RELEAP: Reinforcement-Enhanced Label-Efficient Active Phenotyping for Electronic Health Records

Yang Yang, BS[1], Kathryn I. Pollak, PhD[2,3], Bibhas Chakraborty, PhD[1,4,5,6], Molei Liu, PhD*[7,8], Doudou Zhou, PhD*[6], Chuan Hong, PhD*[1]

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA
[2]Cancer Prevention and Control Research Program, Duke Cancer Institute, Durham, NC, USA
[3]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, USA
[4]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore
[5]Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore
[6]Department of Statistics and Data Science, National University of Singapore, Singapore
[7]Department of Biostatistics, Peking University Health Science Center, Beijing, China
[8]Beijing International Center for Mathematical Research, Peking University, Beijing, China

Corresponding Author:
Chuan Hong

*Co-last authors contributed equally to this work.

ABSTRACT

Objective: Electronic health record (EHR) phenotyping often relies on noisy proxy labels, which undermine the reliability of downstream risk prediction. Active learning can reduce annotation costs, but most rely on fixed heuristics and do not ensure that phenotype refinement improves prediction performance. Our goal was to develop a framework that directly uses downstream prediction performance as feedback to guide phenotype correction and sample selection under constrained labeling budgets.

Materials and Methods: We propose Reinforcement-Enhanced Label-Efficient Active Phenotyping (RELEAP), a reinforcement learning–based active learning framework. RELEAP adaptively integrates multiple querying strategies and, unlike prior methods, updates its policy based on feedback from downstream models. We evaluated RELEAP on a de-identified Duke University Health System (DUHS) cohort (2014–2024) for incident lung cancer risk prediction, using logistic regression and penalized Cox survival models. Performance was benchmarked against noisy-label baselines and single-strategy active learning.

Results: RELEAP consistently outperformed all baselines. Logistic AUC increased from 0.774 to 0.805 and survival C-index from 0.718 to 0.752. Using downstream performance as feedback, RELEAP produced smoother and more stable gains than heuristic methods under the same labeling budget.

Discussion: By linking phenotype refinement to prediction outcomes, RELEAP learns which samples most improve downstream discrimination and calibration, offering a more principled alternative to fixed active learning rules.

Conclusion: RELEAP optimizes phenotype correction through downstream feedback, offering a scalable, label-efficient paradigm that reduces manual chart review and enhances the reliability of EHR-based risk prediction.

Keywords: Reinforcement Learning; Active Learning; Phenotyping; Risk Prediction; Lung Cancer.

1 INTRODUCTION

Phenotyping refers to the process of deriving clinically meaningful traits, such as diseases, behaviors, or risk factors, from raw health data using structured codes, laboratory results, medications, and unstructured clinical text. Intermediate phenotypes play a central role in biomedical prediction models, serving as covariates or mediators that connect underlying patient characteristics to clinical outcomes. Their accuracy directly shapes the performance, interpretability, and fairness of downstream risk prediction. Within electronic health records (EHRs), however, such phenotypes are rarely observed directly. Instead, they must be inferred from raw data sources, including diagnoses, procedures, laboratory results, medications, and clinical text, using computable phenotypes, algorithmically defined constructs based on EHR data (e.g., diagnosis codes, medications, laboratory values, or unstructured clinical text) that approximate clinical conditions or behaviors.[1–3]

Developing high-quality computable phenotypes at scale remains a major bottleneck. While unsupervised approaches like clustering and rule-based heuristics avoid the need for labels, they often lack specificity and clinical validity.[4, 5] Supervised learning offers higher accuracy but depends on manually curated labels derived from chart review, which is time-intensive, costly, and unscalable. These challenges have sparked growing interest in label-efficient phenotyping strategies, which aim to reduce annotation burden without compromising phenotype quality. Approaches such as weak supervision,[6] surrogate labeling,[7] and active learning exemplify this shift toward minimizing reliance on gold-standard labels.[8]

The accuracy of label-efficient phenotyping can be significantly enhanced by leveraging multiple data modalities. Structured EHR elements, such as International Classification of Diseases (ICD) codes,[9] are often used as proxies for behavioral phenotypes like smoking, but their sensitivity is low and subject to substantial label noise.[10] For instance, studies have shown that relying solely on structured codes captures only a small subset of true smokers and systematically underrepresents smoking prevalence.[11] Incorporating unstructured clinical text through natural language processing (NLP) and integrating self-reported data from social history or patient intake forms provides a more complete and accurate phenotype. Such multi-modal integration improves coverage and reduces misclassification, forming a more robust foundation for label-efficient learning.[12]

A further complication is that while small validation cohorts from chart review can be used to benchmark algorithms, such labels are insufficient for monitoring performance across diverse populations or for guiding iterative refinement. This limitation motivates the use of downstream prediction performance to guide active learning rather than relying exclusively on scarce gold-standard labels. A related line of work has sought to directly connect phenotyping with risk prediction. For example, Hong et al. proposed a semi-supervised framework that jointly validates multiple surrogate phenotypes while estimating associations with genetic risk factors.[13] By combining the phenotyping step with the downstream risk genetic association study, their method

improved both phenotype accuracy and the reliability of subsequent genetic association analyses. This illustrates the value of integrating phenotype refinement with predictive modeling, though existing approaches have largely focused on semi-supervised inference rather than adaptive sample selection.

To facilitate consistent terminology, we define the **automatic reference phenotype** as a high-fidelity label that integrates structured self-reported smoking information with large language model-based extraction from clinical notes. This multimodal phenotype serves as the most accurate available reference for downstream evaluation and is denoted as S_{true} throughout this paper.

Building on this definition, we present **Reinforcement-Enhanced Label-Efficient Active Phenotyping (RELEAP)**, a reinforcement learning-driven active learning agent designed to improve phenotype correction under constrained labeling budgets.[14, 15] Unlike prior label-efficient approaches, RELEAP directly uses downstream risk prediction performance as a feedback signal for phenotype refinement. By leveraging automatically constructed reference labels instead of manual chart review, RELEAP is both scalable and cost-efficient. Building on prior work such as Hong et al.,[13] we benchmark RELEAP against noisy-label and single-strategy baselines and quantify label efficiency by the labeling budget required to approach near-oracle performance.

2 MATERIALS AND METHODS

2.1 Motivation Example: Lung Cancer Prediction with Imprecise Smoking Phenotypes

Lung cancer remains a leading cause of cancer-related mortality, and risk prediction models are increasingly being developed to support earlier detection and preventive care. Recent machine learning models trained on large-scale EHR data have demonstrated strong performance in identifying high-risk patients, achieving area under the receiver operating characteristic curve (AUC) values above 0.75 across diverse real-world settings.[16] In these models, smoking behavior consistently emerges as one of the most important predictors, alongside age, race, and chronic respiratory conditions such as chronic obstructive pulmonary disease (COPD).[17]

However, accurately capturing smoking status in EHR data is challenging. In routine care, smoking behavior is often recorded inconsistently, if at all. Proxy variables (e.g., diagnosis codes related to tobacco use) are commonly used for computable phenotyping, but their reliability is questionable. For example, in our analysis of EHR data from the Duke University Health System (DUHS), we found that the ICD-based proxy smoking phenotype (S^*) is markedly incomplete when compared to self-reported smoking status. While structured self-report fields in the vital signs table indicate that approximately 30% of patients have a documented history of smoking, only about 10% of patients carry any smoking-related ICD-9/10 code (e.g., `305.1`, `F17.*`, `Z87.891`). This large discrepancy highlights the limitations of relying on structured diagnosis

codes as a proxy for smoking behavior. The sparse and imbalanced coverage of ICD codes fails to capture the true prevalence of smoking in the population and introduces substantial misclassification bias when used in downstream risk prediction models.

