ECONOMIC STATISTICS CENTRE OF EXCELLENCE

A COLLABORATION WITH Office for National Statistics

Machine Learning for Estimating Catastrophic Health Spending in Disaster-Affected, Data-Scarce Settings

Rozana Himaz, Dimitra Salmanidou, Saman Ghaffarian

ESCoE Discussion Paper 2026-05

March 2026

ISSN 2515-4664

DISCUSSION PAPER

Machine Learning for Estimating Catastrophic Health Spending in Disaster-Affected, Data-Scarce Settings
Rozana Himaz, Dimitra Salmanidou, Saman Ghaffarian
ESCoE Discussion Paper No. 2026-05
March 2026

Abstract

Natural hazard events can increase out-of-pocket health costs and push vulnerable households into poverty. Mitigation measures require understanding changes in health spending patterns using pre- and post-event data, but such data are often unavailable in disaster-affected settings. This represents a fundamental measurement challenge: the absence of pre-event baseline data makes it impossible to construct the counterfactual quantities needed for welfare analysis. To address this measurement problem, we develop a hybrid machine learning approach to estimate unobserved household health spending using longitudinal survey data from Indonesia. We first develop a model around the 2006 Yogyakarta earthquake, for which complete data are available. The model learns spending patterns across income, hazard intensity, and other characteristics, achieving >70% accuracy in a noisy and complex domain. After testing the model for transportability, we apply it to post-2004 Indian Ocean tsunami survey data in Indonesia, to predict plausible baseline health spending. These predictions are used to evaluate the impact of the tsunami on health spending to reveal that without targeted aid, catastrophic health spending would have increased from 4.5% to 29.4% and that moderately damaged households experienced more cost increases than heavily damaged ones. By combining artificial intelligence with 2 household survey data, our framework is a proof-of-concept, for addressing data gaps in official economic statistics, demonstrating how machine learning can enable counterfactual welfare measurement where conventional data collection is absent or incomplete.

Keywords: Natural hazards, catastrophic health spending, disaster risk reduction, tsunami, earthquake, Indonesia, machine learning

JEL classification: C45, C51, C52, C53, I19, H84, O13, Q54

Rozana Himaz, UCL
r.himaz@ucl.ac.uk

Published by:
Economic Statistics Centre of Excellence
King's College London
Strand
London
WC2R 2LS
United Kingdom
www.escoe.ac.uk

Machine Learning for Estimating Catastrophic Health Spending in Disaster-Affected, Data-Scarce Settings

Rozana Himaz*[1], Dimitra Salmanidou[2], Saman Ghaffarian[1]

[1]Department of Risk and Disaster Reduction, University College London, United Kingdom

[2]Advanced Research Computing Centre, University College London, United Kingdom

*Corresponding author email: r.himaz@ucl.ac.uk

Abstract

Natural hazard events can increase out-of-pocket health costs and push vulnerable households into poverty. Mitigation measures require understanding changes in health spending patterns using pre- and post-event data, but such data are often unavailable in disaster-affected settings. This represents a fundamental measurement challenge: the absence of pre-event baseline data makes it impossible to construct the counterfactual quantities needed for welfare analysis. To address this measurement problem, we develop a hybrid machine learning approach to estimate unobserved household health spending using longitudinal survey data from Indonesia. We first develop a model around the 2006 Yogyakarta earthquake, for which complete data are available. The model learns spending patterns across income, hazard intensity, and other characteristics, achieving >70% accuracy in a noisy and complex domain. After testing the model for transportability, we apply it to post-2004 Indian Ocean tsunami survey data in Indonesia, to predict plausible baseline health spending. These predictions are used to evaluate the impact of the tsunami on health spending to reveal that without targeted aid, catastrophic health spending would have increased from 4.5% to 29.4% and that moderately damaged households experienced more cost increases than heavily damaged ones. By combining artificial intelligence with

household survey data, our framework is a proof-of-concept, for addressing data gaps in official economic statistics, demonstrating how machine learning can enable counterfactual welfare measurement where conventional data collection is absent or incomplete.

Keywords: Natural hazards, catastrophic health spending, disaster risk reduction, tsunami, earthquake, Indonesia, machine learning

JEL Classification codes: C45, C51, C52, C53, I19, H84, O13, Q54

Acknowledgements

This work is supported by a Research Collaborations grant, number 1203760081, under the International Science Partnerships Fund. The grant is funded by the UK Department of Science Innovation and Technology in partnership with the British Council. Find out more on the UK Government website https://www.ukri.org/what-we-do/browse-our-areas-of-investment-and-support/international-science-partnerships-fund/. The authors thank the two anonymous referees from the Economic Statistics Centre of Excellence (ESCoE) discussion paper series and participants at the IFS-UCL-LSE/STICERD Development work in progress seminar (October 2025) for their insightful comments. All potential errors are our own.

1. Introduction

Natural hazards such as tsunamis and earthquakes cause significant loss to mental and physical health through injury, illness, and trauma (Bartels & VanRooyen, 2012; Frankenberg et al., 2011). This affects out-of-pocket health costs which can reach catastrophic thresholds – a definition of which is when health spending is above 10% of consumption or income (Wagstaff et al., 2018; Xu et al., 2003) – pushing households into poverty and making it difficult to escape the condition (Dercon, 2004; Hallegatte et al., 2020). A hazard can turn into a disaster when combined with vulnerability – the propensity of people, societies, and ecosystems to be harmed (Blaikie et al., 2014; O'Keefe et al., 1976). Understanding how out-of-pocket health spending changes in response to a hazard event is critical for designing effective social protection and disaster mitigation strategies. But in many low- and middle-income countries, pre- and post-event household health spending data are scarce, limiting the ability to quantify these impacts accurately.

Can advances in machine learning support the estimation of unobserved pre-event health spending? This study addresses this question using advanced machine learning and explainable artificial intelligence combined with longitudinal household survey data from Indonesia. We train and apply deep/machine learning models to predict pre-event health costs using the Indonesia Family Life Survey (IFLS) and then apply an explainable artificial intelligence workflow to analyse and understand the key drivers behind the machine learning process. Some households in the dataset were affected by the 27 May 2006 Yogyakarta earthquake, while all have complete pre- and post-event data. We focus on a subsample forming a quasi-natural experiment, with affected households as the treatment group and a comparable set of unaffected households as the control group. This design enables the models to learn household spending patterns across hazard exposure, wealth, and other characteristics, facilitating the application of the most robust model to the Study of Tsunami Aftermath and Recovery (STAR) dataset to estimate previously unobserved baseline health expenditures. STAR is a tsunami-specific survey conducted following the 26 December 2004

Indian Ocean tsunami and captures a quasi-natural experiment setting by design. The predicted baseline is then used to quantify the causal impact of the tsunami on health spending.

Our approach assumes that household health spending patterns follow broadly similar trends after high-impact, low-probability natural hazards such as earthquakes and tsunamis in Indonesia. The 2004 Indian Ocean tsunami and the 2006 Yogyakarta earthquake provide contrasting but complementary cases. The tsunami caused massive mortality (~160,000) and displacement (~500,000), relatively few injuries, but high psychological trauma (Frankenberg et al., 2008). The earthquake caused fewer deaths (5,700) but far more injuries (138,000) and displaced 700,000 (Bappenas, 2006). The hazard contexts differed in geography, population density, aid intensity and insurgency-exposure with more interior areas of Aceh being exposed to a decades long civil conflict between the Free Aceh Movement and the Indonesian government which was not present in Java. These contextual differences are accounted for, to some extent, using rich data from comparable datasets. Despite differences, both contexts exhibited similar structural vulnerabilities including non-resilient housing, limited early warning, and low household disaster preparedness. Considering that both were high-impact, low-probability events and that the model accounts for key contextual differences, these similarities support the assumption that health spending patterns learned from earthquake data may approximate those following the tsunami. To formally assess this transportability assumption, we employ standardized mean difference analysis and propensity score overlap tests—established methods for evaluating whether a predictive model trained in one context can be validly applied in another (Pearl & Bareinboim, 2014; Stuart, 2010).

The key objectives of the paper are to:

1. Develop and validate explainable machine learning workflows that can predict health share (i.e., out-of-pocket health spending as a proportion of income that we refer to as 'health share' in the rest of this paper) using IFLS longitudinal data for 2000 and 2007.
2. Apply the most robust trained model to STAR 2005 data to estimate unobserved baseline health share. Validate estimates using distributional plausibility, internal consistency, comparison with independent data, triangulation with relevant literature and tests for transportability. In the context of this paper, transportability refers to the validity of applying a predictive model − trained on IFLS household data from an earthquake-affected context − to estimate unobserved health share in the STAR tsunami-affected context. Formally, this requires both that STAR households fall within the covariate support of the IFLS training data, and that the conditional relationship between household characteristics and health spending is sufficiently stable across the two disaster settings. In the context of our paper, transportability refers to the extent to which the distributions in the two contexts overlap, making model transfer credible[1].
3. Quantify the impact of 2004 tsunami on changes to catastrophic health spending.

[1] In the machine learning literature, the definition of transportability is different, referring to the ability to apply a model trained on one population or setting to a new population or setting (target) while maintaining predictive performance. It is related to but distinct from external validity in the econometric sense, which asks whether a treatment effect or finding generalises to a new population. To measure strength of transportability as defined in typical machine learning contexts, metrics such as the Brier score and Integrated Calibration Index are used. These metrics cannot be applied in our setting because they require observed outcomes in the target context, which are precisely what we are trying to predict. Transportability as we define in this paper is the stricter, more formal condition that the conditional relationship between covariates and the outcome, P(Y|X), is stable across contexts Pearl, J., & Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, *29*(4), 579-595. https://doi.org/10.1214/14-sts486 .

The study makes four key contributions. First, it addresses a core economic measurement problem: the absence of pre-event baseline data that makes counterfactual welfare analysis impossible in disaster-affected settings. This data scarcity challenge is recognised in the poverty measurement literature (Dang & Lanjouw, 2023; Dang et al., 2019) but has received little attention in disaster economics. We introduce a methodological innovation merging explainable deep learning with longitudinal household survey data to reconstruct unobserved economic quantities — specifically pre-disaster household health spending. While this approach extends the cross-survey imputation tradition in economic measurement (Dang et al., 2025; Elbers et al., 2003), it addresses a distinct and more demanding problem: imputing a variable for a pre-event baseline for a population that was only surveyed after the disaster, where no directly comparable pre-event survey exists[2].

Availability of pre-hazard data enables counterfactual analysis informing evidence-based shock-responsive policy design. It also allows the creation of social vulnerability functions (also known as damage functions) that explicitly incorporate microeconomic data to quantify disaster risk. Currently, probabilistic disaster risk quantification using tools such as catastrophe modelling focus on damages to infrastructure and property. This is partly due to the lack of socioeconomic data in disaster contexts. The method we introduce is an

[2] Dang et al. (2025) use a linear model with cluster random effects and simulation-based prediction to impute consumption data never collected within an existing survey, validating against true values from comparable surveys at the aggregate level. Our approach differs in three ways. (1) We use a non-parametric hybrid machine learning method that imposes no distributional assumptions (2) We predict individual household-level outcomes rather than aggregate poverty rates. (3) As validation against true pre-event values is impossible for the tsunami context we employ alternative plausibility checks including distributional assessment, comparison with independent survey estimates, and formal transportability testing.

important stepping stone allowing the development of generic vulnerability functions to support innovative risk quantification found in early work such as Salmanidou et al. (2021).

