Collecting data when missingness is unknown: a method for improving model performance given under-reporting in patient populations

Kevin Wu KEVINYWU@STANFORD.EDU
Stanford University, Optum Labs

Dominik Dahlem DOMINIK_DAHLEM@OPTUM.COM
Optum Labs

Christopher Hane CHRISTOPHER.HANE@OPTUM.COM
Optum Labs

Eran Halperin ERAN.HALPERIN@UHG.COM
Optum Labs

James Zou JAMESZ@STANFORD.EDU
Stanford University, Optum Labs

Abstract

Machine learning models for healthcare commonly use binary indicator variables to represent the diagnosis of specific health conditions in medical records. However, in populations with significant under-reporting, the absence of a recorded diagnosis does not rule out the presence of a condition, making it difficult to distinguish between negative and missing values. This effect, which we refer to as latent missingness, may lead to model degradation and perpetuate existing biases in healthcare. To address this issue, we propose that healthcare providers and payers allocate a budget towards data collection (eg. subsidies for check-ups or lab tests). However, given finite resources, only a subset of data points can be collected. Additionally, most models are unable to be retrained after deployment. In this paper, we propose a method for efficient data collection in order to maximize a fixed model's performance on a given population. Through simulated and real-world data, we demonstrate the potential value of targeted data collection to address model degradation.

1. Introduction

ML algorithms that use electronic health records (EHR) aim to capture a patient's health status through data like ICD codes, clinical notes, and lab values (Goldstein et al. (2017); Rhee and Klompas (2020); Seymour et al. (2016); Henry et al. (2019)). However, in practice, medical records are prone to missingness due to recording issues and low healthcare utilization (Goldstein et al. (2017); Haneuse and Daniels (2016)). For example, one study found that 89% of acute psychiatric services were not captured in EHR (Madden et al. (2016)), while another found that upwards of 80% of lab values were missing per patient (Tan et al. (2023)). Under high rates of missing data, ML algorithms can perpetuate existing disparities in healthcare, as healthcare utilization can be confounded with protected attributes like race and socioeconomic status (Pierson et al. (2021); Rajkomar et al. (2018); Goodman et al. (2018); Ferryman and Pitcan (2018); Nordling (2019); Vyas et al. (2020)). In this study, we are particularly interested in data missingness in a target population that a model is deployed on. First, while data may be carefully curated during the training and evaluation phase, there are fewer guarantees of data quality once deployed at new institutions (Wu et al. (2021)). Second, while patterns of missingness can be learned in training and initial evaluation phases, models are often deployed to multiple institutions, making this problem difficult to address in model development alone (Futoma et al. (2020)).

Latent missingness Typically, patients with missing features either have their values imputed or may

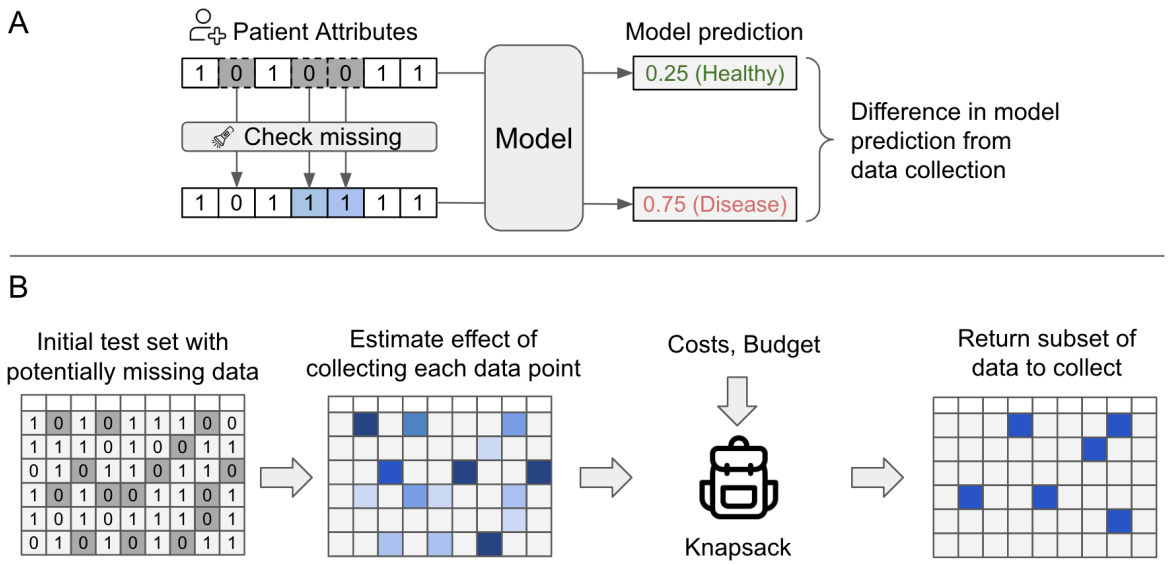

Figure 1: (1A) A schematic of how data collection on potentially missing binary features influences model predictions. For a given patient's attributes, zero values which may mask missing features are checked through an intervention (eg. check-up, lab test, or survey). Changes in the patient values resulting from data collection in turn influence the model's prediction. (1B) A diagram of the steps of our algorithm. Given a set of potentially missing data points, we estimate the value of collecting each one. The costs, budget, and values are inputs to the 0-1 knapsack algorithm, returning the subset of data that maximizes model performance while under budget.

be dropped from the dataset altogether (Hu and Du (2020); Wells et al. (2013)). However, missingness is not always indicated in medical records. EHR-based models commonly use binary features that denote the presence of a health condition, where the positive value *1* indicates the condition is found in medical records, and the negative value *0* indicates otherwise. This binary encoding scheme is prone to *latent missingness*, where the negative value *0* can either refer to the absence of a condition or the failure to capture the diagnosis. As an example, suppose a model considers patients to have had sepsis if the ICD code *A41.9* is found in their hospital encounters. Previous work has found that clinicians were less likely to document severe sepsis when patients were less ill (Whittaker et al. (2013)). In this situation, it would be unclear whether patients with negative values actually had sepsis, masking the missingness in data. In theory, missingness can be indicated when explicit negative diagnoses are not found (eg. NaN values). However, in practice, negative diagnoses may be missing for a high percentage of patients when not recorded for billing purposes (eg. ICD or CPT codes) (Romano and Mark (1994); Fette et al. (2018); De Achaval et al. (2013)). As such, methods like data imputation or deletion would be infeasible.

