Rural Medical Centers Struggle to Produce Well-Calibrated Clinical Prediction Models: Data Augmentation Can Help

Katherine E. Brown, PhD[1], Bradley A. Malin, PhD[1], Sharon E. Davis, PhD[1]
[1]Vanderbilt University Medical Center, Nashville, TN

Abstract

Machine learning models support many clinical tasks; however, challenges arise with the transportability of these models across a network of healthcare sites. While there are guidelines for updating models to account for local context, we hypothesize that not all healthcare organizations, especially those in smaller and rural communities, have the necessary patient volumes to facilitate local fine tuning to ensure models are reliable for their populations. To investigate these challenges, we conducted an experiment using data from a real network of hospitals to predict 30-day unplanned hospital readmission and a simulation study using data from a multi-site ICU dataset to evaluate the utility of synthetic data generation (SDG) to augment local data volumes. Several factors associated with rurality were correlated with model miscalibration and rural sites failed to meet sample size requirements for local recalibration. Our results indicate that deep learning approaches to SDG produced the best local classifiers.

Introduction

Artificial intelligence (AI) tools, including machine learning models, have increased in their adoption for a variety of clinical tasks, including predicting the likelihood of disease[1] or efficacy of medication[2] as well as assessing the likelihood of hospital discharge[3] or readmission[4]. Still, there are non-trivial challenges regarding the transportability of these models across sites[5–7]. A model trained at one site (i.e., hospital, floor, or ward) can deteriorate in overall reliability – measured as the calibration of model outputs to true likelihood of experiencing the outcome – when applied at other sites[8,9]. Various guidelines recommend updating machine learning models using data local to a site to account for this lack of transportability[10–12]; however, we hypothesize that not all healthcare organizations, especially those in smaller and rural communities, have the patient volume necessary to facilitate a meaningful update.

A significant portion of the US population resides in a rural area. As documented in the American Community Survey (ACS, 2018-2022)[13], approximately 56,000,000 people (~18% of the US population) reside in a ZIP Code Tabulation Area (ZCTA) designated as rural or borderline rural by the Centers for Medicare and Medicaid Services (CMS)[14]. Austen and colleagues[15] propose a framework to evaluate clinical prediction models across multiple geographic sites, there has been little investigation into the impact of rurality on model performance or the capabilities of smaller medical centers to update clinical prediction models for their local populations. As such, there is an urgent need to characterize the impact and applicability of model updating and recalibration guidelines on the smaller medical centers that serve these patients. While Austen and colleagues[15] propose a framework to evaluate clinical prediction models across multiple geographic sites, there has been little investigation into the impact of rurality on model performance or the capabilities of smaller medical centers to update clinical prediction models to fine-tune to their local populations.

Thus, in this work, we illustrate a proof-of-concept regarding the existing capabilities of rural medical centers to update clinical prediction models to ensure models are sufficiently well-calibrated for their patients. Additionally, in a simulation study, we evaluate the possibility of synthetic data generation[16,17] to augment existing local data as an option to mitigate sample size concerns, extending previous studies utilizing data augmentation to mitigate equity concerns in AI models[18,19]. We specifically investigate the following research questions (RQs): **RQ 1:** What is the impact of rurality on model discrimination and calibration? **RQ 2:** How well can satellite healthcare facilities update clinical prediction models based on existing guidelines? **RQ 3:** How well can data augmentation mitigate sample size concerns when updating clinical models at smaller/rural sites compared to global data available across all sites?

Methods

This study was approved by the Vanderbilt University Medical Center IRB (#240126).

Let us assume hospital H_0 represents the main hospital or academic medical center associated with a network of hospitals, while hospitals $H_1, H_2, \ldots, H_n$ are satellite hospitals focused on the care of local communities (Figure 1). Let T_{train_i} represent the set of patients at H_i during the model's training period and T_{eval_i} be the set of patients at H_i during the model's evaluation period. Let m be the model trained on T_{train_0}.

The model's prediction on a patient x, denoted a $m(x)$, can be further considered as the composition of $l(f(x))$. In this composition, f denotes a machine learning algorithm, such as a gradient-boosted tree or neural network, trained to produce a value corresponding to the logit of the likelihood of the event of interest. This logit is then provided to a calibrator model $l(\cdot)$ that produces the calibrated probability of the event of interest. Commonly, l is a logistic regression model. While the machine learning model f remains static after initial training, we can compose different calibrator models depending on the data used. For example, a calibrator derived from data from T_{train_i} (e.g., H_i) will be denoted as $l_i(\cdot)$, and a calibrator derived from combining the training data from all sites (e.g., $\bigcup_{i=0}^{n} T_{train_i}$) will be denoted as $l_A(\cdot)$. Riley et al.[10,11] derived several guidelines to determine the sample sizes needed to estimate recalibration parameters l_i.

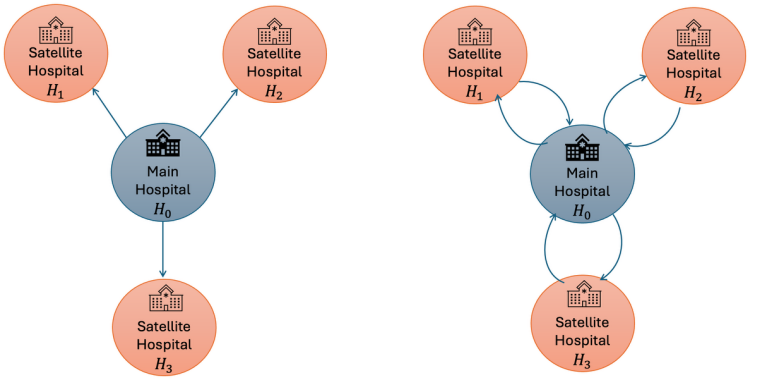

Figure 1. Schematic depicting the scenarios for synthetic data generation.