Second, the paper contributes to the growing literature on transparency and trustworthiness in economic measurement (Coyle & Nakamura, 2019). It applies explainable artificial intelligence methods for economic analysis in disaster contexts. Conventional models predict results, but the drivers of such predictions are an unknown “black box”. The use of explainable artificial intelligence unpacks the main drivers of the predictions allowing checks for plausibility and reliability of the model a requirement that is critical not only for academic validity but for any method intended to inform official statistics or policy in high-stakes settings.

Third, it broadens the scope of machine learning in disaster science by shifting the focus from hazard mapping and post-disaster response using remote sensing toward socioeconomic impacts, high-impact low-probability events, and the measurement of pre-hazard event baselines through secondary household surveys to support welfare measurement and to enable ex-ante risk reduction as much as it does ex-post coping and recovery.

Finally, it provides the first model-based measurement of pre- 2004 tsunami out-of-pocket health spending in Aceh and North Sumatra.

This is a scoping and proof-of-concept study that can be built on and expanded with further training on heterogeneous data sources toward approaches analogous to foundation models in machine learning — large-scale models capable of generalising across contexts — that could predict unobserved welfare outcomes across diverse disaster settings with higher accuracy.

2. Framework and data

2.1 A quasi-experimental framework

Natural hazards such as earthquakes and tsunamis strike populations largely at random, without targeting specific individuals or households, sometimes with little warning. However, the exposure of populations to their effects is not random — socioeconomic circumstances, historical settlement patterns, property values, and discrimination can concentrate more disadvantaged populations in more vulnerable locations. This means that exposed and unexposed groups may differ systematically in ways that confound causal inference. To address this, we do not simply compare affected and unaffected households. Instead, we use carefully constructed treatment groups and comparable control groups. When combined with pre- and post-event data, this enables causal inference on the impact of hazard events on household health spending using methods such as difference-in-differences analysis or propensity score-weighted regressions (Angrist & Pischke, 2009; Craig et al., 2017; Hirano & Imbens, 2004). Such analysis assumes that the parallel trends assumption holds, meaning that in the absence of the hazard, treated and control households would have followed similar trajectories in health spending over time. This allows changes observed after the event to be attributed to the hazard itself rather than pre-existing differences between groups.

In the case of the 2006 earthquake that struck near the major city of Yogyakarta on the island of Java, the treatment group is defined as households located in areas that experienced significant ground shaking, as measured by the instrumental Modified Mercalli Intensity scale. The control group of households comes from unaffected urban areas that are within 5km of cities with a population of more than 300000 people in Java. This classification comes from the careful methodology developed in Kirchberger (2017). This quasi-experimental subset of IFLS data underpins our explainable machine learning models, guiding them to learn patterns specifically associated with high-impact, low-probability hazard exposure and household characteristics relevant to health spending. It is important to

note that the machine learning models are purely predictive, capturing associations rather than causal impacts.

Using a subset of the IFLS data that fits a quasi-experimental setting rather than the full IFLS dataset reduces the noise and complexity of the sample, especially as we hope to apply (or 'transport') the trained model to a different context. Thus, even the 2004 Indian Ocean tsunami data from STAR includes only treatment and control group households that were exposed to heavy, moderate or no/light damage. This construction was inbuilt to the survey by the STAR team, who used satellite images corroborated by field-verification.

2.1.1. Indonesia Family Life Survey (IFLS) – pre- and post-earthquake data

The IFLS is a nationally representative longitudinal study initiated in 1993, covering 11 of 15 Indonesian provinces and approximately 83% of the population. We utilize data from the 2000 and 2007 waves to capture pre- and post-earthquake conditions for the 27 May 2006 Yogyakarta earthquake that recorded a 6.6 magnitude and Medvedev— Sponheuer–Karnik scale of VIII (damaging). The treatment group comprises 430 households directly affected, and the control group includes 2151 households. Respondents were interviewed 6–14 months after the earthquake struck. Corresponding earthquake intensity is based on the work of Kirchberger (2017) who uses the United States Geological Survey's ShakeMap measure of instrumental intensity that maps ground motion parameters onto the Modified Mercalli Intensity scale.

2.1.2. Study of the Tsunami Aftermath and Recovery (STAR) Survey – post-tsunami data

The STAR survey sampled households from areas that were severely damaged along the coast of Aceh and North Sumatra as well as households with little or no damage from areas slightly more interior. The survey was conducted 5–12 months after the event, across 13 districts (kabupaten) and the sampling framework was based on the 2004 National

Socioeconomic Survey , which included 585 enumeration areas within 525 villages (desa). The treatment group comprises 3,651 households with complete information, while the control group comprises 994 households. Treatment intensity is based on a household- level exposure measure constructed using 18 self-reported survey items capturing physical proximity (e.g., felt earthquake, saw tsunami, swept away), direct harm (e.g., injuries, displacement), social impacts (e.g., witnessing death or disappearance of relatives/neighbours), and psychological stress (e.g., fear of death or serious injury). These intensities map onto the STAR team's own broader classification of heavy, medium and no/light damage-based tsunami physical damage and destruction, captured using satellite imagery at a resolution of 0.6 square kilometres, validated by field observations. Figure 1 shows the location of the households based on the geographic coordinates disclosed in the surveys: for IFLS at the subdistrict level and for STAR at the region level.

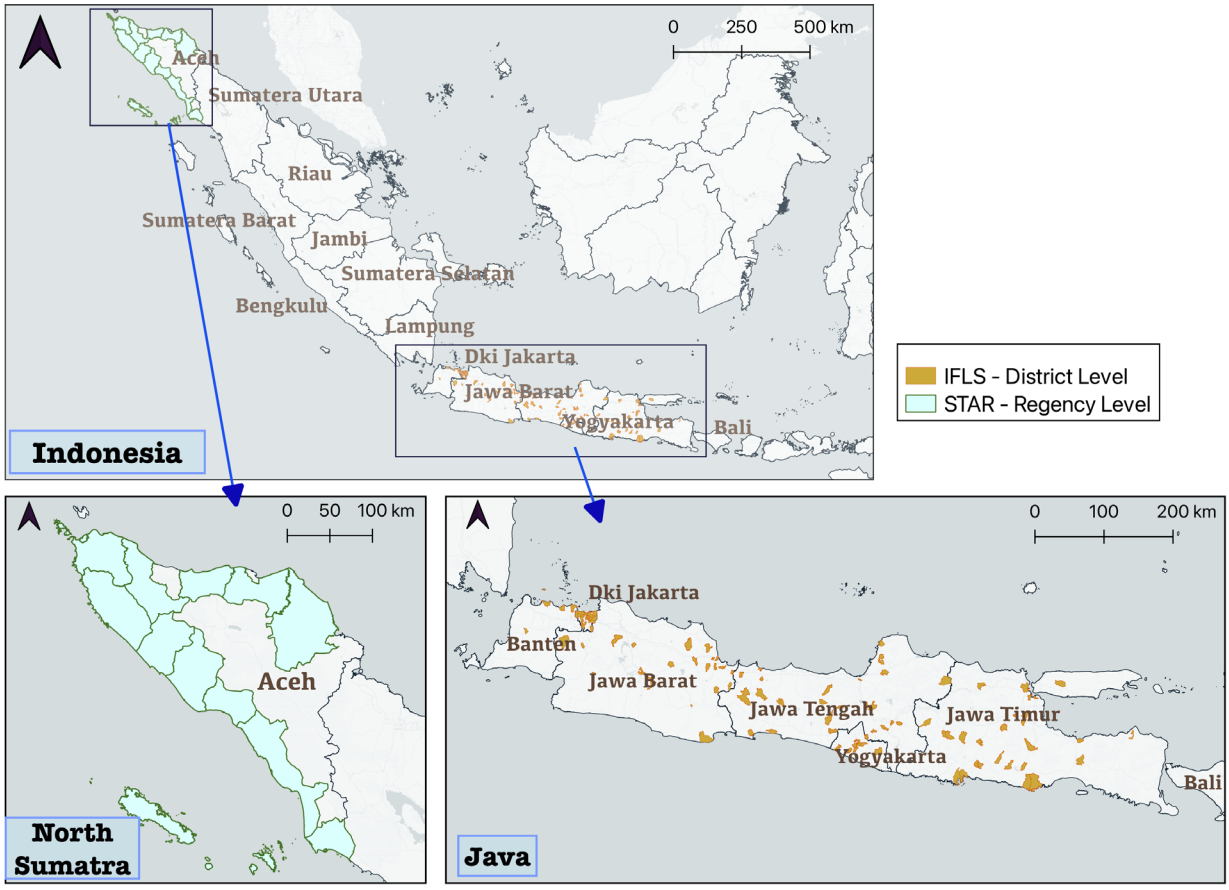
Both surveys include comparable variables for household health spending, demographics, and socio-economic characteristics. IFLS has complete data on pre and post-hazard event characteristics, but the STAR contains mainly post-event data. However, it asks respondents a few retrospective questions about pre-tsunami characteristics such as wealth and income but not about health spending. The available retrospective data is used to construct baseline variables to support a more informed machine learning model. More details on variables used, including summary statistics, are provided in Appendix Table A1.

2.2. Features that influence household health spending share

High-impact, low-probability events such as earthquakes and tsunamis fundamentally disrupt household economic decisions, particularly regarding health expenditure allocation. Figure 2 illustrates how hazard, exposure, and vulnerability interact to generate both health losses and asset destruction, which together drive demand for health services. This demand may be met through public healthcare, insurance, or aid, but when these prove inadequate,

households resort to out-of-pocket spending. Understanding how hazards affect the household budget share of health spending requires examining both demand and supply mechanisms operating through these interconnected channels.

Figure 1: Spatial extent of the two Indonesian surveys used in this study. The Study of Tsunami Aftermath and Recovery (STAR) data spread over 13 districts in Aceh and North Sumatra. The Indonesia Family Life Survey (IFLS) data spread across different provinces in Java.



Demand-Side Features

Natural hazards generate immediate and persistent changes in health service demand. Physical injuries, disease outbreaks, and mental health deterioration following displacement

create direct pressure for healthcare utilisation (Watson et al., 2007). However, simultaneous asset destruction and livelihood disruption constrain households' financial capacity to seek care (García-Gómez et al., 2013). When income falls sharply while health needs rise, the household budget share allocated to health may increase, decrease, or remain stable depending on which effect dominates. Mental wellbeing proves particularly important as poor psychological health can reduce labour force participation and workplace productivity, thereby suppressing earnings (Bubonya et al., 2017; Loughran & Heaton, 2013). Conversely, employment losses and economic shocks can precipitate depressive symptoms, creating bidirectional relationships between economic circumstances and mental health (Frijters et al., 2014).

