Unsupervised Coverage Sampling to Enhance Clinical Chart Review Coverage for Computable Phenotype Development

Zigui Wang, Jillian H. Hurst, Chuan Hong, Benjamin Alan Goldstein

Submitted to: JMIR Medical Informatics
on: February 03, 2025

Table of Contents

Unsupervised Coverage Sampling to Enhance Clinical Chart Review Coverage for Computable Phenotype Development

Zigui Wang[1] BS; Jillian H. Hurst[2] PhD; Chuan Hong[1] PhD; Benjamin Alan Goldstein[1, 2] PhD

[1]Department of Biostatistics and Bioinformatics Duke University School of Medicine Duke University Durham US
[2]Department of Pediatrics Duke University School of Medicine Duke University Durham US

Corresponding Author:
Benjamin Alan Goldstein PhD
Department of Biostatistics and Bioinformatics
Duke University School of Medicine
Duke University
2424 Erwin Road Ste 902
9023 Hock Plaza
Durham
US

Abstract

Background: Developing computable phenotypes (CP) based on electronic health records (EHR) data requires "gold-standard" labels of patient charts obtained from clinicians. Charts are most often sampled randomly, but random sampling may fail to capture the diversity of a given patient population, which may lead to bias of the CP.

Objective: We proposed an unsupervised sampling approach designed to better capture a diverse patient cohort and improve the information coverage of chart review samples.

Methods: Our coverage sampling method utilizes clustering and stratified sampling to ensure diverse representation in chart review samples. We use simulations and a real-world data example to compare the performance of our method with random sampling. The performance of the samples was evaluated based on the information coverage and area under the receiver operator characteristic curve (AUROC).

Results: Our simulation studies demonstrate that our unsupervised approach provided better coverage of patient populations and equal or improved CP performance compared to random samples, especially in scenarios where minority sub-groups were present. In the real-world application, the method also outperformed random sampling, yielding more representative samples and enhancing CP performance.

Conclusions: The proposed coverage sampling method enhances the coverage of chart review samples, leading to the development of CPs that can capture outcomes of interest in a diverse patient population. This approach is particularly beneficial in cohorts with complex or minority sub-groups, providing a robust alternative to random sampling in EHR-based research.

(JMIR Preprints 03/02/2025:72068)
DOI: https://doi.org/10.2196/preprints.72068

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Unsupervised Coverage Sampling to Enhance Clinical Chart Review Coverage for Computable Phenotype Development

Zigui Wang, B.S.[1], Jillian H. Hurst, Ph.D.[2], Chuan Hong, Ph.D.[1], Benjamin A. Goldstein, Ph.D.[1,2]

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

[2]Department of Pediatrics, Duke University School of Medicine, Durham, NC, USA

Abstract

Background: Developing computable phenotypes (CP) based on electronic health records (EHR) data requires "gold-standard" labels of patient charts obtained from clinicians. Charts are most often sampled randomly, but random sampling may fail to capture the diversity of a given patient population, which may lead to bias of the CP.
Objective: We proposed an unsupervised sampling approach designed to better capture a diverse patient cohort and improve the information coverage of chart review samples.
Methods: Our coverage sampling method utilizes clustering and stratified sampling to ensure diverse representation in chart review samples. We use simulations and a real-world data example to compare the performance of our method with random sampling. The performance of the samples was evaluated based on the information coverage and area under the receiver operator characteristic curve (AUROC).
Results: Our simulation studies demonstrate that our unsupervised approach provided better coverage of patient populations and equal or improved CP performance compared to random samples, especially in scenarios where minority sub-groups were present. In the real-world application, the method also outperformed random sampling, yielding more representative samples and enhancing CP performance.

Conclusions: The proposed coverage sampling method enhances the coverage of chart review samples, leading to the development of CPs that can capture outcomes of interest in a diverse patient population. This approach is particularly beneficial in cohorts with complex or minority sub-groups, providing a robust alternative to random sampling in EHR-based research.

1. Introduction

Electronic health records (EHR) data are widely used in clinical research. While they contain dense, often granular information on a patient's health status, they also pose challenges for clinical studies since they lack explicit documentation for either a patient's clinical condition (e.g., diabetes) or reason for the healthcare encounter (e.g., admission due to infection). In principle, the problem list, which provides a historical listing of previous health problems, can be used to identify chronic conditions, though it is often unreliable (1,2). Similarly, fields such as discharge diagnosis may not accurately represent the reason a patient had a visit. Instead, information from diagnosis codes,

laboratory test results and prescriptions or administered medications are used to indicate the presence of a specific clinical condition (3–5). This is a well-known challenge in working with EHR data and has led to the growth of computable phenotypes (CPs). CPs are algorithms, typically Boolean, though sometimes probabilistic, that utilize multiple sources of clinical data—such as diagnoses, laboratory results, and medication records—to infer the clinical condition of a patient or the reason for a visit (6–8).

Creating CPs is a multiphase process that often requires significant collaborative effort from clinicians and informaticians (5,9). One of the key components in CP development is the creation of a set of "gold standard" outcome labels. The outcome labels are typically generated based on manual review of a subset eligible patient charts, which can require significant time (6–8,10). The set of charts that are used to develop these gold-standard labels are usually sampled randomly (11). While random sampling will, on average, produce a representative view of the cohort of interest, since one usually wants to review only a small number of charts, random sampling may not adequately represent the complete range of disease presentations or patient demographics. In the scenario shown in **Figure 1**, sub-groups that have a smaller representation within the larger data set (e.g., rarer presentations of the disease of interest, disease presentation in minority sub-groups), are less likely to be adequately represented based on random sampling. In this case, much larger sample sets are necessary to find a meaningful number of charts from people from these sub-groups (9). In such scenarios, random sampling strategies might not be effective in generating a representative sample for chart review purposes and result in a CP that does not accurately capture the heterogeneity of the condition of interest. This can lead to a CP that performs worse for those sub-populations.