Targeted data collection As a response, we propose an approach where a monetary budget is spent towards targeted data collection. Possible examples of budget-driven data collection include:

1. Partial or complete subsidies for patient check-ups, screening exams, or lab panels.
2. Data cleaning measures (eg. reconciling medical records across multiple institutions, patient questionnaires for self-reported data).
3. Subjecting individual patient data to retrospective chart review by physicians, which have been shown to add sensitivity when compared to ICD-based assessments (Campbell et al. (2011)).

Healthcare providers and payers who deploy ML algorithms face competing interests. First, deployed

models need to be fair with respect to protected patient subgroups. Beyond moral imperatives, federal and state regulatory agencies increasingly regulating AI algorithms to address bias (Pesapane et al. (2018)). Second, healthcare delivery is always resource-constrained, with cost-savings as a major factor in ML adoption (Callahan and Shah (2017); Sujith et al. (2022); Escudero et al. (2012)). In order to balance these two opposing forces, providers need methods to maximize the efficiency of data collection. A data-centric approach to addressing underreporting in data can be used in parallel with model development efforts to prevent model degradation.

In our paper, we provide the following analyses: (1) we formulate the cost-performance tradeoff as a combinatoric optimization problem, (2) we propose a Shapley value-based method for estimating the value of collecting individual data points, and (3) we show that model-guided data collection strategies can significantly outperform random data collection without knowledge of ground-truth labels.

2. Related Works

Latent missingness The concept of unknown missing variables is found in positive and unlabeled (PU) learning (Bekker and Davis (2020); Wawrzeńczyk and Mielniczuk (2022)), where models only have access to positive and unlabeled data. In PU learning, only positive labels for y are known and the negative labels are dispersed within the remaining labels. Previous work has also studied the effects of biases when establishing ground-truth labels for healthcare algorithms (Obermeyer et al. (2019); Chang et al. (2022)). In contrast, we focus on unknown missingness when applied to binary features within X. Previously, Zhou et al. (2022) uses the synonymous term *missingness without indication* to describe cases where missingness is not known. The paper studies the effects of distribution shifts in patterns of missingness in X between training and target populations, and proposes an analytic adjustment for linear models. We extend this work to the case where the model cannot be re-trained and a budget can be used to reveal missing features.

Active learning Budget-constrained data collection is a common theme in active learning, which has the goal of identifying a subset of points to label in order to improve model training (Settles (2009)). This approach has been used specifically on EHR data in order to efficiently label outcome conditions (Chen et al. (2013); Nissim et al. (2016); Ji et al. (2019). Active learning methods commonly use model gradients to guide data selection (Ash et al. (2019); Bouneffouf (2016); You et al. (2014); Settles and Craven (2008)) and recently, a Shapley-value based framework has been proposed as well (Ghorbani et al. (2022)). Work by Natarajan et al. (2018) focuses on active feature elicitation, where the goal is to find an optimal set of examples where collecting missing features for those given examples can best improve classifier performance. This is related to our work in that missing features are collected in reference to the objective of model performance. In contrast, our work focuses on a fixed model, and we focus on the case where missingness is also unknown. Kanani and Melville (2008) focuses on prediction-time active feature-value acquisition, where a subset of a model's features can be acquired for a given cost during inference. This work shares the theme of cost-constraint data collection during inference, but differs from our work in that (1) the set of missing features are fixed and known, and (2) they use a classifier's uncertainty rather than feature sensitivity to guide data acquisition.

Missing features There has been extensive previous work on handling missing features in EHR through data imputation (Hu and Du (2020); Wells et al. (2013)), autoencoders, and representation learning (Malone et al. (2018); Beaulieu-Jones et al. (2017); Xu et al. (2021)), and discarding values (Little and Rubin (2019)). However, such approaches assume that missingness is made known. Unknown missingness is also referred to in survey literature as false negative responses to questions (eg. maternal smoking in epidemiological studies) (Sechidis et al. (2017)). Prior work in this space has involved estimating rates of missingness by incorporating prior knowledge and statistical tests (Sechidis et al. (2017); Lazkecka et al. (2021)) with the goal of inference, rather than prediction.

3. Preliminaries

Notation: We let (X, Y, Z) be a random vector that refers to the observed data, the labels, and the true data, respectively. The support of these distributions are $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \subseteq \{0,1\}^{d_X} \times \mathbb{R} \times \{0,1\}^{d_X}$ for some $d_X \in \mathbb{N}$. Additionally, we use a binary matrix $O \in \{0,1\}^{N \times d_X}$ to indicate whether data are observed, such that $O_{ij} = \mathbb{1}(x_{ij}$ is observed). Fi-

nally, $C \in \mathbb{R}^{N \times d_X}$ is a cost matrix representing the marginal cost of collecting each data point in X. For data points already collected, (ie. $o_{ij} = 1$), the marginal cost is 0. The model's performance at a data $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is given by $\mathcal{L}(h(x), y)$ for some loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and a fixed, pre-trained model $h : \mathcal{X} \to \mathcal{Y}$.

Assumption 1 (Latent missingness) *For a given event $z_{ij} \in \{0, 1\}$, let $o_{ij} \in \{0, 1\}$ be a latent variable indicating whether z_{ij} was observed. A feature x_{ij} with latent missingness is encoded as follows:*

$$x_{ij} = \begin{cases} z_{ij}, & \text{if } o_{ij} = 1 \\ 0, & \text{if } o_{ij} = 0 \end{cases}$$

In particular, we do not make any assumptions about the type of missingness in the dataset (eg. Missing Completely at Random (MCAR) vs. Missing Not at Random (MNAR)).

Definition 1 (Data collection) *We refer to data collection as a secondary check on existing data, as opposed to the process of gathering new data points. For example, if the data state that a given patient does not have hypertension, data collection would refer to an intervention (such as a check-up) that verifies this value. The new data point would then be updated in the EHR for the model to use. In our notation, z_{ij} would refer to the true value of patient i's feature j, whereas x_{ij} refers to the initially observed (potentially masked) value. The indicator matrix O_{ij} refers to all points which are not potentially missing. In the most naive setting, we can set $O = X$, where all zeros are considered potentially missing. In practice, a subset of zeros may have verified negative tests. We assume that the true value of a binary variable is always able to be unmasked through data collection.*

4. Missing Feature Discovery

The goal of our method is to identify the subset of potentially missing data to collect in an evaluation set that will yield the greatest improvement in model performance on the whole set. For example, a healthcare provider or payer may be faced with the decision to either subsidize metabolic panels for 100 high-risk patients or general check-ups for 200 low-risk patients. In this case, the model's improvement from accessing true lab values may be greater than the potential conditions revealed in the 200 check-ups. However, the lab tests may be more expensive than the check-ups, posing a trade-off between budget and model performance.