To address this gap, we constructed an automatic reference phenotype by combining structured self-reported smoking information with unstructured smoking mentions extracted from clinical notes using large language models prior to the prediction index date. This multimodal definition provides a more complete and accurate depiction of patient smoking history and serves as a reference label for evaluation. The discrepancy between S^* and S_{true} highlights a budget-constrained decision problem: given limited access to high-fidelity reference labels, determining which patients should be upgraded from noisy proxies in order to maximize downstream prediction performance. This decision problem motivates the design of RELEAP, which adaptively allocates scarce label queries rather than relying on fixed heuristics.

2.2 Proposed RELEAP agent

The RELEAP workflow (Figure 1) begins with a proxy phenotype (S^*) derived from structured data (X_1), anchored by a small seed set labeled with the automatic reference phenotype (S_{true}), constructed by combining structured self-reported smoking status with large language model (LLM)-extracted note evidence prior to the index date. A downstream risk prediction model for incident lung cancer (Y; time-to-event T) provides feedback throughout the process. At each iteration, candidate unlabeled patients are scored using multiple active learning heuristics (uncertainty, diversity, and query-by-committee, QBC), as described in Section 2.2.3. A reinforcement learning (RL) agent then adaptively learns how to weight these strategies to select the most informative samples, with detailed reward formulations provided in Supplement S4. Throughout training, RELEAP maintains proxy labels (S^*) for all patients; only selected samples have their S^* replaced with S_{true} in each iteration. Newly acquired labels overwrite the corresponding proxies, iteratively refining the training set. The downstream model is retrained after every iteration, and a shaped reward based on predictive performance is used to update the RL policy. Training episodes terminate when the labeling budget is exhausted. This closed feedback loop allows RELEAP to balance exploration and exploitation,[14] progressively aligning corrected phenotypes with true patient characteristics and improving predictive accuracy and calibration.

2.2.1 Aphrodite-based baseline

As a pragmatic comparator, we implemented the Aphrodite method,[18] a noisy-label baseline that relies on structured ICD codes and NLP-derived Concept Unique Identifiers (CUIs) without adaptive querying. Positive examples were seeded by the presence of smoking-related ICD codes (e.g., V15.82, 305.1, F17.*, Z72.0, Z87.891, Z71.6) and/or affirmed NLP CUI mentions, whereas negatives required no smoking-related ICD codes, no affirmed CUI mentions, and zero pre-index notes. Using these seeds, we trained a

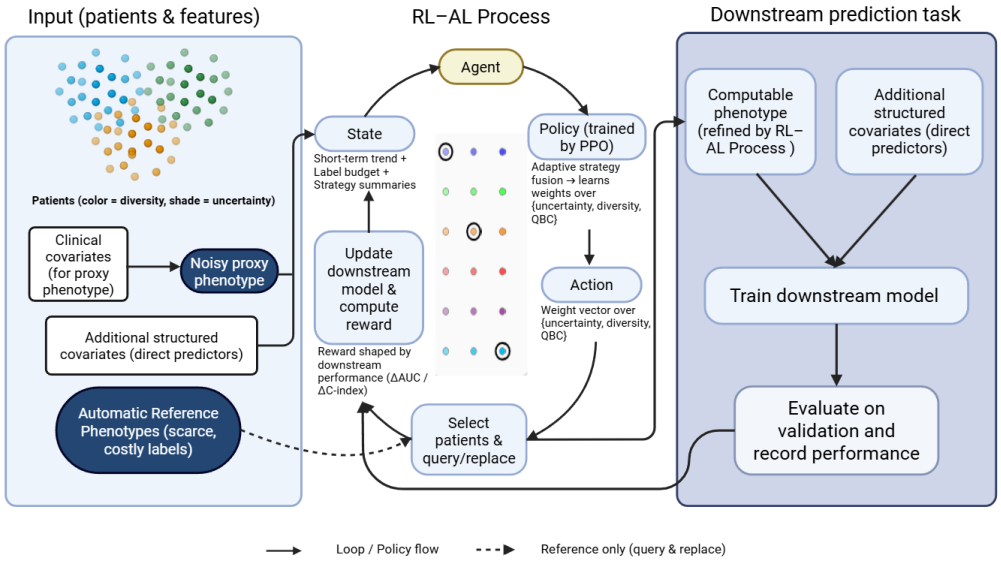

Figure 1: RELEAP workflow. Patient-level inputs include clinical covariates for constructing proxy phenotypes, additional structured predictors for risk models, and a limited pool of costly reference labels. The reinforcement learning-active learning (RL-AL) process adaptively selects patients to correct noisy proxy phenotypes into refined computable phenotypes, guided by a policy trained with proximal policy optimization (PPO). The state vector summarizes short-term downstream performance trends, labeling budget, and strategy-level statistics. The policy outputs weights across active learning strategies (uncertainty, diversity, and query-by-committee, QBC), which drive the action of selecting patients for relabeling. Corrected phenotypes are then combined with additional structured predictors to train downstream risk prediction models. Model performance on validation data (e.g., area under the receiver operating characteristic curve [AUC] or concordance index [C-index]) is logged as the final outcome and also fed back as the reward signal, closing the loop between downstream prediction and the RL-AL process.

regularized logistic regression model on per-code counts (ICD/CUI) augmented with a note-count feature, and then scored all patients. This noisy but scalable baseline follows the Aphrodite framework for weakly supervised phenotype labeling in EHR data.

2.2.2 Active Learning Strategies

We consider three active learning strategies (uncertainty, diversity, and QBC) as the *action basis* for the RELEAP agent. At each iteration, candidate patients are scored by these strategies, normalized, and then combined by the agent to select the top-k samples. Random sampling is included as a comparator but is not part of the RL action space.

All features are standardized as z-scores computed on the *current labeled training set*. For labeled patients, we use the automatic reference phenotype label S_{true}, whereas for unlabeled patients we rely on the continuous proxy score S^*. To ensure comparability, all strategy scores are min–max normalized to the range $[0, 1]$ in each iteration. Ties are resolved by adding a small random jitter ($\epsilon \approx 10^{-6}$). Once selected, patients are labeled without replacement and are never re-queried.

The strategies capture complementary notions of informativeness:

1. **Uncertainty.** This strategy prioritizes patients whose outcomes are most ambiguous under the current model. For each unlabeled patient i, uncertainty is quantified by the entropy of the predicted probability $\hat{p}_{t,i}$:

$$H(\hat{p}_{t,i}) = -\hat{p}_{t,i} \log \hat{p}_{t,i} - (1 - \hat{p}_{t,i}) \log\big(1 - \hat{p}_{t,i}\big),$$

   favoring samples for which the model is least confident.[19]

2. **Diversity.** While uncertainty targets ambiguous cases, diversity promotes coverage across the feature space. Distances are computed between each unlabeled patient and its nearest labeled neighbors, with higher scores assigned to patients farther from the labeled set. A variance adjustment is applied for numerical stability (see Supplement S2).[20]

3. **Query-by-Committee (QBC).** To capture model disagreement, we train a committee of $M = 7$ logistic regression models on bootstrap resamples of the labeled set. Each model applies mild perturbations, including small random jitter in the L2 penalty and feature dropout with probability 0.1. The QBC score reflects disagreement across committee models, measured by the prediction variance with an entropy-based stabilizer (see Supplement S2). This strategy favors patients where competing models disagree most.[21]

4. **Random (baseline).** Finally, to provide a reference benchmark, we include random sampling, which selects patients uniformly from the unlabeled pool. This strategy is excluded from the RL-driven action space.

2.2.3 Reinforcement Learning Integration

We model phenotype correction under a label budget as a partially observable Markov decision process (POMDP). At each iteration, the agent observes a summary state, outputs a nonnegative weight vector, receives a shaped reward from the downstream model's performance, and transitions to the next state. Episodes terminate when the labeling budget is exhausted.

State. The state s_t summarizes the current learning context, including per-strategy score summaries on the unlabeled pool, composition of the labeled set, the recent trajectory of downstream performance, and the remaining budget (see Supplement S3 for full details).

Action. Based on the observed state, the agent outputs a nonnegative **weight vector** over the three informative strategies (uncertainty, diversity, and QBC):

$$\mathbf{w}_t = \left(w^{\text{unc}}, w^{\text{div}}, w^{\text{qbc}}\right), \quad w_i \geq 0, \quad \sum_i w_i = 1.$$

This vector combines the normalized strategy scores into a single ranking function that determines which patients are selected from the unlabeled pool.

