Uncertainty-aware health risk prediction

Edmund Eric Gah · gahedmund146@gmail.com

1 · PROBLEM

what and why

Machine-learning models can achieve excellent predictive performance during development but become unreliable when the data distribution changes. In healthcare, this problem is particularly important because patient populations, disease prevalence, healthcare practices, and data-collection processes can differ across time, geography, and demographic groups. A model may therefore produce a highly confident prediction for a patient who is substantially different from the populations represented in its training data. Traditional evaluation metrics such as accuracy and AUROC may fail to reveal this problem because they measure aggregate performance rather than whether the model knows when its predictions are unreliable. Research question: When a health-risk prediction model encounters distribution shift, can uncertainty estimation reliably identify predictions that are likely to be incorrect, and can adaptation methods restore reliable performance? Central hypothesis: Uncertainty-aware models will be better able to identify unreliable predictions under distribution shift than conventional models, while appropriate adaptation and recalibration methods can recover a meaningful portion of lost predictive performance and calibration. Sub-questions:

- How severely does distribution shift degrade predictive performance and calibration?
- Can uncertainty estimates identify when the model is likely to fail?
- Can adaptation or recalibration methods improve reliability after the distribution changes?

2 · APPROACH

model / method, and why this one

Stage 1 — Establish Strong Baselines Develop and compare Logistic Regression, Random Forest, and XGBoost models. Logistic Regression provides a transparent statistical baseline; Random Forest provides a nonlinear ensemble comparison; and XGBoost provides a strong modern approach for structured/tabular health data. Stage 2 — Introduce Realistic Distribution Shifts • Temporal shift: train on earlier survey years and evaluate on later years.

distributions. The goal is to simulate deployment conditions in which a system encounters data that differs from its development distribution. Stage 3 — Uncertainty Estimation

A key question is whether high uncertainty actually corresponds to a higher probability of prediction error, especially under shifted conditions. Stage 4 — Adaptation and Recovery • Recalibration

Methods will be compared empirically rather than assuming that one technique is universally optimal. The analysis will examine the trade-off between predictive performance, calibration, uncertainty quality, and robustness under shift.

- Geographic shift: train on one set of geographic regions and evaluate on another where sample sizes permit.
- Demographic/population shift: evaluate performance across populations with different demographic and health-feature
- Predictive probability calibration
- Bootstrap or ensemble-based uncertainty where computationally appropriate
- Monte Carlo-based uncertainty where applicable
- Conformal prediction

- Selective prediction/abstention
- Importance/reweighting methods
- Threshold adaptation
- Fine-tuning or retraining using newly available data
- Distribution-aware adaptation

3 · DATASET

source, size, preprocessing

Primary Dataset: CDC Behavioral Risk Factor Surveillance System (BRFSS) The primary experimental dataset will be the Behavioral Risk Factor Surveillance System (BRFSS), a large-scale U.S. health survey containing demographic, behavioral, and health-related variables collected across multiple years. Its multi-year structure makes it useful for investigating temporal distribution shift rather than relying solely on random train/test splits. A subset or set of annual releases will be selected based on consistent availability of relevant variables, adequate sample sizes, comparable target definitions, and sufficient temporal separation between training and evaluation periods. The initial task will focus on diabetes-risk prediction, while the framework can be extended to another health outcome if mentor guidance suggests it. Proposed Data Structure Training: earlier survey years → Validation: intermediate years → Held-out test: later years Preprocessing • Data-quality assessment and missing-value analysis

Preprocessing will be designed so that information from future test distributions does not leak into model development. Subgroup sample sizes will also be examined before drawing conclusions about uncertainty or performance.

- Removal of irrelevant identifiers
- Feature selection based on the research question
- Appropriate encoding of categorical variables
- Standardization where required
- Class-imbalance analysis
- Temporal/geographic splitting
- Leakage prevention
- Reproducible train/validation/test construction

4 · METRICS

how you'll know it worked

Predictive Performance • AUROC

Calibration • Brier Score

A well-calibrated model should produce probabilities that correspond meaningfully to observed outcome frequencies. Uncertainty Quality • Coverage of prediction intervals/sets

A central experiment will test whether observations for which the model expresses high uncertainty are actually more likely to contain prediction errors. Robustness Under Distribution Shift Performance degradation = Performance(in-distribution) − Performance(shifted) Primary success criterion: uncertainty-aware methods should meaningfully distinguish reliable from unreliable predictions under distribution shift, while adaptation or recalibration methods recover a significant portion of lost performance without producing misleading confidence.

