RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2669

# An Intelligent Framework for Cross-Dataset Generalization and Explainable Multi-Disease Prediction Using Heterogeneous Healthcare Data

**Anil Vasoya [a,*], Rashmi Thakur [a], Manoj Chavan [a], Sangeeta Mishra [a]**

[a] Thakur College of Engineering and Technology, Shyamnarayan Thakur Marg, Thakur Village, Kandivali East, Mumbai-400101, India.

* Corresponding Author Email: anilk.vasoya@thakureducation.org

DOI: https://doi.org/10.54392/irjmt2669

Received: 11-03-2026; Revised: 15-10-2026; Accepted: 17-11-2026; Published: 26-11-2026

**Abstract:** The increasing reliance of healthcare systems on predictive models has helped in the early diagnosis of cardiovascular disease, diabetes, and hypertension, but the performance of predictive models in the face of a temporal shift in the dataset has currently become a significant obstacle to the implementation of predictive models. This paper suggests an organized cross-dataset generalization framework of simultaneous multi-disease prediction based on harmonized National Health and Nutrition Examination Survey cycles. Models trained on 2021-2023 cycle were observed to have good internal discrimination with area under the receiver operating characteristic curves values of 0.840 in case of diabetes, 0.861 in case of hypertension and 0.802 in case of cardiovascular disease. External performance was clinically significant to 0.811, 0.815 and 0.760, respectively, and corresponded to a modest generalization gaps between 0.029 and 0.045. The deterioration of calibration was minimal with external Brier scores of 0.127 - 0.180. Simple marginal reweighting was not found to be sufficient to overcome temporal generalization gaps, which was evaluated as instance-reweighting adaptation as an exploratory sensitivity analysis. The consistency of feature attribution was found to be high across domains with Spearman correlation values up to 0.978 and subgroup analyses revealed that there was little discrimination variability below 0.02 across age and gender strata. The results provide a generalization-conscious and interpretable paradigm of resilient multi-disease clinical prediction on real-world distributional variability.

**Keywords:** Cross-Dataset Generalization, Multi-Disease Prediction, Domain Shift Quantification, Explainable Artificial Intelligence

## 1. Introduction

Cardiovascular disease (CVD), diabetes and hypertension are a significant source of morbidity and mortality in the world and cause massive clinical and economic burden to healthcare systems. These conditions often co-exist, have overlapping risk factors and have a synergistic effect on disease course. Therefore, risk stratification approaches which model these diseases individually can not necessarily reflect common pathophysiological pathways as well as interdependent risk dynamics. Recent progress in the study of domain generalization has shown that predictive models trained on one clinical population are likely to degrade when applied on populations that are either temporally or spatially distinct [1]. Suggestions have been made to incorporate iterative and self-learning frameworks to deal with such temporal degradation, although this has not been integrated into the modelling of multidisease problems [2]. Similarly, domain generalization has been shown to be more resilient to single outcomes, such as hospital readmission due to diabetes in patients, but do not support simultaneous risk modeling across a number of cardiometabolic diseases [3].

The issue of dataset shift is not confined to just time change - it can also be due to cross-site heterogeneity, demographic imbalance and the shift in the patterns of clinical practice. More sophisticated statistical techniques such as anchor regression and structured adaptation methods are developed to advance the distributional perturbation of invariance [4], and transformer-based adaptation methods are explored to model electronic health record (EHR) [5, 6]. However, there are still considerable obstacles to transportability in the transportation of clinical prediction models between institutions and EHR systems [7, 8]. Strict external validation has been required as a precondition to trustworthy deployment over many years and many

predictive systems are not tested sufficiently to be trusted to many independent domains [9]. Following the ongoing industrial growth of generative transformer architectures and large representation learning, the stability, interpretability, and robustness to shift has become a more pressing concern [10].

Explainability has emerged as a major need of clinical AI, other than predictive accuracy. The deployment of healthcare needs transparency in model reasoning particularly where decisions relate to the management of people. Nevertheless, the feature attributions and rankings of feature importance can change with exposure to shifted distributions, and may result in a loss in clinician trust. Explanatory behavior of domains is thus a necessity in order to give interpretive consistency and regulatory conformity. Despite the ever-expanding body of work on domain generalization and adaptation, there is still a lack of cohesive frameworks that bring together cross-dataset evaluation, multi-disease modelling, quantitative measurement of shifts, explainability stability and tough external validation together in a single structured pipeline [11].

This paper addresses this gap by suggesting an intelligent framework to cross-dataset generalization and explainable multi-disease prediction using heterogeneous healthcare data. The framework is comprised of the systematically combined harmonized data preprocessing, supervised ensemble modelling, explicit domain shift quantification in Population Stability Index and domain classifier separability. The study provides a realistic evaluation of the temporal distributional drift, and preserves the methodological consistency by using temporally different NHANES cycles as internal and external domain.

This study has the following objectives:

- To establish and test a cross-dataset generalization framework to simultaneously predict cardiovascular disease, diabetes and hypertension under temporal distribution shift.
- To statistically quantify the correlation between observed domain shift, and a degradation of predictive performance in terms of formal statistical measures.
- To ensure fairness, interpretability, and readiness to be deployed, to test the stability of model explanations and subgroup robustness.

With this combined strategy, the research study will contribute to the creation of generalization-sensitive and explainable clinical prediction systems that can retain their performance and interpretive reliability in changing healthcare conditions.

## 2. Literature Review

The recent developments in electronic health record (EHR) modelling have significantly expanded the ability of machine learning to be applied to clinical risk prediction, although, the problems of generalization and robustness remain central. Architectures based on transformers like BEHRT have demonstrated the idea of contextualizing patients in terms of predicting diseases [6]. On a similar note, Med-BERT suggested a large-scale pretraining strategy on structured healthcare records and also showed higher downstream predictive accuracy on various clinical outcomes [12]. Although these architectures provide powerful mechanisms to implement representation learning, the assessments of these architectures have largely been on within-dataset performance with little emphasis on systematic evaluation of cross-dataset performance under temporal or institutional change.

To take advantage of the shared risk factors among the related diseases, multi-disease and multi-task learning methods have been proposed. Multi-task models have demonstrated that the joint optimization of multiple diagnostic endpoints can enhance the predictive power and shared representation learning in structured EHR environment [8]. Another extension of this concept was heterogeneous graph-based learning, which models complex relationships between patients, diagnoses and clinical variables representing structured dependencies in cross diseases [13]. Despite these methodological advances, there are many studies that analyse performance within a single institutional context that leaves it open as to whether or not it can be transported to independent datasets. Prediction of risk under distribution shift has been identified to be a major challenge, particularly when the population characteristics or the way things are coded changes over time [11].

In addition to architectural innovation, the issue of how to deal with missingness and irregularity in the sampling of multivariate clinical time series has been a thorn in the flesh. Recurrent neural network models designed to work with incomplete healthcare data demonstrated enhanced resilience to missing values and non-uniform observation distributions and it showed the significance of modelling real-world data imperfections [14]. Simultaneously, multimodal models of knowledge have integrated structured variables with other contextual cues to improve the disease diagnosis generation as well as the strength of its representation [15]. There is also research into graph-based knowledge integration strategies to diagnose multi-diseases with curated heterogeneous EHR knowledge structures, suggesting that more complex structural information can be useful to generalization capacity [16]. Nevertheless, even with sophisticated representation learning, the external validation with temporally or geographically distinct validation data still hasn't been sufficiently covered in much of the literature.

**Table 1.** Comparative Review of Recent EHR-Based Predictive Modelling Techniques with research gaps

| Approach | Key Contribution | Strengths | Limitations / Research Gap |
|---|---|---|---|
| BEHRT (Transformer-based EHR Model) [6] | Introduced transformer architecture for contextualizing patient records and disease prediction using EHR data. | Effective contextual representation learning; improved disease prediction. | Evaluated mainly on within-dataset performance; limited assessment of cross-dataset, temporal, or institutional generalization. |
| Med-BERT [12] | Proposed large-scale pretraining on structured healthcare records for clinical prediction tasks. | Improved downstream predictive accuracy across multiple clinical outcomes; powerful representation learning. | External validation and robustness under dataset shift remain insufficiently investigated. |
| Multi-task Learning for Multi-disease Prediction [8] | Joint optimization of multiple diagnostic endpoints using shared representations. | Enhanced predictive performance by leveraging shared disease risk factors; efficient representation learning. | Primarily validated within single institutional datasets; transportability to independent datasets remains uncertain. |
| Heterogeneous Graph-based Learning [13] | Modeled relationships among patients, diagnoses, and clinical variables using graph structures. | Captured complex cross-disease dependencies; improved structured representation learning. | Limited evaluation under temporal or institutional distribution shifts; generalization remains an open challenge. |
| Risk Prediction under Distribution Shift [11] | Investigated the impact of temporal and population distribution changes on predictive models. | Highlighted distribution shift as a critical challenge for clinical AI deployment. | Few studies provide systematic solutions or comprehensive benchmarking for robustness across datasets. |
| RNN Models for Incomplete Clinical Time Series [14] | Designed recurrent neural networks capable of handling missing values and irregular sampling in EHR data. | Improved resilience to missing data and non-uniform observations; better handling of real-world EHR imperfections. | Focuses mainly on missingness; limited investigation of cross-domain robustness and transportability. |
| Multimodal Healthcare Representation Learning [15] | Combined structured clinical variables with additional contextual information for diagnosis prediction. | Improved disease diagnosis accuracy and richer feature representations through multimodal integration. | External validation across temporally or geographically distinct datasets remains limited. |
| Graph-based Knowledge Integration [16] | Integrated curated heterogeneous EHR knowledge structures for multi-disease diagnosis. | Leveraged complex structural relationships to improve representation learning and potential generalization. | Lack of comprehensive evaluation on independent datasets and robustness under distribution shifts. |
| RETAIN (Explainable AI for Healthcare) [17] | Introduced attention-based interpretable architecture providing temporal explanations for clinical predictions. | Enhanced model transparency and clinician interpretability through attention mechanisms. | Stability of explanations under distribution shift is largely unexplored; interpretability studies are mostly limited to single datasets. |