Reward. The agent receives a shaped reward derived from the downstream model's performance (e.g., AUC in classification or C-index in survival). Stabilization is achieved via a short moving baseline and mild budget-aware scaling; detailed formulas are provided in Supplement S3.

RL algorithm. The policy $\pi_\theta(a \mid s)$ is optimized with Proximal Policy Optimization (PPO), maximizing the expected discounted return $\mathbb{E}[\sum_{t=1}^{T} \gamma^{t-1} R_t]$ (see Supplement S3).

Iteration loop. Each iteration follows an agent-environment cycle: the agent observes s_t, selects action $\mathbf{w}t$, receives reward r_t, and transitions to $st + 1$. The practical implementation proceeds as follows:

1. Compute strategy scores on the unlabeled pool using uncertainty, diversity, and QBC (with random sampling used only as a baseline).

2. Observe the current state s_t, including strategy summaries, recent trajectory, budget fraction, labeled-set composition, and downstream performance.

3. Select an action $\mathbf{w}_t$ via the PPO policy and rank the unlabeled pool accordingly.

4. Acquire labels by querying S_{true} for the top-k selected patients, overwriting S^* with S_{true}, and marking them as labeled. After this update, the phenotype field S refers to the full training set, where labeled patients use S_{true} and all remaining patients continue to contribute through their S^* values.

5. Retrain the downstream prediction model (logistic or penalized Cox, depending on mode) and evaluate validation performance; full details on feature processing and additional metrics are in Supplement S3.

6. Compute the reward R_t, store (s_t, a_t, R_t, s_{t+1}), and update the PPO agent periodically.

Details of the simulation setup and synthetic data experiments are provided in Supplement S1.

2.3 Evaluation of RELEAP on Real-World EHR Data

We evaluated RELEAP using de-identified EHR data from the DUHS spanning **2014–2024**. This dataset provides a representative testbed for assessing phenotype correction in the context of lung cancer risk prediction, where accurate characterization of smoking behavior is particularly critical. As noted earlier, ICD-only proxies (S^*) tend to underreport smoking compared to the multimodal automatic reference phenotype (S_{true}), motivating evaluation in this setting.

Cohort and study population. The study cohort included adults aged 35–70 years with at least 365 days of prior EHR history before the index date, defined as the qualifying outpatient encounter. Duplicate patient records were removed. Patients with evidence of prevalent lung cancer prior to the index date were excluded. Records with missing demographic information (e.g., sex, race) were retained and handled using predefined preprocessing rules.

Variable construction and notation. Study variables were derived from both structured and unstructured EHR elements. Smoking-related ICD-9/10 codes were extracted within a one-year look-back window prior to the index date and encoded as features (X_1). A proxy phenotype (S^*) was estimated using a logistic regression model trained on X_1. The automatic reference phenotype (S_{true}) was defined by combining structured self-reported smoking status with LLM-extracted mentions from clinical notes prior to index, prioritizing structured entries in case of disagreement. Additional structured covariates (X_2) included demographics, chronic obstructive pulmonary disease (COPD; binary indicator and one-year pre-index count), and a Hispanic ethnicity indicator, all measured at baseline. The outcome (Y) was incident lung cancer after the index date, defined as $\geq$2 primary lung cancer ICD codes occurring within 60 days of each other before the administrative censoring date. Time-to-event (T) was measured from the index date to the first qualifying code; patients without an event were censored at the truncation date.

Automatic reference phenotype definitions and variants. To assess robustness, we considered several definitions of the automatic reference phenotype. Unless otherwise noted, the primary definition integrates structured self-reported smoking status with smoking-related mentions extracted from clinical notes using large language models. For sensitivity analyses, we also evaluated three alternative definitions: (1) combining structured self-reported smoking status with concept identifiers extracted by traditional NLP methods; (2) using only concept identifiers extracted by traditional NLP without relying on self-reported information; and (3) using only smoking-related mentions identified by LLMs without self-reported data. Comparative results for these variants are summarized in Supplement S5.

Preprocessing. Categorical variables were one-hot encoded with an explicit "Unknown" category for missing values. Continuous variables were standardized, and missing values were imputed using the training-set median. All features and S_{true} were constructed strictly from pre-index data to prevent data leakage. A descriptive summary of the final cohort, including demographic characteristics, smoking phenotype distributions (based on S_{true}), outcome rates, COPD prevalence, and coverage of smoking-related ICD codes, is provided in Table 1. Figure 2 further illustrates the proportion of patients with at least one smoking-related ICD code.

Table 1: Dataset summary of the DUHS lung cancer cohort.

Characteristic	Value
Overall	
N	238,119
Age, mean (SD)	52.94 (10.23)
Age, median [IQR]	53 [44–62]
Sex	
Female	138,415 (58.1%)
Male	99,702 (41.9%)
No information	1 (0.0%)
Unknown	1 (0.0%)
Race	
American Indian or Alaska Native	1,512 (0.6%)
Asian	8,357 (3.5%)
Black or African American	57,014 (23.9%)
Multiple race	926 (0.4%)
Native Hawaiian or Other Pacific Islander	227 (0.1%)
No information	318 (0.1%)
Other	4,184 (1.8%)
Refuse to answer	6,227 (2.6%)
White	159,354 (66.9%)
Hispanic/Latino	
Declined to report	8,844 (3.7%)
No	219,828 (92.3%)
Unavailable	187 (0.1%)
Unknown	24 (0.0%)
Yes	9,236 (3.9%)
Smoking	
Never smoked	125,556 (52.7%)
No information	14,940 (6.3%)
Smoked in the past	50,301 (21.1%)
Smokes daily	19,717 (8.3%)
Smokes daily (heavy)	204 (0.1%)
Smokes daily (light)	1,345 (0.6%)
Smokes some days	3,863 (1.6%)
Smoking status unknown	89 (0.0%)
Unknown	22,104 (9.3%)
Outcome	

(continued on next page)

Characteristic	Value
Lung cancer, n (%)	1,846 (0.78%)
Smoking ICD coverage	
No smoking ICD	216,614 (91.0%)
Has $\geq$ 1 smoking ICD	21,505 (9.0%)
COPD	
No	226,055 (94.9%)
Yes	12,064 (5.1%)

Notes. *SEX* values are expanded to full names; *RACE* uses single-code mapping (01–07, NI, UN, OT); *HISPANIC* is reported separately (Y, N, R, UN, NI). *Smoking behavior* was derived from vitals labels in column `smoking_status` and presented as behavioral descriptors (Smokes daily / Smokes some days / Smoked in the past / Never smoked / Unknown). *Heavy* and *light* correspond to vitals labels (e.g., 'heavy tobacco smoker' and 'light tobacco smoker'); no quantitative cigarettes-per-day definition was available in this dataset. *Smoking ICD coverage* is defined as having $\geq$ 1 smoking-related ICD codes.