4.1. Optimization problem

We pose this trade-off as the following combinatoric optimization problem:

$$\max_{o_{ij}} \sum_{i}^{N} \sum_{j}^{d_X} V_{ij} o_{ij} \tag{1}$$

$$\text{s.t.} \sum_{i}^{N} \sum_{j}^{d_X} C_{ij} o_{ij} < b \tag{2}$$

Here, o_{ij} is an indicator variable for whether we perform data collection on x_{ij}, and C and b are the costs and total budget, respectively. Finally, V is our proxy for the performance improvement from collecting x_{ij}, which we define below. This combinatoric optimization problem is solvable with the 0-1 Knapsack algorithm, which identifies the maximum value which can satisfy a given budget constraint given V.

Definition 2 *The value of the i^{th} individual's j^{th} feature is defined as the difference in model performance after collecting x_{ij}, where $\Delta_{ij} \triangleq z_{ij} - x_{ij}$ and $\mathbf{e}_j$ is the unit vector with zeros everywhere except the j^{th} element:*

$$V_{ij} \triangleq \mathcal{L}(y_i, h(x_i)) - \mathcal{L}(y_i, h(x_i + \mathbf{e}_j \Delta_{ij}))$$

Our approach makes the simplifying assumption of linearity in model improvement through $\sum_i^N \sum_j^{d_X} V_{ij} o_{ij}$. In practice, the model's improvement from one data point collection can be dependent on other data points. We aim to address this effect through an observation-dependent Shapley value, which we describe below.

Value of data collection depends on feature sensitivity The change in loss from data collection can be approximated using a first-order Taylor expansion into the following three parts: The data collection's effect on x_i, the feature sensitivity, and the loss sensitivity:

$$V_{ij} \approx \underbrace{\Delta_{ij}}_{\text{Change in } x_i} \times \underbrace{h(x_i + \mathbf{e}_j) - h(x_i)}_{\text{Feature sensitivity}} \times \underbrace{\frac{\partial \mathcal{L}(y_i, h(x_i))}{\partial h(x_i)}}_{\text{Loss sensitivity}}$$

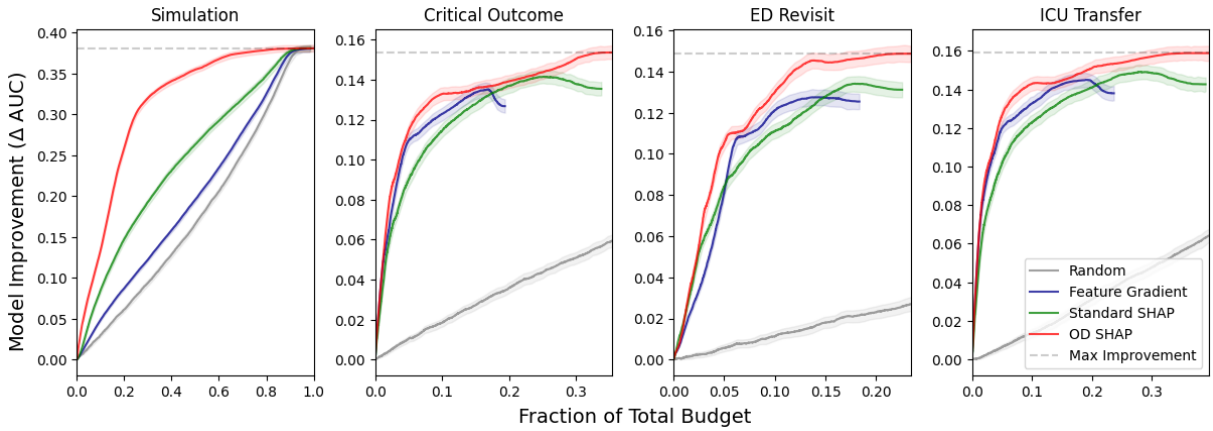

Figure 2: Cost curves displaying the efficiency of each method in improving model performance across data points collected. The x-axis represents the fraction of the total number of potentially missing data points collected as guided by each method, while the y-axis represents the improvement in model performance after those data points have been collected. Improvement is measured by evaluating the model's performance (AUC) after N data points have been collected. Each curve's solid line represents the average performance across 500 runs, while the shaded region represents the 95% confidence interval at each time step. The dotted horizontal line refers to the maximum model performance after collecting all potentially missing data points. For visual clarity, the x-axis is cropped to fit curves for the three non-random methods, so the curves for random allocation may extend beyond the graphs.

Intuitively, this means that the estimated change in loss is composed of (A) whether data collection will change the data, (B) how much the model's prediction changes as the data changes, and (C) how much the loss changes as the model's prediction changes. Direct calculation of the change in x_i and the loss sensitivity requires knowledge of the true value of features, z_{ij}, and the ground-truth label y_i, which are not available during model deployment. However, the feature sensitivity, $h(x_i+\mathbf{e}_j)-h(x_i)$, is available with only access to model inference. As such, our analysis focuses on ways to estimate feature sensitivity.

Estimating feature sensitivity when other features change As data is collected sequentially, the feature sensitivity at time t depends on the order of observations $\{O^{(0)}, ...O^{(t)}\}$. In non-linear models, interactions between features mean that the sensitivity is non-stationary as data is collected. For example, if two potentially missing lab test features interact in a model, the model's sensitivity for lab test A depends on the current value of lab test B, which may or may not be missing. To account for possible changes in other features when computing feature sensitivity, we instead aim to estimate the average marginal change in model prediction over all permutations of possible changes due to data collection. This can be formulated as the Shapley value of each feature over a subset of unobserved features:

Definition 3 (Observation Dependent SHAP) *Fix $i \in \{1,\dots,n\}$, let $G_i := \{j : o_{ij} = 0\}$. Then, for $j \in G_i$, the observation-dependent Shapley value (OD SHAP) for each data point is:*

$$\phi_{ij} = \sum_{S \subseteq G_i \setminus \{j\}} \frac{|S|!(|G_i| - |S| - 1)!}{|G_i|!} [f_i(S \cup \{j\}) - f_i(S)] \tag{3}$$

We use the notation $f_i(S) = h(x_i^*(S))$, where $x_i^*(S) := x_i \odot \mathbb{1}_{S^c} + \mathbb{1}_S$ and $\mathbb{1}_S$ is a d_X-dimensional vector whose element is one if its index is S, zero otherwise. G_i represents the set of indices for all features which have not been observed for patient i, and π_{G_i} is the set of all permutations of those indices. Intuitively, f_i is a wrapper function for h that counterfactually sets all features in S to 1, while preserving the value of all the other features not in S. Finally, we use the absolute value of ϕ_{ij} in our calculation of V_{ij} since we aim to measure the magnitude of feature importance rather than the direction.