Household characteristics capturing demand-side pressures include total income, reflecting overall economic capacity and budget constraints; wealth indices and asset holdings, indicating longer-term economic position and shock absorption capacity; household size, affecting health expenditure needs; and health status variables, directly measuring medical need. Educational attainment—stratified across primary, secondary, and tertiary levels—signals both earning potential and health knowledge influencing care-seeking behaviour. Farming household status indicates livelihood vulnerability to environmental shocks and income volatility.

Supply-Side Features

The post-event healthcare landscape transforms through both deterioration and reconstruction. Infrastructure damage reduces healthcare accessibility, particularly in rural areas where facilities are sparse and transportation networks are disrupted. Simultaneously, humanitarian assistance often delivers substantial short-to-medium term support through medical aid, subsidized services, and cash-for-work programs, temporarily reducing out-of-pocket spending despite elevated health needs. Reconstruction activities may increase

wages in certain sectors, potentially improving household financial capacity (Kirchberger, 2017).

Features capturing supply-side factors include geographic location—rural versus urban—which proxies healthcare accessibility, infrastructure quality, and supply constraints. Aid and transfers including pensions and scholarships capture external resource flows that buffer disaster impacts and improve service access.

Together, these demand and supply-side features capture the pressures, constraints, and resources determining how households allocate budgets toward health in hazard-affected contexts. The machine learning approach uses these multidimensional predictors to learn relationships between household characteristics and the budget share of health expenditure allocation applicable to predicting unobserved baseline spending in the tsunami context.

Figure 2: Conceptual pathways from high-impact, low-probability events to household health spending share Adapted framework from Syukriyah and Himaz (2024), *showing how hazard, exposure, and vulnerability generate health and asset losses that drive demand for health services, which may be financed through public systems, insurance aid. or out-of-pocket spending.*

3. Empirical Methods

3.1. Deep-learning methods to predict unobserved data

We train and test nine different machine learning models from various architectures: Linear models (regression), tree-based (Random Forest, XGBoost, Light Gradient Boosting Machine (LGBM)), neural networks (Feedforward Neural Networks (FNN)) and ensembles (Hybrid FNN-LGBM, Hybrid TabNet-LightGBM). Due to space considerations, we explain below the empirical method with regard to our model of choice, Hybrid FNN-LGBM, as decided based on the results described in more detail in the next section.

The hybrid Feedforward Neural Network–Light Gradient Boosting Machine (FNN–LGBM) model is designed to exploit the complementary strengths of deep neural networks and gradient boosted decision trees in predictive modelling for structured tabular data. The FNN component serves as a powerful, supervised feature generator capable of extracting non-linear and high-order feature interactions, while the LightGBM component provides efficient, robust, and interpretable tree-based regression. By integrating these two paradigms in a stacked generalisation framework, the hybrid model achieves both representational richness and predictive stability, making it well-suited for complex tasks such as estimating pre-event household health spending in data-scarce settings.

Feedforward neural networks (FNNs) constitute one of the foundational architectures in deep learning. Inspired by biological neural systems, artificial neurons receive input signals, apply weighted transformations, and propagate activations through non-linear functions.

Let us denote the input vector by $\mathbf{x} \in \mathbb{R}^d$, where d is the dimensionality of the feature space. The output of a single neuron is given by:

$$f(x; W, b) = \sigma(Wx + b)$$

Here, $\mathbf{W} \in \mathbb{R}^{n \times d}$ represents the weight matrix, $\mathbf{b} \in \mathbb{R}^{n}$ the bias vector, and $\sigma(\cdot)$ a non-linear activation function. Activation functions are essential for introducing non-linearity, enabling the network to model complex, non-linear relationships. In our implementation, rectified linear units (ReLU) were employed in the hidden layers for their computational efficiency and gradient stability, while a sigmoid activation was applied at the output layer to constrain predictions to the interval [0, 1], consistent with the bounded nature of the target variable.

Let us define the layer-wise transformation for a network comprising *L* layers. For each hidden layer *l = 1, …, L*, the output is computed as:

$$\mathbf{h}^{(l)} = \sigma(\mathbf{W}^{(l)} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)})$$

with $\mathbf{h}^{(0)} = \mathbf{x}$, $\mathbf{W}^{(l)} \in \mathbb{R}^{n^{(l)} \times n^{(l-1)}}$, and $\mathbf{b}^{(l)} \in \mathbb{R}^{n^{(l)}}$. In our configuration, the FNN comprises two hidden layers with 128 and 64 units, respectively. Each layer is followed by dropout regularisation (rate = 0.30) to mitigate overfitting.

Supervised Embedding and Feature Augmentation

The output of the final hidden layer, denoted $\mathbf{z}\_\text{FNN} \in \mathbb{R}^{m}$, serves as a dense, task-specific embedding of the input data, where *m = 64* corresponds to the dimensionality of the last hidden layer. To prevent information leakage and ensure unbiased embedding generation, we employed five-fold cross-validation. For each fold, the FNN was trained on the training subset and used to generate embeddings for the corresponding validation subset. The concatenation of these out-of-fold embeddings yields a meta-feature per instance.

Let us define the augmented feature vector as:

$$\mathbf{x}\_\text{aug} = [\mathbf{x}; \mathbf{z}\_\text{FNN}] \in \mathbb{R}^{d+m}$$

where $[\cdot;\cdot]$ denotes vector concatenation.

Gradient Boosting Regression

The second stage of the hybrid model applies LightGBM regression to the extracted feature set. LightGBM constructs an ensemble of decision trees in a sequential manner, where each tree is trained to minimise the residuals of the current model. Let $f_t(x)$ denote the prediction of the t-th tree, and $F_t(x)$ the cumulative model after t iterations. The model update rule is given by:

$$F_t(x) = F_{t-1}(x) + \eta f_t(x)$$

where η is the learning rate. The objective function to be minimised is:

$$\mathcal{L}(\theta) = \Sigma_i \, \ell(y_i, F(x\_aug_i; \theta)) + \Omega(F)$$

Here, $\ell(\cdot)$ denotes the mean squared error loss, θ the parameters of the trees, and $\Omega(F)$ a regularisation term controlling model complexity.

In our implementation, LightGBM was configured with the following hyperparameters: *n_estimators* = *500*, *max_depth* = *10*, *learning_rate* = *0.01*, and a fixed random seed (42) to ensure reproducibility.

The proposed hybrid FNN–LightGBM framework combines the expressive capacity of deep neural embeddings with the predictive stability and interpretability of gradient boosted trees. The FNN component extracts high-level representations tailored to the prediction task, while LightGBM leverages both original covariates and learned embeddings to enhance accuracy. This architecture is particularly well-suited to noisy, heterogeneous domains, as demonstrated by its robust performance in our health expenditure prediction task.

3.2. Transportability of earthquake-based predictions to tsunami context

We conduct formal statistical tests to assess whether the earthquake-trained model can validly predict tsunami outcomes—a question of transportability. First, we calculate standardised mean differences for each covariate—the difference between sample

means for IFLS and STAR divided by the pooled standard deviation. A standard mean difference < 0.25 indicates acceptable balance for observational comparisons (Stuart, 2010). This scale-free metric reveals whether populations differ on observed characteristics affecting health expenditure. Second, we estimate propensity scores—the predicted probability of being in the STAR dataset given household characteristics—via logistic regression (Rosenbaum & Rubin 1983). The propensity score is a balancing score that collapses multidimensional covariate information into a single dimension. We assess common support, the overlapping region where both datasets have observations (Crump et al., 2009). High overlap indicates STAR households have comparable IFLS matches in the training data, enabling prediction without extrapolation. Third, we examine trimmed overlap (propensity scores 0.1–0.9), excluding extreme cases where predictions are least certain (Imbens, 2015). This conservative metric identifies the subset for which transportability is strongest.

3.3. Econometric methods to estimate impact of tsunami on health outcomes

Once the unobserved baseline data are predicted using machine learning methods, the following model is used to estimate the impact of the tsunami on health outcomes. Pre-tsunami data is denoted as t=1 and post-tsunami data collected 5–18 months after the event as t=2.

$$y_{i2} = \alpha + \beta H_i + \beta M_i + \gamma X'_{i1} + \varepsilon_{i1} \qquad (1)$$

for individuals $i = 1 \ldots N$ where

y_{i2} is household monthly income share of out-of-pocket health spending (continuous variable) or catastrophic health spending status (=1 if health spending share > 10%, 0 otherwise). The variable H_i and M_i, represent tsunami damage that can be heavy or moderate (with no damage omitted). X_{i1} is a vector of pre-tsunami controls including predicted health share or baseline catastrophic health spending based on the machine

learning predictions. Given imbalances in pre-tsunami characteristics across damage groups, regressions are weighted using propensity scores for tsunami exposure (Rosenbaum & Rubin, 1983; Syukriyah & Himaz, 2024). The 2004 Indian Ocean tsunami struck without warning, with the first waves reaching Aceh approximately 20 minutes after the undersea earthquake, leaving no time for preparation. This lack of forewarning supports treating the timing and occurrence of the tsunami as exogenous, ruling out anticipatory behavioural responses such as relocation.

However, the exposure of households to tsunami damage reflects both proximity to the coast and underlying socioeconomic characteristics that may affect health spending independently. Indeed, pre-tsunami characteristics were not balanced between the treatment and control groups. We therefore weight all regressions using propensity scores for tsunami exposure based on observable household characteristics (Rosenbaum & Rubin, 1983; Syukriyah & Himaz, 2024), reducing but not eliminating the risk that omitted factors drive results. The identification assumption $E(\varepsilon_{j1}|M_j, H_j, X_{j1}) = 0$ should therefore be understood as conditional on the observable controls and propensity score weighting.