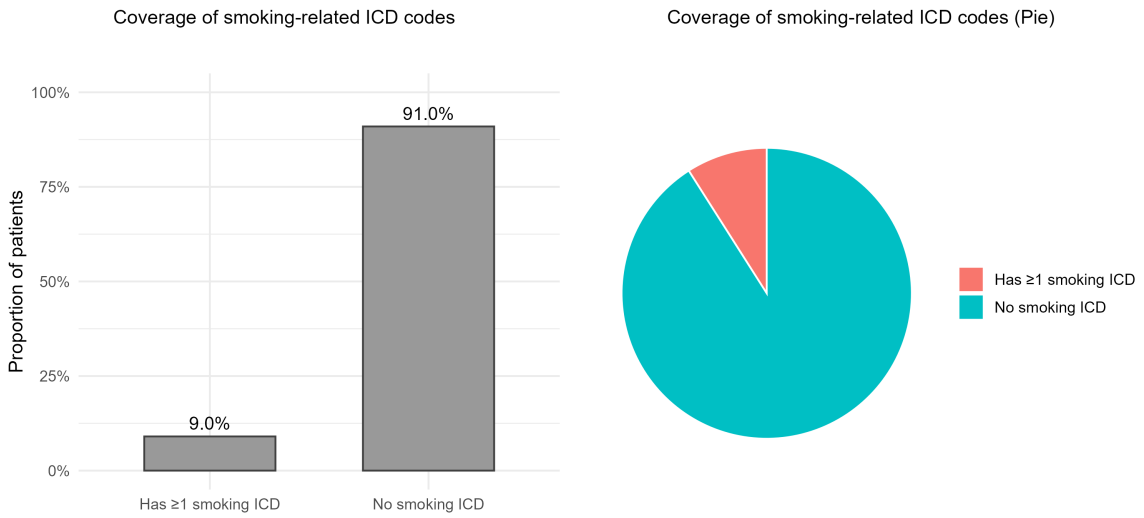

Figure 2: Coverage of smoking-related ICD codes (presence of $\geq$ 1 smoking-related ICD per patient).

Experimental setup. Experiments followed a training-validation framework to simulate real-world phenotype correction. Patients were randomly split into 80% training and 20% validation subsets, stratified by the outcome to preserve event rates. Active learning began with a small, balanced seed set drawn from the training partition, where the noisy proxy labels (S^*) were replaced by the automatic reference labels (S_{true}) to initialize the labeled set. All remaining training patients continued to contribute through their proxy labels, ensuring that the entire training pool was used for model updates. At each iteration, the RELEAP agent generated a set of weights across candidate sampling strategies to guide patient selection. For evaluation, the same weighting scheme was mirrored on the validation set so that training and validation evolved under

comparable replacement dynamics. Downstream prediction models were retrained at every iteration, and their predictive performance was tracked in parallel.

Planned analyses. The primary classification task predicted incident lung cancer using the combined feature set of the phenotype estimate (S) and structured covariates (X_2). Model performance was primarily evaluated by the AUC. A secondary analysis used a penalized Cox proportional hazards model to estimate time-to-event outcomes, with predictive accuracy summarized by the C-index. Performance across iterations was summarized as the mean value with 95% confidence intervals across repeated simulations. Final-round comparisons included measures of discrimination (AUC or C-index), calibration (mean squared error [MSE] of predicted probabilities), and threshold-based metrics such as the F1 score, true positive rate (TPR), and positive predictive value (PPV) at a fixed false positive rate (FPR) of 0.1. We also conducted subgroup analyses stratified by sex, not for clinical inference but to illustrate how different active learning strategies can behave differently under varying data distributions.

Comparators. We benchmarked RELEAP against a hierarchy of baseline models:

(i) **Proxy-only** (`S_star_baseline`): a zero-label baseline trained directly on the noisy proxy phenotype (S^*) without label correction.

(ii) **Aphrodite-based baseline** (`Aphrodite`): a weakly supervised method relying on structured ICD codes and NLP-derived CUIs without adaptive querying.

(iii) **Single-strategy active learning**: `random`, `uncertainty`, `diversity`, `qbc`.

(iv) **RELEAP**: a performance-driven fusion of multiple sampling strategies.

(v) **Oracle** (`S_true_oracle`): a fully supervised model trained with all high-fidelity automatic reference phenotype labels (S_{true}).

For completeness, we also compared RELEAP across multiple automatic reference phenotype definitions described in Section 2.3 and summarized results in Supplement S5.

3 RESULTS

We evaluated RELEAP on the DUHS lung cancer cohort to assess its effectiveness in correcting noisy proxy phenotypes. Results are organized into two parts: (1) overall performance across all patients, considering both logistic classification and survival analyses, and (2) subgroup analyses stratified by sex to illustrate how different active learning strategies behave under varying data distributions.

3.1 Overall Performance Across All Patients

RELEAP was evaluated under two settings, logistic classification and survival analysis. In both modes, RELEAP substantially improved predictive performance over the noisy proxy phenotype baseline (S^*) and approached the oracle trained with all automatic reference phenotype labels (S_{true}). Sensitivity analyses (Supplement S5) showed that the combined definition using structured self-reports and LLM-extracted notes achieved higher AUC and lower MSE than the NLP-based variant, and both SR-based definitions outperformed their NoSR counterparts.

Logistic classification mode (evaluated by AUC). Models were trained on labeled $[S, X_2]$ features and evaluated on the held-out validation set. Each experiment used a total labeling budget of 63,000 patients, queried in batches of 3,000 across 20 iterations, with an initial balanced seed set of 3,000. All methods were repeated over ten independent replications. Figure 3 shows mean $\pm$ 95% confidence interval trajectories for discrimination (AUC, F1 score, TPR, PPV) and calibration (probability MSE).

At convergence, discrimination improved markedly: mean AUC rose from 0.774 ± 0.010 for the noisy proxy baseline to 0.805 with RELEAP and 0.807 with uncertainty sampling, closely approaching oracle-level performance (0.807). F1 and TPR showed comparable gains, while PPV increased modestly from 0.034 to 0.037. Calibration also improved, with MSE decreasing from 0.192 (baseline) to 0.186 under RELEAP, approaching the oracle's 0.180.

Strategy behaviors diverged across metrics. **RELEAP** consistently ranked among the top performers with smoother, more stable trajectories. *Uncertainty* sampling achieved the highest short-term gains but with greater variability. *Diversity* favored calibration but lagged slightly in discrimination, while *QBC* improved F1 score early before stabilizing. *Random* sampling improved gradually but remained inferior overall. RELEAP's advantage lay in adaptively balancing these complementary heuristics.

Survival mode (evaluated by C-index). In the survival setting (penalized Cox regression; Z constructed from S and X_2 with feature screening), experiments used a total labeling budget of 84,000 patients, queried in batches of 4,000 over 20 iterations, with an initial balanced seed set of 4,000. Each method was repeated ten times. Figure 4 demonstrates that all strategies markedly outperformed the noisy proxy baseline (C-index 0.718, 95% CI: 0.713–0.723). RELEAP achieved 0.752 (0.747–0.757), converging toward the oracle (0.753, 0.748–0.758).

Uncertainty and *QBC* reached this plateau fastest, followed closely by **RELEAP**, while *Diversity* and *Random* trailed slightly.

Overall, RELEAP provided the most balanced and reliable trajectory, achieving robust survival prediction with convergence, mirroring the efficiency observed in the logistic classification experiments.