Comparison to standard SHAP The OD SHAP value ϕ_{ij} corresponds to the counterfactual model

prediction when $x_{ij} = 1$. In particular, OD SHAP only considers permutations in features that have not been observed, whereas the standard implementation of SHAP (Lundberg and Lee (2017)) value considers permutations in all features in x_i. Under OD SHAP, $\phi_{ij} = 0$ for all j where $o_{ij} = 1$. Additionally, the intercept, $\phi_{i\emptyset}$, is equal to the model's current prediction on x_i under OD SHAP, whereas the intercept for standard SHAP is equal to the model prediction when all features are equal to 0 (when all features are binary). For example, we assume a patient has a known diagnosis of hypertension, but an unknown value for diabetes. A standard SHAP calculation would take into consideration both the possibility of positive and negative hypertension as part of the calculation for the Shapley value for diabetes. In the observation-dependent version, we would only include the positive value since we already know the patient has a positive diagnosis.

4.2. Algorithm

Our algorithm estimates how each data point will change the model's prediction, and then prioritizes collecting data points that will yield the largest change.

Given b budget and C costs:

1. Initialize O as the indicator matrix for data that are not potentially missing.
2. Compute $V_{ij} = |\phi_{ij}|$ according to Equation 3.
3. Run 0-1 Knapsack algorithm with budget b, costs C, and weights V to obtain subset of data indices $I = \{(i, j) \mid i \in [N], j \in [d_X]\}$
4. Sort indices in I according to their value in V, from greatest to least:
   $I^{ordered} = \text{sort}(I, [V_{ij} \forall (i, j) \in I], desc)$

The model user then takes list I and performs data collection in order. In practice, ϕ_{ij} may assign zero values for features that do not change model prediction. As such, our algorithm only recommends data collection for the subset of data points which have non-zero values.

5. Evaluation

Metric Let $A = \{O^{(0)}, ..., O^{(T)}\}$ be a sequence of data observations made, where $O^{(t)}$ is an indicator matrix for the set of data collected after time t. We define reward $r(O^{(t)})$ as the model's performance after t data are collected:

$$r(O^{(t)}) = \frac{1}{N} \sum \mathcal{L}(y_i, h(x_i + \mathbf{e}_j o_{ij}^{(t)} z_{ij}))$$

We evaluate an entire sequence of choices by taking the average model improvement over all time steps:

$$\text{Avg. Model Improvement} := \frac{1}{T} \sum_{t=0}^{T} (r(O^{(t)}) - r(O^{(0)})) \tag{4}$$

Baselines We compare our method against three baselines:

1. **Random** We collect potentially missing data points at random, which represents a potential default policy.
$$V_{ij}^{\text{random}} = \begin{cases} Unif(0, 1), & \text{if } o_{ij} = 0 \\ 0, & \text{otherwise} \end{cases}$$
2. **Feature Gradient** We use a naive feature importance approach that takes the model performance difference at each step as an indicator of which feature to choose. This approach is naive in that it does not take into account possible feature interactions that occur during data collection.
$$V_{ij}^{\text{grad}} = \begin{cases} |h(x_i + \mathbf{e}_j) - h(x_i)|, & \text{if } o_{ij} = 0 \\ 0, & \text{otherwise} \end{cases}$$
3. **Standard SHAP** We also evaluate the SHAP method for feature attribution as originally proposed by Lundberg and Lee (2017). The formula for ϕ_{ij} in the standard case is the same as Equation 9 but uses the notation $G_i = \{1, ..., d_X\}$ and $x_i^*(S) = \mathbb{1}_S$.

Metrics We use negative log loss in our estimation of V_{ij}, while using AUC (Area Under the ROC Curve) to measure overall performance with respect to a set data collection.

5.1. Data

We use both simulated and real-world data to evaluate our method.

Simulated data We aim to simulate a classification algorithm with EHR-derived features. First, we generate true data Z as independent binary random variables: $Z_j \sim Bernoulli(p_j)$, $p_j \sim Unif(0.5, 1)$. The range $[0.5, 1]$ reflects a high degree of missingness where data collection would be necessary. Next, a masking variable M is generated as binary random variables as well, with $M_j \sim Bernoulli(q_j)$, where $q_j \sim Unif(0.5, 1)$. The initial observed dataset X is the elementwise dot product of Z and M, ie. $X = Z \odot (1 - M)$. The labels are generated as Bernoulli random variables parameterized by the linear combination of Z and $\boldsymbol{\beta} = \{\beta_1, ..., \beta_{d_X}\}, \beta_j \sim Exp(10)$, such that $y \sim Bernoulli(\sigma(Z^T\boldsymbol{\beta}))$, where σ is the logistic function. We generate 1000 total synthetic patients with 20 features and leave 200 rows to be used for the evaluation set. In our experiments, we use a Gradient Boosting classifier, implemented in Python's Scikit-Learn package (Pedregosa et al. (2011)).

MIMIC-IV ICU data To evaluate our method on EHR, we use the MIMIC-IV dataset (Johnson et al. (2023)). Our dataset preprocessing uses benchmarking from Xie et al. (2022), which provides an open-source implementation for data extraction, cleaning, and filtering on several ICU-related outcomes in MIMIC-IV. In particular, we use the following outcomes for prediction tasks: *critical outcomes*, *ED revisits*, *ICU transfers*. ICU transfers refer to critically ill patients who are moved to the ICU within 12 hours. Critical outcomes are a broader category of ICU transfers that also include patient mortalities within 12 hours. ED revisits refer to patients who return to the emergency department within 72 hours after their previous discharge. There are 39 total features used in each prediction task, which fall into the categories of previous hospital stay durations, triage values, chief complaints, and comorbidities. To better evaluate the effects of binary missing variables, we binarize continuous features by encoding values either above or below their mean values. In total, 353150 patients are used in the training set.

Due to the nature of latent missingness, we cannot access naturally occurring missingness labels in MIMIC-IV. As a proxy, we choose a subset of patients with greater rates of positive features and apply a missingness mask over these data. In effect, we aim to capture the effect of masked features while preserving the natural correlation between features and outcomes. For our evaluation set, we sample 500 patients that have a high rate of positive features and apply the same masking procedure as described in our simulated data.

5.2. Experiments

We compare OD SHAP to the three baseline methods in three experiments: (1) cost curves, (2) average model improvement, and (3) missingness rates.

Cost curves The methods are evaluated first by producing a curve where the x-axis is the fraction of potentially missing data points collected and the y-axis is the model improvement (in AUC). Each curve represents the model's performance after observing the true values of the data points selected by each method at each time step t.