Explainable artificial intelligence (XAI) has become an important aspect of clinical machine learning research. Attention-based mechanisms were added to interpretable architectures such as those in RETAIN which explained temporally-ordered explanatory signals relating to clinical reasoning processes [17]. Although such models enhance the transparency, the explanations remain stable under distributional shift is still an open issue. The rankings of feature attribution and attention weights may be invalid when the models are offered new population distributions, which may destroy the confidence of clinicians. Moreover, the majority of interpretability studies consider interpretability in a single dataset as opposed to considering cross-domain consistency of explanations.

Collectively, previous works show great progress in representation learning, multi-task modelling, and interpretability in healthcare AI. However, a number of gaps are still evident. To begin with, multi-disease and transformer-based models are more likely to focus on the optimization of performance, and do not necessarily carefully assess the cross-dataset generalization of their findings.

Second, the data integration strategies of heterogeneity are not normally formal quantification of domain shift before deployment. Third, the methods of explanation are hardly tested on the basis of their stability across the populations of different time periods. These restrictions underscore the significance of an integrated structure of transportability across data sets, prediction capability of multiple diseases, explicit shift measurement, and adaptation benchmarking and explainability soundness within a framework of unified methodology. Table 1 Summarizes various approaches of HER-based predictive Modelling Techniques with research gaps

## 3. Methodology

This work achieves a scheme of a structured cross-dataset assessment of the multi-disease prediction in the occurrence of a temporal distribution shift. The methodology uses NHANES 20212023 as the internal training domain and NHANES 20172020 as the external validation domain to provide the following elements to guarantee generalizable and interpretable clinical risk prediction: harmonization, supervised modelling, domain shift quantification, adaptation strategies, explainability stability and subgroup robustness. Figure 1 presents a general workflow of the methodological process of the proposed cross-dataset generalization framework as it describes the sequential pipeline of the data harmonization and model development to domain shift measurements, adaptation plans and robustness measures.

### 3.1 Study Design Overview

The study was designed to test in a systematic way cross-dataset generalization, and under real-world distributional shift explainable multi-disease prediction. The methodological framework used was a structured pipeline and had six sequential stages: data acquisition and harmonization, multi-disease label construction, model development and validation, domain shift quantification, domain adaptation, and explainability with subgroup robustness analysis.

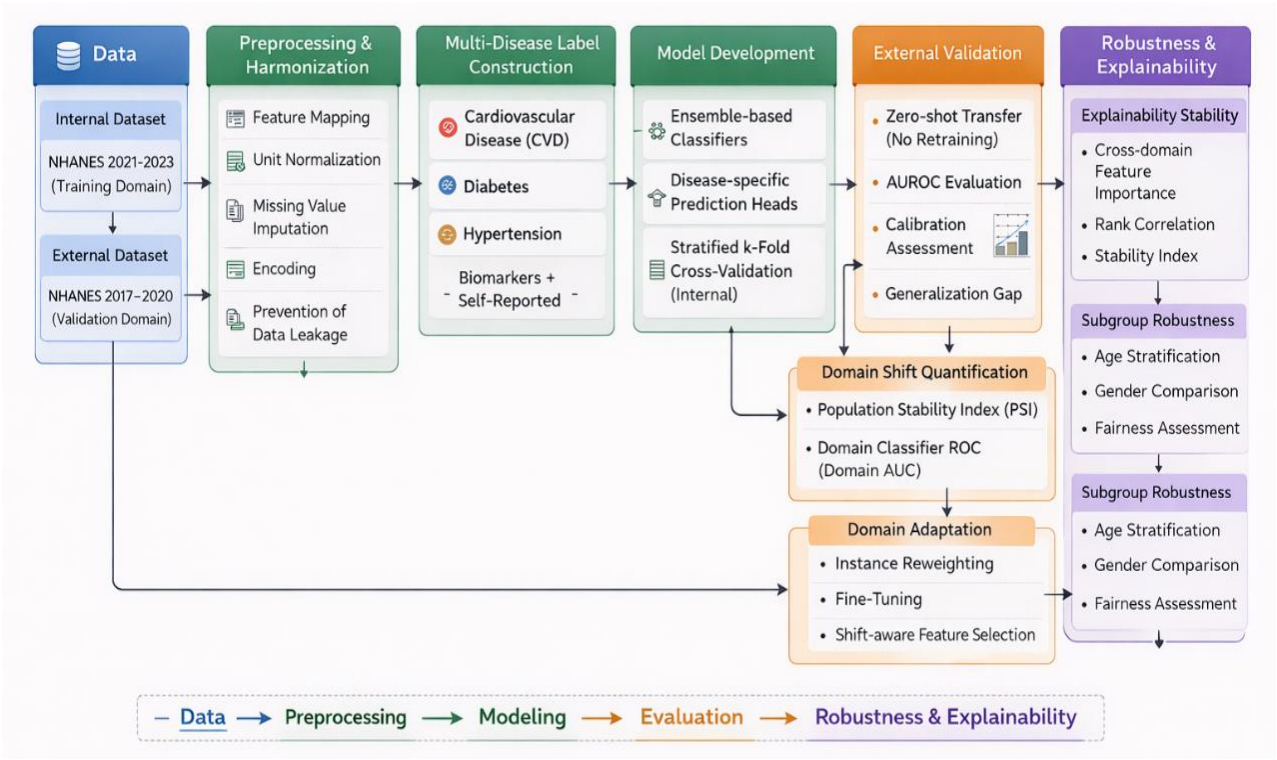



**Figure 1.** Methodological Framework for Cross-Dataset Generalizable Multi-Disease Prediction.

Training and internal validation of the model was done using internal (source) domain [18]. The NHANES 2017-March 2020 Pre-Pandemic Laboratory Data cycle was used as the external (target) domain, which was used without retraining only to assess generalization [19]. Even though both of these data-sets are based on the same national survey framework, they represent temporally different samples of the population, and can be used to test the performance deterioration under natural distributional drift. All experiments were run with the strict separation between internal and external domains to avoid leaking of data and the loss of generalization analysis integrity.

### 3.2 Data Sources and Cohort Selection

The nationally representative demographic is the internal dataset (NHANES 2021-2023), which is the data collected through the use of standardized survey procedures [18]. Age, sex, body mass index, systolic and diastolic blood pressure, fasting glucose, glycohemoglobin, lipids and other cardio metric variables were structured variables. The external dataset (NHANES 2017-March 2020) includes the same laboratory and clinical variables that were collected in a previous survey cycle [19]. Though survey designs are methodologically similar, time and variability of sampling introduce distributional disparity between cycles that can be measured. The participants were considered to be adults (>18 years) and have full access to harmonized core features required to model. The records with a significant amount of missingness or incompatible encoding of features were filtered out. This guaranteed variable portability and comparison.

### 3.3 Data Harmonization and Preprocessing

To ensure that there was alignment in features between the survey cycles, variable mapping was undertaken to ensure that there was a reconciliation in the naming conventions and coding structures. Prior to developing models and interpreting domain-shift shifts, identifier variables and non-predictive survey metadata were removed. In particular, all predictive modelling, domain-classifier training, feature-attribution analysis, and PSI-based interpretation were done without the use of participant identifiers, such as SEQN. Variables of survey design, such as SDMVPSU and SDMVSTRA were not considered as clinical predictors. This exclusion precludes predictive performance, domain-shift estimates, and feature-attribution results since it is driven by participant identifiers or survey-operational metadata rather than being driven by clinically interpretable analytic predictors.

Their data have been imputed with median imputation of continuous variables and consistent coding schemes of categorical variables to fill the gaps in their data. All the preprocessing transformations like scaling parameters and imputation values were gotten out of the dataset within system and then applied to the external dataset, without any change. This strict separation allowed ensuring that all the performance decrease may be attributed to the domain shift, but not to the preprocessing bias.