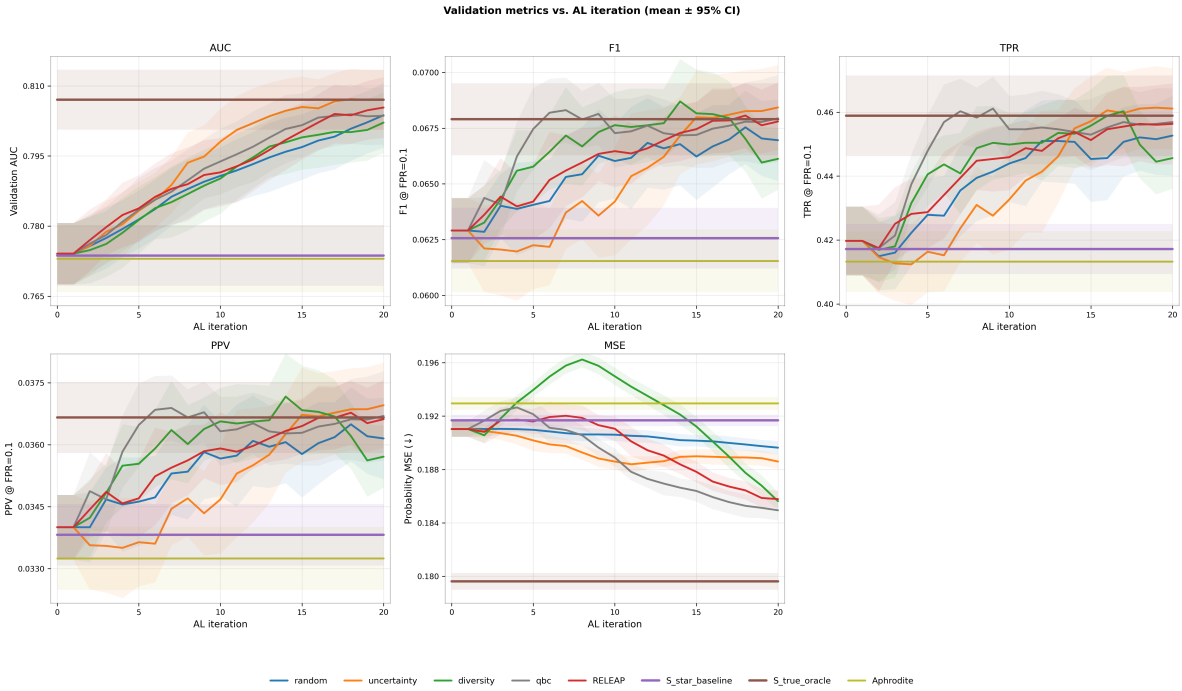

Figure 3: Validation Metrics vs. AL Iteration in the Logistic Mode (mean ± 95% CI over ten replications).

3.2 Subgroup Analyses: Male vs. Female

To further examine population heterogeneity, we stratified the analyses by sex. Figure 5 presents validation trajectories (mean ± 95% CI over ten replications) for AUC, F1 score, TPR, PPV, and probability MSE in male and female subgroups.

Male subgroup. All active learning strategies improved steadily over the noisy proxy baseline (AUC 0.766 ± 0.019) and approached the automatic reference phenotype oracle (0.803 ± 0.018). Final AUC values clustered around 0.798–0.803 across methods, with RELEAP (0.802 ± 0.018) and uncertainty sampling (0.803 ± 0.018) leading. F1 increased modestly from 0.057 ± 0.006 to about 0.063, with RELEAP and QBC showing smoother convergence. TPR rose from 0.464 ± 0.034 to nearly 0.495 (uncertainty), while PPV improved slightly from 0.030 to 0.033. Calibration gains were limited: MSE decreased marginally from 0.214 ± 0.002 to about 0.213 under RELEAP and QBC, remaining above the oracle (0.199 ± 0.002). Overall, all strategies converged to similar discrimination and calibration levels, with RELEAP offering smoother and more stable trajectories rather than absolute superiority.

Female subgroup. In contrast, female subgroup results revealed clearer separation across strategies. Baseline AUC (0.779 ± 0.015) improved to around 0.808–0.809 under RELEAP (0.808 ± 0.012), uncertainty (0.809 ± 0.013), and QBC (0.806 ± 0.012), closely

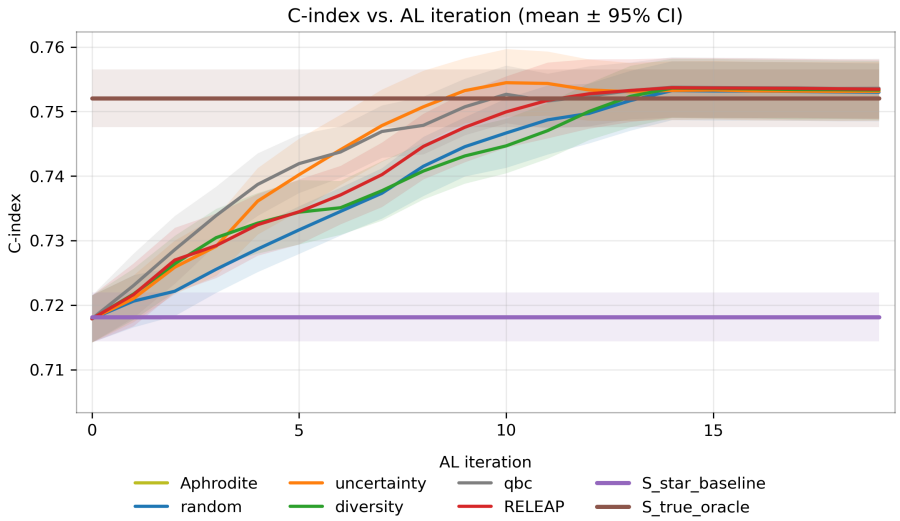

Figure 4: C-index vs. AL Iteration in the Survival (Cox) Mode (mean ± 95% CI over ten replications).

matching the oracle (0.809±0.012). F1 increased from 0.071±0.007 to 0.074–0.075, with RELEAP and QBC showing smoother trajectories, while diversity exhibited sharper fluctuations. TPR rose from 0.380 ± 0.026 to above 0.430 for QBC, uncertainty, and RELEAP, approaching the oracle (0.445 ± 0.036). PPV improved from 0.039 ± 0.004 to approximately 0.041 under RELEAP, QBC, and uncertainty. Calibration gains were most pronounced in females: MSE decreased from 0.175 ± 0.002 to near-oracle levels (0.165 ± 0.002) under RELEAP, QBC, and diversity, while random and uncertainty plateaued slightly higher (≈ 0.168).

Summary. In summary, while male subgroup results showed comparable convergence across strategies with limited calibration improvement, female subgroup results demonstrated clearer benefits from RELEAP. The RELEAP agent achieved balanced gains across discrimination, threshold, and calibration metrics, with the most stable convergence dynamics.

