# Multiaccuracy for Subpopulation Calibration Over Distribution Shift in Medical Prediction Models

**Daniel Kapash** DANIELKPSH@GMAIL.COM
*Software and Information Systems Engineering, Ben Gurion University of the Negev, Be'er Sheva, Israel*

**Ran Balicer** RBALICER@CLALIT.ORG.IL
*Clalit Research Institute, Innovation Division, Clalit Health Services, Tel Aviv, Israel*
*Epidemiology, Biostatistics, and Community Health Sciences, Ben Gurion University of the Negev, Be'er Sheva, Israel*
*The Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at Harvard Medical School and Clalit Research Institute, Boston, US and Tel-Aviv, Israel*

**Omer Reingold** REINGOLD@CS.STANFORD.EDU
*Computer Science, Stanford University, Palo Alto, US*

**Noa Dagan** * NOADAGAN@BGU.AC.IL
*Software and Information Systems Engineering, Ben Gurion University of the Negev, Be'er Sheva, Israel*
*Clalit Research Institute, Innovation Division, Clalit Health Services, Tel Aviv, Israel*
*The Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at Harvard Medical School and Clalit Research Institute, Boston, US and Tel-Aviv, Israel*

**Noam Barda** * NOAM.BARDA@SHEBA.HEALTH.GOV.IL
*Clalit Research Institute, Innovation Division, Clalit Health Services, Tel Aviv, Israel*
*Epidemiology, Biostatistics, and Community Health Sciences, Ben Gurion University of the Negev, Be'er Sheva, Israel*
*ARC Innovation Center, Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel.*

## Abstract

Multiaccuracy was previously demonstrated to improve subpopulation calibration in medical prediction models, ensuring fairness towards subpopulations. Medical prediction models often experience degraded performance due to distribution shifts (e.g. changes in input data resulting from changes in space or time), but the effectiveness of multiaccuracy in ensuring medical predictors' fairness under these circumstances was suggested theoretically but has yet to be studied empirically. To explore this, we trained prediction models using real-world data, applied an adaptation of multiaccuracy as a post-processing step to intersecting subpopulations defined by combinations of protected features such as age, gender, and socioeconomic status, and tested the performance of the models on target test sets from distributions different than the development cohorts. The results demonstrated that the improvement in subpopulation calibration achieved by multiaccuracy was maintained in the target distribution over two experiments, simulating spatial-temporal and migration-induced distribution shifts. On average, over the two experiments, Calibration in the Large mean error and variance measures were reduced by 71.8% and 70.7% on the target distributions after applying multiaccuracy, respectively. These findings highlight the potential of post-processing for multiaccuracy as a tool for enhancing the fairness and reliability of medical prediction models across diverse populations, even under circumstances of major distribution shifts.

**Data and Code Availability** This paper uses the Clalit Health Services (CHS) dataset, a large-scale medical dataset that includes information from approximately 4.9 million insured patients in Israel. Due to patient privacy concerns and legal restrictions, the data cannot be made publicly available. Multiaccuracy implementation code is available.

**Institutional Review Board (IRB)** This study was approved by the Clalit Institutional Review

* Equal contribution as senior authors

Board (Helsinki approval number 0161-19-COM1). Informed consent was waived by the IRB, as all identifying details of the participants were removed before computational analysis.

## 1. Introduction

Prediction models have become increasingly important tools in medicine for predicting the probability of certain events (Chen and Asch, 2017), such as the risk of developing a particular disease. These models are developed using statistical and machine learning techniques, and are trained on large datasets of patient data and other relevant information. They can be used to inform clinical decision making (Leeuwenberg and Schuit, 2020), to guide prevention and screening strategies (Dagan et al., 2020), and to help prioritize and allocate resources (Lerner et al., 2017). While prediction models have the potential to improve patient care, it is important to carefully evaluate their performance and limitations, as well as to consider any potential biases in the predictions made by these models.

Model calibration (Niculescu-Mizil and Caruana, 2005), the agreement between predicted probabilities and observed outcomes, is a crucial performance metric for prediction models. Well-calibrated models should accurately estimate event probabilities overall and within specific subpopulations. However, miscalibration can occur when predicted probabilities do not match observed event frequencies in certain subpopulations (Van Calster et al., 2019), leading to biased predictions and potentially inappropriate clinical decisions (Paulus and Kent, 2020). To promote health equity and avoid unfairness and unintended consequences, such as over-treatment or under-treatment of specific groups (Rajkomar et al., 2018), it is essential to ensure prediction models are well-calibrated across diverse patient subpopulations.

Multicalibration (Hebert-Johnson et al., 2018) is an algorithm designed to ensure that a prediction model is calibrated on a set of intersecting subpopulations defined by a set of features. This is achieved through a post-processing algorithm that iteratively mitigates the calibration error for each of the defined subpopulations, promoting algorithmic fairness. Multiaccuracy (Hebert-Johnson et al., 2018; Kim et al., 2019) is a more relaxed version of multicalibration, where the post-processing algorithm enforces calibration-in-the-large (CITL) (Huang et al., 2020) on the set of subpopulations. This means that within each subpopulation, the average predicted probability of an event matches its observed frequency. Empirical tests on large medical datasets have demonstrated that multiaccuracy can successfully achieve CITL on hundreds of subpopulations while preserving the model's discriminative performance (Barda et al., 2021).

Robustness to distribution shift is a crucial factor to consider when developing and evaluating prediction models. Distribution shift refers to changes in the characteristics of the population or the context in which the model is applied (Roland et al., 2022). For instance, a model developed and validated using data from a particular geographic region may not perform as well when applied to a different region with distinct healthcare systems or patient demographics, a phenomenon often observed in spatial shifts (Cabanillas Silva et al., 2024; Austin et al., 2017). Similarly, a model performance might degrade as input data evolves over time due to changes in clinical practices, population aging, or disease prevalence—common in temporal shifts (Austin et al., 2017). These spatial-temporal and migration-induced shifts are particularly relevant in healthcare, where models trained on one population (e.g., a northern region in 2008) may be deployed a decade later in a different region, or where immigration introduces diverse patient cohorts with distinct risk profiles (Salinas et al., 2012; Lê-Scherban et al., 2016). Although methods like domain adaptation (Zhou et al., 2022) and domain generalization (Wilson and Cook, 2020) have been proposed to address these issues, they have shown limited effectiveness in handling complex, evolving changes in medical records (Guo et al., 2022), where both patient demographics and clinical practices change over space and time.