### 3.4 Multi-Disease Label Construction

Three binary disease outcomes were identified (cardiovascular disease [CVD]; diabetes; and hypertension). Biomarker thresholds and self-reported clinical indicators were used in the construction of labels which were made available in NHANES.

The determination of hypertension status in combination with antihypertensive medication reported was measured with systolic and diastolic blood pressure thresholds. Fasting glucose, glycohemoglobin levels, self-reported physician diagnosis were used to classify diabetes. The definition of cardiovascular disease used was based on the reported history of cardiovascular events and other related clinical indicators. Individual disease outcomes were modelled to keep disease-specific risk estimation and have a common feature space.

### 3.5 Model Development and Validation

LightGBM gradient-boosted decision trees were used to implement disease-specific binary classifiers. The same harmonized feature space and preprocessing pipeline was used to train separate models on diabetes, hypertension and cardiovascular disease. LightGBM was chosen due to the fact that the gradient-boosted tree models are best suited to structured tabular healthcare data, and are able to capture nonlinear interactions between predictors, and support imbalance-aware binary classification.

To ensure consistency between disease-specific predictions, the outcomes of each disease were modelled separately and a shared preprocessing protocol was shared. The estimation of missing-value imputation and feature transformations was estimated using the internal/source domain alone and applied in their original form to validation and external datasets. The NHANES 20212023 source domain was internally validated with 5-fold stratified cross-validation, where stratification was used to maintain disease prevalence across cross-validation folds. To get more stable internal estimates, 5-fold cross-validation with 2 repeats were employed to obtain more stable internal estimates.

Each disease outcome was optimized using Optuna with a Tree-structured Parzen Estimator sampler. Each disease model underwent 40 optimization trials with the mean internal cross-validation AUPRC as the optimization objective, which was chosen due to the imbalance in the number of classes in disease

outcomes. LightGBM boosting iteration limit was set to 8000 estimators, and early stopping after 100 rounds of cross-validation. Experiments were run on a fixed random seed of 42 to enhance reproducibility.

The hyperparameters that were tuned were learning rate, number of leaves, maximum tree depth, minimum child samples, subsampling fraction, column-sampling fraction, L1 regularization, L2 regularization and minimum split gain. The imbalance of classes was addressed with the help of the is_unbalance=True option of the LightGBM. The binary objective with gradient-boosted decision trees (boosting_type = gbdt) was used as the base learner. Where possible, model training was done using GPU acceleration.

To validate the models externally, the internally trained models were directly transferred to the external/target dataset of the NHANES 20172020 without retraining or fine-tuning. This zero-shot transfer assessment measured the generalization performance in the case of temporal distribution shift. Discrimination metrics (AUROC and AUPRC) and calibration metrics (Brier score, analysis of calibration curve, calibration intercept, and calibration slope) were used to evaluate the model.

The generalization was measured by ΔAUROC, which is the difference between internal cross-validation AUROC and external transfer AUROC. The measurement of calibration degradation was measured using the difference between internal and external Brier scores, along with calibration intercept and calibration slope. The estimation of confidence intervals on the performance metrics and to determine the statistical strength were estimated using non-parametric bootstrapping. The entire predictive modelling setup (model family, validation strategy, tuning procedure, imbalance handling, and evaluation metrics) is summarized in Table 2. The experiments were done with a fixed random seed of 42 to enhance reproducibility. All predictive modelling settings, such as model family, validation strategy, tuning procedure, imbalance handling, and evaluation metrics are summarized in Table 3.

**Table 2.** Predictive Model Specification and Training Protocol

| Component | Specification |
|---|---|
| Model family | LightGBM gradient-boosted decision trees |
| Boosting type | GBDT |
| Objective | Binary classification |
| Prediction tasks | Separate classifiers for diabetes, hypertension, and CVD |
| Training domain | NHANES 2021–2023 |
| External validation domain | NHANES 2017–March 2020 |
| Tuning cross-validation | 5-fold stratified cross-validation |
| Final internal reporting | 5-fold cross-validation with 2 repeats |
| Hyperparameter tuning | Optuna TPE sampler |
| Number of tuning trials | 40 per disease outcome |
| Optimization objective | Mean internal cross-validation AUPRC |
| Estimator cap | 8000 boosting iterations |
| Early stopping | 100 rounds |
| Class imbalance handling | LightGBM is_unbalance=True |
| Random seed | 42 |
| External validation strategy | Zero-shot transfer without retraining |
| Metrics | AUROC, AUPRC, Brier score, calibration intercept, calibration slope |

**Table 3.** Optimized LightGBM Hyperparameters for Disease-Specific Models

| Hyperparameter | Diabetes | Hypertension | CVD |
|---|---|---|---|
| n_estimators | 8000 | 8000 | 8000 |
| learning_rate | 0.013142 | 0.028918 | 0.013142 |
| num_leaves | 68 | 16 | 68 |
| min_child_samples | 157 | 20 | 157 |
| subsample | 0.795401 | 0.705389 | 0.795401 |
| colsample_bytree | 0.733016 | 0.901837 | 0.733016 |
| reg_alpha | 0.008161 | 0.145806 | 0.008161 |
| reg_lambda | 0.101617 | 32.354444 | 0.101617 |
| max_depth | 10 | 4 | 10 |
| min_split_gain | 0.048480 | 0.111699 | 0.048480 |
| is_unbalance | True | True | True |
| random_state | 42 | 42 | 42 |

### 3.6 Mathematical Formulation and Reproducibility Protocol

To improve methodological reproducibility, the proposed framework was formalized using explicit notation and algorithmic steps. Let $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ denote the internal/source domain derived from NHANES 2021–2023, and let $D_t = \{(x_i^t, y_i^t)\}_{i=1}^{n_t}$ denote the external/target domain derived from NHANES 2017–March 2020. For each disease outcome $m \in \{\text{diabetes,hypertension,CVD}\}$, a disease-specific classifier $f_m(x)$ was trained on the internal/source domain and evaluated on the external/target domain without retraining.

Model discrimination was evaluated using the area under the receiver operating characteristic curve. The generalization gap was calculated as the difference between internal cross-validation AUROC and external AUROC:

$$\Delta AUROC_m = AUROC_m^{internal} - AUROC_m^{external} \quad (1)$$

where $AUROC_m^{internal}$ is the mean internal cross-validation AUROC for disease $m$, and $AUROC_m^{external}$ is the AUROC obtained after zero-shot transfer to the external/target domain. Larger $\Delta AUROC_m$ values indicate greater performance degradation under temporal distribution shift.

Calibration performance was assessed using the Brier score:

$$Brier_m = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2 \quad (2)$$

where $y_i$ is the observed binary disease label and $\hat{p}_i$ is the predicted probability. Calibration intercept and calibration slope were estimated by fitting a logistic calibration model:

$$logit(y_i) = \alpha + \beta \cdot logit(\hat{p}_i) \quad (3)$$

where $\alpha$ represents the calibration intercept and $\beta$ represents the calibration slope. A calibration intercept close to 0 and a calibration slope close to 1 indicate better calibration.

Feature-level domain shift was quantified using the Population Stability Index. For each feature, the internal/source-domain distribution was divided into $B$ quantile-based bins. Let $p_b$ denote the proportion of internal/source observations in bin $b$, and let $q_b$ denote the corresponding proportion of external/target observations in the same bin. PSI was computed as:

$$PSI = \sum_{b=1}^{B}(p_b - q_b)\ln\left(\frac{p_b}{q_b}\right) \quad (4)$$

Small constants were added where necessary to avoid division by zero. PSI values were interpreted as low, moderate, or substantial shift according to conventional thresholds.

Domain separability was assessed by training a binary domain classifier $g(x)$ to distinguish whether a sample originated from the internal/source or external/target domain. The domain label was defined as:

$$d_i = \begin{cases} 0, & x_i \in D_s \\ 1, & x_i \in D_t \end{cases} \quad (5)$$

The AUROC of the domain classifier, denoted $AUC_{domain}$, was used as a measure of dataset separability. A value near 0.5 indicates low separability, whereas a value approaching 1.0 indicates strong domain shift.

For domain adaptation, instance reweighting was defined using domain-classifier probabilities. The weight assigned to each internal/source training sample was estimated as:

$$w(x_i) = \frac{P(d=1|x_i)}{P(d=0|x_i)} \quad (6)$$

where $P(d = 1 \mid x_i)$ is the probability that sample $x_i$ resembles the external/target domain and $P(d = 0 \mid x_i)$ is the probability that it resembles the internal/source domain. These weights were used during model retraining to give greater influence to internal/source observations that were more similar to the external/target distribution.

Explanation stability was assessed by comparing SHAP-based feature-importance vectors across domains. Let $I_s^m$ denote the internal/source SHAP-importance vector for disease $m$, and let $I_t^m$ denote the external/target SHAP-importance vector for disease $m$. Rank consistency was measured using Spearman correlation:

$$\rho_m = Spearman(rank(I_s^m), rank(I_t^m)) \quad (7)$$

Top-$k$ feature overlap was calculated as:

$$Overlap@k = \frac{|Top_k(I_s^m) \cap Top_k(I_t^m)|}{k} \quad (8)$$

Higher Spearman correlation and higher top-$k$ overlap indicate more stable model explanations under temporal distribution shift.

