Active Learning as Decision Support for EHR Cohort Construction

Anonymous Authors[1]

Abstract

Building disease diagnosis models from electronic health records (EHRs) requires deciding which patients to include in the training cohort, a decision that shapes predictive accuracy, fairness, and clinical utility. We formulate cohort construction as a sequential decision problem and evaluate unsupervised active learning (AL) strategies as patient selection policies. We benchmark five canonical AL strategies for cold-start EHR disease diagnosis using MOTOR foundation model embeddings, across 21 disease tasks on MIMIC-IV with three classifiers (logistic regression, MLP, XGBoost). Entropy sampling is the only strategy that consistently outperforms random across all classifiers, with gains largest for low-prevalence diseases. Diversity-based strategies perform similarly or worse than random because they systematically under-enroll positive-class patients in imbalanced cohorts, regardless of class separability in the embedding space.

1. Introduction

Building a disease diagnosis model from EHRs requires deciding which patients to include in the training cohort. In prospective biomedical studies, this decision faces hard constraints of limited budget, clinical capacity, and rare conditions that require deliberate over-sampling. Cohorts skewed toward majority demographics produce models that fail silently on under-represented groups (Obermeyer et al., 2019; Chen et al., 2021). Data quality has been found to dominate over data quantity (Zha et al., 2023; Sorscher et al., 2022).

AL formalizes this as a sequential selection problem where a strategy selects patients under a limited enrollment budget so that the resulting cohort is maximally informative. Unlike post-hoc data pruning (Sorscher et al., 2022; Mirzasoleiman et al., 2020), AL operates before data are collected, making it directly applicable to prospective study design. Existing AL benchmarks cover single tasks or models (Sener & Savarese, 2018; Ash et al., 2020), focus on imaging or NLP (Gal et al., 2017), or use balanced benchmarks that do not reflect clinical class-imbalance. To our knowledge, no prior study has quantified how AL strategy choice affects prospective patient selection in EHR settings.

Our contributions are threefold. (1) A reproducible benchmark across 21 tasks (prevalence 2.7%–82.9%), five AL strategies, three classifiers, and three clinical metrics (AUROC, normalized AUPRC, Brier score). (2) Establishing Entropy sampling as the only strategy that consistently outperforms random, with gains largest for low-prevalence diseases. (3) Identification of the mechanism whereby diversity strategies systematically under-enroll positive cases, while entropy targets the decision boundary regardless of class separability.

2. Related Work

2.1. Data Quality over Quantity

Training set composition, rather than model capacity, is the primary bottleneck in many practical regimes (Zha et al., 2023). Data pruning can improve generalization by removing redundant or harmful examples (Sorscher et al., 2022; Mirzasoleiman et al., 2020), but requires access to the full dataset and is not applicable to prospective cohort construction. Active selection of informative examples is more suitable. Mindermann et al. (2022) show that focusing on points that are simultaneously learnable, worth learning, and not yet learnt achieves 18× fewer training steps. SemDeDup uses pre-trained embeddings to remove semantically redundant pairs, halving dataset size without degrading performance (Abbas et al., 2023), and similar informativeness criteria cut data budgets by 70% (Agarwal et al., 2025) while matching full-data performance with 13× fewer iterations (Lin

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

et al., 2024; Evans et al., 2024).

2.2. Active Learning

Active learning (AL) formalizes sequential data selection under a labeling budget (Settles, 2009). The literature identifies three principal strategy families, uncertainty sampling, query-by-committee, and information-theoretic methods. Deep Bayesian active learning (Gal et al., 2017) uses Monte Carlo dropout to estimate predictive uncertainty. Coreset (Sener & Savarese, 2018) minimizes the maximum distance from the current labeled set. BADGE (Ash et al., 2020) combines uncertainty and diversity via gradient embeddings; BatchBALD (Kirsch et al., 2019) maximizes mutual information under Bayesian networks. Diverse mini-batch AL (Zhdanov, 2019) combines both signals via K-means partitioning. Hacohen et al. (2022) establish a phase-transition theory where typical examples are best at low budgets and uncertain examples at large budgets. TypiClust, introduced by the same authors, is designed for the low-budget regime. SelectAL (Hacohen & Weinshall, 2023) dynamically switches between typicality and uncertainty strategies. Class imbalance, the norm in clinical EHR data, is a known source of AL instability (Mindermann et al., 2022), and none of these benchmarks evaluate strategies in imbalanced prospective cohort settings.