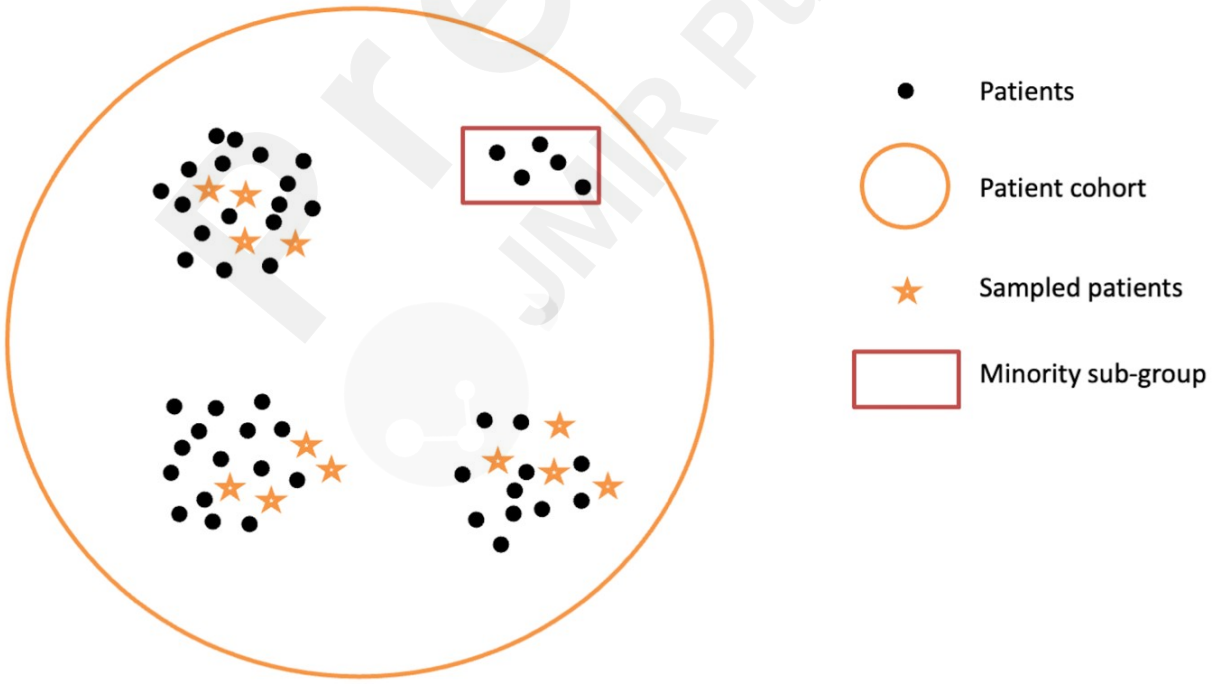

Figure 1:Example of the impact of random sampling on representation of patient sub-groups. The black dots represent unsampled patients; stars represent the sampled patients. The red box represents a sub-group that was missed by random sampling.

In recent years, various methods have been developed to enhance the representativeness of labeled data. Active learning is one such approach that iteratively selects the most informative samples for labeling, aiming to optimize model performance with minimal amount of data (12,13). However, active learning typically requires an initial set of labeled data to train the model and guide the selection process (14). Moreover, active learning methods are typically focused on identifying the samples that will be provide the most leverage on the final model, as opposed to the ones that would best capture the diversity of the patient cohort (15,16).

In this paper, we propose a process for selecting medical charts for review to generate gold standard labelling when constructing CPs. The goal of this method is to ensure that our selection captures the diversity of the full patient cohort. To achieve this, we propose a clustering-based process to generate potential sample pools. We then introduce a novel metric to identify the most representative sample pool that should be used for label generation. By enhancing the information coverage of the training sample, our approach is expected to yield a better performing CP for both sub-groups and the full patient cohort. To illustrate this approach, we employ simulation methods coupled with a real-world data example to demonstrate how this novel sampling approach can match or even surpass the performance of random sampling.

2. Materials and Methods

2.1. Sampling Approach

The overall methodological approach is illustrated in **Figure 2**. We start by considering a study cohort for whom we want to create labels for the presence or absence of a condition of interest (e.g., diabetes, cause-specific admission). We presume that our study cohort is large enough that we do not want to review and label all patient charts. Instead, we want to generate a *sample pool,* from which we will develop or "learn" a CP. In this paper, our analytic task is to determine how to best identify that sample pool. We propose that the best sample pool is one that maxmizes *coverage* of the cohort, providing information about all of the sub-groups that compose the cohort (**Figure 2A)**. In other words, the sample pool should be equally representative of each sub-group, rather than merely reflecting the source population distribution. To assess coverage, we define a novel metric, described below. A variety of methods can be used to generate the sample pool. In this study, we propose using stratified sampling framework, in other words, clustering the data and then sampling from these clusters (**Figure 2B**). By identifying and then sampling from clusters, we hypothesize that we will be able to represent different patient sub-groups, making the chart review sample more reflective of the entire patient cohort.

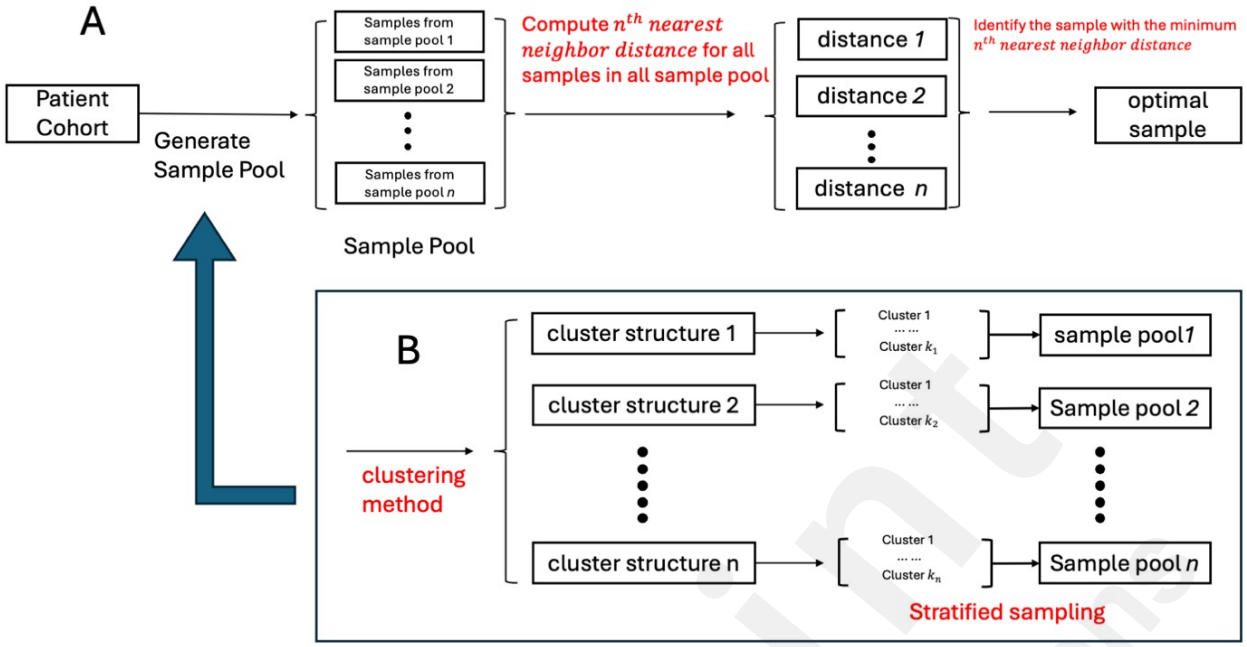

Figure 2: Diagram illustrating our sampling approach. A: procedures for general coverage sampling; B: procedures for generating the sample pool

2.2. Sample pool generation: After defining a cohort of interest, we start by clustering the individual patient records. As illustrated in our real data example, our groupings are driven by clinical factors, so we only use clinical features (i.e., not demographic factors) to conduct the clustering. We then sample records randomly from each of the clusters. For example, if we pre-specify that we want to review 100 charts, and we generate 4 clusters, we would sample 25 records from each cluster. While a variety of clustering algorithms can be used, we suggest hierarchical clustering. The nested cluster structure provided by hierarchical clustering is reflective of our proposed interpretation that the cohort consists of patient sub-groups. For comparison, we also present results from K-means clustering. Notably, with a sufficiently large number of replications, we expect different clustering methods to generate similar sample pools, resulting in comparable representative samples.

2.3. Coverage assessment: A primary step is assessing data coverage. To do so, we propose a novel metric that measures the coverage of the sample for the full data cohort. We define the n^{th} *nearest neighbor distance*, as:

$$n^{th}\,nearest\,neighbor\,distance = \sum_{i=1}^{N} d_i^n$$

For each person, i, in the study cohort of size N, we calculate the Euclidean distance, d, to each person from the sampled set. d_i^n is the distance between the i^{th} person in the cohort, to the n^{th} nearest sampled person. For example, the 5^{th} nearest neighbor refers to the distance to the 5^{th} closest person. After generating the sample pool, we calculate the distance for each individual in the patient cohort to the n^{th} nearest sampled person. We choose the sample pool with the lowest n^{th} nearest neighbor distance. The intuition for the n^{th} nearest neighbor distance is to ensure that for each person in the

full cohort, there is someone in the sample pool that is “near” or representative of them. This should result in greater coverage for under-represented sub-groups and phenotypes compared to random sampling. For example, for a disease that can present clinically in a variety of ways (e.g., diabetes), if a rarer presentation is not represented in the chart sample, then for person i from this minority group, the closest sampled individuals will be in other subgroups, resulting in a larger n^{th} nearest neighbor distance. Moreover, sampling to minimize the the n^{th} nearest neighbor distance will not adversely impact the majority presentation since group members will still have representative samples.