- AUPRC
- Accuracy

- Precision
- Recall/Sensitivity
- Specificity
- F1-score
- Expected Calibration Error (ECE)
- Reliability diagrams
- Calibration slope/intercept where appropriate
- Selective risk
- Risk-coverage curves
- Uncertainty-error correlation
- Out-of-distribution detection performance where applicable
- Failure-detection AUROC/AUPRC

5 · COMPUTE

`training time, memory, dependencies`

Expected environment: Python 3.x, pandas, NumPy, scikit-learn, XGBoost, SciPy, Matplotlib, Seaborn, SHAP where useful, Fairlearn where subgroup analysis is required, a conformal-prediction implementation such as MAPIE, Jupyter, and Git/GitHub. Estimated hardware: 4+ CPU cores; 8–16 GB RAM; GPU not required for core experiments; approximately 5–10 GB storage depending on dataset versions and experiment artifacts. Most tabular baseline experiments should run within minutes. More computationally intensive ensemble or repeated uncertainty experiments may require longer runtimes. Each experiment will record random seeds, model configuration, hyperparameters, training time, dataset version, software dependencies, and evaluation results. MLflow or a lightweight experiment-tracking system may be used.

6 · DELIVERABLES

`model, paper, demo`

1. Reproducible Research Repository • Data-processing pipeline

2. Experimental Model Suite • Conventional baseline models

3. Research Paper • Background and literature review

- Model-training code
- Distribution-shift experiments
- Uncertainty estimation methods
- Calibration analysis
- Evaluation scripts
- Experiment configurations
- Reproducibility documentation
- Calibrated models
- Uncertainty-aware models
- Distribution-adapted models
- Comparative evaluation artifacts

- Research questions and hypotheses
- Dataset and preprocessing
- Experimental methodology
- Distribution-shift construction
- Uncertainty methodology
- Results and statistical analysis
- Failure cases

7 · RISK

what breaks first, and plan B

what breaks first, and plan B 4. Interactive Research Demonstration A lightweight dashboard will demonstrate the difference between conventional and uncertainty-aware prediction. It may allow users to inspect predictions, uncertainty, model comparisons, distribution-shift effects, calibration, risk-coverage curves, and examples where the model is confidently wrong. It will be explicitly presented as a research prototype rather than a clinical diagnostic system. Risk 1 — Distribution shift is weaker than expected Plan B: Quantify naturally occurring shifts using statistical distribution-distance measures and construct controlled covariate-shift experiments while clearly distinguishing them from naturally occurring shifts. If necessary, introduce a second public dataset for external validation. Risk 2 — Uncertainty estimates are poorly calibrated Plan B: Compare multiple uncertainty approaches rather than relying on one method. Calibration will be treated as an experimental outcome rather than an assumption. Risk 3 — Conformal prediction loses coverage under distribution shift Plan B: Explicitly investigate how coverage deteriorates under shift and whether adaptive or weighted conformal approaches can improve it. This failure mode may itself become an important research finding. Risk 4 — Adaptation improves accuracy but worsens calibration Plan B: Require adaptation methods to be evaluated simultaneously on discrimination, calibration, and uncertainty rather than accuracy alone. Risk 5 — Dataset limitations prevent clinical interpretation Plan B: Frame the work as a methodological investigation of reliable ML under population and temporal distribution shift, using health-risk prediction as the application domain rather than claiming clinical validation. Findings can later be tested on a second dataset if resources and mentorship permit. Expected Research Contribution The central contribution will be an empirical framework for studying whether machine-learning confidence remains trustworthy when the data changes. Most conventional machine-learning evaluation asks, “How well does the model perform?” This research asks a more consequential question: “When the model encounters data unlike what it has seen before, does it know that it may be wrong?” By combining distribution-shift analysis, uncertainty quantification, calibration, and adaptation, the project aims to investigate a fundamental challenge in trustworthy AI: developing systems that can recognize the boundaries of their own competence. For healthcare applications, this distinction is particularly important. A model that occasionally makes an incorrect prediction may still be useful; a model that is confidently wrong and unable to recognize unfamiliar situations is considerably more difficult to trust. Ultimately, I hope this research contributes toward machine-learning systems that are not merely accurate under controlled benchmarks, but reliable, uncertainty-aware, and appropriately cautious when deployed in changing real-world environments. Prepared by Edmund Eric Gah • Research Proposal for Mentor Matching