Given this framework, we can formally describe our main hypotheses as follows: **Hypothesis 1 (RQ 1)**: For sites $H_1, H_2, \ldots, H_n$, the optimal calibrator for a machine learning algorithm will not be $l_0(\cdot)$ (e.g., the calibrator derived from the training data of the main hospital/site). **Hypothesis 2 (RQ 1)**: There is an association between rurality and model performance. **Hypothesis 3 (RQ 2)**: Hospitals $H_1, H_2, \ldots, H_n$will not have the required amount of data in T_{train_i} to produce the optimal calibrator $l_i(\cdot)$**. Hypothesis 4 (RQ 3)**: Augmenting existing data in T_{train_i} will produce a more optimal $l_i(\cdot)$ compared to $l_A(\cdot)$.

Experiment 1: Evaluation of Real-World Clinical Prediction Model at Rural Sites.
For the remainder of this study, we refer to this experiment as the Real-World Evaluation.

Data and Model. To characterize the ability of local sites to update and recalibrate clinical prediction models, we consider a model currently in production in a network of hospitals in the Middle Tennessee area (see Table 1). This network includes one urban academic medical center and two satellite hospitals in predominantly rural communities. Table 1 presents each site along with the population density of the county in which the site is located, area deprivation index[20] (ADI), and whether or not the site is eligible for Health Resources and Services Administration (HRSA) rural health funding[21]. The model considered is the LACE+ Index[22] which predicts unplanned 30-day readmission or mortality. We collected the highest LACE+ prediction during each index admission (i.e., not an unplanned readmission) for adult patients at each site between August 2021 and December 2024 and whether the patient experienced an unplanned readmission in the 30 days following discharge.

Table 1. Distribution of patient cohorts, generalized area deprivation Index (ADI), population density to prevent identifiability of the sites. A higher ADI indicates higher disadvantage. Note: RH = Rural Health

Site	Population Density (people per sq. mile)	ADI (county-level)	HRSA RH Funding Eligible	Site Distribution			
				2021	2022	2023	2024
H_0	> 1,200	< 100	No	42,034	41,278	43,596	46,135
H_1	< 250	> 100	Yes	1,218	3,998	4,346	4,673
H_2	< 250	> 100	Yes	275	1,151	1,289	1,584

Experimental Methodology. Since we collected model predictions and outcomes from 2021 to 2024, we perform our evaluation temporally. For example, the experiment labeled "21-22" uses data from 2021 for calibration and data from 2022 for validation. Using the previously defined framework, we consider the model $l_0(f(x))$ calibrated on data from H_0. We also consider $l_A(f(x))$ which is derived using aggregate data from all sites. Finally, we consider models

calibrated on data from individual sites $l_i(f(x))$. All calibrator models are implemented using Platt's method[23]. We use Riley's sample size guidelines[10] for models with a binary outcome to determine if each individual site can recalibrate with existing data. If not, we resample the existing data at random to reach sample size requirements. Since our prediction task is binary, we use the C statistic to measure overall model discrimination. We calculate the calibration intercept α and slope β where ideal values are $\alpha = 0$ and $\beta = 1$, respectively.

We also evaluate the correlation, if any, between common factors associated with rurality and model performance based on Pearson's correlation coefficient. For the factors associated with rurality, we considered population density and ADI. Population density is often used as a defining characteristic for determining if a geographic area is rural[24,25]. ADI comprises of measures related to income, housing status, educational, employment status of residents in a geographic area[20] and is frequently associated with health in the rural US[26,27]. In addition to measuring correlation with $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and C, we also consider the magnitude of difference between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to their ideal values (i.e., $|\boldsymbol{\alpha}|$ and $|\boldsymbol{\beta} - \mathbf{1}|$) to determine if miscalibration is associated with rurality.

Experiment 2: Simulation of Data Augmentation to Mitigate Sample Size Concerns.

To characterize the efficacy of data augmentation to mitigate these sample size concerns, we developed a simulation using a publicly available critical care database that contains admissions across multiple sites nationwide[28]. We opted to use this external database since we did not have direct access to the all model inputs in the Real-World Evaluation. We assume a framework where a network of hospitals develops a machine learning model using data from the main site and then optimize the calibration of this model at the smaller community of hospitals within the network. In this experiment, we develop a model $l_0(f(x))$ using data from the main site (H_0). We then evaluate the impact of developing calibrator models l_i at each of the satellite hospitals H_i using data augmentation techniques in the form of oversampling and synthetic data generation[18,19] to achieve necessary data volumes and compare these local calibrators to a global calibrator l_A using data from all sites. For the remainder of this study, we refer to this experiment as the Simulation Study.

Table 2. Distribution of patients per site based on US region.