Our coverage sampling process can be summarized as following:

1. Cluster the dataset based on clinical factors across a range of k, clusters.
2. Conduct stratified sampling with specified sample size across the clusters multiple times to generate the sample pools.
3. Calculate the n^{th} nearest neighbor distance for each sample set in the sample pool and identify the sample with the minimal n^{th} nearest neighbor distance.

The primary tuning parameter is *n.* This can be prespecified by the user, or *n* can be assessed over a range of values and taking the mean distance. As we show below, the approach is not very sensitive to the choice of *n.*

***2.4. Assumptions*:** The primary assumption of this procedure is that we have a broad cohort from which to sample that fully captures all individuals with the condition from which we wish to define a CP. Meaning, our identified patient cohort (i.e., our denominator) has perfect sensitivity for the outcome of interest and the analytic challenge is improving the specificity of the CP.

2.5. Evaluation Criteria: As shown in Figure 3, we assess the quality of a selected sample for chart review in two ways: cohort coverage and CP performance. For cohort coverage, we compare the n^{th} nearest neighbor distance, with samples exhibiting smaller distances considered more representative of the study cohort. For CP performance, we train a classification model using a sample derived from either our proposed coverage sampling or random sampling methods. All the unsampled patients are regarded as the test dataset. We evaluate the efficacy of these models by comparing the Area Under the Receiver Operating Characteristic Curve (AUROC) using the test dataset. Samples that yield models with higher AUROC values are considered to be better.

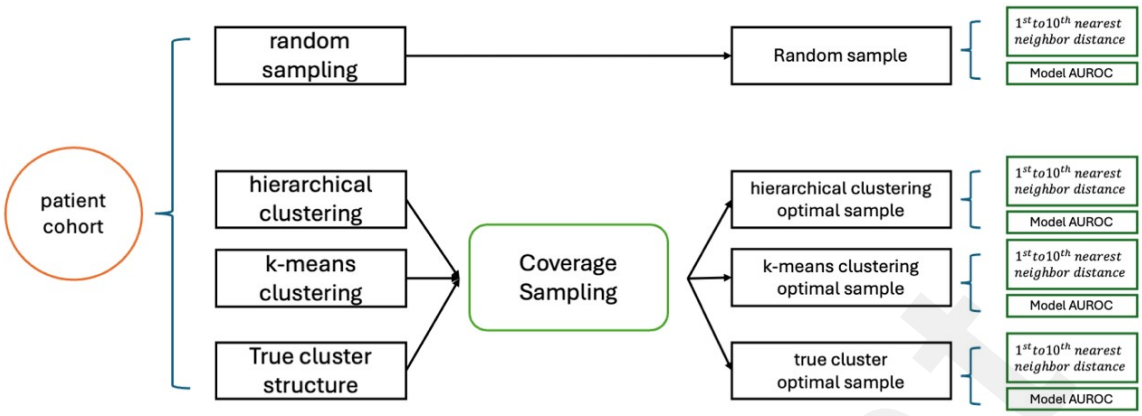

Figure 3: Diagram to show the evaluation criteria of sample quality. The coverage sampling refers to the sampling procedure outlined in Figure 2.

2.6. Simulation Study

We conduct a simulation study to evaluate the efficacy of coverage sampling outlined above. We sample 120 patients from 200 simulated datasets with size of 10,000 and 10 characteristic variables by both random sampling and coverage sampling. Across the datasets, we generate 4 clusters and create a sample with different proportions of each cluster: (simulation set 1: 0.25, 0.25, 0.25, 0.25), (simulation set 2: 0.1, 0.3, 0.3, 0.3), (simulation set 3: 0.1, 0.1, 0.4, 0.4), and (simulation set 4: 0.1, 0.1, 0.1, 0.7). These samples can be interpreted as: equally distributed sub-groups, one minority sub-group two minority sub-groups, and one majority group. The initial simulation set serves as a baseline, wherein all clusters are of equal size. The subsequent simulations (sets 2-4) delve into more complex scenarios, incorporating minority sub-groups to assess their impact on the representation of sub-groups within the samples.

To generate the clustered data, we used the R package *fungible* (17), which employs the following model:

$$X = D_j B + e$$

Where X is the matrix of simulated observations, where each row representing an observation and each column representing a variable; D_j is a matrix of indicator for cluster j, identifying the membership of observation within this cluster; B is a matrix that represents the correlation between the cluster membership and observation scores), and e represents the deviations that generated from a mixture distribution.

To generate outcomes (i.e., phenotypes to be derived), we applythe following model:

$$logit\left(P\left(event=1\right)\right)=\alpha_0+\sum_{j=1}^{k} I\left(cluster_i=j\right)\alpha_0+\sum_{l=1}^{p} X_{il}\beta_l$$

Where $P\left(event=1\right)$ represents the probability of i^{th} encounter outcome equal to 1. $I\left(cluster_i=j\right)$ is the indicator of whether the i^{th} encounter belong to j^{th} cluster or not. X is the design matrix where each row represents one encounter, and each column represents one explanatory variable. α_0, α_i,and β_l are the intercept and main effects corresponding to $I\left(cluster_i=j\right)$ and X. To more accurately reflect real-world conditions, only half of the explanatory variables were incorporated into the generation of the outcome variables, treating the remaining variables as noise (with respect to the outcome).

For each simulated dataset, we assessed information coverage and model performance across three samples. First, using the procedure described in the *Sample pool generation* section, we averaged the 1^{st} to 10^{th} nearest neighbor distances to obtain a hierarchical-cluster-based sample of size 120 and a k-means-cluster-based sample of size 120. For comparison, we selected 120 random samples and 120 that were sampled from the true underlying clusters. In this manuscript, we refer to the four samples as hierarchical, k-means, random, and truth. All data not included in these samples were retained as test data for further analysis. For the model performance comparison, we used each of the four derived samples (*hierarchical cluster coverage, k-means cluster coverage, random, truth)* to fit a logistic regression model to learn a probablitic CP. We computed the area under the receiver operator characteristic (AUROC) to evaluate the model's performance, and averaged the performance over 50 iterations.

2.7. Real-world data application

Our application is motivated from our previous work to develop a CP for a hospital admission due to COVID-19. During the heights of the COVD-19 pandemic, hospitals tested all patients for SARS-CoV-2. Work by us (18) and others (19) has indicated that up to 38% of patients that tested positive for SARS-CoV-2 upon admissions were admitted for reasons other COVID-19. Therefore, a CP for admission due to COVID-19 would need to be more complex than simply a positive SARS-CoV-2 test. Our goal then is define a sample of patients to chart review, in aid of learning a CP for admission due to COVID-19. Since COVID-19 patients could have different presentations, we hypothesize that our coverage sampling approach would be better for learning a CP.

2.7.1. Data Source: We abstracted data from the Duke University Health System (DUHS) EHR system. DUHS consists of three hospitals on a common, EPIC-based, EHR system. The clinical data are organized into a research ready datamart, based on the PCORnet Common Data Model (20).