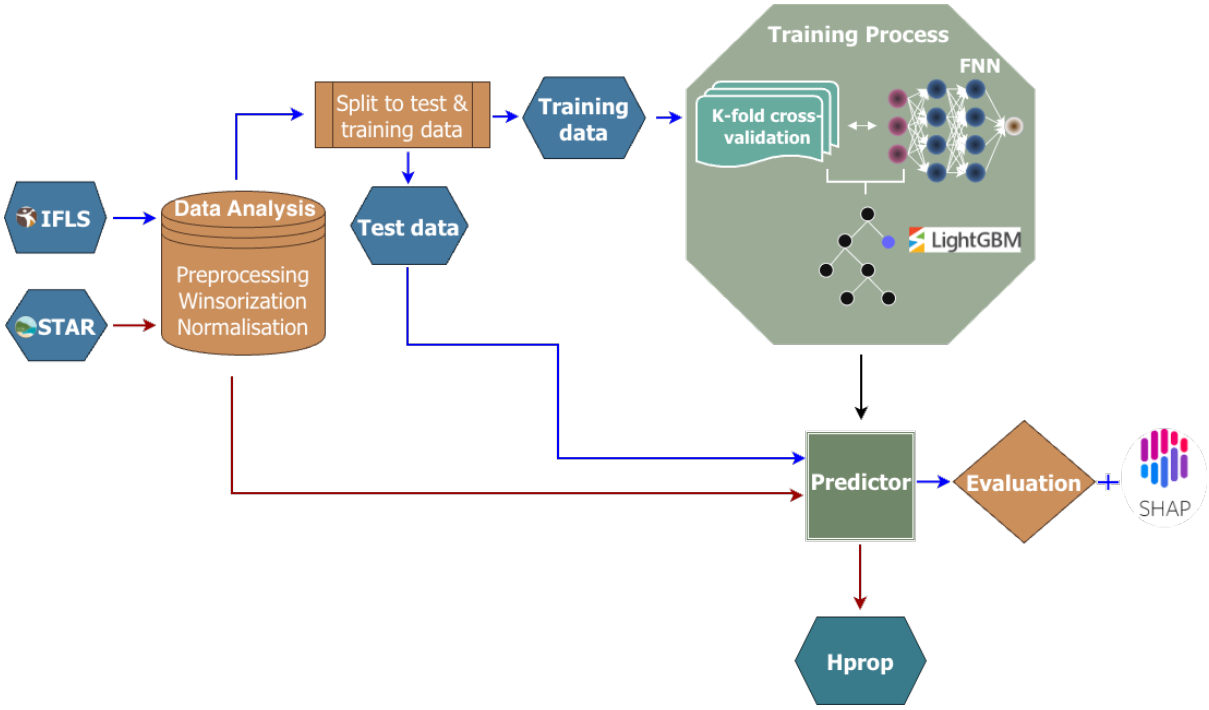
4. Results

4.1 Deep learning model development and validation

To identify the most suitable approach for estimating pre-event household health spending as a share of income (health share) in data-scarce contexts, we developed and validated a diverse set of machine learning and deep learning models (Figure 3). The baseline was a linear regression model, representing a traditional statistical approach that captures only linear relationships between predictors and outcomes. We then implemented XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017), both gradient boosting frameworks that are highly effective for structured data and capable of modelling complex, non-linear patterns. Ensemble tree methods were further extended to a Random Forest (Breiman,

2001), which averages predictions from multiple decision trees to reduce variance and improve generalisation.

Figure 3: The flow chart of the process for training the hybrid Feedforward Neural Network–Light Gradient Boosting Machine (FNN-LGMB) model.



We also tested several deep learning architectures tailored for tabular and sequential data. TabNet (Arik & Pfister, 2021) uses an attention-based mechanism to learn sparse, interpretable feature representations. A feedforward neural network (FNN) was applied as a flexible non-linear function approximator. We further evaluated a Siamese self-attention transformer (Vaswani et al., 2017), designed to learn robust similarity measures between data instances through parallel attention mechanisms.

Finally, we developed two hybrid models that combined deep feature extraction with gradient boosting: the hybrid FNN–LightGBM (FNN-LGBM), which uses an FNN to learn latent representations before LightGBM prediction, and the hybrid TabNet-LightGBM (TabNet-LGBM), which integrates LightGBM's efficiency with TabNet's multi-step attention layers. All

models were trained and evaluated on the same pre-processed dataset using k-fold cross-validation and consistent performance metrics to ensure fair comparison. Across all experiments, the hybrid FNN-LGBM achieved the highest predictive accuracy, with the hybrid LGBM-TabNet performing a very close second. Both outperformed the standalone deep learning and gradient boosting approaches, highlighting the value of combining representation learning with strong ensemble methods in structured tabular data. More details on the deep learning method are provided in the Appendix.

4.2. Pre-quake health share estimation using IFLS data

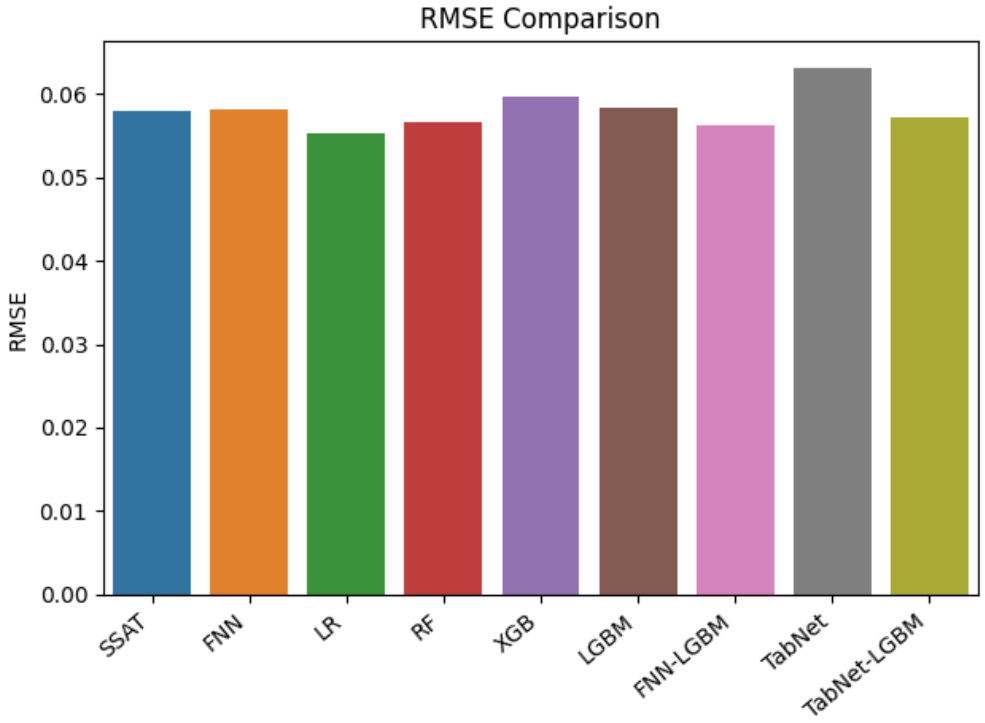
The raw and relative prediction performance metrics in Table 1 show that IFLS pre-quake health spending is best estimated by the FNN-LGBM model. This is based on low Root Mean Squared Error of 0.056, matching or outperforming other models on typical error with Median Absolute Error=0.022 and moderate sensitivity (5.5%) in identifying households with catastrophic spending (health share>0.1). Similar results based on metric evaluation come from Tabnet-LGBM. Figure 4 compares Root Mean Squared Error across selected models. Full prediction performance metrics across all nine models are provided in Appendix Table A2 panels A and B, with Appendix Table A3 providing definitions and calculation methods.

Table 1: Raw prediction metrics for health expenditure proportion (health share) and catastrophic spending detection across models

Model	mean	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	Median Absolute Error	Sensitivity (health share>10%)
Actual	0.039	n/a	n/a	n/a	n/a	n/a
Siamese Self-Attention Transformer	0.021	0.004	0.06	0.019	0.009	0
Feedforward Neural Network	0.027	0.003	0.057	0.026	0.013	0.9
Linear Regression	0.038	0.003	0.055	0.04	0.03	0.9
Random Forrest	0.042	0.003	0.056	0.044	0.015	9.2
XGBoost	0.04	0.004	0.06	0.044	0.024	13.8
Light Gradient Boosting	0.039	0.003	0.058	0.041	0.024	9.2
Feedforward Neural Network –Light Gradient Boosting Machine	0.038	0.003	0.056	0.039	0.022	5.5
Tabnet	0.041	0.004	0.063	0.047	0.018	0.0
TabNet– Light Gradient Boosting Machine	0.038	0.003	0.057	0.039	0.021	4.6

Notes: Mean Squared Error: average squared difference between predicted and observed values; Root Mean Squared Error: square root of Mean Squared Error; Mean Absolute Error: average absolute difference; Median Absolute Error: median absolute difference; Sensitivity: ability of the model to correctly identify individuals spending >10% of income on health.

Figure 4: Machine learning model comparison based on Root Mean Qquared Error (RMSE). SSAT: Siamese Self Attention Transformer; FNN: Feedforward Neural Network; LR: Linear Regression; RF: Random Forrest; XGB: XGBoost; LGBM: Light Gradient Boosting Machine; FNN-LGBM: Feedforward Neural Network –Light Gradient Boosting Machine; TabNet-LGBM: TabNet–Light Gradient Boosting Machine.

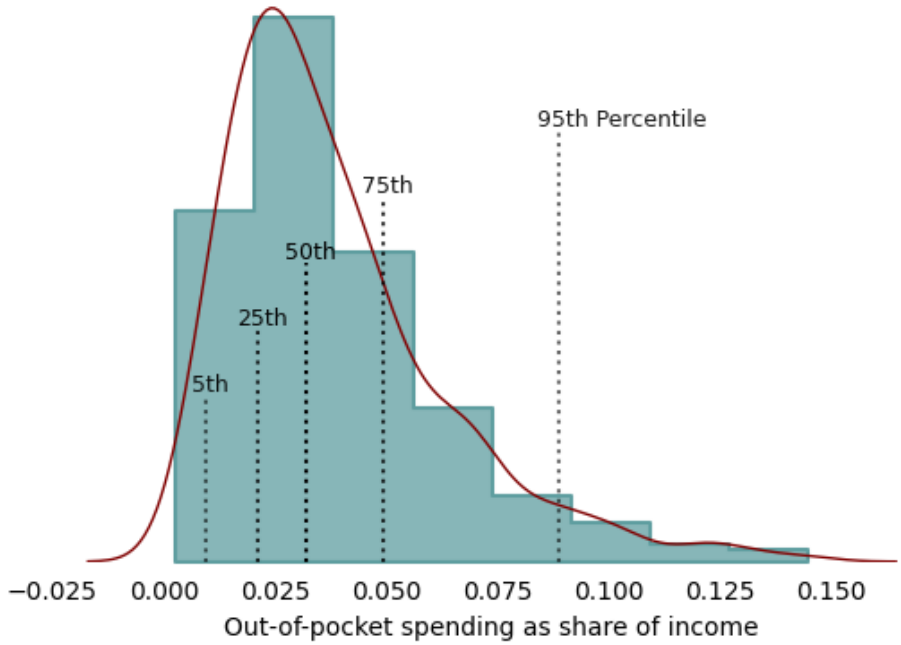


Overall, model performance varied substantially across both raw error metrics and relative evaluation criteria. While linear regression achieved the lowest Root Mean Squared Error (0.055), indicating strong performance in overall error minimisation, it exhibited higher median error (Median Absolute Error = 0.030) and a relatively weak sensitivity (0.9%). This suggests that despite minimising average deviation, linear regression struggled to accurately capture high-burden cases, which are central to the study's aim. Although some tree-based models (Random Forest and XGBoost) showed stronger sensitivity (up to 13.8%) for detecting catastrophic spending, these came at the cost of significantly higher relative Mean Absolute Error, Median Absolute Error, and Quantile Loss values, indicating broader overfitting and noisier performance. Similarly, while the Siamese Transformer model excelled

at low-expenditure precision (lowest $QLoss_{0.25}$ = 0.340), it suffered from minimal sensitivity to catastrophic spenders. In terms of performance across quantiles our model of choice – FNN-LGBM– for applying to STAR data maintained competitive and symmetric performance across quantiles ($QLoss_{0.25}$ = 0.498; $QLoss_{0.75}$ = 0.507). Figure 5 shows health share predictions based on FNN-LGBM for IFLS test data.