	South		**Northeast**	
Site No. (Bed Size)	**2014**	**2015**	**2014**	**2015**
H_0 (> 500)	20144	23344	6040	6042
H_1 (250-499)	12952	15330	1842	2024
H_2 (100-249)	9074	11492	1302	1258
H_3 (< 100)	3242	5840	1158	1136
Total	**45412**	**56006**	**10342**	**10460**

Data. We perform our analysis with data from the eICU Collaborative Research Database[28]. This database contains de-identified ICU admissions from 2014 and 2015 for patients at 208 sites across the US. Each site is affiliated with an aggregate indication of its number of beds (e.g., > 500, 250-499, 100-249, < 100) and US Census region (i.e., South, Northeast, Midwest, West). We combine individual sites based on size (e.g., all hospitals in each region with more than 500 beds are considered as the main site or H_0 for that region). For our simulation, we consider sites in the Northeast and South. In the southern US, there is a wide range in data volumes across hospital sizes; whereas, in the northeastern US, there is less variation in the number of patients by hospital size. Table 2 provides the size of each cohort based on hospital size and region. Patients with a discharge year of 2014 were partitioned into T_{train_i} and patients with a discharge year of 2015 were partitioned into T_{eval_i}. The eICU database contains features necessary to compute APACHE IV (Acute Physiology, Age, and Chronic Health Evaluation) hospital mortality prediction score. Using this features included in this score, the prediction task for our simulation is ICU mortality.

Models. The classification model f in our simulation is a CatBoost gradient-boosted tree[29] trained to predict ICU mortality based on the APACHE features. This model is trained using T_{train_0} – patients from H_0 with a discharge year of 2014 (see Figure 1). We use a random 75% of the data from H_0 in 2014 to train the gradient-boosted tree to serve as our clinical prediction model to mitigate the likelihood of overfitting. The 25% unseen patients from H_0 in 2014 are used to train the calibrator l_0. For sites H_1, H_2, and H_3, we use all data from 2014 to train the calibrator l_1, l_2, and l_3, respectively. The calibrator model l_i is implemented using Platt's method[23]. For H_1, H_2, and H_3, local calibrator models are trained on T_{train_i} – patients from H_i with a discharge year of 2014 – with data augmentation if applicable. The aggregate calibrator model l_A is trained on all available data across sites from 2014.

Data Augmentation. Let t_i be the number of additional samples determined to be necessary to augment T_{train_i} to produce a proper calibrator model at H_i. In the event that a site does not have the requisite number of samples for proper calibration, we hypothesize that data augmentation in the form of oversampling or synthetic data generation may reduce this burden on smaller medical centers and facilitate updating to improve performance.

Table 3. Description of data augmentation methods in the Simulation Study. Note: SDG = Synthetic Data Generation

	Name	Description
Scenario 1: Local sites have computational resources supporting synthetic data generation	SDG	Train a synthetic data generator on T_{train_i}. Generate t instances required to reach minimum sample size requirements.
	SDG + Oversample	Train a synthetic data generator on T_{train_i}. Generate $t/2$ instances from SDG. Oversample T_{train_i} with replacement for the remaining $t/2$ instances.
Scenario 2: Local sites lack computational resources supporting synthetic data generation	Oversample	Oversample T_{train_i} with replacement to reach the minimum sample size required.
	SDG, Conditional	Train a synthetic generator on $\bigcup_{i=0}^{n} T_{train_i}$. Sample t total instances with the desired site.

Modern synthetic data generation is dependent on modeling the univariate and multivariate distributions per feature (e.g., Gaussian Copula[30]) or by deep neural networks such as generative adversarial networks (GANs) and variational autoencoders (VAE)[17,31]. GANs commonly consist of two deep neural networks: a generator network that generates plausible synthetic instances and a discriminator network that determines if a data point is artificially generated. Alternatively, VAEs are built upon the autoencoders, which attempt to reconstruct data instances by reducing the representation to a latent space and expanding the latent representation into the reconstruction. Variational autoencoders build upon this network type by using a probabilistic model in the latent representation to encourage generation of synthetic instances. We choose to use a conditional GAN[32] (CTGAN), Tabular VAE (TVAE)[32], and Gaussian Copula (GaussianCopula)[30] implemented in the Synthetic Data Vault for our analyses. The computational requirements for the proposed synthetic data generation may be difficult for smaller medical centers to realize. Thus, we consider four possible data augmentation scenarios with oversampling and synthetic data generation (Figure 1 and Table 3)[18,19].

First, we consider the scenario when a satellite hospital H_i (with $i \geq 1$) has access to the computational resources necessary to augment their data with synthetic instances (Figure 1, Scenario 1). In this scenario, the local site will train a synthetic data generation model on their original data. Then, the local site could sample the t additional examples needed from the synthetic data generator. As an alternative approach with oversampling, the local site could sample $\lceil t/2 \rceil$ observations from the synthetic data generator and oversample another $\lfloor t/2 \rfloor$ from T_{train_i}(sampling with replacement if necessary). Second, we consider the scenario when a satellite hospital H_i (with $i \geq 1$) does not have access to the computational resources necessary to augment their data with synthetic instances (Figure 1, Scenario 2). Then, the local site would either be required to oversample from T_{train_i} or be dependent on H_0 to generate the necessary synthetic samples. In the latter approach, we assume H_0 has access to data at all sites to train a common synthetic data generation model. Then, we sample from the generator to produce t_i samples from H_i.

Experimental Methodology. We run the proposed simulation twice. Within each execution of the simulation, each data augmentation and respective recalibration is performed five times. This results in 10 measurements of calibration performance. Since our prediction task is binary, we use the C-statistic to measure overall model discrimination and measure calibration with the calibration intercept α and slope β, where ideal values are $\alpha = 0$ and $\beta = 1$, respectively.