Average model improvement We additionally summarize each curve by computing the average model improvement according to Equation 4, which is the average area under each cost curve relative to the model's performance without any data collection. We run each experiment 100 times such that missingness rates are randomized each time. Confidence intervals are produced using bootstrap sampling over the missingness of all the runs.

Missingness We examine the performance of each method under increasing rates of missingness in each feature. Rather than generate missingness across $Unif(0.5, 1)$, we select a fixed missingness rate α in the range $[0.5, 1.0]$ (in intervals of 0.1), and evaluate each method's average model improvement when all features are missing at rate α.

6. Results

	Prediction Task		
Method	Critical Outcome	ED Revisit	ICU Transfer
Feature Gradient	48.1%	55.1%	59.5%
Standard SHAP	50.3%	59.0%	56.2%
OD SHAP	**73.7%**	**90.9%**	**79.5%**

Table 1: Percent improvement over random for each method, which is derived from Figure 4 by dividing values for each method by random for each task.

We present the following four results:

1. Cost curves (Figure 2). We show four cost curves that represent each method's efficiency at discovering valuable missing features. The

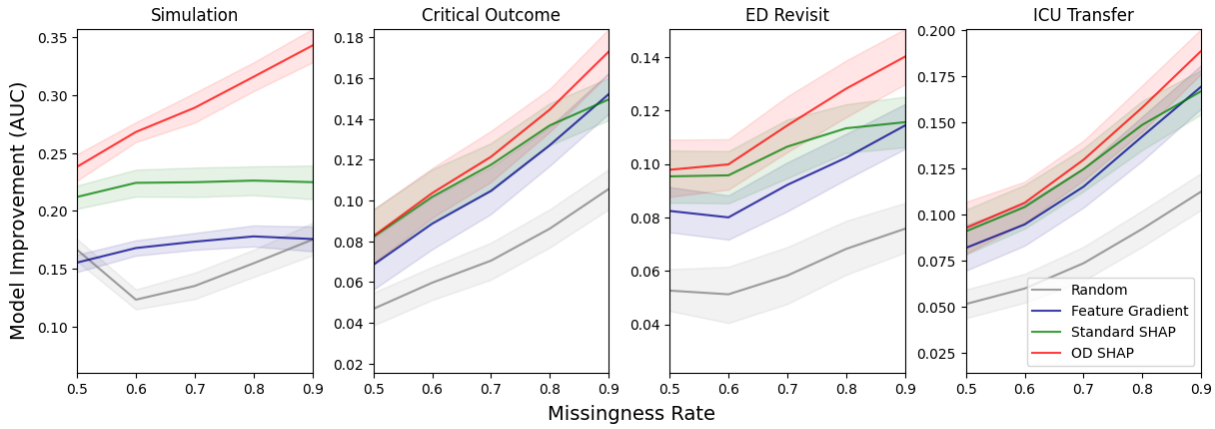

Figure 3: The model improvement from each method across different rates of missingness in each feature. The x-axis is the missingness rate, while the y-axis shows the model improvement in AUC. Each line's solid portion represents the average performance across multiple runs, while the shaded region represents the 95% confidence interval at each time step.

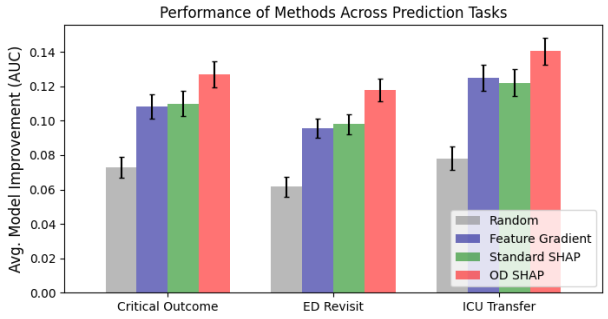

Figure 4: The average improvement for each method (feature gradient, standard SHAP, and OD SHAP) across three prediction tasks (critical outcome, ED revisit, ICU transfer). The methods are compared against a random baseline (in gray). The 95% confidence intervals for each method are provided in the black lines.

top left presents the results from the simulation, while the other three present the results from the MIMIC-IV dataset. The curves show that OD SHAP performs stronger than the three baselines, achieving the maximum performance achieved by collecting all data points at a fraction of the cost. Notably, we find that the standard SHAP method for feature attribution can actually underperform compared to the naive feature gradient, whereas the observation-dependent adjustment made in OD SHAP improves upon both.

2. Performance of Methods Across Prediction Tasks (Figure 4). Each curve is summarized through the average model improvement. Our results show that OD SHAP yields a statistically significant improvement upon the three baselines.

3. Percentage over random (Table 1). We present an alternative interpretation of our results in the form of each method's improvement over the random baseline. We find that OD SHAP can improve upon a random strategy by 73.7%, 90.0%, and 79.5% for critical outcome, ED revisit, and ICU transfer prediction, respectively.

4. Missingness (Figure 3). We find that the relative improvements from using OD SHAP outperform the three baselines as missingness rates increase. We also find that the relative performance of standard SHAP compared to OD SHAP drops as missingness rates increase, indicating that a standard SHAP calculation would not account for potential changes in other covariates at earlier time steps. Importantly, by fixing a single missingness rate, we isolate each method's ability to estimate feature influence alone, rather than simply find features that are missing at high rates.

7. Discussion

Missing data poses a significant hurdle to the deployment of medical AI algorithms. As an increasing number of EHR-derived algorithms are deployed into healthcare settings, it is crucial that healthcare providers can address the negative effects of missingness on model performance. In this paper, we propose a targeted data collection approach to addressing missingness from under-reporting. Our results show that the feature gradient and standard Shapley value do not adequately account for the covariate shift that occurs during data collection. We propose a modification to the Shapley value feature importance, OD SHAP, that addresses this issue.

Implications for model fairness Budget-driven data collection can be specifically directed towards addressing model biases with regard to protected subgroups such as race and socioeconomic status, which can have significantly lower healthcare utilization compared to majority groups (Alegria et al. (2011); Dickman et al. (2022)). Our method allows for healthcare providers to target data collection efforts to specific individuals and features, whereas complementary efforts such as community outreach and public health initiatives may be less specific. Future work in this field can include approaches for optimizing tradeoffs between multiple subgroups as well.