To understand the key features that contributed to the predictions, we use Shapley Additive Explanation ("Shapley") plots (Lundberg & Lee, 2017), Figure 6. The most important features extracted by the FNN model are provided by the figure at the bottom while the predictors are in the main Shapley summary plot on the top. As seen from the summary plot, the top drivers of higher pre-quake health share are pre-quake assets, health condition, wealth, a household head with relatively higher education at completed tertiary or secondary school levels. The top driver of lower health share are household size and transfers. Although income is important in determining health share, it is the denominator of the health share metrics, and its effects are non-linear and heterogenous. Overall, these drivers corroborate findings in the literature that looks at household-level drivers of out-of-pocket health spending or catastrophic spending (Anindya et al., 2021; Deaton & Paxson, 1998; Fattah et al., 2023).

Figure 5: Health share predictions for the Indonesia Family Life Survey (IFLS) test data. The percentiles of the predictions are represented with dashed vertical lines. The kernel density function is estimated.



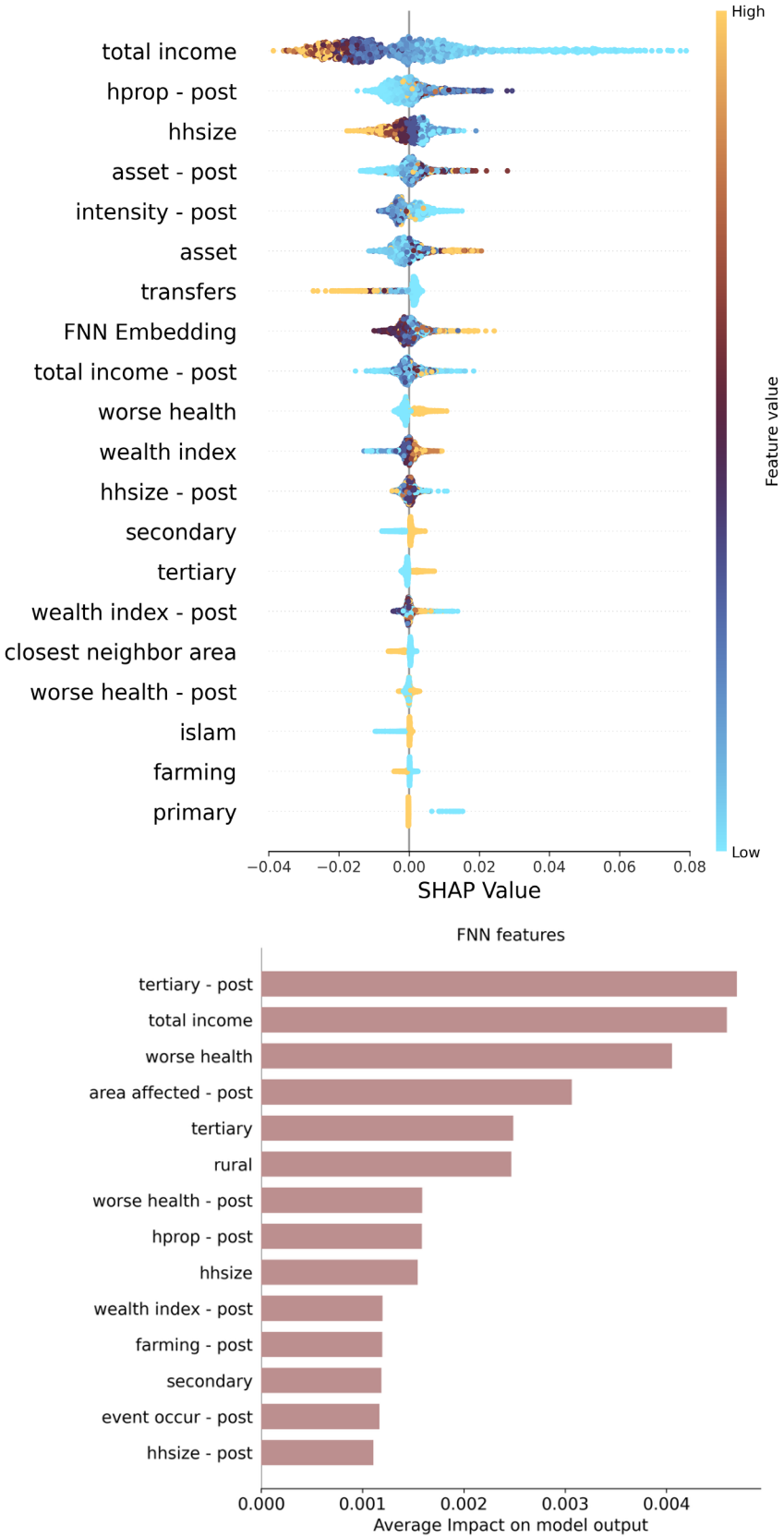


Figure 6: a) Top: Shapley values for the most important features in the training process of the hybrid machine learning model. b) Bottom: The bar plot represents the most important features in the Feedforward Neural Network (FNN) embeddings.

4.3. Transportability assessment

Applying the methods discussed in section 3.2, covariate balance assessment finds contextual differences between settings (64% of covariates with standardised mean differences < 0.25), most notably in rural location (standardised mean differences = 0.81) and farming occupation (standardised mean differences = 0.51). However, propensity score analysis revealed 99.9% common support**,** indicating that despite mean differences, households with similar characteristic profiles exist across both contexts. Theoretically central variables—education, wealth, and health status—demonstrated good balance (all standardised mean differences < 0.21). While approximately 40% of STAR households represent the rural-farming tail of the IFLS distribution, the near-complete propensity score overlap supports reliable prediction for most cases**.** Therefore, while the model demonstrates reasonable validity for transportability, its predictions for households within the rural-farming demographic tail should be considered informed estimates within a plausible range, rather than precise point predictions**.**

Table2: Assessment of Model Transportability from IFLS (Earthquake) to STAR (Tsunami)

Test	Metric	Threshold	Result	Assessment
Covariate balance	% vars Standardised mean difference<0.25	>75%	64% (7/11)	Moderate
Core variables	Education, wealth, health	-	SMD<0.21	Well-balanced
PS overlap (full)	Common support %	>75%	99.9%	Excellent
PS overlap (trimmed)	Within [0.1, 0.9]	>80%	65.6%	Moderate

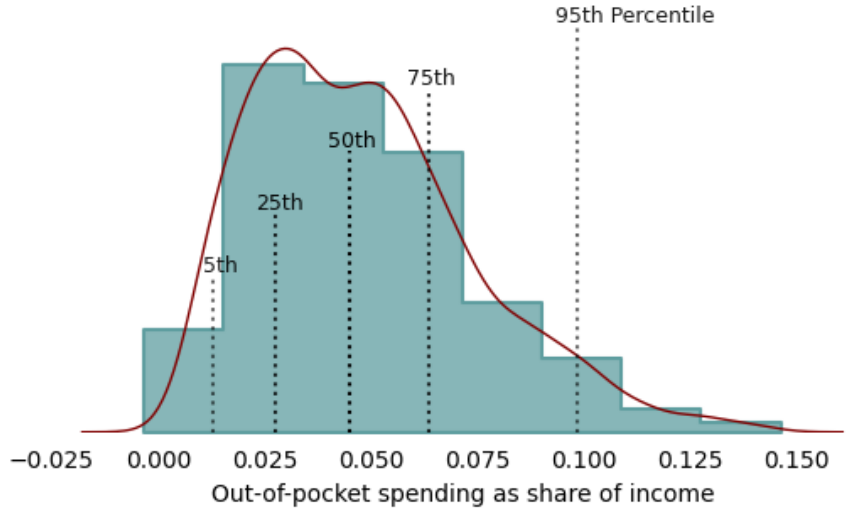
Contextual differences	Rural, farming	-	SMD>0.50	Noted

Note: Propensity scores estimated via logistic regression on household characteristics (rural, farming, household size, income, wealth, assets, health status, education). Common support defined as overlapping propensity score range between datasets. Trimmed overlap restricts to propensity scores between 0.1 and 0.9, excluding extreme cases.

4.4. Application to STAR to estimate pre-tsunami spending

Appendix Table A4 provides summary statistics for pre-tsunami health predictions from all nine models. In the absence of actual data to confirm the validity of the predictions, can we infer how well the FNN-LGBM model –our model of choice– does when applied to a dataset from a different context? To do so, we first examine the distributional plausibility and internal consistency of the estimates. As seen in Figure 7 the model predicts a pre-tsunami mean health share of 0.0481 (standard deviation = 0.0263), which is lower than the observed post-tsunami mean in STAR of 0.0563 (standard deviation = 0.0884), aligning with the hypothesis that catastrophic events like the tsunami increase out-of-pocket spending burdens. Moreover, the predicted distribution spans a realistic range—from near-zero values up to 0.1462, without the implausible extremes observed in some other models (e.g., linear regression's minimum = -0.22; XGBoost = -0.03). The median prediction (0.0448) is also reasonably lower than the post-tsunami median in STAR (0.0655). Compared to alternatives, FNN–LGBM maintains a balanced spread with no degenerate distributions, no heavy truncation at the lower bound, and no negative or highly skewed artefacts. Given its strong out-of-sample performance on IFLS test data and the internal coherence of its STAR predictions, the FNN–LGBM provides a robust and credible baseline for estimating pre-tsunami health share in STAR.

Figure 7: Health share predictions for the STAR data. The percentiles of the predictions are represented with dashed vertical lines. The kernel density function is estimated.



We then examine how well the predictions do when compared with estimates from an independent source- World Bank (2006)'s estimations of health share based on the National Socioeconomic Survey data collected in 2004. The latter is a large-scale nationally representative household survey conducted by *Badan Pusat Statistik*, Indonesia's national statistics agency, for monitoring purposes, representative of the population, unlike STAR that is tsunami-specific and not representative. The survey instruments it used to collect health, spending and income data differ from those used by STAR. Moreover, the World Bank report estimates the share of health spending as a percentage of household spending rather than income. Although STAR and the National Socioeconomic Survey are not strictly comparable, the latter reports average health share of 3.7% while the Hybrid FNN-LGBM model predicts 4.8%. We consider this 1.1 percentage point difference reasonable given contextual differences. The predictions also fit general patterns seen on out-of-pocket health spending in Indonesia at the time with the share higher for rural versus urban households.