Previous theoretical work suggested that multiaccuracy could potentially improve robustness to distribution shifts (Kim et al., 2022), primarily through the lens of ensuring unbiased estimates on entire target populations and target subpopulations. However, empirical validation is necessary to fully understand its effectiveness and limitations in real-world scenarios. In this study, we set out to empirically explore how post-processing for multiaccuracy affects the behavior of medical prediction models under various circumstances of distribution shifts. Specifically, we designed experiments to evaluate the overall and subpopulation calibration under spatial-temporal and migration-induced distribution shifts, with and without multiaccuracy, using large-scale, real-world data.

## 2. Materials and Methods

### 2.1. Data

In this study we use medical data from Clalit Health Services (CHS), Israel's largest healthcare provider. Israel has mandatory health insurance provided by four integrated payer-providers, of which CHS is the largest. The dataset used for this study covers the years 2000-2022 and includes information from approximately 4.9 million insured patients. This extensive dataset incorporates a broad spectrum of medical information, such as patient diagnoses, lab test outcomes, clinical visits, hospitalizations, medication prescriptions, vital sign recordings, and demographic information.

### 2.2. Study Populations and Datasets

We created a separate dataset for each experiment, utilizing a retrospective cohort design. The outcome explored in each dataset was cardiovascular disease (CVD), including percutaneous coronary intervention, atrial fibrillation, heart failure, myocardial infarction, coronary artery bypass, and stroke (cir, 2019), and accordingly, the cohort definitions were defined to create a patient population relevant for the prediction of this outcome. We used various index dates to allow for the various experiments as detailed below. As of each index date, we included patients aged between 30 and 90 years, who had been members of the healthcare provider for at least one year and who had no prior history of CVD. For label assignment - individuals were considered to have experienced a CVD event if they had a diagnosis code for CVD recorded in a 5 or 10 year period after the index date, depending on the specific experiment.

Features considered in each dataset included all medication dispensing, clinic visits, hospitalizations, vital signs, the 300 most common lab results, diagnoses, and demographics data available in the patient's electronic health record. To determine feature values, we used data from a 3-year window prior to the index date, except for diagnoses which were taken from the entire patient history prior to the index date. When more than one data point was available for a specific feature in a specific patient, different aggregation methods were applied depending on the nature of the data. Medication dispensing, clinic visits, and hospitalizations were aggregated by summation within the 3-year window. Continuous values, such as vital signs and lab results, were aggregated using various statistics - minimum, maximum, mean, variance, and regression slope, calculated over the 3-year period.

### 2.3. Model Development

Initially, each dataset consisted of approximately 22,000 features. Feature selection, necessary to improve model performance and reduce the required computation, was performed by training a regularized XGBoost (Chen and Guestrin, 2016) model to predict CVD on a cohort defined as of 1/1/2012 with a 10-year prediction horizon. The model was trained using 80 estimators (trees), a maximum tree depth of 4, a gamma value (which determines the minimum loss reduction required to make a further partition on a leaf node of the tree) of 2, and an alpha value (which determines the L1 regularization term on weights) of 1. These hyperparameters were manually chosen to create a lean model that prioritizes the most informative features while avoiding overfitting. We then selected the features that were used at least once in the model.

Missing data was handled using single stochastic imputation. We performed 8 imputation passes, using the 10 highest correlating features for each missing value. We initiated missing values with median feature values for the first round of imputation.

The final models were trained using L2-regularized ("Ridge") Logistic Regression. We fitted this model on the training set of the source population in each experiment and tested each model on the test set of its source population and on the target population defined for the experiment to assess the impact of the distribution shift on model performance.

### 2.4. Multi-Accuracy Post Processing

The post-processing algorithm we used is an adaptation of the alpha-multi-AE algorithm presented by Hebert-Johnson et al. (2018). The algorithm (Algorithm 1) takes as input an initial predictor $p$, a collection of protected subpopulations $G$, a training set $S$, a violation parameter $\alpha$, and a minimal group size $N$. It iteratively improves the accuracy-in-expectation of the predictor with respect to the protected groups.

In each iteration, the algorithm randomly permutes the groups in $G$ and checks for each group $g$ whether the predictions on $g$ are accurate in expectation up to the allowed violation $\alpha$, considering only groups with size at least $N$. If a violating group is found, the algorithm updates the predictor by adding a term to $p$

**Algorithm 1:** Iterative Multiaccuracy Fit

```
1. $done \leftarrow False$
2. while not $done$ do
    (a) $done \leftarrow True$
    (b) for each $g \in \text{RandomPermutation}(G)$ do
        i. $S_g = \{x \in S | \mathbf{1}_g(x) = 1\}$
        ii. $\Delta_g = \frac{\sum_{(x,y) \in S}((y - p(x)) \cdot \mathbf{1}_g(x))}{|S_g|}$
        iii. if $|\Delta_g| > \alpha$ and $|S_g| \geq N$ then
            A. $p \leftarrow p + \Delta_g \cdot \mathbf{1}_g$
            B. $done \leftarrow False$
            C. break
3. $p \leftarrow \text{Clip}(p, 0, 1)$
4. return $p$
```

that "nudges" the predictions on $g$ in the direction of higher accuracy. This update is performed by adding the violation amount $\Delta_g$ multiplied by the indicator function of $g$ to the current predictor. The iteration continues until no more violating groups are found.

After the iterative updates, the final predictor $p$ is obtained by clipping the predictions to the range $[0, 1]$. The resulting predictor $p$ satisfies $(G, \alpha) - accuracy - in - expectation$ on the training set $S$, ensuring that all protected groups in $G$ of size at least $N$ have predictions that are accurate in expectation up to $\alpha$.

Hyperparameter tuning was conducted to optimize multiaccuracy convergence, model fit, inference runtime, and generalization to the test set, with AUROC as the primary target metric. Smaller $N$ increases $|G|$, potentially lengthening both fit and inference times and mostly risking overfitting, necessitating manual tuning. In our adaptation of the multiaccuracy algorithm, fitting runtime scales as $\mathcal{O}(|G| \cdot (1/\alpha^2))$ (Hebert-Johnson et al., 2018). Inference runtime, required for post-processing perturbations, is $\mathcal{O}(|G|)$, directly tied to $N$. However, this overhead is negligible relative to the inference latency of most commonly used state-of-the-art machine learning models such as deep neural networks or gradient-boosted decision trees.