Regulatory context As EHR-derived algorithms are increasingly subject to government regulation (Food et al. (2019)), deployed models face restrictions with regard to model retraining and adjustments (Gilbert et al. (2021)). This regulatory environment underscores the need for a data-centric approach to improving model performance since models may generally be considered fixed after deployment. Furthermore, algorithms are typically deployed to a large number of sites, challenging notions of a one-size-fits-all approach to algorithmic evaluation (Futoma et al. (2020)). Our method can be flexibly used on a chosen subgroup and only requires inference access to the model. As with any interventional approach, data collection needs to be done so in accordance to ethical and legal considerations so as to prevent further harm.

Deployment considerations Our proposed approach is an extension of existing efforts by healthcare spenders to encourage positive health outcomes through subsidies. Currently, insurance plans commonly offer reduced copayments for preventive care services such as annual physical exams or screenings, which has been shown to have a sizable impact on increasing healthcare utilization in patient subgroups (Gruber (2006)). A specific example is UnitedHealthcare's HouseCalls, where the insurance company provides a free yearly check-in directly to patients' homes. More broadly speaking, healthcare subsidies have been used as a tool to encourage a variety of desired health outcomes for under-represented groups such as farm workers, low-income women, and African-American and Latino males at risk of STIs (Gorter et al. (2003)). Our proposal frames algorithmic performance as the intermediate outcome of interest. In general, such an approach may face practical obstacles in deployment, as active data collection and cleaning pipelines may not be readily available in every organization (Zhang et al. (2022)). However, with regulatory standards for EHR-derived algorithms increasing over time (Ross (2022)), such pipelines will likely be justifiable investments for any organization that deploys such models.

Additionally, in practice, EHR-based algorithms may also include time-sensitive variables such as a daily heart rate reading, diet, or location. In such cases, rather than a one-off data collection scheme, healthcare payers may want to allocate a budget towards ongoing data monitoring. For example, rather than subsidizing a one-time survey of a patient's eating habits, the budget can be allocated towards continuous monitoring. In such cases, an allocated budget can be partitioned given period of time (ie. budget per month or year).

Dependence on feature sensitivity The extent to which feature sensitivity alone can be a reliable indicator of a data point's value depends on (1) the rates of missingness and (2) the model's error with respect to the true (revealed) data distribution. By focusing only on feature sensitivity, we are defaulting to a value of $\Delta_{ij} = 1$, which is the assumption that all possibly missing data are indeed missing. In cases where some features have very low rates of missingness, the gap between feature sensitivity alone and the true value of V_{ij} will be more drastic. A possible remedy can include informed data imputation, where domain experts can measure population-level disease rates and infer whether the current covariate rates are higher or lower than expected. Missingness rates are heavily domain-dependent, but this information can be easily integrated into our approach by multiplying each data point's value by its expected probability as

outlined in the decomposition of the Taylor approximation. Next, our method also assumes equal loss sensitivity for each data point. In situations where the model exhibits much higher error rates for certain data points than others (eg. a bias towards specific patient subgroups), the information gained from loss sensitivity is higher. While not covered in this paper, a potential way to address this may involve using the model's uncertainty as a proxy for loss sensitivity.

A consequence of operating under limited knowledge of groundtruth labels is the fact that not all data collection will result in improved model performance. In expectation, data acquisition improves model performance when the model has lower error under the true distribution vs. the corrupted distribution (ie. when $\mathbb{E}[\mathcal{L}(y_i, h(Z_i))] < \mathbb{E}[\mathcal{L}(y_i, h(X_i))]$. However, when the model exhibits error w.r.t. the true distribution, collecting certain data points may result in a (locally) worse performance. Empirically, we see this effect more apparent in our baseline methods (standard SHAP and feature gradient).

Limitations Due to the nature of latent missingness, the natural distribution of missing data in MIMIC-IV is unknown to us. As such, we simulate missingness in a subgroup of patients with high healthcare utilization. However, additional empirical evaluations of our methods may include real-world evidence where data collection is actually carried out and reported. Furthermore, in practice, data collection costs may not be strictly linear as paying for one type of data collection could change the costs of additional data collection.

While the Shapley value is a commonly used interpretability method, other methods such as LIME also yield the additivity property required for our method (Ribeiro et al. (2016)). In general, there may be limitations in the extent to which the axiomatic properties of Shapley values hold in practice (Fryer et al. (2021)), although our paper only requires that the additivity property holds.

Our paper underscores the viability of a data-centric approach toward improving model performance when deployed across sites with latent missing features. Overall, we find that targeted data collection can be a useful tool to improve the performance of a deployed model. In particular, the cost curves in our results suggest that the efficiency of data collection is convex, such that a small portion of the total budget is enough to achieve the majority of the performance gains. As algorithms are deployed in increasingly complex healthcare settings, providers will need the ability to tradeoff between model performance and costs in order to adhere to regulatory requirements and internal standards.

Data and Code Availability This paper uses the MIMIC-IV dataset (Johnson et al., 2023), which is available on the PhysioNet repository. The code used in this paper will be made available on GitHub.

Institutional Review Board (IRB) Because data was de-identified or a Limited Data Set in compliance with the Health Insurance Portability and Accountability Act and customer requirements, Institutional Review Board approval or waiver of authorization was not required.

References

Margarita Alegria, Nicholas J Carson, Marta Goncalves, and Kristen Keefe. Disparities in treatment for substance use disorders and co-occurring disorders for ethnic/racial minority youth. *Journal of the american academy of Child & adolescent Psychiatry*, 50(1):22–31, 2011.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Brett K Beaulieu-Jones, Jason H Moore, and Pooled Resource Open-Access ALS Clinical Trials Consortium. Missing data imputation in the electronic health record using deeply learned autoencoders. In *Pacific symposium on biocomputing 2017*, pages 207–218. World Scientific, 2017.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109:719–760, 2020.

Djallel Bouneffouf. Exponentiated gradient exploration for active learning. *Computers*, 5(1):1, 2016.

Alison Callahan and Nigam H Shah. Machine learning in healthcare. In *Key Advances in Clinical Informatics*, pages 279–291. Elsevier, 2017.

Peter G Campbell, Jennifer Malone, Sanjay Yadla, Rohan Chitale, Rani Nasser, Mitchell G Maltenfort, Alex Vaccaro, and John K Ratliff. Comparison of icd-9–based, retrospective, and prospective assessments of perioperative complications:

assessment of accuracy in reporting. *Journal of Neurosurgery: Spine*, 14(1):16–22, 2011.

Trenton Chang, Michael W Sjoding, and Jenna Wiens. Disparate censorship & undertesting: A source of label bias in clinical machine learning. August 2022.

Yukun Chen, Robert J Carroll, Eugenia R McPeek Hinz, Anushi Shah, Anne E Eyler, Joshua C Denny, and Hua Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20(e2):e253–e259, 2013.