4.5. Impact of tsunami on health spending

Using the predicted baseline values, we can analyse changes to household health spending and catastrophic spending due to the 2004 tsunami. As Table 3 summarises, mean health spending share (health share) increased from 4.8% pre-tsunami to 17.0% post-tsunami when tsunami-specific aid given to households is excluded. The share of households experiencing catastrophic health spending (defined as health share > 0.1) representing moderate financial stress, rose sharply from 4.5% to 29.4% without aid. Including aid mitigates this to some extent (18.3%), indicating its protective effect. These patterns hold when we look at severe catastrophic spending (defined as health share > 0.25) representing heavy financial stress. Disaggregating by wealth shows stark disparities. Among the poorest 20% of households, nearly one-third were already above the catastrophic threshold pre-tsunami, rising to 46.4% post-tsunami without aid. In contrast, no catastrophic spending was recorded among the richest 20% pre-tsunami, with post-tsunami exposure rising to 17.7% without aid.

Table 3. Summary statistics for health spending burden and incidence of catastrophic spending, before and after the tsunami (with and without aid), overall and by wealth quintile.

	Pre-tsunami	Post-tsunami (with Aid)	Post-tsunami (without aid)
Health spending as a share of income (health share)	0.048	0.056	0.17
Catastrophic spending (health share > 0.1)	4.5%	18.3%	29.4%
Severe Catastrophic spending (health share > 0.25)	0	7.2%	16.1%
Catastrophic spending (>0.1) for poorest 20%	29.4%	23.2%	46.4%

Catastrophic spending (>0.1) for richest 20%	0	11.3%	17.7%

Note: All pre-tsunami values are based on data estimated using the hybrid Feedforward Neural Network –Light Gradient Boosting Machine model.

To estimate causal effects, we apply the empirical model discussed in section 3.3. The results (Table 4) show that aid played a significant role in mitigating out-of-pocket health spending. Without aid, households located in heavily and moderately damaged areas spent 5 and 8 percentage points, respectively, of their income on health compared to households in no damage areas (column 4). However, with aid, these shares fell to being no different from pre-tsunami spending in heavy damage areas and 1 percentage point in moderately damaged areas (column 1). The results are similar with catastrophic spending. Without aid, households located in moderately damaged areas were 6 percentage points more likely to experience catastrophic health spending (>10% of income) compared to households in undamaged areas, holding pre-tsunami controls constant. However, aid directly targeted to households halved this probability. Pre-existing wealth was a strong predictor of protection from catastrophic health spending while a history of high health spending and an agricultural livelihood increased the risk.

Another pattern that emerges is that households experiencing medium tsunami damage consistently faced a higher probability of catastrophic health spending compared to those in heavy damaged areas, with or without aid. However, the contrast should be interpreted with caution because the two groups may not be fully comparable populations post-tsunami. Survivor bias means those surviving in heavily damaged areas may be less injured than those in moderate damage areas, as it has been noted that those with severe injuries have a higher risk of drowning in tsunami floods (Morgan et al., 2005). This selection makes the two groups incomparable in their underlying health need. Moreover, the lower financial burden in heavy damage areas may have been due to unmeasured indirect assistance such as mobile

clinics and free services that were concentrated in the hardest-hit areas in the aftermath of the tsunami. This post-treatment confounding affects the comparability of observed financial outcomes across the two groups. Another point is that the massive influx of aid increased incomes significantly in heavy damage areas as noted in Himaz (2022) and thus inflated the denominator of the health share metric, mechanically reducing health. Together, these mechanisms suggest the observed pattern reflects the realised financial burden on survivors rather than the causal effect of damage on health, and that the mitigation of financial burden was more pronounced in heavy damage areas.

Table 4: Impact of tsunami on out-of-pocket health spending metrics

	(1) Health share (with aid)	(2) Catastrophic spending (>10%) (with aid)	(3) Health share (without aid)	(4) Catastrophic spending (>10%) (without aid)
Medium damage	0.01***	0.03**	0.08***	0.06***
	(0.00)	(0.01)	(0.02)	(0.02)
Heavy damage	-0.01	-0.03*	0.05**	0.01
	(0.00)	(0.02)	(0.03)	(0.02)
Pre-tsunami controls	YES	YES	YES	YES
Observations	4,645	4,645	4,577	4,645

Notes: "Health Share" is defined as out-of-pocket health spending as a share of total household income. "Catastrophic Spending" is an indicator for whether this share exceeds 10% or 25%, respectively. Columns 1–2 use total household income *excluding* aid; Columns 3–4 use income *including* aid. Results in columns 1 and 3 based on Ordinary Least Squares estimator while columns 2 and 4 are based on probit estimator. All regressions inverse probability weighted based on propensity scores for tsunami damage exposure. Robust standard errors in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

5. Discussion

This paper trains and tests machine learning models to predict baseline household health share in the context of the 2006 Yogyakarta earthquake in Indonesia and applied the most suitable model to estimate pre-event health share for Aceh and North Sumatra, Indonesia, in the context of the 2004 tsunami. The empirical findings using estimated pre-event tsunami health share reveal substantial tsunami-induced increases in catastrophic health spending—from 4.5% to 29.4% without aid—with targeted assistance mitigating approximately half this burden. It also shows that those located in moderately damaged areas experienced higher out-of-pocket health costs, possibly due to the concentration of aid in heavy damage areas.

The methodological approach, however, requires careful interpretation regarding cross-context transportability. Applying an earthquake-trained model to predict tsunami impacts across different Indonesian regions and time periods introduces inherent uncertainty. While formal assessments achieved 99.9% propensity score overlap and reasonable balance on theoretically central variables, the 64% overall covariate balance and substantial rural/farming differences (standardised mean difference > 0.50) indicate meaningful contextual variation. This represents a trade-off: accepting prediction uncertainty to enable causal analysis versus having no baseline counterfactual whatsoever.

Advancing this framework requires validation through additional events with complete longitudinal data before and after hazard events. However, such datasets are scarce especially in low- and middle- income settings. Recent advances in machine learning enhance transferability across events and regions through the creation and implementation of large-scale foundation models, trained on heterogeneous data sources (Bodnar et al., 2025). In that context, this work provides a scoping and proof-of-concept study that assesses how machine learning approaches can be leveraged for transferable representations of disaster health vulnerability. Establishing a proof-of-concept forms a necessary step towards future multi-hazard and multi-regional models and future work will

investigate the integration of additional data into the framework. For example, Indonesia's IFLS covers other natural hazards beyond Yogyakarta, especially low-impact, high-probability flooding events, enabling tests of whether earthquake-trained models generalise to other hazards, or whether multi-event training improves tsunami predictions. Such validation could establish further the potential to move from hazard-specific to general applicability.

Transferring results from one type of hazard event, such as an earthquake, to another, such as a flood, may be problematic because of differences in damage profiles and geographical patterns of impact. Future work using the wider IFLS data could incorporate formal uncertainty quantification and develop hazard-specific vulnerability functions to clarify when cross-hazard transportability holds versus when hazard mechanisms fundamentally alter health spending responses. Such work can also exploit administrative data, cross-sectional household surveys (such as the Living Standard Measurement Survey) or the Demographic Health Survey to construct pseudo-panels capturing information before and after the event for the training model, and corresponding data after an event for the context for which transportation is needed. In this case, the unit of assessment may not be the household, but the unit used in pseudo panel construction.

Practical guidance for future researchers seeking to estimate unobserved welfare outcomes in disaster-affected, data-scarce settings.

From the perspective of survey data, the method requires a large longitudinal household survey or pseudo panel with data before and after the hazard to test and train the machine learning model from context A and a comparable dataset from as close as possible context B for the event for which data is scarce. Until a larger pool of work is developed, better learning will be supported by creating quasi-experimental settings for both contexts A and B (i.e., construct valid hazard-affected treatment and control groups). Both contexts A and B require the construction of a wide range of identical variables to maximise model

performance. At least one of these variables needs to be an objective measure of appropriate hazard intensity (e.g., instrumental Mercalli index for earthquakes, flood depth, velocity or satellite-based intensity classification for tsunamis). Context A will have a total of n variables pre-hazard and the same variables measured post-hazard. If estimating a baseline outcome for context B is the goal, context B will have $n - 1$ pre-hazard variables and n variables post-hazard. The model trained and tested on context A will be applied to context B to estimate the missing variable, if the two populations are sufficiently similar on key observed characteristics.

From a survey design perspective, our method suggests the vital importance of household geolocation provision in surveys to recover relevant hazard intensity metrics from other sources such as climate, geophysical or satellite data. It also suggests the importance of including retrospective questions in post-disaster surveys for key characteristics such as assets, wealth, health status, and employment to enable unobserved data estimation.

From a machine learning perspective, the method suggests the use of approaches from a wide range of architectures including linear models, tree-based models, neural networks, and hybrid models. Adopting explainable artificial intelligence methods that identify the primary drivers of predictions allows users to assess model plausibility and reliability using contextual knowledge. Transparent reporting of predictive uncertainty is essential when such estimates inform causal inference and policy decisions.

Overall, this framework represents an initial step toward scalable, multi-hazard modelling of disaster-induced welfare impacts. While uncertainty remains inherent in cross-context transportability, structured validation and expanding multi-event datasets can progressively strengthen the reliability and policy relevance of this approach.

6. Conclusion

This study proposes a methodological framework that combines predictive machine learning with explainability methods and longitudinal household survey data to address a core economic measurement problem: how to reconstruct unobserved pre-event baseline data to enable credible welfare analysis in data-scarce settings.

When training the model, we paid particular attention to transparency and trustworthiness by using post-hoc artificial explainability methods such as Shapley. This allowed us to understand the features that drove the model's predictions. Quite apart from its usefulness in assessing of model reliability, the Shapley results can have a role in directly informing policy. For example, in our model of choice the Hybrid FNN-LGBM model, the drivers of positive out-of-pocket health spending shares were health condition, education levels and households' wealth (i.e., assets) rather than income. The impact of the latter was heterogeneous and non-linear. This heterogeneity in how income relates to health cost burden suggests that interventions prioritising health are best delivered in kind rather than as cash transfers. This is an area for further research.

The modelling was used to estimate baseline health shares for the tsunami context. Using this, we were able to glean some insights that would not otherwise have been possible. First, aid targeted directly to households in the context of wider recovery efforts, even if not directed towards health *per se*, can significantly mitigate out-of-pocket health spending increases post-event. While this reflects the effects of the massive influx of aid to heavy damage areas in the aftermath of the disaster, the consistently negative impacts on health costs under different specifications for moderate damage households highlight a second insight: Better coordination and distribution of aid focusing on areas around those of high devastation could have improved health spending outcomes for the wider affected areas, while leaving households with heavy damage no worse off than they were before the hazard event. The stronger apparent effect in moderate damage areas should, however, be

interpreted cautiously given survivor bias and the concentration of indirect aid in the hardest-hit areas, which limit the direct comparability of the two groups.