2.7.2. Source Cohort: Our study cohort consisted of all patients with an inpatient admission and a positive test for SARS-CoV-2 from March 2020 through March 2023 (when routine testing stopped). This definition has perfect sensitivity, but poor specificity, for capturing admissions due to COVID-19. Following our previous work, we split this cohort into training and testing data. The testing data

consisted of 441 patients admitted from January 16 to 22, 2022 and were already chart reviewed for operational purposes. Additional information regarding the testing data can be found in (18). The training data consisted of the other 7,743 unlabeled patients with positive SARS-CoV-2 tests from 2020-2023.

2.7.3. Features Used: For coverage sampling and CP generation, we used 46 clinically relevant features such as encounter characteristics (encounter type, admitting source, discharge disposition), diagnoses, laboratory tests conducted, and medications administered. **Supplemental Table S1** provides full details on features used. While we extracted demographic characteristics, we did not include these in the sampling or CP development steps.

2.7.4. Sampling and Outcome Labeling: We generated 2 samples of 100 using coverage and random sampling from the training dataset of 7,743 patients. For the coverage sampling, we used hierarchical clustering and identified the cluster struction that minimized the 1^{st} nearest neighbor distance. An infectious disease specialist (JHH) chart reviewed and labeled the encounter as due to COVID-19 and/or related sequela or not.

2.7.5. Method Evaluation: We compare the patient characteristics for the samples that were selected from each sampling approach. Then, using the criteria defined above, we evaluate the coverage of the sample of the full cohort. Finally, we used each sample to learn a probabilistic CP based on a LASSO logistic regression. We evaluated each version on the independent test data.

All analyses were conducted in R version 4.3.2. The source code used in this experiments is available at github. This study was approved and declared exempt by the Duke School of Medicine IRB, protocol Pro00109397 (9/14/2021).

3. RESULTS

3.1. Evaluation of sampling methods using simulated data

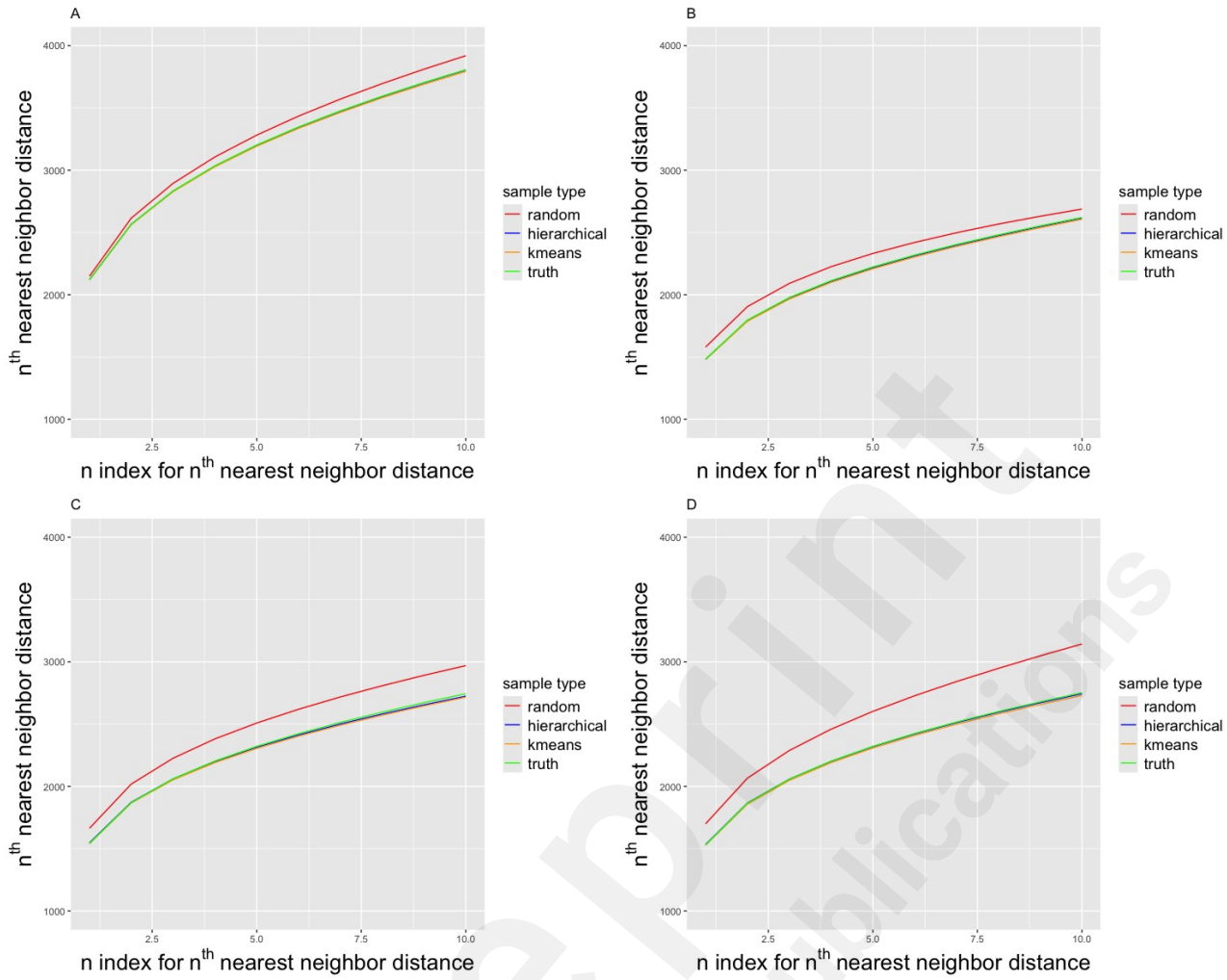

Figure 4: 1st to 10th nearest neighbor distance mean across 200 simulated data samples for 4 cluster ratios. The red line represents the random sample; the blue line represents the coverage sampling based on hierarchical cluster; the orange line represents the coverage sampling based on k-means clustering; the green line represents the coverage sampling based on the true cluster. A: All simulated data follows the baseline cluster ratio (0.25,0.25,0.25,0.25). B: All simulated data follows cluster ratio (0.1,0.3,0.3,0.3). C: All simulated data follows cluster ratio (0.2,0.2,0.4,0.4). D: All simulated data follows cluster ratio (0.1,0.1,0.1,0.7).

Figure 4 presents the mean *1st* to *10th* nearest neighbor distances for both random and coverage sampling methods across four distinct scenarios. Figure 4A illustrates a baseline scenario where all clusters are of equal size, while Figures 4B-D depict scenarios with one, two, and three minority sub-groups, respectively. In each scenario, the coverage samples consistently exhibit smaller nearest neighbor distances compared to those from random samples. As minority sub-groups are incorporated into the simulated cohort, the advantage of the coverage sample over random sample increases. Notably, the distances in the hierarchical- and k-means-clustered coverage samples are closely aligned with those observed in true cluster configurations, indicating similar coverage of the cohort.

After generating the samples, we used the data to learn a probabilistic CP and tested its performance. **Table 1** presents the mean *1st* to *10th* nearest neighbor distance and mean logistic model's AUROC

between a random sample and coverage samples generated using hierarchical clustering, k-means clustering and a true cluster structure of size 120 across 200 simulated datasets. Values highlighted in asterisk (**Table 1**) indicate AUROCs that are significantly higher than that of the random sample at the 0.05 significance level. The visualization of the AUROC results is shown in **Figure S1**. Consistent with coverage results, in the baseline scenario without any minority sub-group, coverage samples exhibit similar AUROC compared to random samples. However, with the introduction of minority sub-group, the coverage samples based on hierarchical, k-means and true cluster structures all produced significantly higher AUROC values compared to random samples. Additionally, we observed that the coverage samples using hierarchical clustering and k-means clustering exhibited similar performance, suggesting that the choice of clustering method has minimal impacts on coverage sampling, provided there are sufficient repetitions.