Results

Impact of Rurality on Model Performance. The Real-World Evaluation assessed the discrimination and calibration of the LACE+ Index at a network of hospitals in the Middle Tennessee area. These hospitals include one urban academic medical center and two rural satellite hospitals. Table 4 presents the C-statistic along with the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of the calibration line of the model developed using data from the main site $\boldsymbol{H_0}$ and evaluated at all three sites $\boldsymbol{H_0}$, $\boldsymbol{H_1}$, and $\boldsymbol{H_2}$. We compute the 95% confidence intervals based on the performance at three evaluation scenarios across a total of four years. We note that C-statistic – which indicates the model's overall ability at identifying the outcome of interest – differs by as much as about 12% across the individual sites and $\boldsymbol{H_0}$ does not have the best discrimination. Model calibration – which is measured by the slope $\boldsymbol{\beta}$ and intercept $\boldsymbol{\alpha}$ of the calibration line formed by the model's

predictions – exhibits more noticeable differences across the sites. When using a calibrator derived from the main site (Table 4, $\boldsymbol{l_0}$), we found that rural sites experienced worse calibration than the primary urban site.

We found that model discrimination and model calibration curve coefficients themselves are not highly correlated with either factor of rurality (Table 5). For discrimination performance, this may be a by-product of the similarity of performance across included sites, and if other models or networks experience larger differences in discrimination across sites, this correlation may change. Interestingly, the magnitude of miscalibration ($|\boldsymbol{\alpha}|$ and $|\boldsymbol{\beta - 1}|$) was highly correlated with both considered factors of rurality. For population density, this is to be expected. Rural areas have lower population and as a result, lower population density. Since there are fewer people in the areas serviced by satellite medical centers, the overall patient population during a given period is expected to be smaller for satellite hospitals compared to their urban counterparts. A simplistic explanation is that fewer patients mean fewer samples that can be used for model calibration and validation, resulting in larger calibration error when a model is applied at satellite hospitals. Alternatively, population density may be a proxy variable for unmeasured or unmeasurable aspects of rurality. Additional, analyses will be necessary to confirm.

Table 4. C-statistics and calibration coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ across each year of calibration/evaluation based on calibrator.

	Site	C (95% CI)	$\boldsymbol{\alpha}$ (95% CI)	$\boldsymbol{\beta}$ (95% CI)
$\boldsymbol{l_0}$	$\boldsymbol{H_0}$	0.63 (0.63, 0.63)	0.1 (0.1, 0.11)	1.05 (1.05, 1.05)
	$\boldsymbol{H_1}$	0.68 (0.68, 0.68)	0.88 (0.88, 0.89)	1.59 (1.59, 1.59)
	$\boldsymbol{H_2}$	0.56 (0.56, 0.56)	-0.76 (-0.77, -0.76)	0.66 (0.66, 0.66)
$\boldsymbol{l_A}$	$\boldsymbol{H_0}$	0.63 (0.63, 0.63)	0.09 (0.09, 0.09)	1.03 (1.03, 1.03)
	$\boldsymbol{H_1}$	0.68 (0.68, 0.68)	0.87 (0.86, 0.87)	1.57 (1.56, 1.57)
	$\boldsymbol{H_2}$	0.56 (0.56, 0.56)	-0.76 (-0.77, -0.76)	0.65 (0.65, 0.66)
$\boldsymbol{l_i}$	$\boldsymbol{H_0}$	0.63 (0.63, 0.63)	0.1 (0.1, 0.11)	1.05 (1.05, 1.05)
	$\boldsymbol{H_1}$	0.68 (0.68, 0.68)	0.62 (0.6, 0.63)	1.25 (1.24, 1.26)
	$\boldsymbol{H_2}$	0.56 (0.56, 0.56)	0.3 (0.28, 0.31)	1.1 (1.09, 1.11)
$\boldsymbol{l_i}$ *OS*	$\boldsymbol{H_0}$	N/A	N/A	N/A
	$\boldsymbol{H_1}$	0.68 (0.68, 0.68)	0.21 (0.2, 0.22)	1.04 (1.04, 1.05)
	$\boldsymbol{H_2}$	0.56 (0.56, 0.56)	-0.3 (-0.33, -0.28)	0.78 (0.77, 0.8)

Table 5. Pearson correlation coefficients for factors associated with rurality (population density and ADI) compared with model discrimination, calibration, and miscalibration.

Factor Associated with Rurality	Year	$\boldsymbol{\alpha}$	$\boldsymbol{\|\alpha\|}$	$\boldsymbol{\beta}$	$\boldsymbol{\|\beta - 1\|}$	C
Population Density (in people/sq. mile)	21-22	0.09	-0.95	0.03	-0.91	0.17
	22-23	0.01	-0.95	-0.13	-0.81	0.11
	23-24	0.09	-0.95	0.03	-0.90	0.21
ADI	21-22	0.17	1	0.23	0.98	0.06
	22-23	0.25	1	0.38	0.94	0.15
	23-24	0.17	1	0.23	0.98	0.06

The correlation with Area Deprivation Index is also expected; however, the explanation for this correlation is not as straightforward**.** We calculated the ADI for the county in which each site is located. Higher ADI (i.e., higher disadvantage) is commonly associated with predominantly rural areas. Since population density is often a defining characteristic of rurality, the correlation between model miscalibration and ADI could be a by-product of existing correlation between ADI and population density ($\boldsymbol{\rho = -0.97}$). However, high ADI is not both necessary and sufficient to define a rural area (i.e., high ADI can occur in a predominantly urban area). Thus, another possible explanation is that sites located in areas with higher ADI (e.g., higher disadvantage) have the propensity to face greater model miscalibration concerns. Factors including accessibility of healthcare, potentially differing protocols or opportunities for follow-up, and different distributions of potential cases could result in models requiring more detailed maintenance than simple recalibration to best serve local populations. Further experiments with nationwide data across multiple models will be necessary to elucidate the nature of the relationship between ADI and model miscalibration.