The approach introduced in this work carries significant implications for global efforts in disaster risk reduction that currently give insufficient weight to social vulnerability and microeconomic insights, especially in risk quantification efforts. Current catastrophe and climate risk models remain predominantly physical science-led, with economic measurement incorporated, if at all, using macro-level metrics. Part of the reason is data scarcity. Being able to estimate unobserved data opens a path to filling this gap.

While demonstrated here for health spending in Indonesia, the framework extends in principle to other welfare outcomes rarely measured with pre-event baselines — such as food security, educational expenditure, or labour market participation — and to contexts beyond high-impact, low-probability events, though realising this broader potential requires building a much larger pool of evidence across diverse hazard types and settings. The approach could equally support prediction of post-disaster outcomes where a pre-event baseline exists, and need not be confined to high-impact, low-probability events. The binding constraint is not the machine learning method itself but the availability of comparable datasets across contexts.

Appendix

Table A1: Variable definitions

Variable name	Definition	Mean, S.D and range	
		IFLS (2000 and 2007)	STAR (2004+2005)
anyquake	IFLS: =1 if district had any housing destruction STAR: =1 if household located in 'heavy' or 'medium' damage area	0.0833, 0.276, (0-1)	0.195, 0.396, (0,1)
Instrumental Modified Mercalli Intensity	IFLS: instrumental intensity using the Modified Mercalli intensity scale taken from Kirchberger (2017) who uses USGS ShakeMap. STAR: Tsunami exposure based on 18 self-reported survey items capturing proximity, direct harm, social impacts and psychological stress. Full list of items available in the STAR codebook.	2.238, 2.396 (0-8)	0.186, 0.1, (0, 1)
hhsize	number of members in the household	4.302, 1.86, (1-16)	4.522, 2.01, (1, 17)

transfers	monthly scholarships, insurance, aid in 2004 prices	1.22+04, 2.895+05, (0-1.917+07)	138014, 782390, (0, 74216664)
worse health	At least one household member reports worsening health over the past 12 months"	0.326, 0.468, (0-1)	0.236, 0.425, (0,1)
Area affected	=1 if Java 0 otherwise	0.613, 0.487, (0,1)	1, 0, (1, 1)
total income	monthly total household income per capita in 2004 prices	379812, 991234.191, (0 , 41770832)	184199.568, 363542.422, (0, 14885496)
primary	=1 if at least one household member educated at primary school level 0 otherwise	0.984, 0.126, (0,1)	0.989, 0.103, (0,1)
secondary	=1 if at least one household member educated at secondary school level 0 otherwise	0.848, 0.359 (0,1)	0.832, 0.373, (0,1)

tertiary	=1 if at least one household member educated at tertiary level 0 otherwise	0.242, 0.428, (0, 1)	0.221, 0.415, (0, 1)
asset	total household value in 2004 Rupiah	64007907.5, 126634114, (0, 2040814080)	33844077, 70301812, (0, 1858499968)
wealth index	Household wealth index	0.427, 0.146, (0, 0.923)	0.439, 0.186, (0, 1)
farming	=1 if household engaged in farming, 0 otherwise	0.233, 0.423, (0, 1)	0.483, 0.499, (0, 1)
health share	monthly health spending per household member in 2004 prices/monthly household income in 2004 prices	0.039, 0.095, (0, 0.989)	only for 2005 data: 0.167, 0.907, (0-31.44)
islam	=1 if Muslim, 0 otherwise	0.9, 0.3, (0, 1)	1, 0, (1, 1)
rural	=1 if rural, 0 otherwise	0.361, 0.48, (0,1)	0.727, 0.445, (0, 1)
area affected	IFLS: =1 if location is Java 0 otherwise STAR: =1 if Aceh or North Sumatra 0 otherwise	0.613, 0.487, (0,1)	1, 0, (1, 1)

closest neighbour	IFLS: =1 if ethnicity is Sunda 0 otherwise STAR: =0	0.205,0.404, (0,1)	0, 0, (0, 0)
furthest neighbour	IFLS: =1 if ethnicity is Acehnese 0 otherwise STAR: =0	0.001, 0.034, (0, 1)	0, 0, (0, 0)

Notes: STAR data for 2004 have been constructed using retrospective questions regarding the relevant variables asked in the 2005 survey. All variables are standardised independently by dataset when used in machine learning models.

Table A2, Panel 1: Raw prediction metrics for health expenditure proportion (health share).

Model	mean	std	min	25%	50%	75%	max	MSE	RMSE	MAE	MedAE
Actual	0.039	0.057	0	0.004	0.013	0.044	0.194	n/a	n/a	n/a	n/a
Siamese Self-Attention Transformer	0.021	0.008	0.002	0.016	0.02	0.024	0.069	0.004	0.06	0.019	0.009
Feedforward Neural Network	0.027	0.012	0.002	0.02	0.025	0.032	0.115	0.003	0.057	0.026	0.013
Linear Regression	0.038	0.013	-0.043	0.032	0.039	0.044	0.117	0.003	0.055	0.04	0.03
Random Forest	0.042	0.025	0.007	0.024	0.036	0.054	0.148	0.003	0.057	0.044	0.026
XGBoost	0.04	0.035	-0.026	0.016	0.033	0.057	0.231	0.004	0.06	0.044	0.024
LGBoost	0.039	0.029	-0.009	0.018	0.033	0.053	0.179	0.003	0.058	0.041	0.024
Hybrid FNN-LGBM	0.038	0.025	0.003	0.021	0.032	0.049	0.145	0.003	0.056	0.039	0.022
TabNet	0.042	0.027	-0.533	0.036	0.041	0.047	0.425	0.004	0.063	0.048	0.036
HybridLGBMTabNet2000	0.038	0.025	0.002	0.02	0.032	0.05	0.136	0.003	0.057	0.039	0.021

Notes: Linear regression achieved the lowest Root Mean Squared Error (RMSE) of 0.055, in health share units, indicating superior overall error minimisation, while the Hybrid FNN-LGBM balanced this with stronger median accuracy (Median Absolute Errot MedAE = 0.022 vs. 0.030 for linear regression). Mean Squared Error (MSE) values are reported in health share² while all other metrics are in health share units. See Panel 2 for quantile-specific performance.; missing values for 'Actual' indicate perfect alignment by definition. See Appendix Table 3 for full definitions and calculation methods for the performance metrics.

Table A2, Panel 2: Relative performance metrics and catastrophic spending detection across models.

Model	Relative MAE	Relative MedAE	Relative_Qloss025	Relative_Qloss075	Sensitivity (health share>10%)
Actual	0.0000	0.0000	0.0000	0.0000	.
Siamese Self-Attention					.
Transformer	0.9171	0.4383	0.3402	0.5769	
Feedforward Neural Network	0.9345	0.5057	0.3924	0.5421	0.9
Linear Regression	1.0387	0.7786	0.5149	0.5238	0.9
Random Forest	1.0542	0.7265	0.5461	0.5081	9.2
XGBoost	1.0886	0.7195	0.5519	0.5367	13.8
LGBoost	1.0505	0.7220	0.5236	0.5269	9.2
Hybrid FNN-LGBM	1.0047	0.6731	0.4977	0.5070	5.5
TabNet	1.1486	0.8832	0.5906	0.5580	0.0
HybridLGBMTabNet2000	1.0182	0.6511	0.5019	0.5162	4.6

Notes: The Siamese Self-Attention Transformer achieved the lowest relative QLoss at the 25th quantile (0.340), indicating superior accuracy for low-spending households, while the Hybrid FNN-LGBM showed balanced performance across quantiles (0.498–0.507). However, RandomForest and XGBoost showed superior sensitivity (9.2–13.8%) in detecting catastrophic spenders (health share>10%), despite higher

relative errors. Relative metrics (MAE/MedAE/QLoss) are unitless ratios (model error relative to baseline); sensitivity is reported as a percentage of correctly identified catastrophic spenders (health share > 10%). Missing values reflect model failures to predict above-threshold spending. All metrics derived from IFLS data. See Appendix Table 3 for full definitions and calculation methods for the performance metrics.

Table A3: Definitions and formulas for raw and relative model evaluation metrics for predicting health spending as a proportion of income.

Metric	Interpretation	Formula/Method of calculation
Raw Metrics		
Mean Squared Error (MSE)	Average squared deviation between predicted and actual health spending proportions	$1/n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Root Mean Squared Error (RMSE)	Typical prediction error in health share. For example, RMSE = 0.056 implies model predictions typically fall within ±5.6% of income. RMSE penalizes large errors more than MAE.	$\sqrt{1/n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$
Mean Absolute Error (MAE)	Average absolute prediction error. For example, MAE = 0.039 implies predictions deviate by 3.9% of income on average.	$1/n \sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$
Median Absolute Error (MedAE)	Typical (median) prediction error. For example, MedAE = 0.022 means half of predictions are within 2.2% of income. Less sensitive to outliers than MAE.	$median(\lvert y_i - \hat{y}_i \rvert)$

ttest of difference between actual and predicted values	Tests whether the average prediction bias is statistically different from 0. H_0: mean difference = 0	$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}$
Relative Metrics		
Relative MAE	Average prediction error as a percent of the full possible health share range (0 to 0.2).	$MAE * \frac{100}{max - min}$
Relative MedAE	Median prediction error as a percent of full health share range.	$MedAE * \frac{100}{max - min}$
Relative Quantile Loss (τ=0.25, 0.5, 0.75)	Evaluates asymmetric error tolerance (e.g., underestimates vs. overestimates). At τ = 0.5, equivalent to MedAE.	$\frac{1}{n\sum_{i=1}^{n}\rho_\tau(y_i - \hat{y}_i)} * \frac{100}{max - min}$ Where $\begin{cases} \tau.u, & if\ u \geq 0 \\ (\tau - 1).u, & if\ u < 0 \end{cases}$
Sensitivity health share>0.1	Ability of the model to correctly identify individuals spending >10% of income on health. Sensitivity = 100 indicates perfect detection of all high spenders.	$\frac{True\ Positives}{True\ Positives + False\ Negatives} * 100$ Where

		A True positive= case where $y_i > 0.1$ and $\hat{y}_i > 0.1$ A False Negative=$y_i > 0.1$ but $\hat{y}_i \leq 0.1$

Notes: $\hat{y}_i$, y_i are predicted and actual values for i^{th} data point, n=sample size; $\bar{d}$= prediction mean bias; s_d= sample standard deviation of difference; outcome range: health share∈[0,0.2]

Table A4: STAR Predictions for pre-tsunami health spending as a proportion of income

	mean	std	min	25%	50%	75%	max
Actual, post-tsunami	0.0563	0.0884	0	0	0.0168	0.0655	0.3275
Siamese Self-Attention Transformer	0.0296	0.0121	0.0003	0.0203	0.0291	0.0392	0.0852
Feedforward Neural Network	0.0315	0.0124	0	0.0239	0.0292	0.0378	0.1457
Linear Regression	0.0395	0.0117	-0.2204	0.0357	0.0403	0.0446	0.1285
Random Forest	0.0525	0.0229	0.0058	0.0345	0.0492	0.0682	0.1307
XGBoost	0.0436	0.0313	-0.0293	0.0213	0.0386	0.0612	0.2202
LGBoost	0.0559	0.0321	-0.0123	0.0302	0.052	0.0779	0.1857
Hybrid FNN-LGBM	0.0481	0.0263	-0.0033	0.0277	0.0448	0.0636	0.1462
TabNet	0.0456	0.012	0.0048	0.036	0.0463	0.054	0.1201
Hybrid LGBM-TabNet	0.0471	0.0254	-0.0009	0.0275	0.042	0.0618	0.148

Table A5: Impact of Tsunami Damage on Household Health Expenditure and the incidence of catastrophic spending in 2005 (full results).