The multiaccuracy post-processing algorithm was applied to the model using the source distribution's training data. The post-processed model was then used to make predictions on both the source distribution's test data and the entire target distribution data.

## 2.5. Experiments

We investigated two types of distribution shifts: spatial-temporal and migration-induced. For the spatial-temporal shift, we defined the source distribution as patients from a northern district of Israel (the city of Haifa), with the index date set to January 1st, 2008, and the target distribution as patients from Israel's South district with the index date set to January 1st, 2018. To allow a larger temporal difference, we used a 5-year prediction window. For the migration-induced shift, we designated the native-born population of Israel as the source distribution and immigrants to Israel as the target distribution, both with the index date set to January 1st, 2012, and a 10-year prediction window.

The source distribution data was split into training and test sets using a 70/30 and 80/20 ratio in the spatial-temporal shift experiment and the migration-induced shift experiment, respectively.

For the spatial-temporal experiment, the protected variables included age, ethnicity, sex, socioeconomic status, and immigration status. Age was stratified into 5 equal-frequency bins within the study range of 30-90 years. Ethnicity was categorized as Jewish, Arab, or Other. Socioeconomic status was discretized into 3 bins using an internal categorization based on the location of the patients' primary care provider.

In the migration-induced shift experiment, ethnicity and immigration status were replaced by geographic districts as protected variables. Immigration status was not relevant since the source and target distributions were partitioned by this variable. Ethnicity was also not relevant for this experiment because immigration status effectively also partitioned this feature's categories between the source and target populations.

## 2.6. Performance Metrics

We evaluated three performance categories in each dataset: global discrimination, global calibration, and subpopulation calibration.

Area Under the Receiver Operating Characteristic curve (AUROC) was used as the main metric for global discrimination performance.

Global calibration was measured using the Integrated Calibration Index (ICI) (Austin and Steyerberg, 2019), defined in Equation (1), quantifies the difference between the observed probabilities and the predicted probabilities, taking into account the distribution of the predicted probabilities. It does this by assigning weights to each observation based on how frequently the corresponding predicted probability occurs in the dataset (i.e., the empirical density function of the predicted probabilities). By incorporating these weights, the ICI effectively emphasizes the calibration of the model in regions where predictions are more common. It equals 0 when the predictor is perfectly calibrated.

$$\text{ICI} = \int_0^1 |x - x_c| \cdot \phi(x)\, dx \tag{1}$$

$x$ denotes a predicted probability in the interval $(0, 1)$ and $x_c$ denotes the value of the calibration curve at $x$. $\phi(x)$ denote the density function of the distribution of predicted probabilities.

$$CITL = \left| \frac{E(\hat{p})}{(1 - E(\hat{p}))} \Big/ \frac{E(y)}{(1 - E(y))} \right| \tag{2}$$

Subpopulation calibration was measured using the calibration-in-the-large (CITL) metric, defined in Equation (2). It measures the odds ratio of the average of the predictions and the average of the outcome (Huang et al., 2020). It equals 1 when the predictor is perfectly calibrated.

$$\begin{aligned} CITL\_ERROR &:= \exp\left(|\log(CITL)|\right) - 1 \\ &= \max\left(CITL, \frac{1}{CITL}\right) - 1 \end{aligned} \tag{3}$$

Our definition of CITL error in Equation (3) quantifies the calibration error while symmetrically handling deviations from perfect calibration (where CITL=1). The logarithm normalizes the ratio, ensuring that both overestimation and underestimation are treated equally through the absolute value. The exponential function is then applied to reverse the effect of the logarithm, restoring the original scale of the error. Subtracting 1 ensures that the error is zero when the model is perfectly calibrated, making this a useful metric for assessing calibration deviations.

# 3. Results

## 3.1. Study population

For the spatial-temporal shift experiment, the source population consisted of 285,251 patients from the Northern District in 2008, divided into 199,675 for the training dataset and 85,576 for the test dataset, while the target population included 210,216 patients from the Southern District in 2018. The outcome proportions for the source training, source test, and target datasets were 3.2%, 3.1%, and 2.5%, respectively.

In the migration-induced shift experiment, the source population comprised 1,106,685 patients who were native-born, with 885,317 in the training set and 221,368 in the test set, whereas the target population included 529,450 patients who immigrated to Israel. The outcome proportions for these datasets were 2.06% for both the source training and source test datasets and 4.57% for the target dataset.

Additional details regarding each of the cohorts, including the distribution of protected variables, are provided in Table 3 and Table 4 in Appendix A. It can be noted that in both experiments, the source train and test populations are very similar to one another, while the target population demonstrates differences (data shifts) across almost all variables.

## 3.2. Modeling and Multiaccuracy

Table 1 presents the feature importance rankings derived from the XGBoost model for the top 12 features based on their importance scores. The table also presents an indication of whether the feature is part of the Pooled Cohort Equations (PCE) model of the American College of Cardiology and the American Heart Association to estimate an individual's 10-year risk of atherosclerotic CVD (cir, 2019), a subset of the broader CVD outcome used for this study. The PCE uses a total of 10 features, and it can be noted that there is a strong correlation with the predictors utilized in the XGBoost and the PCE model, as 8 out of the top 12 features are either identical to or serve as aliases for predictors employed in the PCE model.

The feature selection resulted in 652 features with a positive gain in the XGBoost model. The feature distribution over the different data sources is presented in Appendix B Figure 3.

Hyperparameter search for multiaccuracy resulted in a value of 0.0002 for the violation parameter $\alpha$ and a minimal group size of 3000. These hyperparameters obtained a good balance between multiaccuracy convergence, fit and inference runtime, and generalization to the source test set. After filtering by the minimal allowed group size, the spatial-temporal experiment included 455 protected subpopulations, while the migration-induced experiment consisted of 516 protected subpopulations.