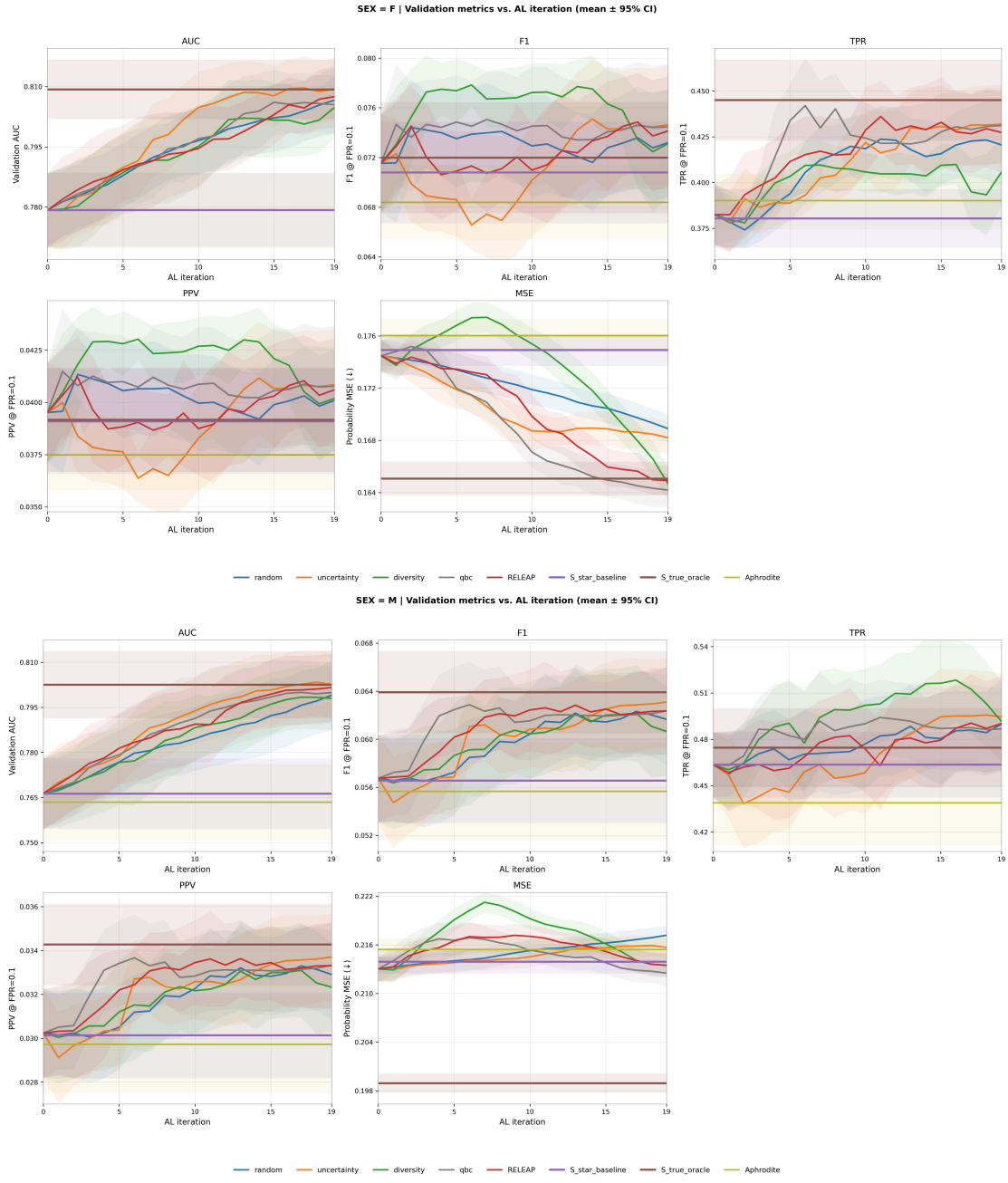

Figure 5: Validation Metrics vs. AL Iteration in the Logistic Mode (mean ± 95% CI over ten replications) within male and female subgroups.

4 DISCUSSION

In this work, we proposed RELEAP, a reinforcement learning-enhanced agent for label-efficient active phenotyping. Unlike traditional active learning approaches that rely on a fixed querying strategy, RELEAP dynamically adjusts the weighting of multiple heuristics based on downstream model performance. By integrating uncertainty, diversity, and committee-based strategies within an RL-guided policy, RELEAP improves the quality of acquired labels and enhances predictive performance under constrained labeling budgets. The use of reinforcement learning further provides natural flexibility to incorporate multi-metric rewards, enabling balanced improvements in both discrimination and calibration.

Our real-world evaluation using the DUHS EHR cohort demonstrated consistent performance gains. Across both logistic classification and survival analyses, RELEAP effectively closed the gap between noisy proxy phenotypes and automatic reference phenotype labels, achieving near-oracle discrimination while also improving threshold-level metrics and calibration. Subgroup analyses further illustrated that while all strategies performed similarly in stable settings (e.g., males), the RL-guided approach provided clear advantages in groups with greater variability (e.g., females), underscoring the value of adaptivity over fixed heuristics. These findings suggest that reinforcement learning can play an important role in guiding label acquisition, especially in clinical domains where proxies are noisy and manual labeling is costly. By leveraging downstream performance as feedback, RELEAP avoids the limitations of single-strategy active learning and demonstrates robust improvements across settings.

This study has several limitations. First, RELEAP was validated only in the context of smoking phenotype correction for lung cancer risk prediction. While smoking is a salient and clinically important phenotype, further evaluation is needed to assess generalizability across other phenotyping tasks and disease contexts. Second, the current implementation focused on EHR data from a single health system, which may limit external validity. Finally, we considered only three active learning strategies; additional heuristics or feature modalities may further enrich the policy space.

Future work should extend RELEAP to a broader set of phenotype correction tasks, including behavioral, genomic, and imaging-derived phenotypes. Beyond phenotyping, the agent could be applied to downstream applications such as risk stratification, treatment effect heterogeneity, and trial eligibility screening, where label efficiency and adaptivity are equally important. Incorporating richer reward functions that balance fairness, interpretability, and cost-effectiveness may also further enhance RELEAP's impact in real-world clinical decision support.

5 CONCLUSION

In conclusion, RELEAP provides a practical framework for adaptive, label-efficient phenotyping. By using downstream model performance as a feedback signal, RELEAP

dynamically refines phenotype labels to enhance risk prediction accuracy and model reliability in real-world EHR settings.

COMPETING INTERESTS

None declared.

DATA AVAILABILITY STATEMENT

Simulation code and data are available upon request. EHR data contain protected health information and are not able to be shared.

References

1. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* 2010; 26:1205–10
2. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat Protoc* 2019; 14:3426–44
3. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2013; 21:221–30. DOI: 10.1136/amiajnl-2013-001935. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3932460/ [Accessed on: 2025 Sep 21]
4. Liao KP, Sun J, Cai TA, et al. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. *J Am Med Inform Assoc* 2019; 26:1255–62
5. Halpern Y, Horng S, Choi Y, et al. Electronic medical record phenotyping using the anchor and learn framework. *J Am Med Inform Assoc* 2016; 23:731–40. DOI: 10.1093/jamia/ocw011. Available from: https://pubmed.ncbi.nlm.nih.gov/27107443/ [Accessed on: 2025 Sep 21]
6. Nogues IE, Wen J, Lin Y, et al. Weakly semi-supervised phenotyping using electronic health records. *J Biomed Inform* 2022; 134:104175
7. Hong C, Wen J, Zhang HG, et al. Label efficient phenotyping for Long COVID using electronic health records. *npj Digit Med* 2025; 8:405
8. Chen Y, Carroll RJ, Hinz ERM, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *J Am Med Inform Assoc* 2013; 20:e253–e259

9. Hong Y and Zeng ML. International classification of diseases (ICD). *KO Knowledge Organization* 2023; 49:496–528

10. Wiley LK, Shah A, Xu H, et al. ICD-9 tobacco use codes are effective identifiers of smoking status. *J Am Med Inform Assoc* 2013; 20:652–8. DOI: 10.1136/amiajnl-2012-001557. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3721171/ [Accessed on: 2025 Sep 21]

11. Ruckdeschel JC, Riley M, Parsatharathy S, et al. Unstructured data are superior to structured data for eliciting quantitative smoking history from the electronic health record. *JCO Clin Cancer Inform* 2023; 7:e2200155

12. Rajendran S and Topaloglu U. Extracting Smoking Status from Electronic Health Records Using NLP and Deep Learning. *AMIA Jt Summits Transl Sci Proc* 2020; 2020:507–16. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7233082/ [Accessed on: 2025 Sep 21]

13. Hong C, Liao KP, and Cai T. Semi-supervised validation of multiple surrogate outcomes with application to electronic medical records phenotyping. *Biometrics* 2019; 75:78–89

14. Settles B. Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012. DOI: 10.2200/S00429ED1V01Y201207AIM018

15. Fang M, Li Y, and Cohn T. Learning how to Active Learn: A Deep Reinforcement Learning Approach. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* 2017 :595–605. DOI: 10.18653/v1/D17-1063. Available from: https://aclanthology.org/D17-1063/ [Accessed on: 2025 Sep 21]

16. Chandran U, Reps J, Yang R, et al. Machine learning and real-world data to predict lung cancer risk in routine care. *Cancer Epidemiol Biomarkers Prev* 2023; 32:337–43

17. US Preventive Services Task Force. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* 2021; 325:962–70. DOI: 10.1001/jama.2021.1117. Available from: https://jamanetwork.com/journals/jama/fullarticle/2777243 [Accessed on: 2025 Sep 21]

18. Banda JM, Halpern Y, Sontag D, et al. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. *AMIA Jt Summits Transl Sci Proc* 2017; 2017:48–57. Available from: https://pubmed.ncbi.nlm.nih.gov/28815104/

19. King G and Zeng L. Logistic Regression in Rare Events Data. *Political Analysis* 2001; 9:137–63. DOI: 10.1093/oxfordjournals.pan.a004868. Available from: https://gking.harvard.edu/files/0s.pdf [Accessed on: 2025 Sep 21]

20. Sener O and Savarese S. Active Learning for Convolutional Neural Networks: A Core-Set Approach. *International Conference on Learning Representations (ICLR)*. 2018 :1–12. Available from: https://openreview.net/forum?id=H1aIuk-RW [Accessed on: 2025 Sep 21]

21. Freund Y, Seung HS, Shamir E, et al. Selective Sampling Using the Query by Committee Algorithm. *Machine Learning* 1997; 28:133–68. DOI: 10.1023/A:1007330508534. Available from: https://link.springer.com/article/10.1023/A:1007330508534 [Accessed on: 2025 Sep 21]