Ability of Satellite Hospitals to Update Clinical Prediction Models. Additionally, in the Real-World Evaluation, we evaluated the capability of real-world data from a multi-hospital network to recalibrate a model at geographically distinct satellite sites. Table 6 provides the number of samples required for local site-specific recalibration, samples available, and the number of samples to be generated, if applicable. First, we consider using all available data at a local site, regardless of if the site has the requisite samples for local recalibration. We examined the C-statistic and

calibration coefficient ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) values of the site-specific calibrators $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$ (Table 4, $\boldsymbol{l_i}$). We see mild improvement in calibration at $\boldsymbol{H_1}$ when using available samples and more noticeable improvement in calibration at $\boldsymbol{H_2}$ when using available samples. Of note, the range of possible values in the 95% confidence interval increases at both sites from 0.00-0.01 to 0.02 to 0.03 indicating that while better calibration occurred at these sites, the variation in calibration performance also increased. This is likely a result of neither $\boldsymbol{H_1}$ or $\boldsymbol{H_2}$ having the recommended number of samples for optimal, local recalibration[10].

Table 6. Sample size requirements for each site in the Real-World Evaluation, aggregated across time. Number of samples required to generate is denoted #RG

Site	Samples Required (95% CI)	Samples Available (95% CI)	#RG (95% CI)
$\boldsymbol{H_0}$	14244 (14224, 14263)	42303 (42260, 42345)	-28059 (-28121, -27997)
$\boldsymbol{H_1}$	16461 (16350, 16572)	3187 (3125, 3249)	13274 (13103, 13445)
$\boldsymbol{H_2}$	15299 (15234, 15364)	905 (885, 925)	14394 (14315, 14474)

Thus, we consider a scenario in which available data across all sites are aggregated to form a single, multi-site calibrator (Table 4, $\boldsymbol{l_A}$). This approach resulted in calibration very closely aligned to using the calibrator derived from the main site, $\boldsymbol{l_0}$. This is likely due to the fact that samples from site $\boldsymbol{H_0}$ composes between 88% and 97% of available samples for recalibration for each year in which the data was available. This results in rural sites $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$ having a low prevalence among samples used for recalibration. Ultimately, this indicates that global recalibration may be ill-suited when patients at rural sites have low prevalence in the overall sample space for recalibration.

Finally, we considered developing site-specific calibrators, $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$, by oversampling (OS) available data to reach minimum sample size requirements (Table 4, $\boldsymbol{l_i\ OS}$). We find that for site $\boldsymbol{H_1}$, oversampling to reach sample size requirements was sufficient to successfully calibrate the local calibrator $\boldsymbol{l_1}$. For site $\boldsymbol{H_2}$, oversampling available data does not result in the same improvement in calibration as seen by $\boldsymbol{H_1}$. We suspect this is a result of the fact that $\boldsymbol{H_2}$ consists of less than 3% of all samples – the smallest site in our evaluated cohort. Thus, it is reasonable that the vast amount of oversampling required to reach minimum sample size requirements ultimately caused the calibrator to overfit.

Table 7. C-statistic and 95% confidence interval of model on unseen data in the Simulation Study. Calibrator used is the main site calibrator, l_0

	South			Northeast		
Site No.	C (95% CI)	$\boldsymbol{\alpha}$ (95% CI)	$\boldsymbol{\beta}$ (95% CI)	C (95% CI)	$\boldsymbol{\alpha}$ (95% CI)	$\boldsymbol{\beta}$ (95% CI)
H_0	0.85 (0.85, 0.85)	-0.38 (-0.48, -0.28)	0.79 (0.76, 0.81)	0.81 (0.81, 0.82)	-0.57 (-0.58, -0.56)	0.57 (0.57, 0.58)
H_1	0.85 (0.85, 0.85)	-0.45 (-0.53, -0.37)	0.79 (0.77, 0.81)	0.81 (0.81, 0.82)	-0.84 (-0.85, -0.84)	0.51 (0.5, 0.51)
H_2	0.84 (0.84, 0.84)	-0.41 (-0.47, -0.35)	0.85 (0.83, 0.86)	0.81 (0.8, 0.82)	-1.09 (-1.18, -1)	0.62 (0.6, 0.65)
H_3	0.84 (0.84, 0.84)	-0.23 (-0.35, -0.11)	0.84 (0.81, 0.87)	0.79 (0.79, 0.8)	-0.87 (-0.92, -0.82)	0.54 (0.53, 0.56)

Data Augmentation to Mitigate Sample Size Concerns. As evidenced in our findings from the Real-World Evaluation, it can be seen that satellite hospitals within a network may experience poor calibration when applying models calibrated on data from a primary hospital and oversampling to reach minimum sample size requirements may be insufficient for recalibration. To better emulate a local network of hospitals, we combined individual sites based on size (see Table 2). Additionally, we note in Table 2 that the northeast US has fewer patients in this database compared to the southern US. One satellite site in the southern US and all three satellite sites in the northeast US required data augmentation to reach the minimum sample size for recalibration.