### 3.3. Experiment Results

The global discrimination performance (AUROC) of the models for both distribution shift experiments before and after applying multiaccuracy on both source and target populations are provided in Table 2. We observed that the distribution shift indeed resulted in reduced AUROC as anticipated. In the spatial-temporal shift experiment, the AUROC decreased from 0.815 in the source population to 0.784 in the target population. Similarly, in the migration-induced shift experiment, the AUROC dropped from 0.801 in the source population to 0.742 in the target population. These results demonstrate the expected performance degradation due to distribution shift. Notably, multiaccuracy, which is focused on improving calibration, did not harm the AUROC values, which presented similar values in the source test set before and after the application of the post-processing algorithm. In addition, the multiaccuracy process did not alter the AUROC after the distribution shift.

The global calibration results of the Integrated Calibration Index (ICI) measure before and after applying multiaccuracy are presented in Table 2. The results suggest that multiaccuracy had a minimal impact on global calibration, as measured by the ICI, across both source and target distributions in the two experiments. In the spatial-temporal shift experiment, the ICI slightly improved from 0.006323 to 0.004865 in the source population and from 0.004821 to 0.004625 in the target population after applying multiaccuracy. For the migration-induced shift experiment, the ICI improved from 0.002933 to 0.002009 in the source population but slightly worsened from 0.006868 to 0.007388 in the target population after applying multiaccuracy. Similarly, examining the CITL values for the entire population, also presented in Table 2, reveals less substantial changes before and after applying multiaccuracy. This stability in global CITL is expected, as the initial models were already calibrated-in-the-large on the entire population, that is, they were accurate in expectation at the population level, thus leaving little room for improvement through the multiaccuracy process.

Subpopulations calibration-in-the-large (CITL) mean error and variance before and after applying multiaccuracy in both spatial-temporal and migration-induced distribution shift experiments is presented in Table 2. Figure 1 and Figure 2 show CITL error of subpopulations ordered by size, before and after applying multiaccuracy on both source and target populations.

In the spatial-temporal shift experiment, the mean CITL error decreased from 0.275 to 0.191 in the source population and from 0.396 to 0.226 in the target population after applying multiaccuracy. The variance in CITL also reduced substantially, from 0.3399 to 0.0496 in the source population and from 0.5060 to 0.1643 in the target population. For the migration-induced shift experiment, similar improvements were observed. The mean CITL error decreased from 0.651 to 0.214 in the source population and from 0.431 to 0.108 in the target population. The variance in CITL was reduced from 1.6234 to 0.3255 in the source population and from 0.7126 to 0.1848 in the target population. On average over the two experiments, the CITL mean error was reduced by 67.1% on the source populations and by 71.8% on the target distributions after applying multiaccuracy. The CITL variance was reduced on average by 82.7% on the source populations and by 70.7% on the target populations.

To provide insight into subpopulation-specific performance, we include supplemental PDF, detailing CITL errors for each protected subpopulation in the temporal-spatial and migration-induced shift experiments, respectively. These tables report errors on both source and target distributions, before and after applying multiaccuracy (MA), demonstrating consistent error reductions across most subpopulations.

## 4. Discussion

In this study, we investigated the effects of post-processing for multiaccuracy in medical prediction models under various circumstances of distribution shifts. Specifically, we explored model behavior in terms of overall calibration and of calibration fairness across subpopulations using experiments over temporal-spatial and migration-induced distribution shifts. We demonstrated that multiaccuracy successfully improved subpopulation calibration on the source population, and this improvement was largely