	(1) Health share(with aid)	(2) Catastrophic (>10%) (with aid)	(3) Catastrophic (>25%) (with aid)	(4) Health share (no aid)	(5) Catastrophic (>10%) (no aid)	(6) Catastrophic (>25%) (no aid)
Medium damage	0.01***	0.03**	0.03***	0.08***	0.06***	0.05***
	(0.00)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Heavy damage	-0.01	-0.03*	0.01	0.05**	0.01	0.03
	(0.00)	(0.02)	(0.01)	(0.03)	(0.02)	(0.02)
Pre-tsunami controls						
Wealth	-0.03***	-0.13***	-0.08***	-0.03	-0.09**	-0.10***
	(0.01)	(0.04)	(0.03)	(0.07)	(0.04)	(0.04)
Rural residence	-0.01	-0.01	-0.01	0.03	0.02	0.02
	(0.00)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
Agricultural livelihood	0.00	0.02	0.02*	0.02	0.04**	0.04**

	(0.00)	(0.02)	(0.01)	(0.03)	(0.02)	(0.02)
Health spending share	0.45***			2.32***		
	(0.07)			(0.37)		
Catastrophic spending (>0.1)		0.06	0.03		0.10**	0.10**
		(0.04)	(0.03)		(0.04)	(0.04)
Constant	0.05***			-0.01		
	(0.01)			(0.03)		
Observations	4,645	4,645	4,645	4,577	4,645	4,645

Notes: Columns (1) and (4) report OLS results with inverse probability weighting based on propensity scores for tsunami damage exposure. Columns (2), (3), (5), and (6) report probit marginal effects from inverse probability weighted probit estimations. Health share refers to out-of-pocket health spending as a share of total household income. Catastrophic spending indicates health expenditure exceeding 10% or 25% of income. Columns 1–3 use income including aid; columns 4–6 exclude aid. All models control for pre-tsunami wealth, rural residence, and farming status; columns (1) and (4) also adjust for baseline health share, while columns (2), (3), (5), and (6) adjust for baseline catastrophic spending (>0.1). Robust standard errors in parentheses. Significance levels: *p<0.10, **p<0.05, ***p<0.01.

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Anindya, K., Ng, N., Atun, R., Marthias, T., Zhao, Y., McPake, B., van Heusden, A., Pan, T., & Lee, J. T. (2021). Effect of multimorbidity on utilisation and out-of-pocket expenditure in Indonesia: quantile regression analysis. *BMC health services research*, *21*(1), 427.

Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(8), 6679-6687. https://doi.org/10.1609/aaai.v35i8.16826

Bappenas, (2006). Yogyakarta and Central Java Natural Disaster: A Joint Report of BAPPENAS, the Provincial and Local Governments of DI Yogyakarta, the Provincial and Local Governments of Central Java, and International Partners. The 15th Meeting of the Consultative Group on Indonesia (CGI) Jakarta,

Bartels, S. A., & VanRooyen, M. J. (2012). Medical complications associated with earthquakes. *The lancet.*, *379*(9817), 748-757. https://doi.org/10.1016/S0140-6736(11)60887-8

Blaikie, P., Cannon, T., Davis, I., & Wisner, B. (2014). *At risk: natural hazards, people's vulnerability and disasters*. Routledge.

Bodnar, C., Bruinsma, W.P., Lucic, A. *et al.* A foundation model for the Earth system. *Nature* **641**, 1180–1187 (2025). https://doi.org/10.1038/s41586-025-09005-y

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/a:1010933404324

Bubonya, M., Cobb-Clark, D. A., & Wooden, M. (2017). Mental health and productivity at work: Does what you do matter? *Labour economics.*, *46*, 150-165. https://doi.org/10.1016/j.labeco.2017.05.001

Chen, T., & Guestrin, C. (2016, 2016). XGBoost.

Coyle, D., & Nakamura, L. I. (2019). Toward a framework for time use, welfare, and household centric economic measurement.

Craig, P., Katikireddi, S. V., Leyland, A., & Popham, F. (2017). Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. *Annual review of public health*, *38*(1), 39-56.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187-199. https://doi.org/10.1093/biomet/asn055

Dang, H.A., Jolliffe, D. and Carletto, C., 2019. Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments. *Journal of Economic Surveys*, *33*(3), pp.757-797.

Dang, H.-A. H., & Lanjouw, P. F. (2023). Regression-based imputation for poverty measurement in data-scarce settings. In *Research handbook on measuring poverty and deprivation* (pp. 141-150). Edward Elgar Publishing.

Dang, H. A. H., Kilic, T., Abanokova, K., & Carletto, C. (2025). Poverty Imputation in Contexts Without Consumption Data: A Revisit With Further Refinements. *Review of Income and Wealth*, *71*(1). https://doi.org/10.1111/roiw.12714

Deaton, A., & Paxson, C. (1998). Economies of scale, household size, and the demand for food. *Journal of political economy*, *106*(5), 897-930.

Dercon, S. (2004). *Insurance Against Poverty* (1 ed.). Oxford University Press. https://doi.org/10.1093/0199276838.001.0001

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality. *Econometrica*, *71*(1), 355-364. http://www.jstor.org/stable/3082050

Fattah, R. A., Cheng, Q., Thabrany, H., Susilo, D., Satrya, A., Haemmerli, M., Kosen, S., Novitasari, D., Puteri, G. C., Adawiyah, E., Hayen, A., Gilson, L., Mills, A., Tangcharoensathien, V., Jan, S., Asante, A., & Wiseman, V. (2023). Incidence of catastrophic health spending in Indonesia: insights from a Household Panel Study 2018–2019. *International Journal for Equity in Health*, *22*(1). https://doi.org/10.1186/s12939-023-01980-w

Frankenberg, E., Friedman, J., Gillespie, T., Ingwersen, N., Pynoos, R., Rifai, I. U., Sikoki, B., Steinberg, A., Sumantri, C., Suriastini, W., & Thomas, D. (2008). Mental Health in Sumatra After the Tsunami. *American Journal of Public Health*, *98*(9), 1671-1677. https://doi.org/10.2105/ajph.2007.120915

Frankenberg, E., Gillespie, T., Preston, S., Sikoki, B., & Thomas, D. (2011). MORTALITY, THE FAMILY AND THE INDIAN OCEAN TSUNAMI. *Economic Journal*, *121*(554), F162-F182. https://doi.org/10.1111/j.1468-0297.2011.02446.x

Frijters, P., Johnston, D. W., & Shields, M. A. (2014). The effect of mental health on employment: Evidence from Australian panel data. *Health economics.*, *23*(9), 1058-1071. https://doi.org/10.1002/hec.3083

García-Gómez, P., Kippersluis, H. v., O'Donnell, O., & Doorslaer, E. v. (2013). Long-Term and Spillover Effects of Health Shocks on Employment and Income. *The Journal of Human Resources.*, *48*(4), 873-909. https://doi.org/10.1353/jhr.2013.0031

Hallegatte, S., Vogt-Schilb, A., Rozenberg, J., Bangalore, M., & Beaudet, C. (2020). From Poverty to Disaster and Back: a Review of the Literature. *Economics of Disasters and Climate Change*, *4*(1), 223-247. https://doi.org/10.1007/s41885-020-00060-5

Himaz, R. (2022). Business recovery in Aceh and North Sumatra following the Indian Ocean Tsunami. *International journal of disaster risk reduction : IJDRR.*, *73*, 102868. https://doi.org/10.1016/j.ijdrr.2022.102868

Hirano, K., & Imbens, G. W. (2004). *The Propensity Score with Continuous Treatments*. Wiley. https://doi.org/10.1002/0470090456.ch7

Imbens, G. a. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kirchberger, M. (2017). Natural disasters and labor markets. *Journal of Development Economics*, *125*, 40-58. https://doi.org/10.1016/j.jdeveco.2016.11.002

Loughran, D. S., & Heaton, P. (2013). *Post-traumatic stress disorder and the earnings of military reservists : technical report*. RAND Corporation.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Morgan, O., Ahern, M. and Cairncross, S., 2005. Revisiting the tsunami: health consequences of flooding. *PLoS medicine*, *2*(6), p.e184.

O'Keefe, P., Westgate, K., & Wisner, B. (1976). Taking the naturalness out of natural disasters. *Nature*, *260*(5552), 566-567. https://doi.org/10.1038/260566a0

Pearl, J., & Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, *29*(4), 579-595. https://doi.org/10.1214/14-sts486

Rosenbaum, P. R., & Rubin, D. B. (1983). THE CENTRAL ROLE OF THE PROPENSITY SCORE IN OBSERVATIONAL STUDIES FOR CAUSAL EFFECTS. *Biometrika*, *70*(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Salmanidou, D. M., Ehara, A., Himaz, R., Heidarzadeh, M., & Guillas, S. (2021). Impact of future tsunamis from the Java trench on household welfare: Merging geophysics and economics through catastrophe modelling. *International Journal of Disaster Risk Reduction*, *61*, 102291.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, *25*(1), 1-21. https://doi.org/10.1214/09-sts313

Syukriyah, D., & Himaz, R. (2024). Short and medium-run effects of the Indian Ocean tsunami on health costs in Indonesia. *World development.*, *180*, 106648. https://doi.org/10.1016/j.worlddev.2024.106648

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Watson, J. T., Gayer, M., & Connolly, M. A. (2007). Epidemics after Natural Disasters. *Emerging Infectious Diseases*, *13*(1), 1-5. https://doi.org/10.3201/eid1301.060779

Wagstaff, A., Flores, G., Hsu, J., Smitz, M.-F., Chepynoga, K., Buisman, L. R., Van Wilgenburg, K., & Eozenou, P. (2018). Progress on catastrophic health spending in 133 countries: a retrospective observational study. *The Lancet Global Health*, *6*(2), e169-e179. https://doi.org/10.1016/s2214-109x(17)30429-1

World Bank. (2006). *Aceh Public Expenditure Analysis: Spending for Reconstruction and Poverty Reduction*. https://www.gfdrr.org/sites/default/files/publication/APEA.pdf

Xu, K., Evans, D. B., Kawabata, K., Zeramdini, R., Klavus, J., & Murray, C. J. L. (2003). Household catastrophic health expenditure: a multicountry analysis. *The lancet.*, *362*(9378), 111-117. https://doi.org/10.1016/S0140-6736(03)13861-5