Table 7 presents the C-statistic and 95% confidence interval when $\boldsymbol{H_0}$ trains and calibrates a clinical prediction model on $\boldsymbol{T_{train_0}}$ and provides this model to $\boldsymbol{H_1}$, $\boldsymbol{H_2}$, and $\boldsymbol{H_3}$ without local recalibration. We note that there are minor differences in C-statistics across sites; however, we note moderate variability in calibration across different sites. Given the observed differences in calibration, we calculated the minimum sample size required for estimating recalibration parameters according to Riley and colleagues[10] (Table 8). We note that for sites in the southern US, only $\boldsymbol{H_3}$ requires data augmentation to meet sample size requirements, but for the northeast US, all three satellite sites $\boldsymbol{H_1}$, $\boldsymbol{H_2}$, and $\boldsymbol{H_3}$ require data augmentation. We note that generating each of these sites required combining data from hospitals of similar sizes within each region into $\boldsymbol{H_0}$, $\boldsymbol{H_1}$, $\boldsymbol{H_2}$, and $\boldsymbol{H_3}$. Even in this artificial scenario, there are

instances where recalibration cannot be adequately performed with the available sample sizes. This further motivates the need to understand the impact of data augmentation for recalibrating models at sites with limited sample sizes.

Figure 2 presents a two-dimensional plot of the calibration intercept $\boldsymbol{\alpha}$ on the x-axis and calibration slope $\boldsymbol{\beta}$ on the y-axis. The closeness of the point $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is to (0,1) indicates the success of the recalibration.

Table 8. Sample size requirements for each site in the Simulation Study, aggregated across time. Rows in bold indicate sites in which data augmentation is necessary. Number of samples required to generate is denoted #RG

	South			Northeast		
Site	**Samples Required**	**Samples Available**	**#RG**	**Samples Required**	**Samples Available**	**#RG**
H_0	6007	8810	N/A	3395	2586	809
H_1	6961	12952	N/A	3555	1302	2253
H_2	6854	9074	N/A	4309	1842	2467
H_3	5787	3242	2545	4645	1158	3487

For the APACHE model in the southern US, sites $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$ had the requisite samples for recalibration, and we found that using locally available data for these sites yielded the best calibration on unseen data. Of note, $\boldsymbol{H_2}$ performance displayed higher variance than $\boldsymbol{H_1}$ and has a smaller number of available instances for recalibration and evaluation. This could be a result of an overall smaller pool of data used to evaluate calibration at this site. Only $\boldsymbol{H_3}$ required data augmentation to reach sample size requirements, and we find that training CTGAN on data from all sites then conditionally sampling to retrieve instances for $\boldsymbol{H_3}$ to yield the most calibrated predictions across data augmentation strategies.

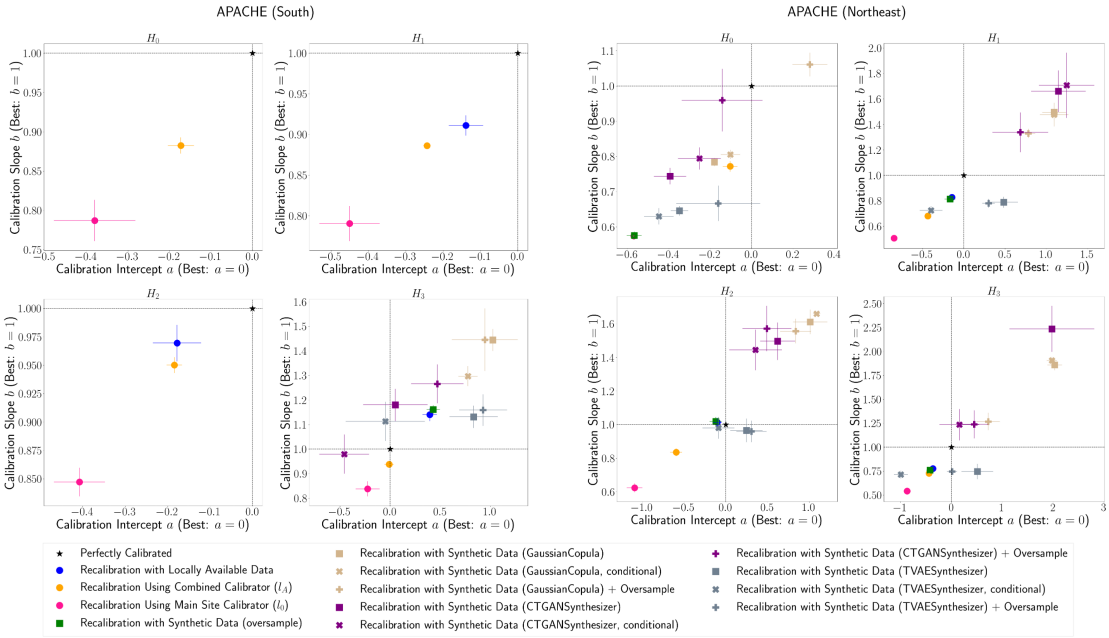

Figure 2. Plots of slope/intercept of the calibration curve for each data augmentation method.

For the APACHE model in the northeast US, we find that all three satellite sites, $\boldsymbol{H_1}$, $\boldsymbol{H_2}$, and $\boldsymbol{H_3}$ required data augmentation to meet the sample size requirements for recalibration. For sites $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$, oversampling or using locally available data resulted in the best performance, with conditionally sampled TVAE-generated data resulting in the best calibrator using synthetic data generation. For $\boldsymbol{H_3}$, the combination approach using synthetically generated data and oversampled local data produced the best local calibrator. We note for this simulation scenario there is generally less available data than for the APACHE model in the northeastern US. This lack of data could result in the synthetic data generators not producing adequate data for recalibration and as a result recalibration efforts benefitting more from a combined approach of synthetic data and oversampling.

Both geographic regions had differing capabilities at the primary site $\boldsymbol{H_0}$. In the southern US, the main site had the requisite samples for recalibration; however, calibration performance using the main site calibrator $\boldsymbol{l_0}$ had higher variation than the combined calibrator $\boldsymbol{l_A}$. For the northeast US, the main site did not have the requisite number of

samples for local recalibration; however, we see that combining synthetically generated data and oversampling the existing data resulted in a better calibrated classifier than the combined calibrator $\boldsymbol{l_A}$.