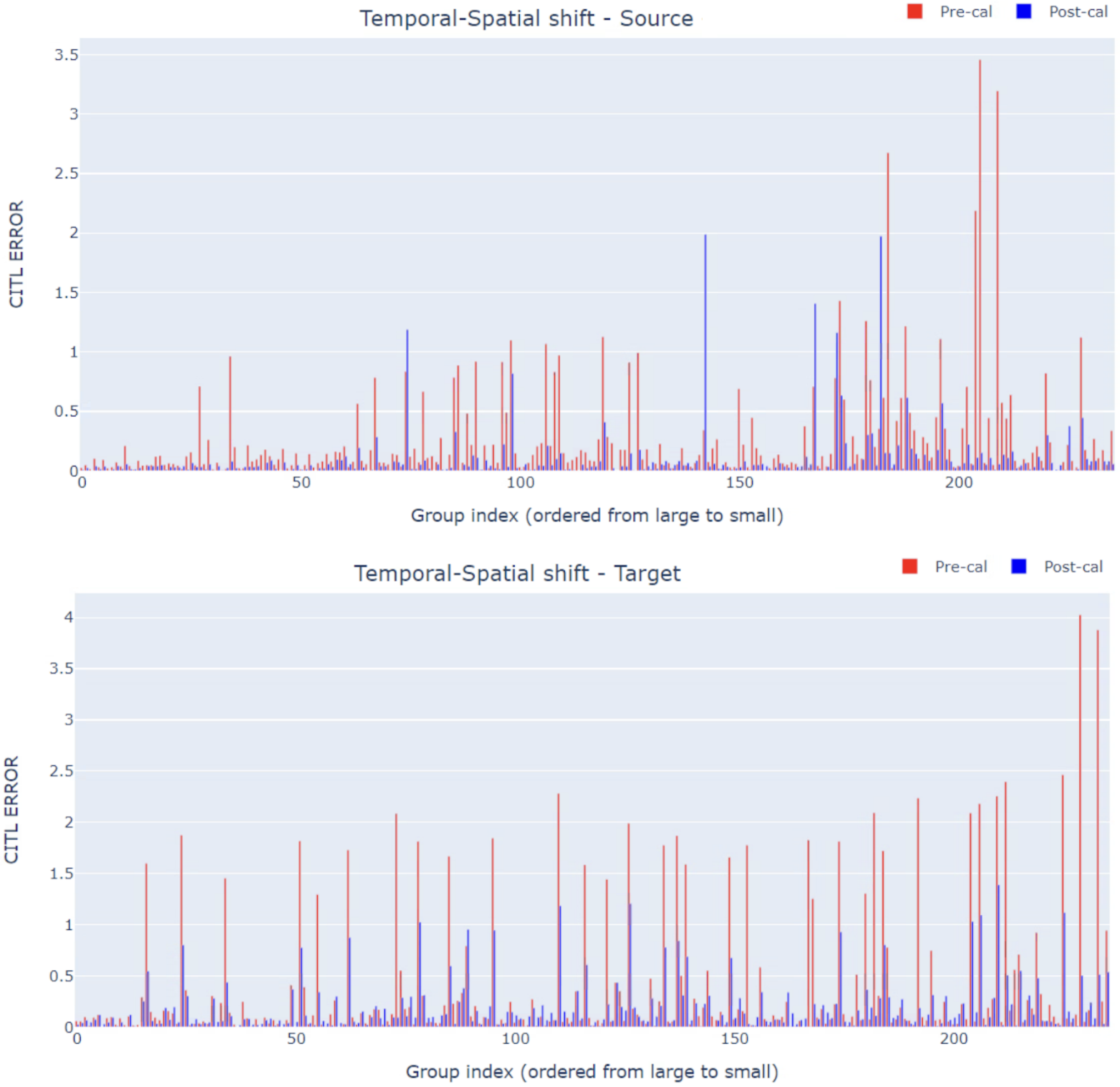

Figure 1: Calibration-in-the-large error scores for subpopulations before and after Multiaccuracy in source and target distributions of the Spatial-temporal distribution shift experiment. Groups are ordered from largest to smallest, so the left-most group represents the entire population. The graph presents the 236 groups with size at least N=3000 in the source test set.

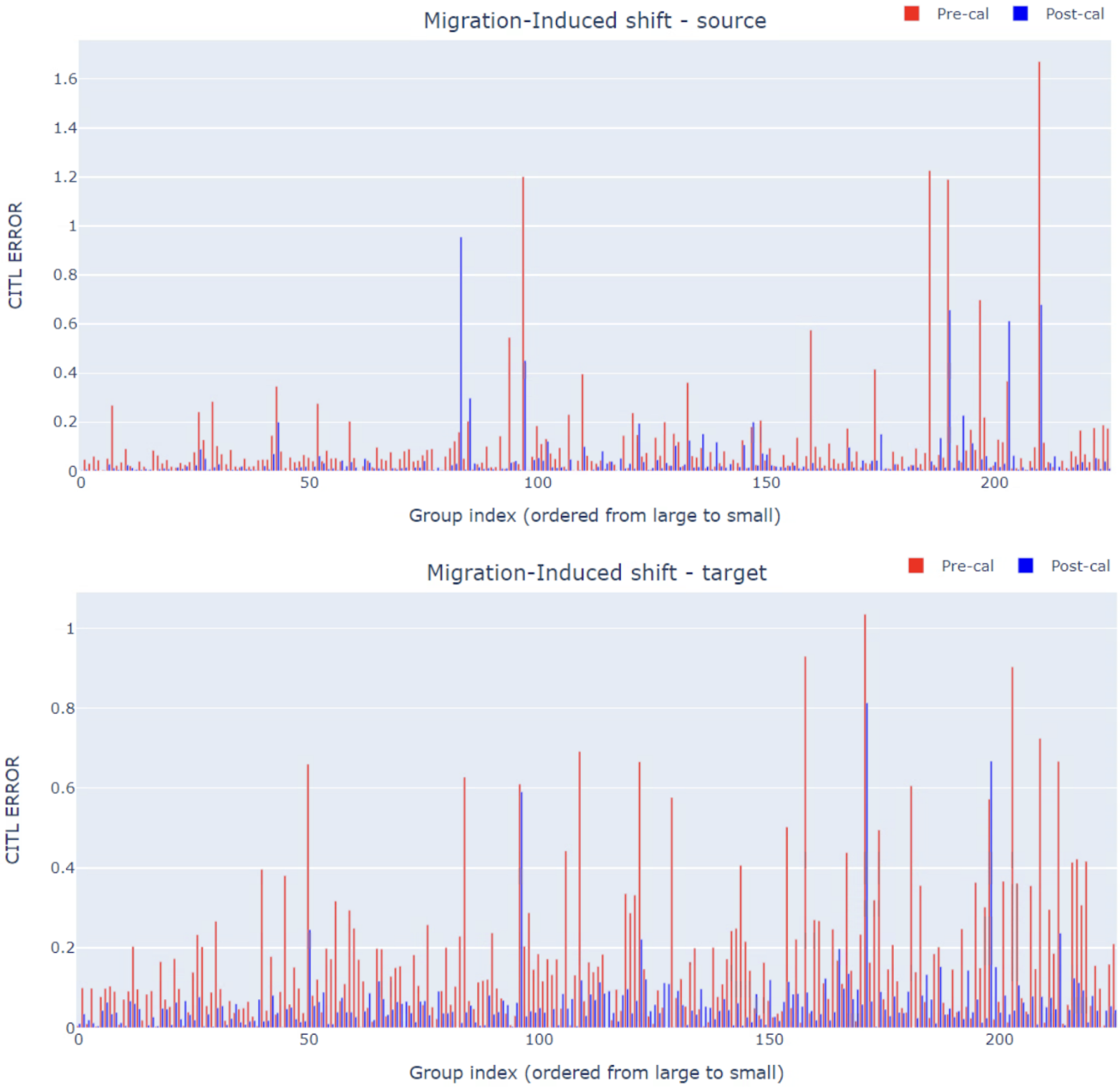

Figure 2: Calibration-in-the-large error scores for subpopulations before and after Multiaccuracy in source and target distributions of the migration-induced shift experiment. Groups are ordered from largest to smallest, so the left-most group represents the entire population. The graph presents the 226 groups with size at least N=3000 in the source test set.

Table 1: Feature importance ranking based on XGBoost gain and their usage status in the PCE model.

| Predictor | In PCE model | Importance |
|---|---|---|
| **Age** | **Yes** | **53718** |
| **Sex** | **Yes** | **8333** |
| **Systolic Blood Pressure** | **Yes** | **5557** |
| **Diabetes** | **Yes** | **3614** |
| **Smoking Status** | **Yes** | **2685** |
| **Cholesterol Medication (Statins)** | **Alias** | **1642** |
| Hyperlipidemia Diagnosis | No | 1252 |
| **Hypertension Diagnosis (Sulfonylureas)** | **Alias** | **959** |
| HCT Blood Test | No | 588 |
| Gynecologist Visits | No | 587 |
| **Diabetes Medication** | **Alias** | **554** |
| White Blood Cell Count | No | 532 |

maintained on the target population. This finding indicates that post-processing using multiaccuracy for ensuring fairness in prediction in the source population remains effective even under substantial distribution shifts of various kinds. Specifically, we observed an average reduction of 70.7% in the variance of CITL values across subpopulations in the target distributions after applying multiaccuracy.