2.3. Foundation Models and Data Efficiency

Pre-trained representations substantially reduce the task-specific data needed for clinical prediction. The EHRSHOT benchmark (Wornow et al., 2023) shows that a model pre-trained on 2.57M patient records generalizes to 42 unseen clinical tasks with only a handful of labeled examples. EHR foundation models span health-system language models (Jiang et al., 2023), generalist medical AI (Moor et al., 2023), and structured-record encoders (Rasmy et al., 2021). MOTOR (Steinberg et al., 2021) encodes a patient's full structured history into a fixed-length embedding capturing population-level clinical structure, and remains the only open-source EHR foundation model to our knowledge. Tamkin et al. (2022) show that pre-trained models require up to 5× fewer labels when combined with uncertainty-based AL; Vysogorets & Gopal (2024) demonstrate that frozen pre-trained features are sufficient for efficient AL in text classification. Lu et al. (2025) confirm that uncertainty sampling retains its edge over diversity strategies on tabular datasets when model and query strategy are compatible.

3. Benchmark

3.1. Data and Tasks

We use MIMIC-IV (Johnson et al., 2023) and define 21 binary disease diagnosis tasks spanning cancers, cardiometabolic diseases, and acute conditions (Appendix A), covering a prevalence range from celiac disease (2.7%) to hypertension (82.9%).

3.2. Patient Representation

This work builds upon existing works showing that pre-trained representations substantially reduce annotation cost in AL (Tamkin et al., 2022; Vysogorets & Gopal, 2024). Each patient is represented by a 768-dimensional embedding from MOTOR (Steinberg et al., 2021), a time-to-event foundation model pre-trained on large-scale longitudinal EHR data. MOTOR is task-agnostic and not trained on any of our 21 tasks. Given structured records up to a fixed prediction time t^*, it encodes the full history of diagnoses, procedures, medications, and laboratory events. Each task is a binary classification at admission: the label is whether a patient has the disease at admission t^*. For cases, t^* is the date of first qualifying diagnosis admission; controls are matched patients admitted without the qualifying diagnosis. Both cases and controls are required to have at least one prior admission before t^*, ensuring MOTOR encodes a meaningful clinical history for every patient. We simulate cold-start enrollment. At each round, the AL strategy selects B=2 patients from the unlabeled pool $\mathcal{U}$, their labels are revealed, and the classifier is retrained from scratch on the updated cohort $\mathcal{L}$. We initialize with $|\mathcal{L}_0|$=20 stratified patients and run 500 rounds (up to 1 020 patients total), with five-fold cross-validation. We evaluate AUROC, AUPRC normalized by prevalence (AUPRC_n), and Brier score (Brier, 1950) at every round and summarize these metrics across rounds using the Area Under the Learning Curve (AULC) as the primary summary statistic.

3.3. Query Strategies

Besides the random baseline, we evaluate four strategies spanning two families: model-free strategies (Coreset, K-Means, TypiClust), which select based solely on embedding geometry and require no labeled data, and a model-based strategy (Entropy), which score candidates using classifier predictions. To ensure strategy comparability across classifiers, gradient-based methods such as BADGE (Ash et al., 2020), which require differentiable representations, were excluded. Random is our primary baseline. Entropy queries $i^* =$

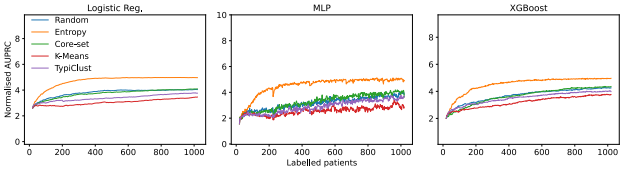

Figure 1. Normalized AUPRC learning curves across 21 tasks (MOTOR). Entropy diverges from random from round one; K-Means and TypiClust fall below random. Shaded bands: ±1 std across tasks.

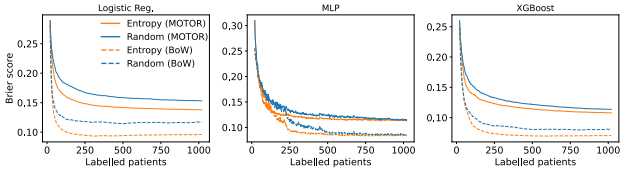

Figure 2. Brier score learning curves, entropy vs. random on MOTOR (solid) and BoW (dashed). Entropy improves calibration under both feature sets; the BoW improvement is at least as large as the MOTOR one.

$\arg\max_{i \in \mathcal{U}} H(\hat{p}_i)$, where H is the binary predictive entropy. Coreset (Sener & Savarese, 2018) applies greedy k-center in embedding space. K-Means (Zhdanov, 2019) enrolls the nearest patient to each cluster centroid. TypiClust (Hacohen et al., 2022) selects high-density, typical examples. We evaluate logistic regression (L2), MLP (two hidden layers), and XGBoost (Chen & Guestrin, 2016). Full hyperparameters are in Appendix B.

4. Results

Table 1 reports AULC across all metrics, strategies, and classifiers. Across metrics and classifiers, entropy emerges as the only strategy that significantly outperforms random while all diversity-based strategies are neutral or harmful. These observations corroborate prior evidence that uncertainty sampling retains its edge over diversity strategies on tabular data (Lu et al., 2025).