Table 1: comparison of mean distance and AUC between random sample, and coverage samples based on hierarchical clustering, k-means clustering, and true cluster

Cluster Ratio	Sample Type	Mean of 1^{st} - 10^{th} nearest neighbor Distance	Overall AUROC (95% CI)	Sub-group 1 AUROC (95% CI)	Sub-group 2 AUROC (95% CI)	Sub-group 3 AUROC (95% CI)	Sub-group 4 AUROC (95% CI)
(0.25,0.25,0.25,0.25)	Random	3257.498	0.751 (0.747, 0.755)	0.735 (0.732, 0.738)	0.739 (0.735, 0.743)	0.731 (0.728, 0.735)	0.733 (0.730, 0.736)
	Hierarchical	3170.708	0.751 (0.747, 0.756)	0.734 (0.731, 0.738)	0.739 (0.735, 0.742)	0.731 (0.727, 0.734)	0.732 (0.729, 0.735)
	K-means	3163.75	0.752 (0.748, 0.757)	0.735 (0.732, 0.738)	0.738 (0.734, 0.742)	0.731 (0.727, 0.734)	0.732 (0.729, 0.735)
	Truth	3173.694	0.748 (0.744, 0.753)	0.731 (0.727, 0.735)	0.735 (0.731, 0.738)	0.727 (0.724, 0.731)	0.729 (0.725, 0.732)
(0.1,0.3,0.3,0.3)	Random	2293.954	0.691 (0.678, 0.703)	0.706 (0.694, 0.717)	0.607 (0.596, 0.618)	0.609 (0.599, 0.618)	0.601 (0.593, 0.610)
	Hierarchical	2192.928	0.742✶ (0.731, 0.753)	0.750✶ (0.741, 0.760)	0.633✶ (0.623, 0.643)	0.638✶ (0.630, 0.647)	0.629✶ (0.620, 0.638)
	K-means	2183.826	0.740✶	0.746✶	0.632✶	0.636✶	0.628✶

			(0.729, 0.752)	(0.737, 0.755)	(0.623, 0.642)	(0.628, 0.644)	(0.620, 0.636)
	Truth	2197.208	0.739∗ (0.727, 0.751)	0.747∗ (0.736, 0.757)	0.631∗ (0.621, 0.642)	0.634 (0.625, 0.643∗)	0.622∗ (0.613, 0.632)
(0.1,0.1,0.4,0.4)	Random	2478.755	0.747 (0.737, 0.757)	0.746 (0.737, 0.756)	0.743 (0.734, 0.753)	0.619 (0.612, 0.626)	0.617 (0.609, 0.626)
	Hierarchical	2286.317	0.778∗ (0.769, 0.788)	0.774∗ (0.767, 0.781)	0.769∗ (0.762, 0.776)	0.636∗ (0.630, 0.641)	0.635∗ (0.628, 0.642)
	K-means	2276.429	0.775∗ (0.764, 0.786)	0.774∗ (0.767, 0.781)	0.771∗ (0.764, 0.778)	0.633∗ (0.628, 0.639)	0.634∗ (0.627, 0.640)
	Truth	2292.119	0.782∗ (0.773, 0.791)	0.778∗ (0.772, 0.784)	0.773∗ (0.767, 0.779)	0.639∗ (0.633, 0.644)	0.637∗ (0.630, 0.644)
(0.1,0.1,0.1,0.7)	Random	2584.656	0.731 (0.718, 0.745)	0.740 (0.730, 0.750)	0.743 (0.735, 0.752)	0.739 (0.730, 0.747)	0.586 (0.580, 0.592)
	Hierarchical	2287.864	0.769∗ (0.756, 0.782)	0.776∗ (0.771, 0.782)	0.776∗ (0.770, 0.781)	0.772∗ (0.767, 0.777)	0.601∗ (0.595, 0.607)
	K-means	2276.252	0.775∗ (0.762, 0.788)	0.777∗ (0.771, 0.783)	0.774∗ (0.769, 0.780)	0.771∗ (0.765, 0.776)	0.600∗ (0.594, 0.606)
	Truth	2290.974	0.772∗ (0.759, 0.785)	0.780∗ (0.774, 0.786)	0.778∗ (0.773, 0.783)	0.774∗ (0.769, 0.780)	0.601∗ (0.596, 0.607)

3.2. Evaluation of sampling methods using real-world data

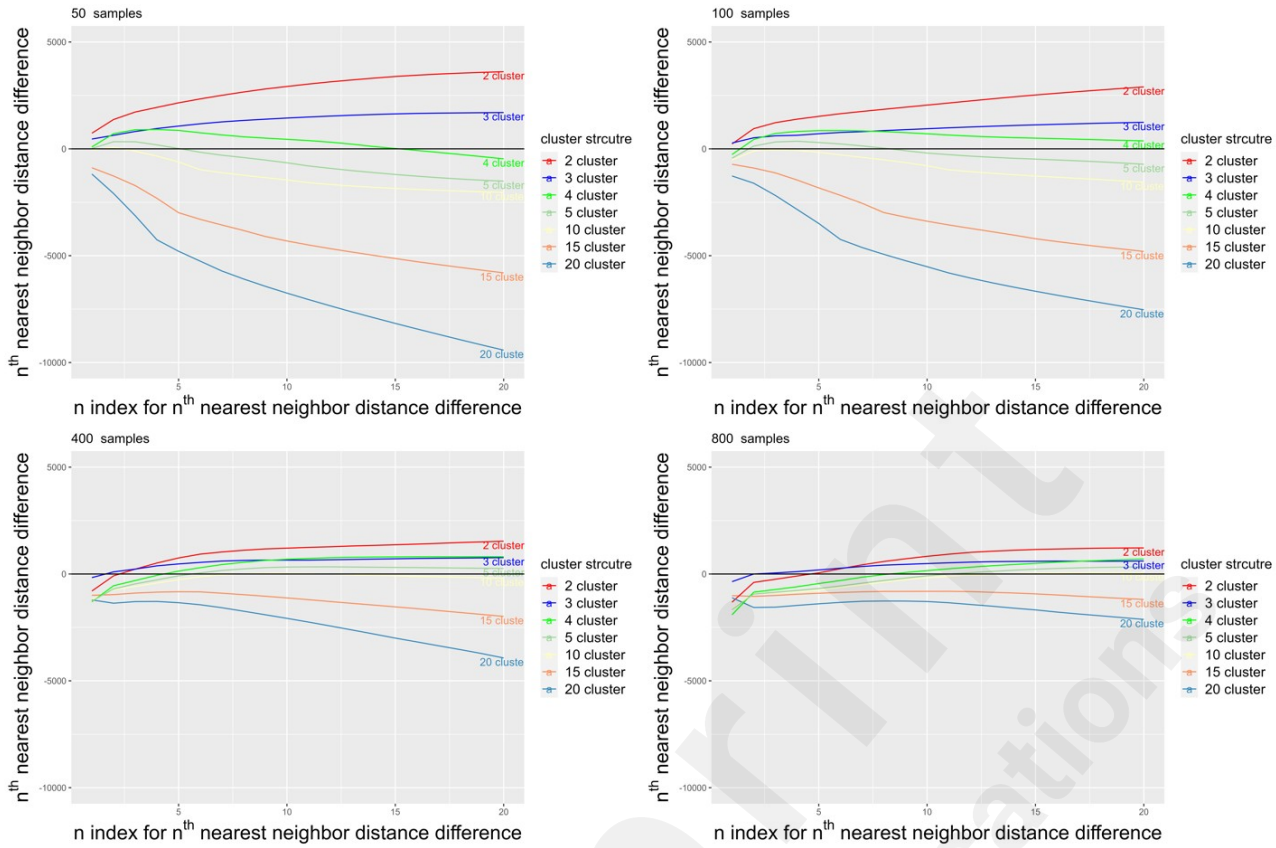

Figure 5: Mean n^{th} nearest neighbor distance difference (random sample distance – coverage sample distance) over 100 replications for real-world data.