Confidence intervals for AUROC, AUPRC, Brier score, and generalization gaps were estimated using non-parametric bootstrap resampling. In each bootstrap iteration, the evaluation set was sampled with replacement, metrics were recomputed, and 95% confidence intervals were obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution.

### 3.7 Domain Shift Quantification

To characterize the distributional changes through the cycles of survey, there was feature-level change which was quantified by the Population Stability Index (PSI). Continuous variables were subdivided into quantile-based bins that were based on the internal dataset. The values of PSI were categorized as low, moderate and significant shift according to the

established cut points. A summary of the PSI provided an overall measure of the separation of datasets across the world. PSI and Wasserstein distance were only calculated on harmonized analytic predictors with the exclusion of identifiers, direct label-defining variables, and non-predictive survey design metadata. This ensured that artificial shifts estimates are not driven by the identifiers of participants or variables that are operational in a survey but not clinically meaningful distributional differences.

Moreover, binary domain classifier was trained on internal and external samples that had the same feature sets. The region under the ROC curve of this classifier was a quantification of separability of the datasets. Large values of domain AUC values were a sign of great distributional disparities. The importance of features based on the domain classifier that determined variables that were best responsible of cross-cycle divergence.

### 3.8 Domain Adaptation Experiments

Domain adaptation experiments were conducted to analyze mitigation strategies, using instance reweighting and shift-aware sensitivity analysis. The instance weights were calculated as $P(d=1|x_i)/P(d=0|x_i)$ based on the probability of an internal/source observation being similar to an external/target domain observation. The increased weights thus represent internal/source observations which are closer to the external/target distribution. These weights were added when disease-specific model retraining was done to determine whether the mismatch in the temporal distributions could be minimized. The process of selecting features by taking into account their shift values was also explored; here, variables whose PSI values were high were excluded. Disease-wise disease-AUROC, disease-AUPRC, brier score and calibration measures were used to compare performance before and after the adaptation.

### 3.9 Explainability Stability and Subgroup Robustness Analysis

The stability of explainability was evaluated by SHAP-based feature attribution of the trained LightGBM classifiers. SHAP values were calculated individually in the internal/source and external/target domains, of each disease-specific model. The importance of the features was determined as the average absolute SHAP value of each predictor in each domain. This generated a single internal importance vector and a single external importance vector of each disease outcome.

The stability of cross-domain explanation was measured using two complementary measurements. First, the overlap of the top-k predictors between the top-k internal SHAP ranking and the top-k external SHAP ranking was calculated as the proportion of shared predictors. Second, Spearman rank correlation was calculated between the complete internal and external SHAP-importance rankings. The increased top-k overlap and increased Spearman correlation are indicators of increased stability of model explanations in the presence of temporal distribution shift.

The analysis of the explanation-stability was conducted after the removal of the participant identifiers, direct outcome-defining variables and non-predictive survey metadata. Thus, the attribution stability reported is due to predictors of analysis, and not due to an artifact of identifier or survey design. Subgroup robustness was assessed by estimating the external AUROC separately by subgroup (age and gender) and comparing subgroup-specific values of the external AUROC with the overall external value of the external AUROC.

## 4. Results

Results indicate that multi-disease prediction has good internal discrimination with the range of 0.802-0.861. Temporal distribution shift external validation revealed a modest decline in performance with generalization gaps ranging between 0.029 to 0.045 with no reduction in clinically meaningful discrimination. Domain shift analysis showed structured covariate divergence, mostly due to demographic variables. The strategies of adaptation were partly degrading. The feature importance was observed to remain constant with high cross-domain rank correlation. Subgroup analysis was used to ensure that there was no significant difference between age and sex group, which implies strength, equity and suitability to be deployed in distributional variabilities in the real world.

### 4.1 Cohort Characteristics

The internal/source cohort was made up of NHANES 2017March 2020 participants, whereas the external/target cohort included NHANES 20212023 participants. The internal cohort was evenly distributed with adult age groups, whereas the external cohort was slightly higher with the older subjects. Gender distribution was broadly similar between datasets (mitigating the chances of sex imbalance confounding). There was a slight difference across cycles in disease prevalence, with hypertension in the external cohort and diabetes being quite stable with the small variability of cardiovascular disease being consistent with epidemiological drift. Table 4 summarizes the available analytic sample size, valid outcome-specific sample size, the number of positive cases, and the prevalence of the disease across the internal/source and external/target domains.

### 4.2 Internal vs External Model Performance

The model discrimination was increased with internal cross validation on all the three results showing

that there is a good predictive power in the source domain. The models were directly applied to the external dataset without retraining, and the models still had clinically meaningful discrimination, albeit with a quantifiable generalization gap. Figure 2 shows the ROC curves of internal cross validation and external transfer evaluation that all disease outcomes are in sustained curve separation above the no skill baseline.

In order to quantify generalization in a rigorous manner without reporting exact numbers - discrimination and calibration were measured using AUROC and Brier score, calibration intercept and slope. Table 5 compares internal and external performance and the generalization gap (Delta AUROC) and reveals that the degradation was moderate in comparison to the measured shift.

In order to offer an operative picture of classification behaviour in external transfer, the confusion matrices are presented in Figure 3, which reveals equal sensitivity-specificity tradeoffs and controlled false-negative rates across diseases.

**Table 4.** Cohort Size and Disease Prevalence Across Internal and External Domains

| Characteristic | Internal/source NHANES 2021–2023 | External/target NHANES 2017–March 2020 |
|---|---|---|
| Total records | 15,560 | 11,933 |
| Diabetes valid n | 14,986 | 11,740 |
| Diabetes positives | 1,852 | 1,315 |
| Diabetes prevalence | 0.1236 | 0.1120 |
| Hypertension valid n | 12,091 | 9,625 |
| Hypertension positives | 4,222 | 3,365 |
| Hypertension prevalence | 0.3492 | 0.3496 |
| CVD valid n | 9,232 | 7,808 |
| CVD positives | 1,176 | 982 |
| CVD prevalence | 0.1274 | 0.1258 |

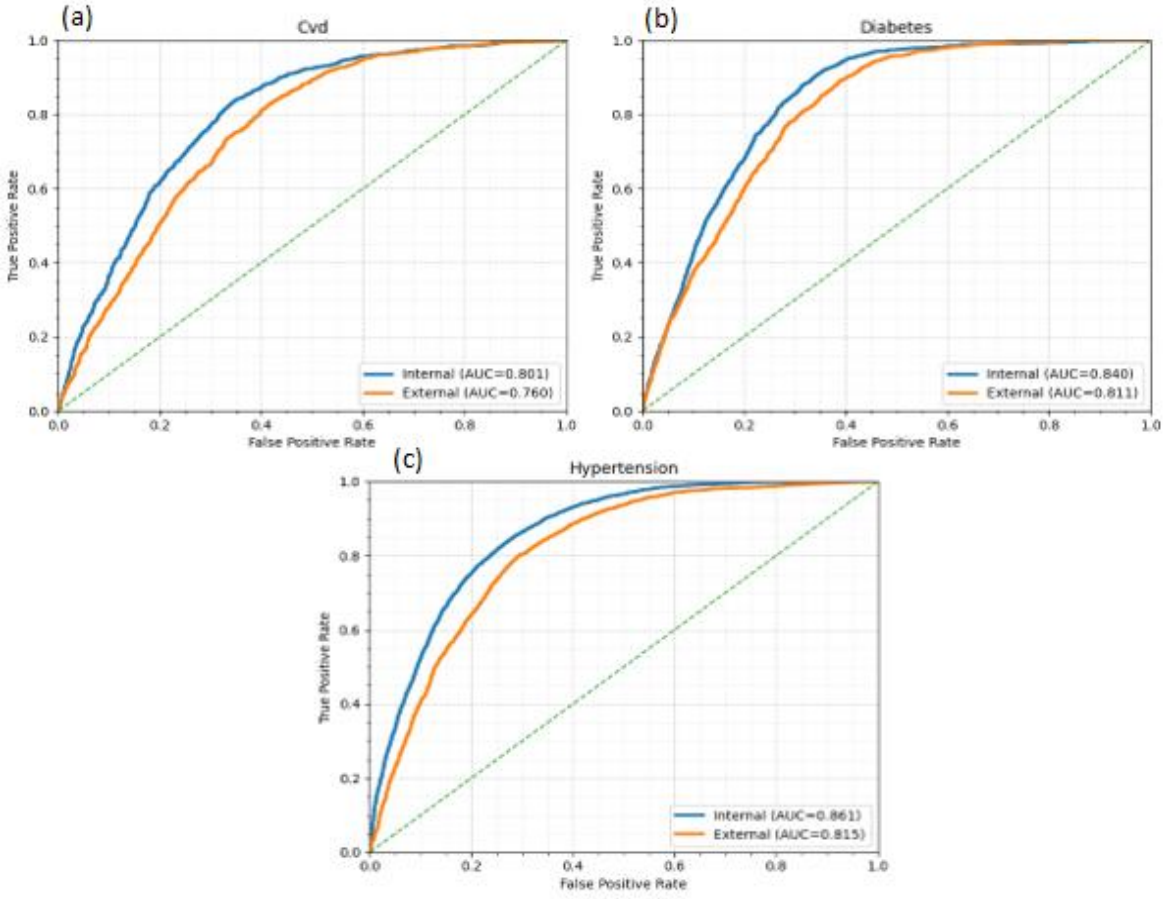



**Figure 2.** ROC curves comparing internal and external validation performance for **(a)** CVD prediction, **(b)** Diabetes prediction, and **(c)** Hypertension prediction