Sofia De Achaval, Chris Feudtner, Shana Palla, and Maria E Suarez-Almazor. Validation of icd-9-cm codes for identification of acetaminophen-related emergency department visits in a large pediatric hospital. *BMC health services research*, 13:1–8, 2013.

Samuel L Dickman, Adam Gaffney, Alecia McGregor, David U Himmelstein, Danny McCormick, David H Bor, and Steffie Woolhandler. Trends in health care use among black and white persons in the us, 1963-2019. *JAMA network open*, 5(6): e2217383–e2217383, 2022.

Javier Escudero, Emmanuel Ifeachor, John P Zajicek, Colin Green, James Shearer, and Stephen Pearson. Machine learning-based method for personalized and cost-effective detection of alzheimer's disease. *IEEE transactions on biomedical engineering*, 60(1):164–168, 2012.

Kadija Ferryman and Mikaela Pitcan. Fairness in precision medicine. 2018.

Georg Fette, Markus Krug, Mathias Kaspar, Leon Liman, Georg Dietrich, Maximilian Ertl, Jonathan Krebs, Stefan Störk, and Frank Puppe. Estimating a bias in icd encodings for billing purposes. In *MIE*, pages 141–145, 2018.

Food, Drug Administration, et al. Proposed regulatory framework for modifications to artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd). 2019.

Daniel Fryer, Inga Strümke, and Hien Nguyen. Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, 9:144352–144360, 2021.

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020.

Amirata Ghorbani, James Zou, and Andre Esteva. Data shapley valuation for efficient batch active learning. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pages 1456–1462. IEEE, 2022.

Stephen Gilbert, Matthew Fenech, Martin Hirsch, Shubhanan Upadhyay, Andrea Biasiucci, and Johannes Starlinger. Algorithm change protocols in the regulation of adaptive machine learning–based medical devices. *Journal of Medical Internet Research*, 23(10):e30545, 2021.

Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.*, 24(1):198–208, January 2017.

Steven N Goodman, Sharad Goel, and Mark R Cullen. Machine learning, health disparities, and causal reasoning. *Annals of internal medicine*, 169 (12):883–884, 2018.

Anna Gorter, Peter Sandiford, Zillyham Rojas, and Micol Salvetto. Competitive voucher schemes for health background paper. *Instituto Centro Americano de la Salud (ICAS)*, 2003.

Jonathan Gruber. *The role of consumer copayments for health care: lessons from the RAND health insurance experiment and beyond*, volume 7566. Citeseer, 2006.

Sebastien Haneuse and Michael Daniels. A general framework for considering selection bias in ehr-based studies: what data are observed and why? *EGEMs*, 4(1), 2016.

Katharine E Henry, David N Hager, Tiffany M Osborn, Albert W Wu, and Suchi Saria. Comparison of automated sepsis identification methods and electronic health record–based sepsis phenotyping: improving case identification accuracy by accounting for confounding comorbid conditions. *Critical care explorations*, 1(10), 2019.

Zhiyong Hu and Dongping Du. A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction. *PLoS One*, 15(9):e0237724, 2020.

Zongcheng Ji, Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. Cost-sensitive active learning for phenotyping of electronic health records. *AMIA Summits on Translational Science Proceedings*, 2019:829, 2019.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-wei H Lehman, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10 (1):1, 2023.

Pallika Kanani and Prem Melville. Prediction-time active feature-value acquisition for cost-effective customer targeting. *Advances in neural information processing systems (NIPS)*, 2008.

Malgorzata Lazkecka, Jan Mielniczuk, and Pawel Teisseyre. Estimating the class prior for positive and unlabelled data via logistic regression. *Advances in Data Analysis and Classification*, 15(4): 1039–1068, 2021.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Jeanne M Madden, Matthew D Lakoma, Donna Rusinak, Christine Y Lu, and Stephen B Soumerai. Missing clinical and behavioral health data in a large electronic health record (ehr) system. *Journal of the American Medical Informatics Association*, 23(6):1143–1149, 2016.

Brandon Malone, Alberto Garcia-Duran, and Mathias Niepert. Learning representations of missing data for predicting patient outcomes. *arXiv preprint arXiv:1811.04752*, 2018.

Sriraam Natarajan, Srijita Das, Nandini Ramanan, Gautam Kunapuli, and Predrag Radivojac. On whom should i perform this lab test next? an active feature elicitation approach. In *IJCAI*, pages 3498–3505, 2018.

Nir Nissim, Mary Regina Boland, Nicholas P Tatonetti, Yuval Elovici, George Hripcsak, Yuval Shahar, and Robert Moskovitch. Improving condition severity classification with an efficient active learning based framework. *Journal of biomedical informatics*, 61:44–54, 2016.

Linda Nordling. A fairer way forward for ai in health care. *Nature*, 573(7775):S103–S103, 2019.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Filippo Pesapane, Caterina Volonté, Marina Codari, and Francesco Sardanelli. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in europe and the united states. *Insights into imaging*, 9:745–753, 2018.

Emma Pierson, David M Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, 2021.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

Chanu Rhee and Michael Klompas. Sepsis trends: increasing incidence and decreasing mortality, or changing denominator? *Journal of Thoracic Disease*, 12(Suppl 1):S89, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

Patrick S Romano and David H Mark. Bias in the coding of hospital discharge data and its implications for quality assessment. *Medical care*, 32(1): 81–90, 1994.

Casey Ross. In new guidance, FDA says AI tools to warn of sepsis should be regulated as devices — statnews.com. https://www.statnews.com/2022/09/27/health-fda-artificial-intelligence-guidance-sepsis/, 2022. [Accessed 21-Mar-2023].

Konstantinos Sechidis, Matthew Sperrin, Emily S Petherick, Mikel Luján, and Gavin Brown. Dealing with under-reported variables: An information theoretic solution. *International Journal of Approximate Reasoning*, 85:159–177, 2017.

Burr Settles. Active learning literature survey. 2009.

Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008.

Christopher W Seymour, Vincent X Liu, Theodore J Iwashyna, Frank M Brunkhorst, Thomas D Rea, André Scherag, Gordon Rubenfeld, Jeremy M Kahn, Manu Shankar-Hari, Mervyn Singer, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):762–774, 2016.

AVLN Sujith, Guna Sekhar Sajja, V Mahalakshmi, Shibili Nuhmani, and B Prasanalakshmi. Systematic review of smart health monitoring using deep learning and artificial intelligence. *Neuroscience Informatics*, 2(3):100028, 2022.

Amelia LM Tan, Emily J Getzen, Meghan R Hutch, Zachary H Strasser, Alba Gutiérrez-Sacristán, Trang T Le, Arianna Dagliati, Michele Morris, David A Hanauer, Bertrand Moal, et al. Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *Journal of Biomedical Informatics*, page 104306, 2023.

Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

Adam Wawrzeńczyk and Jan Mielniczuk. Revisiting strategies for fitting logistic regression for positive and unlabeled data. *International Journal of Applied Mathematics and Computer Science*, 32(2): 299–309, 2022.

Brian J Wells, Kevin M Chagin, Amy S Nowacki, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3), 2013.

Stacey-Ann Whittaker, Mark E Mikkelsen, David F Gaieski, Sherine Koshy, Craig Kean, and Barry D Fuchs. Severe sepsis cohorts derived from claims-based strategies appear to be biased towards a more severely ill patient population. *Critical care medicine*, 41(4), 2013.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E Ho, and James Zou. How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals. *Nature Medicine*, 27(4):582–584, 2021.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9 (1):658, 2022.

Yiming Xu, Xiaohong Liu, Liyan Pan, Xiaojian Mao, Huiying Liang, Guangyu Wang, and Ting Chen. Explainable dynamic multimodal variational autoencoder for the prediction of patients with suspected central precocious puberty. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1362–1373, 2021.

Xinge You, Ruxin Wang, and Dacheng Tao. Diverse expected gradient active learning for relative attributes. *IEEE transactions on image processing*, 23(7):3203–3217, 2014.

Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, pages 1–16, 2022.

Helen Zhou, Sivaraman Balakrishnan, and Zachary C Lipton. Domain adaptation under missingness shift. *arXiv preprint arXiv:2211.02093*, 2022.

Appendix A. Additional Experimentation

Varying costs We additionally explore how varying costs of the features affect our data collection strategy. We synthetically generate feature costs both a uniform distribution ($C_j \sim Unif(0,1)$) and exponential distribution ($C_j \sim Exp(1)$). In total, we sample costs across 500 random seeds and report the average cost curves across the fraction of the total budget spent. In our experiments, we use a fractional knapsack algorithm for computational feasibility.

The results from these cost structures are shown in Figure 5. In general, we find that our method (OD SHAP) is robust to different cost distributions and that adding in a variable cost structure improves the relative performance of OD SHAP compared to a setting with equal costs. We also find that for exponentially distributed costs, each method requires a higher fraction of the total budget to achieve maximum model performance. As the costs are more heavy-tailed, more valuable data points can be downweighted by their costs, such that other less valuable but cheaper data points may be chosen first.

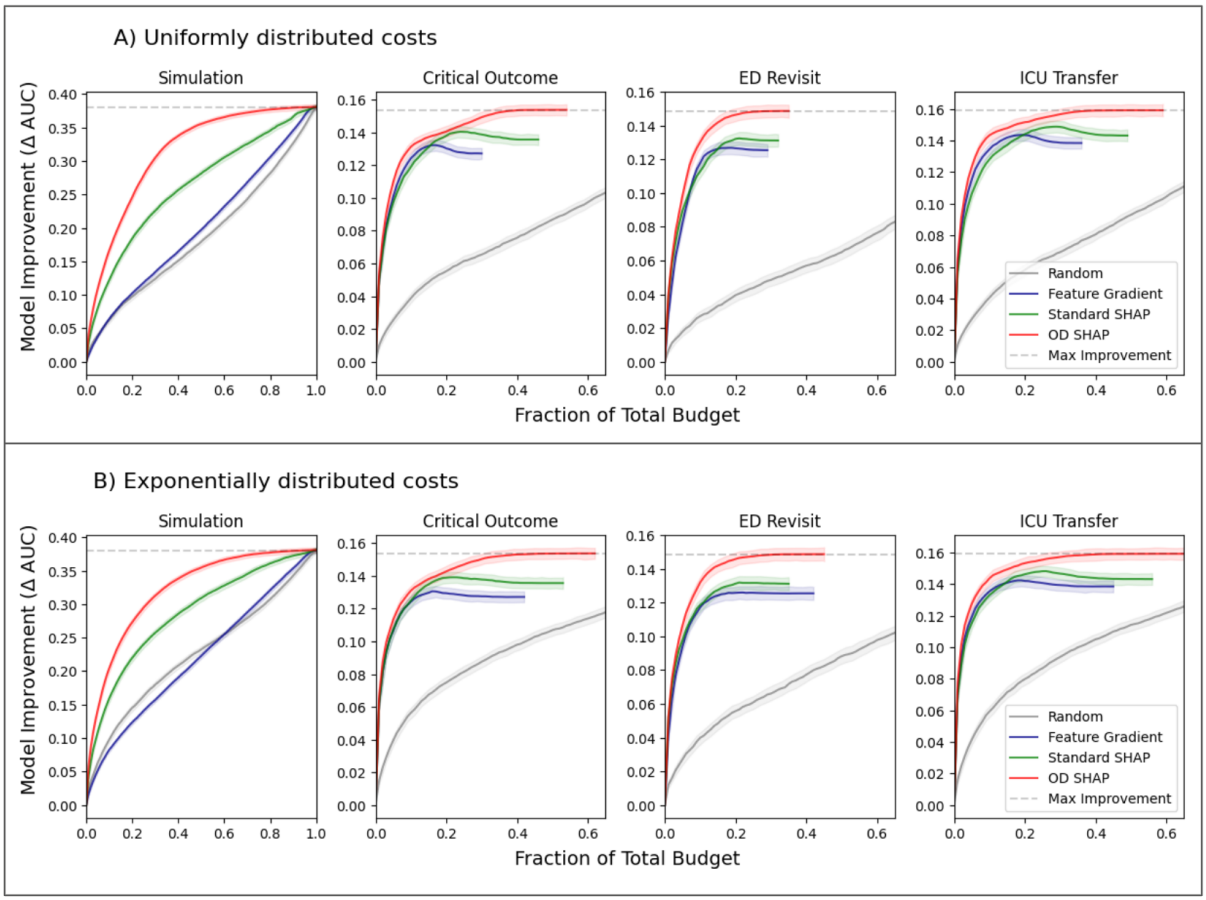

Figure 5: Cost curves displaying the efficiency of each method in improving model performance across data points collected. The x-axis represents the fraction of the total budget required to collect all potentially missing data points, while the y-axis represents the improvement in model performance after those data points have been collected. Improvement is measured by evaluating the model's performance (AUC) after N data points have been collected. Each curve's solid line represents the average performance across 500 runs, while the shaded region represents the 95% confidence interval at each time step. The dotted horizontal line refers to the maximum model performance after collecting all potentially missing data points. In subplot (A), we generate feature costs according to a uniform distribution, whereas in subplot (B), we generate them according to an exponential distribution.