Figure Legends

Figure 1: RELEAP workflow. Patient-level inputs include clinical covariates for constructing proxy phenotypes, additional structured predictors for risk models, and a limited pool of costly reference labels. The RL-AL process adaptively selects patients to correct noisy proxy phenotypes into refined computable phenotypes, guided by a policy trained with PPO. The state vector summarizes short-term downstream performance trends, labeling budget, and strategy-level statistics. The policy outputs weights across active learning strategies (uncertainty, diversity, QBC), which drive the action of selecting patients for relabeling. Corrected phenotypes are then combined with additional structured predictors to train downstream risk prediction models. Model performance on validation data (e.g., AUC or C-index) is both logged as final results and fed back as the reward signal, closing the loop between downstream prediction and the RL-AL process.
Figure 2: Coverage of smoking-related ICD codes (presence of ≥ 1 smoking-related ICD per patient).
Figure 3: Validation metrics vs. AL iteration in the logistic mode (mean $\pm$ 95% CI over ten replications).
Figure 4: C-index vs. AL iteration in the survival (Cox) mode (mean $\pm$ 95% CI over ten replications).
Figure 5: Validation metrics vs. AL iteration in the logistic mode (mean $\pm$ 95% CI over ten replications) within male and female subgroups.

S1. Simulation Study

Simulation Setting

We evaluated the finite-sample performance of the proposed reinforcement learning–active learning (RL–AL) framework using synthetic data generated under the minimal causal structure $X_1 \to S_{\text{true}} \to S^*$ and $(S_{\text{true}}, X_2) \to Y$. The simulation setting is as follows.

- **Sample size n.** For each replicate, we generate a cohort of $n = 1000$, split into 80% training and 20% validation sets stratified by Y, under a fixed labeling budget and batch size for active learning.
- **Predictors X_1 and X_2.** Draw $X_1 \in \mathbb{R}^{d_{X1}}$ and $X_2 \in \mathbb{R}^{d_{X2}}$ independently from standard Gaussian distributions, providing phenotype-related and outcome-related signals, respectively.
- **Automatic reference phenotype S_{true}.** Generate S_{true} via a logistic link from X_1, with Gaussian perturbation applied to the linear predictor before passing through a sigmoid.
- **Proxy phenotype S^* (continuous).** Construct S^* by adding Gaussian noise to S_{true} and applying a sigmoid, mimicking a noisy ICD-based proxy label.
- **Outcome Y.** Generate the binary outcome from a logistic model with a fixed coefficient for S_{true} and random coefficients on X_2.

We compared random sampling, uncertainty, diversity, query-by-committee (QBC), and the proposed RELEAP under identical labeling budgets.

Simulation Results

Figure 1 summarizes validation metrics across active-learning iterations ($n = 100$ runs). All active-learning strategies improved over the noisy S^* baseline and converged toward the oracle S_{true} reference.

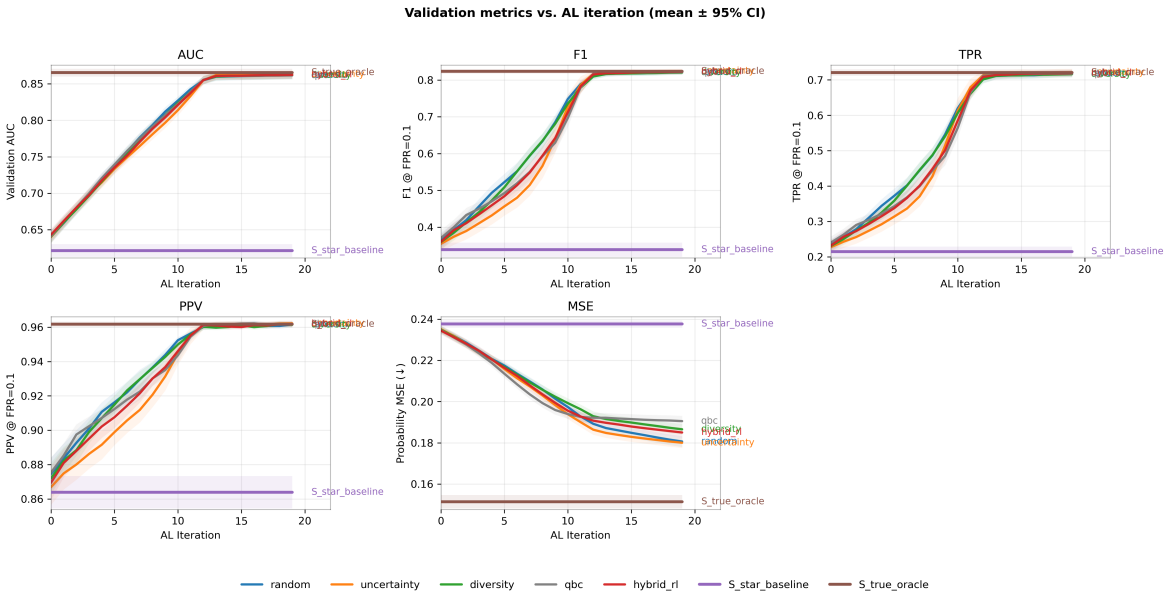

Figure 1: Simulation study: validation metrics across active-learning iterations (mean ± 95% CI, $n = 100$ runs).

S2. Active Learning Strategies

We evaluated several active learning (AL) strategies used in RELEAP for selecting the most informative unlabeled patients to query during each iteration t. Let $\mathcal{L}_t$ denote the labeled set, $\mathcal{U}_t$ the unlabeled pool, and Y the clinical outcome (e.g., lung cancer occurrence). Each sample $i \in \mathcal{U}_t$ is represented by a feature vector composed of its phenotype proxy S and structured covariates X_2.

Uncertainty

At iteration t, a logistic model f_t is trained on $\mathcal{L}_t$ to predict the probability $\hat{p}_{t,i} = f_t([S_i, X_{2,i}]) = P(Y_i = 1 \mid S_i, X_{2,i})$ for each unlabeled patient i. The uncertainty score $H(\hat{p}_{t,i})$ quantifies the model's confidence and is defined as the Shannon entropy:

$$H(\hat{p}_{t,i}) = -\hat{p}_{t,i} \log \hat{p}_{t,i} - (1 - \hat{p}_{t,i}) \log(1 - \hat{p}_{t,i}),$$

where higher values indicate greater uncertainty (i.e., probabilities closer to 0.5). Patients with the highest uncertainty are prioritized for labeling. Here, S_i denotes the current phenotype estimate, $X_{2,i}$ the structured covariate vector, and $\hat{p}_{t,i}$ the predicted event probability for patient i.

Diversity

To ensure the labeled set covers diverse regions of the feature space, we compute pairwise cosine distances between each unlabeled patient i and its $k = 10$ nearest

labeled neighbors in the standardized $[S, X_2]$ space. Let d_{ij} denote the cosine distance between patient i and labeled neighbor j. The diversity score for patient i is defined as:

$$D_i = \text{mean}(d_{ij}) + \lambda \times \text{std}(d_{ij}), \quad \lambda = 0.5,$$

where the first term encourages global coverage and the second term favors patients in underrepresented regions. Patients with the highest D_i values are selected.

Query-by-Committee (QBC)

To capture model disagreement, a committee of $M = 7$ logistic models $\{f_t^{(m)}\}_{m=1}^{M}$ is trained on bootstrap resamples of $\mathcal{L}_t$. Each model applies mild perturbations, including random variation in the L_2 regularization term and feature dropout with probability $p = 0.1$. For each patient i, the QBC score quantifies the disagreement among models:

$$Q_i = \text{Var}_m\left[\hat{p}_{t,i}^{(m)}\right],$$

where $\hat{p}_{t,i}^{(m)}$ is the predicted probability of $Y_i = 1$ from model m. A small entropy stabilizer (weight 0.1) is added to prevent degenerate variance when predictions are extreme.

