | 128 | 8.08 (6.78–9.53) | |
| Residence | | | | < 0.001 | 50,001–100,000 | 1154 | 58 | 5.03 (3.84–6.45) | |
| Rural | 3551 | 392 | 11.04 (10.03–12.12) | | 100,001–200,000 | 348 | 12 | 3.45 (1.79–5.95) | |
| Urban | 4449 | 383 | 8.61 (7.80–9.47) | | 200,001–300,000 | 78 | 3 | 3.85 (0.80–10.83) | |
| Education | | | | < 0.001 | >300,000 | 72 | 4 | 5.56 (1.53–13.62) | |
| Primary school or below | 2364 | 372 | 15.74 (14.29–17.27) | | Health status | | | | < 0.001 |
| Junior/senior high school | 3956 | 336 | 8.49 (7.64–9.41) | | Healthy | 6190 | 409 | 6.61 (6.00–7.26) | |
| College or bachelor's degree | 1587 | 64 | 4.03 (3.12–5.12) | | Fair | 692 | 98 | 14.16 (11.65–16.98) | |
| Master's degree or above | 93 | 3 | 3.23 (0.67–9.14) | | Unhealthy | 1118 | 268 | 23.97 (21.50–26.59) | |
| Marital status | | | | < 0.001 | Health insurance | | | | < 0.001 |
| Never married | 764 | 27 | 3.53 (2.34–5.10) | | No protection | 530 | 49 | 9.25 (6.92–12.04) | |
| Married/cohabiting | 6479 | 627 | 9.68 (8.97–10.42) | | Low protection | 5531 | 576 | 10.41 (9.62–11.25) | |
| Divorced or widowed | 757 | 121 | 15.98 (13.44–18.79) | | High protection | 1939 | 150 | 7.74 (6.59–9.02) | |
| Housesize | | | | < 0.001 | Hospitalization | | | | |
| Small | 4580 | 557 | 12.16 (11.23–13.14) | | No | 7131 | 510 | 7.15 (6.56–7.77) | |
| Medium | 2959 | 192 | 6.49 (5.63–7.44) | | Yes | 869 | 265 | 30.49 (27.45–33.68) | |
| Large | 461 | 26 | 5.64 (3.72–8.15) | | | | | | |

Note: CHE = catastrophic health expenditure; CI = confidence interval. Rates are presented as percentages. Income is measured in CNY.

- **Best model**：XGBoost achieved the best discrimination (AUROC=0.806), outperforming RF(0.797) and DT(0.772)(Figure 1).
- **Interpretation：** SHAP ranked hospitalization (0.381), age(0.301), household composition(0.277), self-reported health(0.223), and income (0.178) as top contributors. Category-level SHAP suggested increased risk associated with low income, advanced age, poor health, hospitalized and having older household members. Findings were robust across alternative thresholds (Fig. 2).

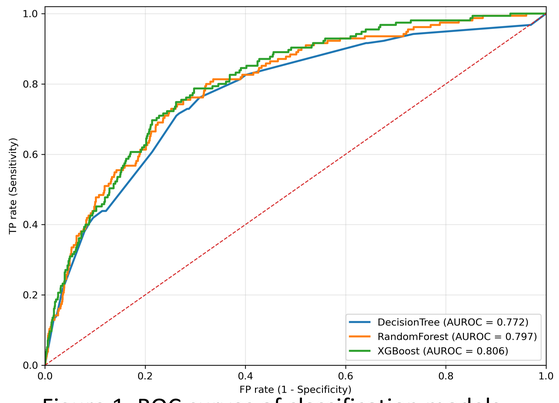


Figure 1. ROC curves of classification models

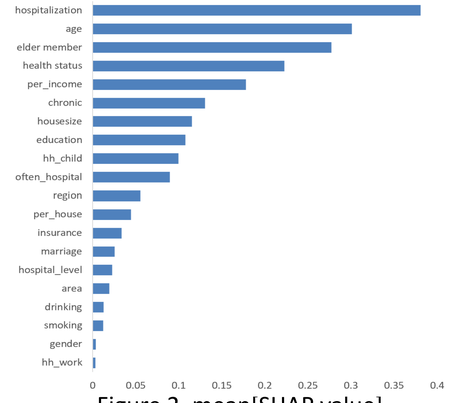


Figure 2. mean[SHAP value]

## CONCLUSION

- **The Burden:** CHE remains a severe financial threat in China, affecting 1 in 10 households, with significantly higher risks concentrated among the elderly, hospitalized, and low-income populations.
- **The Predictors:** Interpretable machine learning identified hospitalization, age, household composition, and income as the top predictors of CHE risk, outperforming traditional statistical correlations by capturing complex interactions.
- **The Pathway Forward:**
  **--Precision Financial Protection**: Moving beyond uniform benefit packages, our findings support a shift toward risk-based, proactive interventions:
  **--Early Warning System**: Integrate CHE risk scores (based on age, hospitalization history, and income) into health insurance databases to flag high-risk households.
  **--Targeted Assistance**: Prioritize medical financial assistance and cap co-payments for households identified as high-risk by the model, especially those with hospitalized elderly members.
  **--Policy Responsiveness:** Adjust reimbursement schemes dynamically for specific vulnerable profiles rather than applying blanket increases.

## CONTACT INFORMATION

- *Correspondence: Min Hu, PhD, Professor
- E-mail: humin@fudan.edu.cn
- Founding：The National Natural Science Foundation of China (No. 72474053)

Presented ISPOR 2026 | Philadelphia, PA, US | May 17-20, 2026