In real-world data, the true number of clusters or patient sub-groups is unknown. We therefore explored a range of potential cluster structures, including 2, 3, 4, 5, 10, 15 and 20 cluster structures. Based on simulated data results, with enough replications, the choice of clustering method does not impact our sampling approach; thus, we only evaluated hierarchical clustering in the real-world data analysis. **Figure 5** shows the difference in the mean 1^{st} to 20^{th} nearest neighbor distances in samples generated using the coverage and random samples (i.e., coverage sample n^{th} nearest neighbor distance subtracted from random sample n^{th} nearest neighbor distance). The results demonstrate that for smaller sample sizes (50 and 100), samples drawn from structures with 2, 3, and 4 clusters provide a more accurate representation than their random counterparts. However, as the chart review sample size increases from 400 to 800, the n^{th} nearest neighbor distance of the coverage samples aligns more closely with that of random samples. It is noteworthy that samples derived from 10, 15, and 20 cluster-structures perform less effectively across all sample sizes.

We conducted a chart review of the 100 random samples and 100 coverage samples. Our coverage sampling approach selected a sapmle pool based on 1^{st} nearest neighbor distance. **Table 2** presents the demographic characteristics of the full 7743 patient cohort, as well as the demographics of the coverage and random samples. The standardized mean differences (smd) of the random sample and coverage sample are also shown. Notably, the coverage sample includes a higher percentage of young adults (22%) compared to the random sample (5%), with other demographic variables showing similar prevalence patterns in both samples. After generating labels, we fit a lasso logistic regression model to learn and test a CP. **Table 3** presents the 1^{st} nearest neighbor distance for both

coverage sample and random sample, as well as the AUROC for the learned CP. Additionally, we plot the 1^{st} to 20^{th} nearest neighbor distances of coverage and random sample in the **Figure S2**. Given that the true cluster structure is unknown, we report the AUROC at the demographic variable level. The 1^{st} nearest neighbor distance and the AUROC results indicate that the coverage sample slightly outperforms the random sample. This pattern is also observed at the demographic feature level; however, these differences did not reach statistical significance. Nonetheless, in terms of magnitude, coverage samples demonstrate a notable improvement in the coverage of young adults compared to random samples. Of note, we found that increased coverage of demographic groups is not directly correlated with model performance. For example, the coverage and random samples both have similar proportions of males and females. However, the coverage sampling performs nominally better within each sex group. This supports not including demographics in the clustering step, and relying on clinical drivers of differentiation.

Table 2: Demographic characteristics of real-world data full sample, coverage sample and random sample

Characteristics	***Full Sample***	***Coverage sample***	***Random sample***	***Random vs Cluster SMD***
Sample size	*7743*	*100*	*100*	
Male sex, n (%)	*3737 (48.3%)*	*43(43.0%)*	*47(47.0%)*	*0.080*
Age, n (%)				*0.391*
Children (0 – 18)	*307 (4.0%)*	*3 (3.0%)*	*2 (2.0%)*	
Young adult (18 – 35)	*995 (12.9%)*	*24 (24.0%)*	*11 (10.0%)*	
Middle adult (35 – 65)	*3015 (38.9%)*	*38 (38.0%)*	*38 (38.0%)*	
Older adult (>65)	*3426 (44.2%)*	*35 (35.0%)*	*49 (49.0%)*	
Race and Ethnicity, n (%)				*0.131*
Hispanic	*833 (10.8%)*	*13 (13.0%)*	*15 (15.0%)*	
Non-Hispanic Black	*2999 (38.7%)*	*44 (44.0%)*	*39 (39.0%)*	
Non-Hispanic white	*3572 (46.1%)*	*36 (36.0%)*	*40 (40.0%)*	
Non-Hispanic Asian	*110 (1.4%)*	*3 (3.0%)*	*2 (2.0%)*	
Other races	*229 (3.0%)*	*4 (4.0%)*	*4 (4.0%)*	
Group primary payment				*0.239*
Private	*3564 (46.0%)*	*48 (48.0%)*	*50 (50.0%)*	

Public	*3215 (41.5%)*	*42 (42.0%)*	*35 (34.0%)*	
Self-pay	*307 (4.0%)*	*4 (4.0%)*	*3 (3.0%)*	
Others	*657 (8.5%)*	*6 (6.0%)*	*12 (12.0%)*	

Table 3: Mean AUC comparison between coverage sample and random sample on real-world data

Characteristics (n)	***Random Sample AUROC (95% CI)***	***Cluster Sample AUROC (95% CI)***
1st ***nearest neighbor distance***	*57777.82*	*54213.58*
Overall	*0.726 (0.680,0.772)*	*0.747 (0.701,0.793)*
Sex		
Female (n=220)	*0.724 (0.659,0.784)*	*0.763 (0.695,0.824)*
Male [(=221)	*0.725 (0.656,0.789)*	*0.730 (0.663,0.797)*
Age, n (%)		
Children (0 – 18) [(n=2)	*0.609 (0.312,0.875)*	*0.656 (0.312,0.937)*
Young adult (18 – 35) (n=65)	*0.789 (0.685, 0.886)*	*0.867 (0.774,0.947)*
Middle adult (35 – 65) (n=179)	*0.723 (0.652, 0.793)*	*0.727 (0.647,0.797)*
Older adult (>65) [185]	*0.669 (0.587,0.749)*	*0.673 (0.588,0.754)*
Race and Ethnicity, n (%)		
Hispanic (n=30)	*0.828 (0.674,0.963)*	*0.850 (0.692,0.973)*
Non-Hispanic Black (n=211)	*0.730 (0.660,0.791)*	*0.736 (0.664,0.802)*
Non-Hispanic white (n=187)	*0.703 (0.629,0.778)*	*0.725 (0.658,0.798)*
Non-Hispanic Asian (n=1)	*NA*	*NA*
Other races (n=12)	*0.611 (0.222,0.944)*	*0.925 (0.778,1)*
Group primary payment		
Private (n=216)	*0.698 (0.626,0.764)*	*0.737 (0.667,0.805)*
Public (n=176)	*0.743 (0.672,0.811)*	*0.736 (0.661,0.809)*
Self-pay (n=22)	*0.642 (0.423,0.857)*	*0.733 (0.485,0.923)*

Others (n=27)	*0.820 (0.641,0.961)*	*0.805 (0.623,0.950)*

4. DISCUSSION

CPs are a key component of secondary research with EHR data (21,22). A required step in CP development is conducting a manual chart review to establish a set of "gold-standard" labels to identify patients with and without the condition or outcome of interest. This manual review can be highly time-consuming (23,24). Little work has been conducted on how to optimally select charts for review, with investigators most often using random chart selection (11). This can lead to inefficiencies as potentially informative or edge cases can be missed. To address this concern, we have proposed a sampling strategy to select charts for review that captures the diversity of a population of interest. The key aspect of our method is identifying the optimal sample using a new metric that we have termed the n^{th} nearest neighbor distance. We assessed our method using both simulated and real-world data, evaluating both the information coverage and CP performance. Our findings indicate that our sampling strategy consistently outperforms random sampling in both aspects.