While multiaccuracy improved subpopulation calibration, it did not substantially alter global calibration on the target distributions, as measured by the Integrated Calibration Index (ICI) and CITL. This suggests that its benefits are mainly limited to subpopulations in our experimental setup.

Consistent with previous findings (Barda et al., 2021), we showed that the multiaccuracy process did not harm the discriminatory performance of the model (as measured by AUROC). In addition, we provide new evidence that the multiaccuracy also did not harm the discriminatory performance under distribution shift (beyond the AUROC drop caused by the distribution shift itself).

Our findings build upon and extend previous research in this area. The ability of multiaccuracy to improve fairness in the source population was previously demonstrated by Barda et al. (2021). Our study first validated these results and then extended them by demonstrating that this fairness is maintained under distribution shift, which, to the best of our knowledge, is the first empirical demonstration of this effect. This aligns with the theoretical work of Kim et al. (2022), who suggested that multiaccuracy could potentially improve robustness to distribution shifts. However, our results contrast with some domain adaptation techniques that aim to improve overall model performance under shift (Zhou et al., 2022; Wilson and Cook, 2020; Guo et al., 2022), as we found that multiaccuracy primarily benefits subpopulation calibration rather than global performance metrics.

A key strength of this study lies in its demonstration of multiaccuracy's ability to ensure fairness towards subpopulations under extreme distribution shift circumstances, such as changes in both place and time, and for dramatically different populations like immigrants and non-immigrants (including a dramatic change in outcome rate). This comprehensive analysis was made possible by access to a unique and extensive dataset from a healthcare provider, covering approximately 4.9 million patients over two decades. The breadth and depth of this dataset allowed us to experiment with realistic distribution shifts and evaluate their impact on various subpopulations. Additionally, the high degree of overlap between our model's top-ranking features and the Pooled Cohort Equations (PCE) predictors (cir, 2019) provides compelling validation of both our dataset's reliability and modeling process.

From a theoretical perspective, multiaccuracy has the potential to also mitigate reductions in discrimination performance when the underlying shift is in the marginal distribution of the input features, $P(X)$, but not in the conditional distribution of the target variable given the input features, $P(Y|X)$ (Roth, 2020). Though we would not expect to see this effect in this experiment, as we only performed multiaccuracy for a small set of subpopulations, aiming

Table 2: Model performance metrics before and after multiaccuracy (MA) in spatial-temporal and migration-induced distribution shift experiments. AUROC, ICI, subpopulation CITL Error mean, and CITL variance, are shown for source and target distributions test sets. Subpopulation statistics are computed for groups with more than $N = 3000$ samples in the source test set, totaling 236 groups in the temporal-spatial shift experiment and 226 groups in the migration-induced shift experiment.

| | | Temporal-Spatial | | | Migration-induced | | |
|---|---|---|---|---|---|---|---|
| | | Pre-MA | Post-MA | Change | Pre-MA | Post-MA | Change |
| AUROC | Source | 0.815 | 0.812 | -0.003 | 0.801 | 0.807 | +0.006 |
| | Target | 0.784 | 0.783 | -0.001 | 0.742 | 0.741 | -0.001 |
| ICI | Source | 0.006323 | 0.004865 | -23.1% | 0.002933 | 0.002009 | -31.5% |
| | Target | 0.004821 | 0.004625 | -4.1% | 0.006868 | 0.007388 | +7.6% |
| Population<br>CITL Error | Source | 0.0244 | 0.0167 | -31.6% | 0.0156 | 0.0090 | -42.3% |
| | Target | 0.0597 | 0.0335 | -43.9% | 0.0194 | 0.0049 | -74.7% |
| Subpopulations<br>CITL Error mean | Source | 0.275 | 0.191 | -59.0% | 0.651 | 0.214 | -75.2% |
| | Target | **0.396** | **0.226** | **-63.7%** | **0.431** | **0.108** | **-79.9%** |
| Subpopulations<br>CITL - Variance | Source | 0.3399 | 0.0496 | -85.4% | 1.6234 | 0.3255 | -80.1% |
| | Target | **0.5060** | **0.1643** | **-67.5%** | **0.7126** | **0.1848** | **-73.9%** |

instead on improving fairness. Further experiments are needed to ascertain the real-world conditions under which this positive effect will be observed.

While our results demonstrate multiaccuracy's utility for subpopulation calibration under distributional shifts, we note that its post-hoc, model-agnostic design differs fundamentally from training-time robustness methods (e.g., domain adaptation (Wilson and Cook, 2020) and domain generalization techniques Zhou et al. (2022)). These approaches often require retraining or source data access, limiting comparability in deployment-focused settings. However, multiaccuracy's explicit guarantees for computationally identifiable subpopulations position it as a complementary tool to address granular fairness degradation under shift. Future work should explore integrating multiaccuracy with training-time interventions to unify subpopulation-level calibration and global robustness, particularly in real-world clinical environments where both data shifts and fairness are critical.

Our study has several limitations. First, the experiments were conducted using data from a single country. To establish the generalizability of our findings, it is essential to validate the results on diverse datasets from different healthcare settings and populations. Second, while our experiments covered temporal, spatial, and migration-induced shifts, there may be other types of distribution shifts that were not explored in this study. Future work should investigate the performance of multiaccuracy under a broader range of distribution shift scenarios, including shifts in data collection methods and clinical practices. Finally, our study focused on a specific medical prediction task, namely the prediction of CVD risk. It would be valuable to assess the effectiveness of multiaccuracy in other medical prediction tasks, such as predicting the risk of other diseases or treatment outcomes, to gain a more comprehensive understanding of its potential benefits and limitations.

## 5. Conclusion

In summary, our study demonstrates that the improvement in calibration fairness across hundreds of subpopulations, achieved through postprocessing of medical prediction models using a multiaccuracy algorithm, is preserved even under circumstances of distribution shifts. In real-world healthcare settings, patient populations often exhibit significant heterogeneity across various demographic, geographic, and temporal shifts. Consequently, models developed on a specific population may not perform equally well for all subpopulations when applied to different populations or time periods. This can lead to biased predictions and unfair treatment of certain patient subpopulations, perpetuating or exacerbating health disparities. By demonstrating that multiaccuracy can maintain improved subpopulation calibration fairness even under distribution shift, our study highlights its potential for promoting fairness in medical decision-making.