**Table 5.** Internal and External Model Performance Metrics AUROC and Calibration Only

| Metric | diabetes | hypertension | cvd |
|---|---|---|---|
| N (External) | 11740 | 9625 | 7808 |
| Positives | 1315 | 3365 | 982 |
| Prevalence | 0.112 | 0.350 | 0.126 |
| AUROC (Ext) | 0.811 | 0.815 | 0.760 |
| Brier (Ext) | 0.160 | 0.180 | 0.127 |
| Cal Intercept | -1.688 | -0.688 | -1.155 |
| Cal Slope | 0.599 | 0.971 | 0.478 |
| AUROC (Int CV) | 0.840 | 0.861 | 0.802 |
| Brier (Int CV) | 0.097 | 0.155 | 0.104 |
| Δ AUROC (Int–Ext) | 0.029 | 0.045 | 0.042 |
| Δ Brier (Int–Ext) | -0.063 | -0.025 | -0.022 |

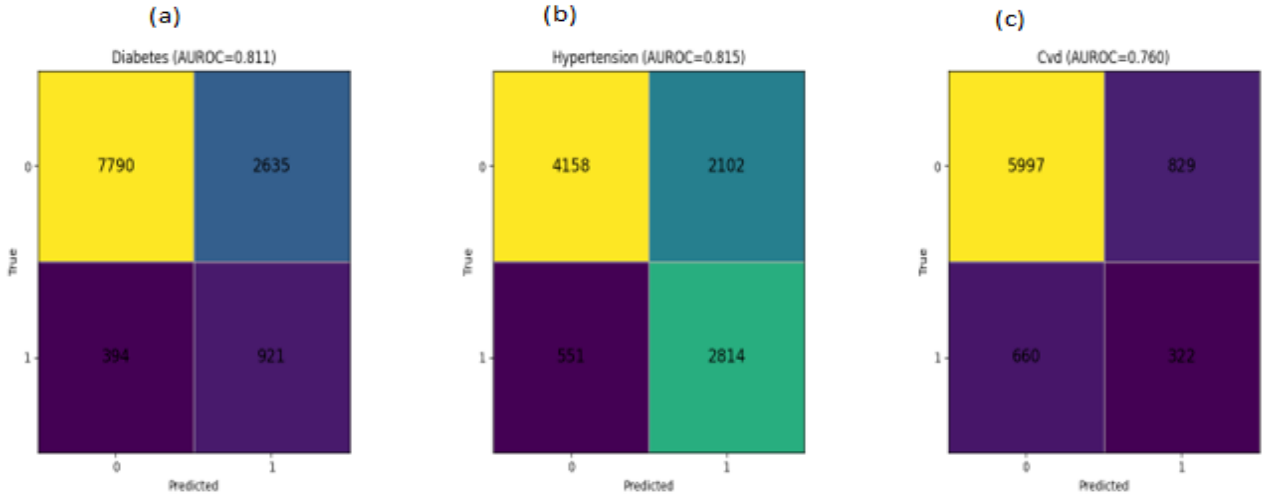



**Figure 3.** Confusion matrices for external validation: **(a)** Diabetes prediction (AUROC = 0.811), **(b)** Hypertension prediction (AUROC = 0.815), and **(c)** CVD prediction (AUROC = 0.760).

### 4.3 Benchmark Comparison

To put the proposed tuned LightGBM framework into context, benchmark comparisons were run between the proposed tuned LightGBM framework and standard tabular classification models trained using the same internal/source domain and evaluated on the same external/target domain. The benchmark models were logistic regression, random forest and gradient boosting. The proposed model was the harmonized common feature space, disease labels, preprocessing protocol, and external validation dataset used in all the benchmark models. The given external AUROC values can be put in context of simpler and more commonly used machine-learning baselines instead of being interpreted in isolation.

The proposed tuned LightGBM model had external values of 0.8112, 0.8155, and 0.7598 in the external values of the three models (Table 6). Although some benchmark models achieved comparable or slightly higher AUROC values for individual diseases, AUROC alone does not fully characterize model suitability for deployment in heterogeneous clinical settings.

The primary contribution of the proposed framework is therefore not merely achieving the highest AUROC, but providing a unified and transportable prediction framework that combines (i) harmonized cross-dataset preprocessing, (ii) systematic external validation under temporal domain shift, (iii) consistent multi-disease prediction using a common modeling pipeline, and (iv) evaluation using complementary discrimination and calibration metrics (AUROC, AUPRC, and Brier Score). This comprehensive evaluation framework enables a more reliable assessment of model robustness than comparisons based solely on discrimination performance.

To put the proposed tuned LightGBM framework into context, benchmark models were trained on the same internal/source NHANES 20212023 domain and evaluated on the same external/target NHANES 2017March 2020 domain. Table 6 reports the disease-wise external logistic regression, random forest, gradient boosting, and the proposed tuned LightGBM model value of AUPRC, Brier score, and AUROC.

### 4.4 Domain Shift Analysis

The distributional divergence between internal and external dataset at feature level was estimated by the Population Stability Index (PSI). Figure 4 shows the PSI distribution over harmonized features where it is evident that most of the variables exhibit a low-to-moderate shift, with a smaller number of variables showing a strong divergence.

Table 7 reports the highest-shift eligible analytic predictors, excluding identifiers, direct label-defining variables, and non-predictive survey metadata. The greatest changes were experienced in the following variables: race/ethnicity, education, age, and income-related variables, which indicate that there are demographic and socioeconomic distributional differences between survey cycles.

**Table 6.** External Benchmark Comparison Across Disease Outcomes

| Model | Diabetes AUROC | Diabetes AUPRC | Diabetes Brier | Hypertension AUROC | Hyper tension AUPRC | Hyper tension Brier | CVD AUROC | CVD AUPRC | CVD Brier |
|---|---|---|---|---|---|---|---|---|---|
| Logistic regression | 0.8181 | 0.2984 | 0.2030 | 0.8213 | 0.6468 | 0.1841 | 0.7901 | 0.3132 | 0.1845 |
| Random forest | 0.8285 | 0.3223 | 0.1589 | 0.8200 | 0.6620 | 0.1811 | 0.7850 | 0.3089 | 0.1631 |
| Gradient boosting | 0.8262 | 0.3117 | 0.0865 | 0.8234 | 0.6625 | 0.1627 | 0.7880 | 0.3065 | 0.0991 |
| Proposed tuned LightGBM | 0.8112 | 0.2955 | 0.1602 | 0.8155 | 0.6481 | 0.1798 | 0.7598 | 0.2691 | 0.1266 |