Discussion and Conclusion

Real-World Evaluation. Our analysis illustrates mild differences in discrimination and moderate deterioration in calibration when applying a model developed at a primary, urban site to satellite, rural sites. This is not unexpected in the context of previous studies[8,9,15]. Austin and colleagues[15] developed validation techniques of prediction models across multiple sites. In their evaluation of a mortality prediction model for heart failure patients, they noticed more pronounced differences in C-statistic than in our study; however, we noticed more variation across sites in both calibration slope and intercept than their study. While this may be an artifact of the different patient cohort and prediction model examined in both studies, both studies highlight the growing need for consideration of local validation for models applied at multiple sites, even within the same network of hospitals.

While differences in the C-statistic were minimal for this model and scenario, further evaluation of other models at these sites and at other urban-rural hospital networks will be necessary to confirm the nature of intra-network performance variation. Conversely, we noticed more extensive calibration differences across geographic sites and suspect that this will be a trend for other models and networked hospitals. Still, additional external analysis is necessary to confirm. Nonetheless, we note that our findings are consistent with prior work that indicates calibration is more sensitive to differences between training and validation cohorts than discrimination[8,9].

Lyons and colleagues[33] conducted an external validation of a proprietary sepsis model across a network of nine hospitals. In their study, the evaluated model was developed by an EHR vendor and provided for immediate use by institutions. The model experienced pronounced differences in discrimination performance across sites. Additionally, the authors observed correlation between C-statistics and multiple variables associated with sepsis diagnosis, including time to diagnosis, sepsis incidence rates, comorbidity burden, and cancer prevalence. These factors, except time to diagnosis, were noted to be inversely correlated to model discrimination performance. Their findings and those from our study highlight the need to investigate clinical and socio-economic factors that could be associated with degradation in model discrimination and calibration.

Simulation Study. In light of concerns posed when evaluating calibration of models across geographic sites, we performed a simulation study with a publicly available multi-site ICU database. We evaluated the efficacy of synthetic data generation using sites in the northeast and southern US. We find that for the southern US, deep learning-based approaches to generate sufficient data for local recalibration resulted in classifiers that were closest to a perfect calibration; however, for the northeast US, using existing data and oversampling was the most advantageous for two of the three satellite sites. Since the southern US had more data overall, it is reasonable to hypothesize that the deep learning approaches may have overfit the data in the northeast US, resulting in less robust synthetic instances. This hypothesis is further validated by Yan et al.[17] who found that using deep learning approaches on samples and concepts with low prevalence may result in statistical properties being inadequately learned by GAN-based synthetic data generation techniques. Overall, though, despite concerns about smaller sample sizes used to train synthetic data generation techniques, deep learning-based approaches typically produced better samples for recalibration than the statistical synthetic data generation technique tested. This indicates that deep learning-based approaches hold more promise for producing samples for recalibration than statistical data generation techniques. Additionally, the benefits of using synthetic data generation to produce additional samples for recalibration is further demonstrated when the main site itself does not have the requisite samples for recalibration, as we observed for the northeast US (Figure 2).

Additionally, we find that recalibration was more successful using data generated by deep learning-based approaches when such generator models were trained on all available data as opposed to at individual sites (e.g., conditional sampling approach). While this would logically follow from our previous hypothesis and be the most accessible for sites with more limited computational or technical staffing resources, it poses potential concerns regarding data access and privacy. This conditional approach requires that the main site has access to data from each satellite site, which may not be true or feasible for some networks of institutions. This could lead to a situation where a site does not have the computational resources to perform their own synthetic data generation, but the main site does not have the necessary access privilege to perform the synthetic data generation. Thus, further research is necessary to understand when these scenarios occur and how to make synthetic data generation and other data augmentation methods more applicable for these sites.

Finally, guidelines for minimum sample size for estimating recalibration parameters also apply to determining the number of samples necessary to perform validation of model discrimination and calibration. Since the southern US had a higher data volume than the northeast US, the southern sites had the minimum sample size for recalibration and validation. The northeast US, however, did not meet the sample size requirements for robust validation (see Tables 4 and 9). Thus, another possible reason as to why synthetic data generation was not the best data augmentation technique could be due to an insufficient data volume for validation to provide a complete picture of model calibration and discrimination. Performing data augmentation for validation was outside of the scope of this work and require additional considerations to prevent data leakage and evaluate whether generated data are robust for model validation.

Limitations and Conclusions. There are several avenues for future work to enhance the robustness of our findings. First, we only consider the scenario in which a machine learning model is trained using main site data. Future work should also consider the scenario with a model trained across all sites and other validation techniques for multi-site models[15]. Additionally, we focused on synthetic data generation with little fine-tuning, nor do we use a synthetic data generation method developed specifically for the purposes of synthetic EHR record generation. Finally, future work should consider larger cohorts like the All of Us Research Program[34] and data from other multi-center hospital networks.

In this work, we explore the transportability of clinical prediction models across geographically diverse sites and the ability of small, rural healthcare centers to locally recalibrate models given available data resources. While clinical prediction tools can help realize increased health outcomes, the urban-rural divide in healthcare in the US also extends to the localization of clinical prediction models to smaller healthcare facilities. Through our analysis, we see that connections to large medical centers is not enough to promote accurate prediction models at all sites within a healthcare system, and data augmentation and synthetic data augmentation may provide the necessary data volumes to enable local recalibration at smaller facilities. Data augmentation approaches, however, may assume unrestricted access to local data volumes or local access to computational resources which may be infeasible for some healthcare networks. Addressing these needs will require novel approaches that balance local and network-supported data augmentation.