## References

2019 acc/aha guideline on the primary prevention of cardiovascular disease: Executive summary: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Circulation*, Aug. 31, 2019. URL https://www.ahajournals.org/doi/10.1161/CIR.0000000000000677.

P. C. Austin and E. W. Steyerberg. The integrated calibration index (ici) and related metrics for quantifying the calibration of logistic regression models. 2019. doi: 10.1002/sim.8281.

P. C. Austin, D. van Klaveren, Y. Vergouwe, D. Nieboer, D. S. Lee, and E. W. Steyerberg. Validation of prediction models: examining temporal and geographic stability of baseline risk and estimated covariate effects. *Diagnostic and Prognostic Research*, 1:12, 2017. doi: 10.1186/s41512-017-0012-3.

N. Barda et al. Addressing bias in prediction models by improving subpopulation calibration. *J. Am. Med. Inform. Assoc. JAMIA*, 28(3):549–558, Mar. 2021. doi: 10.1093/jamia/ocaa283.

P. Cabanillas Silva, H. Sun, M. Rezk, D. Roccaro-Waldmeyer, J. Fliegenschmidt, N. Hulde, V. von Dossow, L. Meesseman, K. Depraetere, J. Stieg, R. Szymanowsky, and F. Dahlweid. Longitudinal model shifts of machine learning–based clinical risk prediction models: Evaluation study of multiple use cases across different hospitals. *J Med Internet Res*, 26:e51409, 2024. doi: 10.2196/51409. URL https://www.jmir.org/2024/1/e51409.

J. H. Chen and S. M. Asch. Machine learning and prediction in medicine — beyond the peak of inflated expectations. *N. Engl. J. Med.*, 376(26):2507–2509, Jun. 2017. doi: 10.1056/NEJMp1702071.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA, Aug. 2016. Association for Computing Machinery. doi: 10.1145/2939672.2939785.

N. Dagan, N. Barda, D. Riesel, I. Grotto, S. Sadetzki, and R. Balicer. A score-based risk model for predicting severe covid-19 infection as a key component of lockdown exit strategy. *medRxiv*, May 23, 2020. doi: 10.1101/2020.05.20.20108571.

L. L. Guo et al. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Sci. Rep.*, 12(1):2726, Feb. 2022. doi: 10.1038/s41598-022-06484-1.

U. Hebert-Johnson, M. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1939–1948. PMLR, Jul. 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Y. Huang, W. Li, F. Macheret, R. A. Gabriel, and L. Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc. JAMIA*, 27(4):621–633, Apr. 2020. doi: 10.1093/jamia/ocz228.

M. P. Kim, A. Ghorbani, and J. Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, New York, NY, USA, Jan. 2019. Association for Computing Machinery. doi: 10.1145/3306618.3314287.

M. P. Kim, C. Kern, S. Goldwasser, F. Kreuter, and O. Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proc. Natl. Acad. Sci.*, 119(4):e2108097119, Jan. 2022. doi: 10.1073/pnas.2108097119.

A. M. Leeuwenberg and E. Schuit. Prediction models for covid-19 clinical decision making. *Lancet Digit. Health*, 2(10):e496–e497, Oct. 2020. doi: 10.1016/S2589-7500(20)30226-0.

B. Lerner, S. Desrochers, and N. Tangri. Risk prediction models in ckd. *Semin. Nephrol.*, 37(2):144–150, Mar. 2017. doi: 10.1016/j.semnephrol.2016.12.004.

F. Lê-Scherban, S. S. Albrecht, A. Bertoni, N. Kandula, N. Mehta, and A. V. Diez Roux. Immigrant status and cardiovascular risk over time: results from the multi-ethnic study of atherosclerosis. *Annals of Epidemiology*, 26(6):429–435.e1, 2016. doi: 10.1016/j.annepidem.2016.04.008.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 625–632, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

J. K. Paulus and D. M. Kent. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *Npj Digit. Med.*, 3(1):1–8, Jul. 2020. doi: 10.1038/s41746-020-0304-9.

A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.*, Dec. 2018. doi: 10.7326/M18-1990. URL https://www.acpjournals.org/doi/10.7326/M18-1990.

T. Roland et al. Domain shifts in machine learning based covid-19 diagnosis from blood tests. *J. Med. Syst.*, 46(5):23, 2022. doi: 10.1007/s10916-022-01807-1.

A. Roth. Uncertain: Modern topics in uncertainty estimation. 2020.

J. J. Salinas, B. Abdelbary, J. Wilson, M. Hossain, S. Fisher-Hoch, and J. McCormick. Using the framingham risk score to evaluate immigrant effect on cardiovascular disease risk in mexican americans. *Journal of Health Care for the Poor and Underserved*, 23(2):666–677, 2012. doi: 10.1353/hpu.2012.0058.

B. Van Calster et al. Calibration: the achilles heel of predictive analytics. *BMC Med.*, 17(1):230, Dec. 2019. doi: 10.1186/s12916-019-1466-7.

G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *arXiv: arXiv:1812.02849*, Feb. 06, 2020. doi: 10.48550/arXiv.1812.02849.

K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–20, 2022. doi: 10.1109/TPAMI.2022.3195549.

## Appendix A. Experiments Population Tables

Table 3: Spatial-temporal shift experiment population characteristics by protected variables. Source population - Northern district 2008, Target population - Southern district 2018.

| **Variable** | **Levels** | **Source Train** | **Source Test** | **Target** |
|---|---|---|---|---|
| | All | 199,675 | 85,576 | 210,216 |
| Outcome (%) | 1 | 6,333 (3.2) | 2,657 (3.1) | 5,335 (2.5) |
| | 0 | 193,342 (96.8) | 82,919 (96.9) | 204,881 (97.5) |
| Gender (%) | Male | 89,810 (45.0) | 38,386 (44.9) | 96,045 (45.7) |
| | Female | 109,865 (55.0) | 47,190 (55.1) | 114,171 (54.3) |
| Age Group (%) | 30–36 | 40,392 (20.2) | 17,250 (20.2) | 50,989 (24.3) |
| | 37–45 | 42,346 (21.2) | 18,340 (21.4) | 51,938 (24.7) |
| | 46–54 | 42,383 (21.2) | 18,064 (21.1) | 37,929 (18.0) |
| | 55–63 | 35,903 (17.8) | 15,264 (17.8) | 33,752 (16.1) |
| | 64–90 | 38,651 (19.3) | 16,658 (19.5) | 35,608 (16.9) |
| Ethnicity (%) | Jewish | 140,459 (70.3) | 60,082 (70.2) | 160,168 (76.2) |
| | Arab | 54,941 (27.5) | 23,624 (27.6) | 39,401 (18.7) |
| | Other | 4,275 (2.1) | 1,870 (2.2) | 10,647 (5.1) |
| Immigrant Status (%) | Yes | 127,193 (63.7) | 54,587 (63.8) | 137,389 (65.4) |
| | No | 72,483 (36.3) | 30,989 (36.2) | 72,827 (34.6) |
| Socioeconomic Status (%) | Low | 76,315 (38.2) | 32,948 (38.5) | 75,076 (35.7) |
| | Medium | 63,032 (31.6) | 26,875 (31.4) | 58,538 (27.8) |
| | High | 54,175 (27.1) | 23,109 (27.0) | 48,858 (23.2) |