One of the motivations for this approach is the presumption that within any group of patients with a particular condition, there are patient sub-groups that may have a different presentation of that condition. For example, while many patients with diabetes will have HgbA1c values > 6.5% there will be some individuals with controlled diabetes and normal HgbA1c values; however, these patients still have diabetes (25). Such scenarios require the creation of complex CPs that can identify patients with diabetes who have a variety of disease presentations (26). If these patient sub-groups are small enough, a random selection of charts may not provide sufficient coverage of these sub-groups to ensure that the CP performs equitably for all patient sub-groups. As our results demonstrate, coverage sampling has its greatest impact in CP performance when minority sub-groups are present. However, the presence of such minority sub-groups is not a requirement for the method to perform well. In scenarios without minority sub-groups, our method performs comparably to random sampling, highlighting the robustness of the approach.

A novel aspect of our approach is the development of a metric, the n^{th} nearest neighbor distance, to measure the coverage of a given sample. Existing metrics, such as Simpson's Diversity Index and Shannon's Entropy, quantify overall variability across multiple demographic features (27). Simpson's Diversity Index measures the probability that two individuals randomly selected from a sample will belong to different categories, thereby emphasizing the dominance or evenness of group representation (28). Shannon's Entropy quantifies diversity by accounting for both the abundance and the evenness of the categories present, using information theory to assess the uncertainty in predicting the category of a randomly chosen individual (29). While these metrics effectively address general diversity measurement goals, they do not directly align with our specific goal of evaluatingthe representation and coverage of minority sub-groups. The n^{th} nearest neighbor distance explicitly evaluates the distance between records in the unsampled group to those in the sampled group. This targeted focus enables a more precise assessment of the extent to which minority sub-groups are included in study samples, thereby avoiding underrepresentation in the set of records used

for CP development.

Although the construction and generation of sample pools for chart review has not been widely discussed in the CP literature, parallel work exists in the active learning literature. Current active learning methods can be categorized as query-acquiring (pool-based) or query-synthesizing (15,30). We focus on query-acquiring active learning, as query-synthesizing methods are not directly analagous with our work. Query-acquiring active learning employs various sampling strategies, including uncertainty sampling or information-theoretic measures, to identify which sampling strategies would be most impactful for continued labeling (13,31). Therefore, the underlying burden for query-acquiring active learning is the same as in our method: efficient need for labelling (32,33). Most existing methods, including uncertainty sampling or information-theoretic measures, focus on identifying the most influential records to enhance the performance of a given prediction task (15,16,34–37). In contrast, our method, is not based on a supervised objective. Further, our method seeks to select records that best capture diversity, rather than records that are most representative. While representativeness and diversity may be related, they are not necessarily equivalent.

To illustrate our approach, we tested our method with the real-world task of identifying hospital encounters due to COVID-19. During the heights of the COVID-19 pandemic (2020 – 2023), all patients admitted to our health system's hospitals were tested for SARS-CoV2. As we and others have noted, approximately 38.2% of patients with a positive SARS-CoV-2 test, were admitted for reasons other than COVID-19 (18,19). Therefore, if one wanted to identify patients admitted due to COVID-19, a positive SARS-CoV-2 test would not be a sufficient CP because of its poor specificity. We compared the performance of chart review sample based on a random selection of charts and our coverage sampling method. Overall, the cluster-based sample yielded a better performing CP.

While we selected charts based solely on clinical data elements, there were meaningful demographic differences between samples derived from randomly selected charts and through through coverage sampling. Specifically, the cluster-based sample included younger patients, a greater number of non-Hispanic Black patients (though fewer Hispanic patients), and more individuals with public insurance compared to the cohort derived from random sampling. This result highlights one of the key opportunities in this approach: deriving a less biased sample on which to build a CP. As others have described, one of the mechanisms of algorithmic bias is having unrepresentative samples used to develop the algorithm (38). For instance, in the context of rare diseases, the typical ratio of patients with a given rare condition to those without the condition is approximately 100:1 (39). In such scenarios, employing random sampling may result in underrepresentation of minority sub-groups in the chart review sample. When the review sample does not accurately reflect the patient population, the resulting CPs can produce biased results. For example, if certain demographic groups are underrepresented in the dataset, the CP may not learn to make accurate predictions for these groups, leading to disparities in performance (40). To account for this, algorithmic solutions have been proposed including data augmentation (41,42), resampling technique (43,44), and algorithmic adjustments (45). However, instead of addressing this problem algorithmically, we propose addressing it via design. As such, by clustering the data and sampling equally from the obtained clusters, we aim for the chart review sample to better represent the patient population. An advantage of our method is that it does not require the researcher to pre-specify

groups. Moreover, as our empirical results show, we are able to capture demographic diversity with just clinical data.

While our approach shows promise, there are some limitations. Firstly, the performance of our method is related to the quality of the cluster analysis. As others have noted, clustering methods can be highly variable (46). This variability may be more obvious in EHR data, which often suffer data quality issues. Because the clustering step is a means to obtaining a representative sample, we address this by generating multiple samples and selecting the one with best coverage. In principle, it is possible to skip the clustering step and directly choose an optimal sample, leading to more robust results. While such an approach is worthy of further exploration, it would be more computationally expensive and would not necessarily yield meaningfully better results. Another potential limitation is that the coverage sample (intentionally) generates a sample that will likely have a different event rate than the true event rate within the full patient population. While this does not present a problem for rank-based metrics like AUROC, it may affect the calibration of other metrics, such as Kullback-Leibler divergence (47). When calibration is a priority, recalibration methods can be employed (48).

5. Conclusions

Overall, our results show that our coverage sampling method can provide a more representative sample than random sampling, especially when the source cohort contains minority sub-groups. This approach can lead to generation of a CP that has better performance in the overall study population as well as within sub-groups. While CP development is a key part of secondary research with EHR data, little work has been done on how best to derive samples for learning CPs. This work addresses this gap and seeks to spur more investigation in this area. Ultimately, this sampling method has the potential to improve future clinical research by making gold-standard chart review labeling a more efficient process.

6. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT in order to improve the language and readability of manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

1. Wang ECH, Wright A. Characterizing outpatient problem list completeness and duplications in the electronic health record. Journal of the American Medical Informatics Association. 2020 Aug 1;27(8):1190–7.

2. Grauer A, Kneifati-Hayek J, Reuland B, Applebaum JR, Adelman JS, Green RA, et al. Indication

alerts to improve problem list documentation. Journal of the American Medical Informatics Association. 2022 May 1;29(5):909–17.

3. Ghirardello S., Garrè M.-L., Rossi A., Maghnie M. The Diagnosis of Children with Central Diabetes Insipidus. 2007;20(3):359–76. Available from: https://doi.org/10.1515/JPEM.2007.20.3.359

4. Miller RL, Grayson MH, Strothman K. Advances in asthma: New understandings of asthma's natural history, risk factors, underlying mechanisms, and clinical management. Journal of Allergy and Clinical Immunology [Internet]. 2021;148(6):1430–41. Available from: https://www.sciencedirect.com/science/article/pii/S0091674921015232

5. Wang L, Olson JE, Bielinski SJ, St. Sauver JL, Fu S, He H, et al. Impact of Diverse Data Sources on Computational Phenotyping. Front Genet. 2020 Jun 3;11.

6. Gearing RE, Mian IA, Barber J, Ickowicz ; Abel. A Methodology for Conducting Retrospective Chart Review Research in Child and Adolescent Psychiatry. Vol. 15, J Can Acad Child Adolesc Psychiatry. 2006.