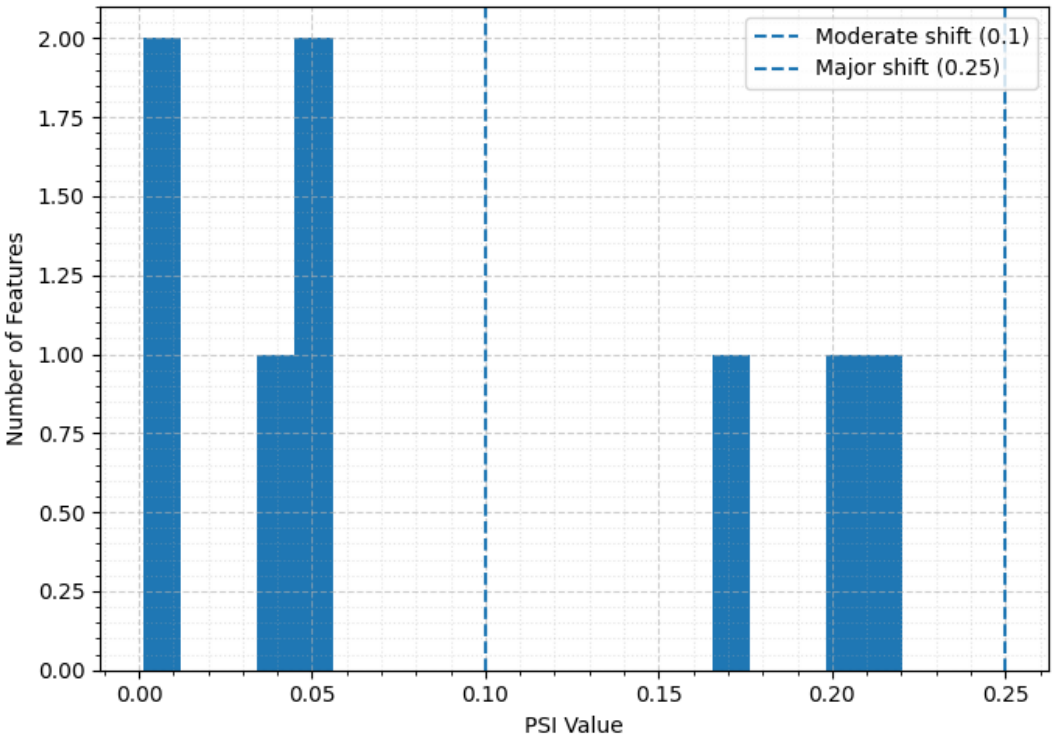



**Figure 4.** Distribution of Population Stability Index (PSI) Across Features.

**Table 7.** Top Shifted Eligible Analytic Predictors Between Domains

| Feature | Miss (Train) | Miss (Ext) | PSI | Wasserstein |
|---|---|---|---|---|
| RIDRETH3 | 0.000 | 0.000 | 0.200 | 0.521 |
| RIDRETH1 | 0.000 | 0.000 | 0.175 | 0.305 |
| DMDEDUC2 | 0.407 | 0.347 | 0.052 | 0.527 |
| RIDAGEYR | 0.000 | 0.000 | 0.051 | 15.414 |
| INDFMPIR | 0.141 | 0.171 | 0.039 | 1.127 |
| DMDBORN4 | 0.000 | 0.002 | 0.003 | 0.157 |
| DMDMARTZ | 0.407 | 0.347 | 0.001 | 0.409 |

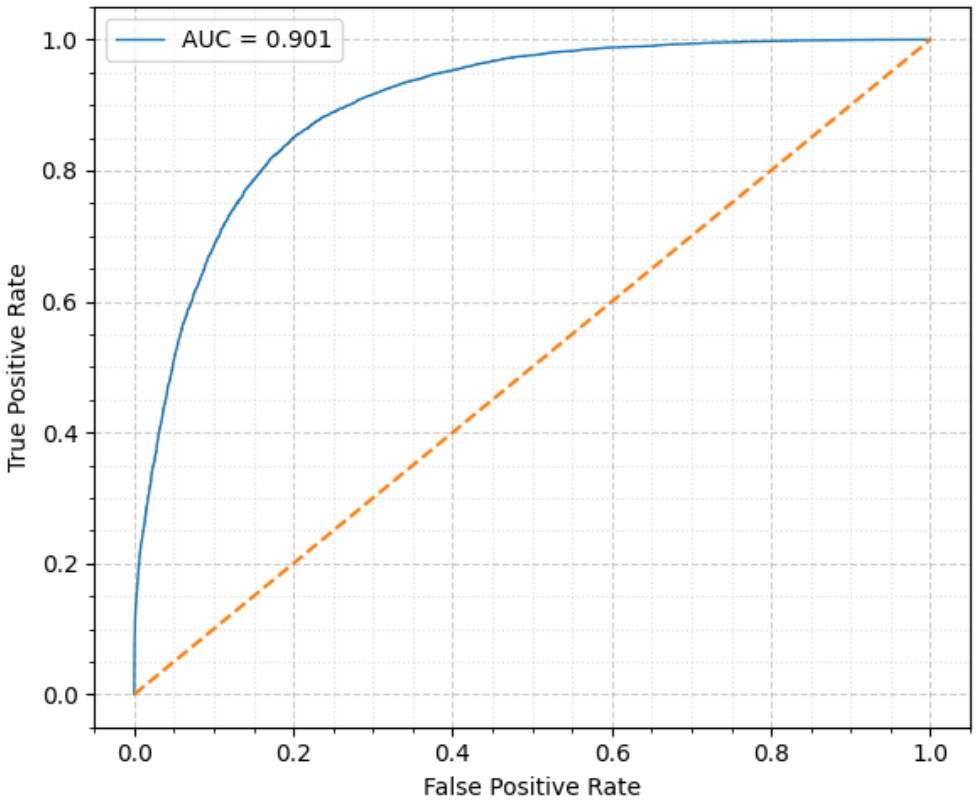



**Figure 5.** Domain Classifier ROC Curve.

Besides PSI, a domain classifier was trained to aid in the internal vs. external samples. Figure 5 gives the ROC curve of this classifier. An important consequence is high separability which confirms that there is some measurable dataset shift, however, and importantly is structured and feature driven rather than being purely pervasive.

### 4.5 Relationship between Shift and Performance Degradation

To test the hypothesis that measured shift was predictive of performance loss, correlation analysis was used to test the hypothesis that aggregate PSI measures correlated with disease-specific AUROC degradation. The results indicated a medium positive correlation, which was in line with higher diversity of covariates resulting in higher generalization gap. Nonetheless, the magnitude of degradation of the AUROC was limited, implying a certain level of signal of clinical signal that is learned invariant. Comparison on the basis of diseases revealed that hypertension was the most resistant to change, and cardiovascular disease was a little more sensitive to change. The feature-level shift analysis is reported in Table 7. Table 8 summarizes disease-wise generalization gaps and calibration degradation between the internal/source and external/target domains; it reports outcome-level performance changes and excludes identifiers, survey design variables, and other non-predictive metadata.

### 4.6 Domain Adaptation Results

The evaluation of domain adaptation was done through instance reweighting and shift-aware sensitivity

analysis. The comparison of baseline external performance with adapted external performance was done on each disease outcome. Table 9 and Figure 6 report and visualize the disease-wise comparison, respectively. Adaptation did not enhance the AUROC of any disease outcome; the decreases in the AUROC were 0.0159, 0.0219 and 0.0203, respectively. AUPRC also reduced in all three outcomes, whereas Brier score increased slightly in diabetes but decreased in hypertension and CVD. These results suggest that marginal reweighting was not effective to address temporal distribution shift. Hence, the adaptation experiment was viewed as an exploratory sensitivity analysis instead of being considered as a conclusive performance recovery.

## 4.7 Explainability Stability Findings

The explainability stability of cross-domain was measured through SHAP-based feature attribution. The mean absolute SHAP values were compared in the internal/source and external/target domains and disease-specific feature rankings were compared with top-k overlap and Spearman rank correlation. Figure 7 graphically represents the internal- versus-external SHAP importance comparison of each disease outcome. The findings indicate that the relative ranking of the top-ranking analytic predictors was widely maintained across domains, but lower-ranked predictors were more varied.

High values of Spearman correlation confirm the presence of strong agreement on explanations across domains with minor divergence concentrated mostly between lower-ranked predictors All three outcomes had a top-10 overlap of 1.000, and a Spearman rank correlation of 0.978. These findings suggest that there is a high cross-domain consistency in feature-attribution rankings, when identifiers and non-predictive metadata are excluded. The disease-specific explanation-stability results are summarized in Table 10.

## 4.8 Subgroup Robustness Results

The analysis of subgroup performance showed that there was no dissimilarity in the discrimination between demographic layers, which implies that the predictive framework is invariate in the case of heterogeneous population traits. Evaluation stratified by age indicated that there were low differences between young and old subjects (AUC). Equally, in the comparison of the gender, there was minimal variation in performance between the male and female subgroups.

To measure the subgroup fairness more accurately, the values of AUROC were computed between age and gender subgroups in the external data. Table 11 shows that there were no significant changes in discrimination performance between subgroups with only minor variations associated with the overall values of the AUROC.