Table 4: Migration-induced shift experiment population characteristics by protected variables. Source population - native-born, Target population - Immigrants.

| **Variable** | **Levels** | **Source Train** | **Source Test** | **Target** |
|---|---|---|---|---|
| | All | 885,317 | 221,368 | 529,450 |
| Outcome (%) | 1 | 18,212 (2.06) | 4,570 (2.06) | 24,179 (4.57) |
| | 0 | 867,105 (97.94) | 216,798 (97.94) | 505,271 (95.43) |
| Gender (%) | Male | 420,926 (47.55) | 105,507 (47.66) | 214,951 (40.60) |
| | Female | 464,391 (52.45) | 115,861 (52.34) | 314,499 (59.40) |
| Age Group (%) | 29–34 | 182,908 (20.66) | 45,450 (20.53) | 35,967 (6.79) |
| | 35–40 | 184,056 (20.79) | 46,253 (20.89) | 41,205 (7.78) |
| | 41–48 | 177,649 (20.07) | 44,517 (20.11) | 53,422 (10.09) |
| | 49–57 | 173,628 (19.61) | 43,635 (19.71) | 90,686 (17.13) |
| | 58–90 | 167,076 (18.87) | 41,513 (18.75) | 308,170 (58.21) |
| District (%) | Tel-Aviv | 68,666 (7.76) | 17,062 (7.71) | 56,659 (10.70) |
| | Center | 105,102 (11.87) | 26,242 (11.85) | 90,102 (17.02) |
| | Eilat | 5,716 (0.65) | 1,469 (0.66) | 4,492 (0.85) |
| | Haifa | 164,739 (18.61) | 41,651 (18.82) | 89,084 (16.83) |
| | Jerusalem | 90,435 (10.21) | 22,653 (10.23) | 38,022 (7.18) |
| | North | 120,887 (13.65) | 30,062 (13.58) | 43,553 (8.23) |
| | Sharon | 147,441 (16.65) | 36,820 (16.63) | 73,289 (13.84) |
| | South | 83,975 (9.49) | 21,100 (9.53) | 74,801 (14.13) |
| | Petah-Tikva | 98,351 (11.11) | 24,309 (10.98) | 58,999 (11.14) |
| Socioeconomic Status (%) | Low | 276,390 (31.22) | 69,284 (31.26) | 131,814 (24.90) |
| | Medium | 234,448 (26.48) | 58,374 (26.37) | 208,605 (39.40) |
| | High | 322,706 (36.45) | 80,868 (36.51) | 167,035 (31.55) |

## Appendix B. Features Selection Distribution

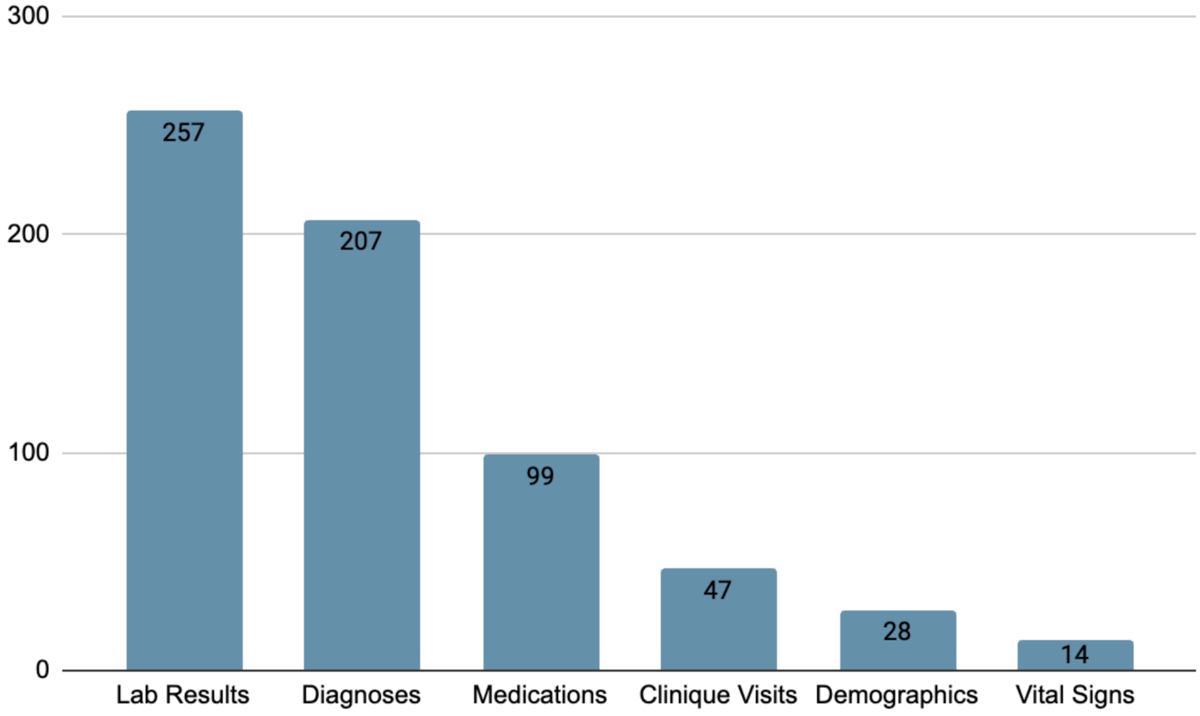

Figure 3: Number of features with positive gain for each data category in the XGBoost feature selection process.