7. Panacek E. Performing Chart Review Studies. Edward A Panacek.

8. McKenzie J, Rajapakshe R, Shen H, Rajapakshe S, Lin A. A semiautomated chart review for assessing the development of radiation pneumonitis using natural language processing: Diagnostic accuracy and feasibility study. JMIR Med Inform. 2021 Nov 1;9(11).

9. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, Malin BA, et al. Limestone: High-throughput candidate phenotype generation via tensor factorization. J Biomed Inform. 2014 Dec 1;52:199–211.

10. Carrell DS, Floyd JS, Gruber S, Hazlehurst BL, Heagerty PJ, Nelson JL, et al. A general framework for developing computable clinical phenotype algorithms. Journal of the American Medical Informatics Association. 2024 Aug 1;

11. Vassar M, Matthew H. The retrospective chart review: important methodological considerations. J Educ Eval Health Prof. 2013 Nov 30;10:12.

12. Miller B, Linder F, Mebane WR. Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches. Political Analysis [Internet]. 2020/04/21. 2020;28(4):532–51. Available from: https://www.cambridge.org/core/product/CF6AAF05F465D5BC688A9548433123C1

13. Cohn DA, Ghahramani Z, Jordan MI. Active Learning with Statistical Models. Vol. 4, Journal of Artiicial Intelligence Research. 1996.

14. Huang SJ, Jin R, Zhou ZH. Active Learning by Querying Informative and Representative Examples.

15. Sinha S, Ebrahimi S, Darrell T. Variational Adversarial Active Learning. 2019 Mar 31; Available from: http://arxiv.org/abs/1904.00370

16. Li X, Guo Y. Adaptive Active Learning for Image Classification. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. 2013. p. 859–66.

17. Waller NG, Underhill JM, Kaiser HA. A method for generating simulated plasmodes and artificial test clusters with user-defined shape, size, and orientation. Multivariate Behav Res. 1999;34(2):123–42.

18. Chang F, Krishnan J, Hurst JH, Yarrington ME, Anderson DJ, O'Brien EC, et al. Comparing Natural Language Processing and Structured Medical Data to Develop a Computable Phenotype for Patients Hospitalized Due to COVID-19: Retrospective Analysis. JMIR Med Inform [Internet]. 2023 Aug 22;11:e46267–e46267. Available from: https://medinform.jmir.org/2023/1/e46267

19. Klann JG, Strasser ZH, Hutch MR, Kennedy CJ, Marwaha JS, Morris M, et al. Distinguishing Admissions Specifically for COVID-19 From Incidental SARS-CoV-2 Admissions: National Retrospective Electronic Health Record Study. J Med Internet Res. 2022 May 1;24(5).

20. Hurst JH, Liu Y, Maxson PJ, Permar SR, Boulware LE, Goldstein BA. Development of an electronic health records datamart to support clinical and population health research. J Clin Transl Sci. 2021;5(1).

21. Pfaff ER, Crosskey M, Morton K, Krishnamurthy A. Clinical annotation research kit (CLARK): Computable phenotyping using machine learning. Vol. 8, JMIR Medical Informatics. JMIR Publications Inc.; 2020.

22. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. Vol. 13, Wiley Interdisciplinary Reviews: Computational Statistics. John Wiley and Sons Inc; 2021.

23. Johnston KM, Lakzadeh P, Donato BMK, Szabo SM. Methods of sample size calculation in descriptive retrospective burden of illness studies. BMC Med Res Methodol [Internet]. 2019;19(1):9. Available from: https://doi.org/10.1186/s12874-018-0657-9

24. Connolly A, Kirwan M, Matthews A. A scoping review of the methodological approaches used in retrospective chart reviews to validate adverse event rates in administrative data [Internet]. Available from: https://academic.oup.com/intqhc/article/36/2/mzae037/7658312

25. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, et al. A comparison of phenotype definitions for diabetes mellitus. Journal of the American Medical Informatics Association. 2013;20(E2).

26. Spratt SE, Pereira K, Granger BB, Batch BC, Phelan M, Pencina M, et al. Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus. Journal of the American Medical Informatics Association. 2017 Apr 1;24(e1):e121–8.

27. Peet RK. The Measurement of Species Diversity [Internet]. Vol. 5, Source: Annual Review of Ecology and Systematics. 1974. Available from: https://www.jstor.org/stable/2096890

28. SIMPSON EH. Measurement of Diversity. Nature [Internet]. 1949;163(4148):688. Available from: https://doi.org/10.1038/163688a0

29. Shannon CE. A Mathematical Theory of Communication. Vol. 27, The Bell System Technical

Journal.

30. Settles B. Active Learning Literature Survey [Internet]. 2009. Available from: http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf

31. Freund Y, Seung HS, Shamir E, Tishby N. Selective Sampling Using the Query by Committee Algorithm. Mach Learn [Internet]. 1997;28(2):133–68. Available from: https://doi.org/10.1023/A:1007330508534

32. Muslea I, Minton S, Knoblock C. Active Semi-Supervised Learning = Robust Multi-View Learning. Vol. 2, ICML. 2002. 435–442 p.

33. McCallum A, Nigam K. Employing EM and Pool-Based Active Learning for Text Classification. In: Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1998. p. 350–8. (ICML '98).

34. Jain P, Kapoor A. Active learning for large multi-class problems. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. p. 762–9.

35. Kapoor A, Grauman K, Urtasun R, Darrell T. Active Learning with Gaussian Processes for Object Categorization. In: 2007 IEEE 11th International Conference on Computer Vision. 2007. p. 1–8.

36. MacKay DJC. Information-Based Objective Functions for Active Data Selection. Neural Comput [Internet]. 1992 Jul 1;4(4):590–604. Available from: https://doi.org/10.1162/neco.1992.4.4.590

37. Tong S, Koller D. Support vector machine active learning with applications to text classification. J Mach Learn Res [Internet]. 2002 Mar;2:45–66. Available from: https://doi.org/10.1162/153244302760185243

38. Suresh H, Guttag J. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In: ACM International Conference Proceeding Series. Association for Computing Machinery; 2021.

39. Hu Y, Chen F, Cai Y, Yuan Y. A random under-sampled deep architecture with medical event embedding: Highly 175 imbalanced rare disease classification with ehr data. Network. 2019;20.

40. Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Friedler SA, Wilson C, editors. Proceedings of the 1st Conference on Fairness, Accountability and Transparency [Internet]. PMLR; 2018. p. 77–91. (Proceedings of Machine Learning Research; vol. 81). Available from: https://proceedings.mlr.press/v81/buolamwini18a.html

41. Kennedy G, Dras M, Gallego B. Augmentation of Electronic Medical Record Data for Deep Learning. In: Studies in Health Technology and Informatics. IOS Press BV; 2022. p. 582–6.

42. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. J Big Data. 2019 Dec 1;6(1).

43. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Vol. 16, Journal of Artificial Intelligence Research. 2002.

44. Kim YT, Kim DK, Kim H, Kim DJ. A Comparison of Oversampling Methods for Constructing a Prognostic Model in the Patient with Heart Failure.

45. Hino H. Active Learning: Problem Settings and Recent Developments. 2020 Dec 8; Available from: http://arxiv.org/abs/2012.04225

46. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett. 2010 Jun 1;31(8):651–66.

47. Hershey JR, Olsen PA. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07. 2007. p. IV-317-IV–320.

48. King G, Zeng L. Logistic Regression in Rare Events Data. Political Analysis. 2001 Jan 4;9(2):137–63.