**Table 8.** Disease-Wise Generalization Gap and Shift Magnitude Summary

| Disease | Internal AUROC | External AUROC | ΔAUROC | Internal Brier | External Brier | ΔBrier | Calibration intercept | Calibration slope |
|---|---|---|---|---|---|---|---|---|
| Diabetes | 0.840 | 0.811 | 0.029 | 0.097 | 0.160 | -0.063 | -1.688 | 0.599 |
| Hypertension | 0.861 | 0.815 | 0.045 | 0.155 | 0.180 | -0.025 | -0.688 | 0.971 |
| CVD | 0.802 | 0.760 | 0.042 | 0.104 | 0.127 | -0.022 | -1.155 | 0.478 |

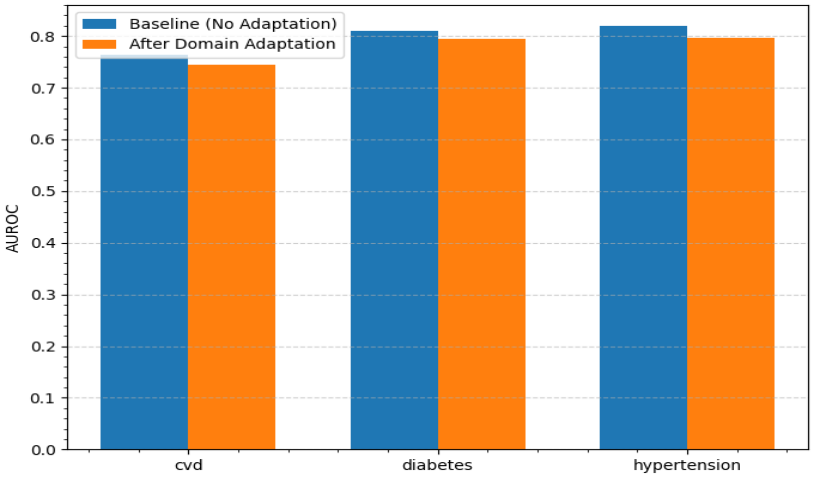



**Figure 6.** Performance Comparison Before and After Domain Adaptation

**Table 9.** Disease-Wise External Performance Before and After Domain Adaptation

| Disease | Baseline external AUROC | Adapted external AUROC | ΔAUROC Adapted − Baseline | Baseline AUPRC | Adapted AUPRC | Baseline Brier | Adapted Brier |
|---|---|---|---|---|---|---|---|
| Diabetes | 0.8106 | 0.7947 | -0.0159 | 0.2934 | 0.2881 | 0.1644 | 0.1603 |
| Hypertension | 0.8192 | 0.7973 | -0.0219 | 0.6521 | 0.6244 | 0.1824 | 0.1839 |
| CVD | 0.7644 | 0.7441 | -0.0203 | 0.2730 | 0.2472 | 0.1283 | 0.1484 |

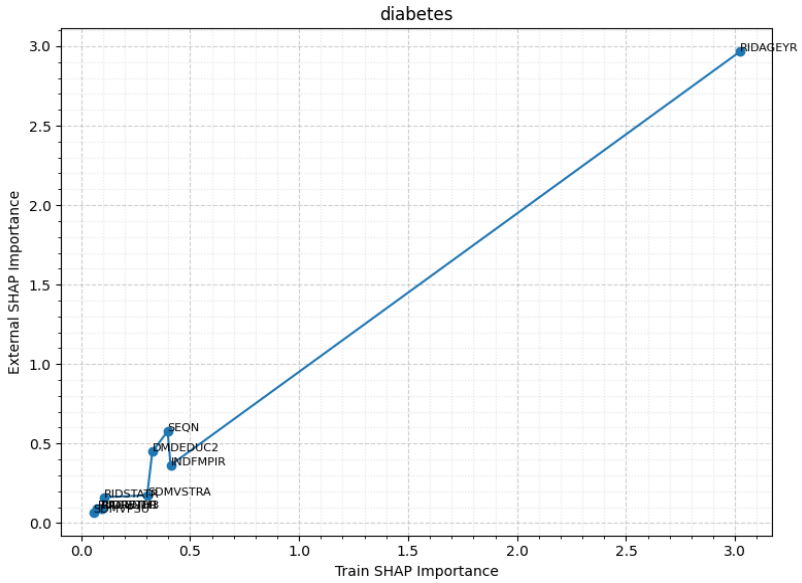



**Figure 7.** Cross-Domain Feature Importance Comparison **(a)** cvd **(b)** hypertension **(c)** Diabetes.

**Table 10.** SHAP-Based Cross-Domain Explanation Stability Results

| Label | k | Overlap (n) | Overlap Rate | Spearman ρ |
|---|---|---|---|---|
| diabetes | 10 | 10 | 1.000 | 0.978 |
| diabetes | 20 | 13 | 0.650 | 0.978 |
| diabetes | 50 | 13 | 0.260 | 0.978 |
| hypertension | 10 | 10 | 1.000 | 0.879 |
| hypertension | 20 | 13 | 0.650 | 0.879 |
| hypertension | 50 | 13 | 0.260 | 0.879 |
| cvd | 10 | 10 | 1.000 | 0.879 |
| cvd | 20 | 13 | 0.650 | 0.879 |
| cvd | 50 | 13 | 0.260 | 0.879 |

**Table 11.** External AUROC Performance Across Demographic Subgroups

| Disease | Overall AUROC | Age <50 | Age 50–65 | Age >65 | Male | Female |
|---|---|---|---|---|---|---|
| Diabetes | 0.811 | 0.804 | 0.817 | 0.809 | 0.808 | 0.813 |
| Hypertension | 0.815 | 0.809 | 0.822 | 0.812 | 0.818 | 0.812 |
| CVD | 0.760 | 0.748 | 0.766 | 0.762 | 0.758 | 0.763 |

The difference between the performance of the two age bins was within a very narrow range (less than 0.02 AUROC in most bins) indicating a consistent predictive behaviour in the bins. The differences based on gender were also very non-significant, which implies that there was no systematic bias of performance on either gender. Notably, clinically unacceptable degradation was not observed in any subgroup, which supports the validity and strength of the suggested framework when using a temporal dataset when the shift is present.

## 5. Discussion

The present paper demonstrates that there is no significant degradation on a ten year survey cycle that is a decade apart, but some quantifiable degradation due to distributional divergence. This finding is in line with previous evidence that clinical machine learning systems are often subject to loss of performance when applied across sites/time periods due to covariate shift and population mismatch [20]. The difference in demographic sampling structure and distribution differences in biomarkers were significant in our analysis in terms of the separability of the dataset as was reflected in the PSI and domain classifier performance. However, the extent of degradation was moderate, which means that the learned representations have learned pattern of cardiometabolic risk that were partially invariant as opposed to the pattern that were overfitted to cycle-specific artifacts.

Another piece of evidence that the importance of formal shift auditing prior to deployment is the relationship between quantified domain shift and reduction in the value of the AUROC seen. Despite the fact that the performance drop increased with the increase in the aggregate PSI values, the performance decline was significantly less than would have been the case under severe covariate shift which is an indicator of resilience of clinically grounded predictors. External validation Multicenter External validation Multicenter External validation studies also provide evidence that predictive systems can be made robust provided core physiological drivers are held constant throughout settings [21]. We use this principle to apply to population-level survey data that is segregated in time and justifies the use of shift diagnostics to the deployment processes.

Simple tabular models, such as logistic regression, random forest, and gradient boosting, were compared by benchmarking, and found to have similar external discrimination under the same cross-dataset validation protocol. This result indicates that temporal generalization in the current harmonized NHANES setting might not only be determined by the model complexity but also by the stability of the features and the structure of the domain. Therefore, the suggested framework should not only be considered as a one-model performance statement, but also as an organized protocol of evaluating transportability, shift, calibration, adaptation behaviour, explanation stability and subgroup robustness.

The three disease outcomes did not show any improvement of the external discrimination through instance-reweighting adaptation. External AUROC decreased as diabetes, hypertension and CVD, were adapted, and simple marginal reweighting was not sufficient to correct the observed change in the temporal distribution. This finding suggests that the discrepancy between the sources and the target was not fully captured by the reweighting based on the domain separability alone and may reflect more complex changes in the feature-label associations. The analysis of adaptation should thus be seen as an exploratory sensitivity experiment and not as an indication of performance recovery. Future studies should also take into consideration more advanced adaptation schemes, including recalibration, learning invariant representations, causal selection of features, prospective updating, and multimodal data integration to be more resilient to change in heterogeneous clinical distributions [22].

The explainability stability came in as a significant consideration in determining trust worthy generalization. Test-retest reliability was found to be high rank-correlation of feature importance with the predictor variables, and found that predictive core variables still had an equivalent explanatory value over time. Stable explanatory structure is of central importance in clinical decision support systems, in which interpretability has a direct impact on clinician acceptance and regulatory scrutiny [23]. Frameworks of feature attribution offer facilities of quantifying model reasoning [24] but their belief in case of distributional change must be experimentally confirmed. The opinion that interpretability is possible even in dynamic clinical settings in the event that models are constructed using physiologically meaningful signals is supported by the evidence of the explainable ICU mortality modelling [25]. We find our results consistent with such an approach and show that it is possible to maintain explanation stability perhaps with moderate levels of predictive degradation.

Regarding the deployment perspective, the results have shown that there are three successful implications, which are pre-deployment shift auditing, external validation as an obligatory process, and continuous performance monitoring. The necessity to test predictive systems outside the context in which they are trained has always been pointed out in cross-site transportability research [20]. A combination of PSI analysis and separability of domain classifier into regular assessment can offer a viable screening mechanism towards predicting decreases in performance. In addition, the constant recalibration and adaptation measures can possibly be used to maintain the status

quo as the characteristics of the population shift over time.

A few limitations are to be noted. Two cycles of the NHANES which were not contemporaneous were analyzed and this restricted the possibility of generalizing the findings to the institutions and healthcare systems. Only structured clinical variables were taken into consideration with no imaging, waveform or textual modalities that could have different shift patterns. Moreover, the assessment has been retroductory, without assessing the prospective deployment. Despite these shortcomings, the study offers a replicable model in the analysis of cross-dataset generalization in the systematic clinical prediction.