Acknowledgements
This research was funded by the National Institutes of Health grants T15LM007450 and U54HG012510.

References

1. Manrai AK. Physicians, Probabilities, and Populations—Estimating the Likelihood of Disease for Common Clinical Scenarios. JAMA Internal Medicine. 2021 Jun 1;181(6):756–7.
2. Lewin-Epstein O, Baruch S, Hadany L, Stein GY, Obolski U. Predicting Antibiotic Resistance in Hospitalized Patients by Applying Machine Learning to Electronic Medical Records. Clinical Infectious Diseases. 2021 Jun 1;72(11):e848–55.
3. Zhang X, Yan C, Malin BA, Patel MB, Chen Y. Predicting next-day discharge via electronic health record access logs. Journal of the American Medical Informatics Association. 2021 Nov 25;28(12):2670–80.
4. Rumshisky A, Ghassemi M, Naumann T, Szolovits P, Castro V, Mccoy T, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. TRANSLATIONAL PSYCHIATRY. 2016 Oct 18;6.
5. Lasko TA, Strobl EV, Stead WW. Why do probabilistic clinical models fail to transport between sites. npj Digit Med. 2024 Mar 1;7(1):1–8.
6. Bennett T, Russell S, King J, Schilling L, Voong C, Rogers N, et al. Accuracy of the Epic Sepsis Prediction Model in a Regional Health System [Internet]. arXiv; 2019 [cited 2024 Jan 15]. Available from: http://arxiv.org/abs/1902.07276
7. Habib AR, Lin AL, Grant RW. The Epic Sepsis Model Falls Short-The Importance of External Validation. JAMA Intern Med. 2021 Aug 1;181(8):1040–1.
8. Davis S. Calibration Drift Among Regression and Machine Learning Models for Hospital Mortality.
9. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. Journal of the American Medical Informatics Association. 2017 Nov 1;24(6):1052–61.
10. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Statistics in Medicine. 2021;40(19):4230–51.

11. Riley RD, Collins GS, Ensor J, Archer L, Booth S, Mozumder SI, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. Statistics in Medicine. 2022;41(7):1280–95.
12. Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. Journal of the American Medical Informatics Association. 2019 Dec 1;26(12):1448–57.
13. Census Bureau Tables [Internet]. [cited 2024 Feb 8]. Available from: https://data.census.gov/table
14. Ambulance Fee Schedule & ZIP Code Files | CMS [Internet]. [cited 2024 Feb 8]. Available from: https://www.cms.gov/medicare/payment/fee-schedules/ambulance
15. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. J Clin Epidemiol. 2016 Nov;79:76–85.
16. Liu Y, Stouffs R, Theng Y. Development of Synthetic Patient Data to Support Urban Planning for Public Health. In: Werner L, Koering D, editors. Nanyang Technological University. 2020. p. 315–22.
17. Yan C, Zhang Z, Nyemba S, Li Z. Generating Synthetic Electronic Health Record Data Using Generative Adversarial Networks: Tutorial. JMIR AI. 2024;3:e52615.
18. Jackson N, Yan C, Malin B. Enhancement of Fairness in AI for Chest X-ray Classification. In: AMIA Annual Symposium Proceedings. 2025.
19. Theodorou B, Danek B, Tummala V, Kumar SP, Malin B, Sun J. Improving medical machine learning models with generative balancing for equity and excellence. npj Digit Med. 2025 Feb 14;8(1):1–11.
20. Singh GK. Area Deprivation and Widening Inequalities in US Mortality, 1969–1998. Am J Public Health. 2003 Jul;93(7):1137–43.
21. Rural Health Grants Eligibility Analyzer [Internet]. [cited 2024 Feb 23]. Available from: https://data.hrsa.gov/tools/rural-health
22. van Walraven C, Wong J, Forster AJ. LACE+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. Open Medicine. 2012;6(3):e80.
23. Platt J, others. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. 1999;10(3):61–74.
24. USDA ERS - What is Rural? [Internet]. [cited 2024 Jan 12]. Available from: https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications/what-is-rural/
25. How We Define Rural | HRSA [Internet]. [cited 2025 Mar 11]. Available from: https://www.hrsa.gov/rural-health/about-us/what-is-rural
26. Balio CP, Galler N, Mathis SM, Francisco MM, Meit MB, Kate E. Use of the area deprivation index and rural applications in the peer-reviewed literature. Rural Health Equity Research Center. 2024;
27. Fairfield KM, Black AW, Ziller EC, Murray K, Lucas FL, Waterston LB, et al. Area deprivation index and rurality in relation to lung cancer prevalence and mortality in a rural state. JNCI cancer spectrum. 2020;4(4):pkaa011.
28. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018 Sep 11;5(1):180178.
29. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support.
30. Benali F, Bodénès D, Labroche N, de Runz C. MTCopula: Synthetic complex data generation using copula. In: 23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP). 2021. p. 51–60.
31. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020 Oct 22;63(11):139–44.
32. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN. In: Advances in Neural Information Processing Systems. 2019.
33. Lyons PG, Hofford MR, Yu SC, Michelson AP, Payne PRO, Hough CL, et al. Factors Associated With Variability in the Performance of a Proprietary Sepsis Prediction Model Across 9 Networked Hospitals in the US. JAMA Internal Medicine. 2023 Jun 1;183(6):611–2.
34. The “All of Us” Research Program. The new england journal of medicine. 2019;