Random Baseline

As a benchmark, we include a random sampling strategy that uniformly selects patients from $\mathcal{U}_t$ without using model-derived scores. This provides a lower-bound reference for evaluating the efficiency of informed querying strategies.

S3. Reinforcement Learning Formulas

The reinforcement learning (RL) component dynamically adjusts the weighting of these strategies. At iteration t, the system observes a state vector s_t, selects an action $\mathbf{w}_t$, and receives a reward r_t based on the downstream model performance.

State

The state s_t aggregates information summarizing both model performance and data composition:

$$s_t = [m_t, \mathbf{z}_t^{\text{strat}}, \mu_t^{(S)}, \sigma_t^{(S)}, \text{slope}(m), \text{var}(m), b_t],$$

where - m_t is the current downstream evaluation metric (AUC for classification or C-index for survival); - $\mathbf{z}_t^{\text{strat}}$ summarizes the per-strategy score distributions on $\mathcal{U}_t$ (median and 80th percentile for each strategy); - $\mu_t^{(S)}$ and $\sigma_t^{(S)}$ are the mean and standard deviation of phenotype values among labeled samples; - slope(m) and var(m) denote the short-term trend and variability of recent metric trajectories; - b_t is the fraction of labeling budget remaining at iteration t.

Action

The policy network outputs a normalized weight vector $\mathbf{w}_t$ assigning importance to each active learning strategy:

$$\mathbf{w}_t = (w_t^{\text{unc}}, w_t^{\text{div}}, w_t^{\text{qbc}}), \quad w_i \geq 0, \ \sum_i w_i = 1.$$

This vector linearly combines normalized scores from the uncertainty, diversity, and QBC strategies to determine which patients to query from $\mathcal{U}_t$.

Reward

The reward r_t measures the improvement in downstream model performance. Let m_t denote the active metric (AUC or C-index) and $\bar{m}_t$ its moving average over the most recent h iterations. The relative performance gain is defined as:

$$g_t = \frac{m_t - \bar{m}_t}{1 - \bar{m}_t + \varepsilon},$$

where ε is a small constant for numerical stability. This gain is then scaled by two shaping factors: (1) a progress term $(1 + 2\,\text{prog}_t)$, where prog_t is the fraction of budget already used, and (2) a trajectory term $\tau_t \in \{1.0, 1.2\}$ that upweights improving trends. The shaped reward is:

$$R_t^{\text{raw}} = g_t(1 + 2\,\text{prog}_t)\tau_t,$$

which is normalized online:

$$r_t = \frac{R_t^{\text{raw}} - \mu_t}{\sigma_t + \varepsilon},$$

where μ_t and σ_t are the running mean and standard deviation of rewards.

RL Algorithm

We optimize the policy using Proximal Policy Optimization (PPO), which learns a stochastic mapping $\pi_\theta(a \mid s)$ from states to actions by maximizing the expected discounted return:

$$\max_\theta \ \mathbb{E}\Big[\sum_{t=1}^{T} \gamma^{t-1} R_t\Big],$$

where $\gamma \in (0, 1)$ is the discount factor. Experience tuples (s_t, a_t, R_t, s_{t+1}) are collected during active learning and used to update both the policy (actor) and value (critic) networks periodically.

S4. Mathematical Formulation

Let $\mathcal{A} = \{\text{uncertainty}, \text{diversity}, \text{QBC}\}$ denote the candidate strategy set. At iteration t, the agent observes state

$$s_t = [\mathbf{z}_{\text{strat}}(t), p_+^{\text{lab}}(t), \text{slope}(m), \text{var}(m), b_t],$$

where $p_+^{\text{lab}}(t)$ is the proportion of positive samples in the labeled set. The action $\mathbf{w}_t \in \Delta^2$ represents the simplex weight vector over strategies. The shaped cumulative reward is:

$$R_t = \alpha(m_t - m_{t-1}) + (1-\alpha)\sum_{k=1}^{K} \gamma^{k-1}(m_{t+k-1} - m_{t+k-2}),$$

where $\alpha \in [0, 1]$ controls the balance between immediate and long-term performance gains, and γ is the temporal discount factor.

S5. Automatic Reference Phenotype Variants

We compared several alternative definitions of the automatic reference phenotype (S_{true}), which serves as the refined phenotype label used to correct the noisy proxy (S^*). Each variant reflects a different combination of structured self-reported smoking information, traditional natural language processing (NLP)–derived concept identifiers, and large language model (LLM)–extracted mentions from clinical notes:

- **Self-report + NLP CUIs:** Combines structured self-reported smoking status with Concept Unique Identifiers(CUIs) extracted from clinical notes using traditional NLP pipelines such as rule-based or dictionary-based methods.
- **Self-report + LLM mentions:** Combines structured self-reported smoking status with smoking-related mentions extracted from clinical notes using a large language model (LLM).
- **NLP CUIs only:** Uses concept identifiers extracted by traditional NLP methods without incorporating any structured self-reported information.
- **LLM mentions only:** Uses smoking-related mentions identified by a large language model without relying on structured self-reported data.

Figure 2 summarizes validation performance across discrimination (area under the receiver operating characteristic curve, **AUC**), threshold-based metrics (F1 score, true positive rate [**TPR**], and positive predictive value [**PPV**]), and calibration (mean squared error of predicted probabilities, **MSE**).

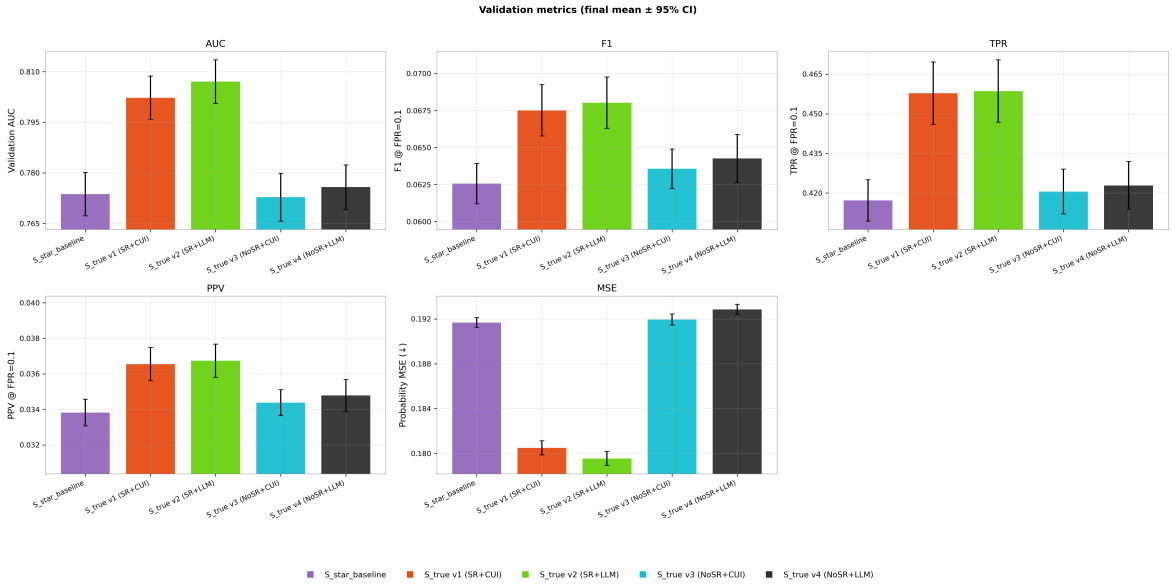

Figure 2: Comparison of Automatic Reference Phenotype variants (S_{true} definitions) against the noisy proxy baseline (S^*). Bars show mean validation performance (AUC, F1 score, TPR, PPV, MSE) with 95% CI error bars.

Across all evaluation metrics, the definitions that integrated structured self-reported information with either NLP- or LLM-derived text features achieved the highest predictive performance relative to the noisy proxy phenotype (S^*). By contrast, definitions relying solely on unstructured text—whether extracted by traditional NLP or LLMs—showed reduced accuracy and calibration.