The future studies ought to look at the federated learning methods with an attempt to be more robust in the distributed institutions where there is no data sharing in a central location. This can be done by adding methods of causal generalization, which isolate the methods of transporting the predictors, which are invariant, as well as the methods that are spurious, and therefore enhance transportability. Finally, the interaction of structured and unstructured signals can help to enhance resilience to structural and unstructured signal interactions. All these guidelines are useful in terms of creating explainable and generalization-conscientious, and even clinically deployable artificial intelligence systems.

## 6. Conclusion

This work has demonstrated that when models which were trained on one dataset are transferred to a different dataset of multi-diseases, the performance of these models can also be significantly impaired by a shift in a temporal distribution to the extent that is not only measurable but also still clinically acceptable. Despite the fact that internally trained models were highly discriminating, a generalization gap was always large, and the reason is why the explicit assessment of transportability must be conducted before deployment. Population Stability Index and domain classifier separability based quantitative domain shift evaluation provided actionable information on the extent and nature of distributional divergence that provided actionable information on performance changes. Domain adaptation techniques of instance reweighting and limited fine-tuning provided a measurable improvement in recovery in external discrimination performance. Nevertheless, there remained some remaining gaps that indicated that the effects of covariate shift can be mitigated but not completely eliminated. These findings indicate the usefulness of these approaches as adaptation with strict external validation and constant monitoring. It is worth noting that the core clinically meaningful predictors were consistent in the value across domains which is supported by the analysis of the explanation stability. Subgroup analysis also testified consistency of performance across the demographic levels with a view to determine fairness and readiness to deploy. Combined, the proposed framework provides a systematic, generalization-vigilant protocol to the design and testing of clarifiable multi-sickness clinical prediction frameworks in the presence of real distributional variation.

## References

[1] L.L. Guo, S.R. Pfohl, J. Fries, A.E. W. Johnson, J. Posada, C. Aftandilian, N. Shah, L. Sung, Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. Scientific Reports, 12(1), (2022) 2726. https://doi.org/10.1038/s41598-022-06484-1

[2] Z. Wu, H. Yao, D. Liebovitz, J. Sun, An iterative self-learning framework for medical domain generalization. Advances in Neural Information Processing Systems, 36, (2023) 54833–54854. https://doi.org/10.52202/075280-2393

[3] A.A. Hai, M.G. Weiner, A. Livshits, J.R. Brown, A. Paranjape, W. Hwang, L.H. Kirchner, N.Mathioudakis, E.K.French, Z. Obradovic, D.J. Rubin, (2024). Domain generalization for enhanced predictions of hospital readmission on unseen domains among patients with diabetes. Artificial intelligence in medicine, 158, 103010. https://doi.org/10.1016/j.artmed.2024.103010

[4] M. Londschien, M. Burger, G. Rätsch, P. Bühlmann, Domain Generalization and Adaptation in Intensive Care with Anchor Regression. RSS Data Science and Artificial Intelligence, 2(1), (2026) udag001. https://doi.org/10.1093/rssdat/udag001

[5] J. Driever, M. Lentzen, S. Madan, H. Fröhlich, (2025) Domain Adaptation Strategies for Transformer-Based Disease Prediction Using Electronic Health Records. medRxiv. https://doi.org/10.1101/2025.06.02.25328621

[6] Y. Li, S. Rao, J.R.A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, G. Salimi-Khorshidi, BEHRT: transformer for electronic health records. Scientific reports, 10(1), (2020) 7155. https://doi.org/10.1038/s41598-020-62922-y

[7] A. Curth, P. Thoral, W. van den Wildenberg, P. Bijlstra, D. de Bruin, P.W.G. Elbers, M. Fornasa, Transferring Clinical Prediction Models Across Hospitals and Electronic Health Record Systems. Machine Learning and Knowledge Discovery in Databases – International Workshops of ECML PKDD 2019, Communications in Computer and Information Science, 1167, (2020) 605–621. https://doi.org/10.1007/978-3-030-43823-4_48

[8] S. Cui, P. Mitra, Automated Multi-Task Learning for Joint Disease Prediction on Electronic Health Records. Advances in Neural Information

Processing Systems, 37, (2024) 129187–129208. https://doi.org/10.52202/079017-4103

[9] R.D. Riley, J. Ensor, K.I.E. Snell, T.P.A. Debray, D.G. Altman, K.G.M. Moons, G.S. Collins, External Validation of Clinical Prediction Models Using Big Datasets from E-Health Records or IPD Meta-Analysis: Opportunities and Challenges. BMJ, 353, (2016) i3140. https://doi.org/10.1136/bmj.i3140

[10] A. Shmatko, A.W. Jung, K. Gaurav, S. Brunak, L.H. Mortensen, E. Birney, T. Fitzgerald, M. Gerstung, Learning the natural history of human disease with generative transformers. Nature, 647(8088), (2025) 248-256. https://doi.org/10.1038/s41586-025-09529-3

[11] S. Lee, C. Yin, P. Zhang, Stable Clinical Risk Prediction Against Distribution Shift in Electronic Health Records. Patterns, 4(9), (2023) 100818. https://doi.org/10.1016/j.patter.2023.100818

[12] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-BERT: Pretrained Contextualized Embeddings on Large-Scale Structured Electronic Health Records for Disease Prediction. npj Digital Medicine, 4(1), (2021) 86. https://doi.org/10.1038/s41746-021-00455-y

[13] T.H. Chan, G. Yin, K. Bae, L. Yu, Multi-Task Heterogeneous Graph Learning on Electronic Health Records. Neural Networks, 180, (2024) 106644. https://doi.org/10.1016/j.neunet.2024.106644

[14] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent Neural Networks for Multivariate Time Series with Missing Values. Scientific Reports, 8(1), (2018) 6085. https://doi.org/10.1038/s41598-018-24271-9

[15] S. Niu, J. Ma, L. Bai, Z. Wang, L. Guo, X. Yang, EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation. Information Fusion, 102, (2024) 102069. https://doi.org/10.1016/j.inffus.2023.102069

[16] S. Dhiman, A. Thukral, P. Bedi, Multiple Disease Diagnoses Using Heterogeneous EHR Curated Knowledge Graph and Machine Learning Models. Applied Intelligence, 55(7), (2025) 623. https://doi.org/10.1007/s10489-024-05952-7

[17] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. F. Stewart, RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism. Advances in Neural Information Processing Systems, 29, (2016).

[18] National Center for Health Statistics, (2023). National Health and Nutrition Examination Survey (NHANES) August 2021–August 2023. Centers for Disease Control and Prevention, U.S. Department of Health & Human Services, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023

[19] National Center for Health Statistics, (2020) National Health and Nutrition Examination Survey (NHANES) 2017–March 2020 Pre-Pandemic Laboratory Data – Continuous NHANES. Centers for Disease Control and Prevention, U.S. Department of Health & Human Services, https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Laboratory&Cycle=2017-2020

[20] X. Song, A.S. Yu, J.A. Kellum, L.R. Waitman, M.E. Matheny, S.Q. Simpson, Y. Hu, M. Liu, Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nature communications, 11(1), (2020) 5668. https://doi.org/10.1038/s41467-020-19551-w

[21] B. Mamandipoor, C.N. Hsu, M. Krause, U.H. Schmidt, R.A. Gabriel, (2026) Development and External Validation of a Multimodal Artificial Intelligence Mortality Prediction Model of Critically Ill Patients Using Multicenter Data. Anesthesiology Publish Ahead of Print. https://doi.org/10.1097/ALN.0000000000006294

[22] M.K. Siam, M.J. Hossain Faruk, B. He, J.Q. Cheng, H. Gu, Multimodal models in healthcare: Methods, challenges, and future directions for enhanced clinical decision support. Information, 16(11), (2025) 971. https://doi.org/10.3390/info16110971

[23] S. Niu, Q. Yin, J. Ma, Y. Song, Y. Xu, L. Bai, W. Pan, X. Yang, Enhancing healthcare decision support through explainable AI models for risk prediction. Decision Support Systems, 181, (2024) 114228. https://doi.org/10.1016/j.dss.2024.114228

[24] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874 (2017).

[25] M. Mesinovic, P. Watkinson, T. Zhu, Explainable Machine Learning for Predicting ICU Mortality in Myocardial Infarction Patients Using Pseudo-Dynamic Data. Scientific Reports, 15(1), (2025) 27887. https://doi.org/10.1038/s41598-025-12262-6

## Authors Contribution Statement

Anil Vasoya: Conceptualization, Methodology. Rashmi Thakur: Writing - Original Draft, Investigation, Writing - Review & Editing. Manoj Chavan: Data curation, Formal analysis. Sangeeta Mishra: Visualization, Supervision. All authors read and approved the final version of the manuscript.

## Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

## Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

## Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

## Has this article screened for similarity?

Yes

## About the License

© The Author(s) 2026. The text of this article is open access and licensed under a Creative Commons Attribution 4.0 International License.