AUPRC directly rewards rare-class ranking. Figure 1 shows entropy diverging from random from round one across all classifiers, with gains approximately 4× larger for low-prevalence tasks. Diversity-based strategies perform poorly: K-Means underperforms random with deficits exceeding 40% relative AUPRC AULC for rare diseases, while Coreset and TypiClust track random. AUROC confirms the same ranking, with entropy gains $\Delta = +0.009$ to $+0.035$. Coreset's neutral AUROC masks a Brier degradation ($\Delta = +0.009$ for logistic regression); full AUROC and Brier curves are in Figures 8–9 and 7.

At the terminal state (Figure 5), entropy exceeds random on 20/21 tasks (mean AUPRC 4.94 vs. 4.05), with the largest advantages for the rarest diseases. The effect holds across the full prevalence range, and is consistent across cohort sizes: entropy's advantage over random is preserved at both small ($n \leq 100$) and large ($n \geq 700$) budgets (Figure 10). Per-task heatmaps and AULC breakdowns are in Figures 11 and Table 3.

5. Ablation

To understand the drivers of this performance gap, we examine three potential mechanisms for the entropy advantage: (1) pre-trained embedding geometry, (2) class separability in the embedding space, and (3) positive-class enrichment.

5.1. Input Representation Effects

Given the performance of entropy sampling on MOTOR embeddings, we investigate whether this advantage is inherent to the strategy or a product of the representation space. Replacing MOTOR with raw bag-of-words concept-count features, we compare the AULC gain of entropy over random (Δ) under both metrics (Figure 6). Considering the AUROC_n, Entropy benefits are substantially reduced on BoW, especially for logistic regression and MLP. For XGBoost, entropy leads to a larger AUPRC_n than using MOTOR embeddings, with $\Delta = +1.47$ over random sampling. However, in terms of AUROC, entropy falls below random for logistic regression ($\Delta = -0.036$) and MLP ($\Delta = -0.014$) on BoW but retains a modest advantage ($\Delta = +0.012$) with XGBoost. Figure 2 shows the Brier score AULC gain of entropy over random under both feature sets. Entropy improves calibration regardless of feature type. Brier gains are $\Delta = -0.017$, -0.007, -0.008 on MOTOR and $\Delta = -0.021$, -0.008, -0.011 on BoW for logistic regression, MLP, and XGBoost respectively.

This explains why entropy leads to higher normalized AUPRC. Better-calibrated probability estimates still reward positive-case recovery, which AUPRC captures. Entropy selects patients that are uncertain under BoW features, but such patients are not necessarily near the true disease boundary, so the global AUROC of all patients does not improve. MOTOR's embedding geometry is what aligns feature-space uncertainty with true clinical uncertainty, making entropy's selections informative for both calibration and discrimination.

Table 1. AULC summary (MOTOR embeddings, 21 tasks, 5 folds). Mean ± std. AUPRC_n: AUPRC normalized by prevalence (>1 = above-chance positive-class ranking). Bold = best. ↓ Brier: lower is better.

	Logistic Reg.			MLP			XGBoost		
Strategy	AUROC	AUPRC_n	Brier↓	AUROC	AUPRC_n	Brier↓	AUROC	AUPRC_n	Brier↓
Random	0.764 ±0.064	3.810 ±3.320	0.165 ±0.086	0.747 ±0.062	3.171 ±2.221	0.131 ±0.065	0.770 ±0.061	3.707 ±2.848	0.128 ±0.071
Entropy	0.794 ±0.063	4.724 ±4.163	0.147 ±0.096	0.783 ±0.062	4.550 ±3.759	0.124 ±0.074	0.779 ±0.059	4.547 ±3.815	0.120 ±0.074
Coreset	0.765 ±0.060	3.740 ±3.203	0.173 ±0.076	0.746 ±0.057	3.280 ±2.407	0.134 ±0.064	0.769 ±0.057	3.681 ±2.906	0.135 ±0.069
K-Means	0.730 ±0.069	3.059 ±2.530	0.221 ±0.077	0.724 ±0.054	2.555 ±1.426	0.145 ±0.068	0.746 ±0.060	3.213 ±2.247	0.141 ±0.067
TypiClust	0.745 ±0.064	3.400 ±2.872	0.182 ±0.086	0.736 ±0.061	2.892 ±1.863	0.137 ±0.066	0.762 ±0.060	3.516 ±2.607	0.132 ±0.071

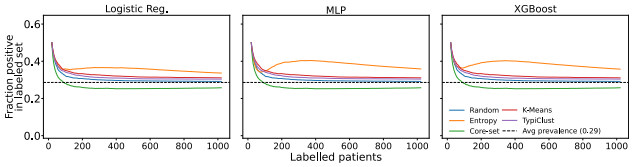

Figure 3. Positive-class enrichment ratio per strategy (21 tasks, 5 folds). Dashed line: background prevalence. Diversity strategies remain at or below prevalence; entropy exceeds it from round one.

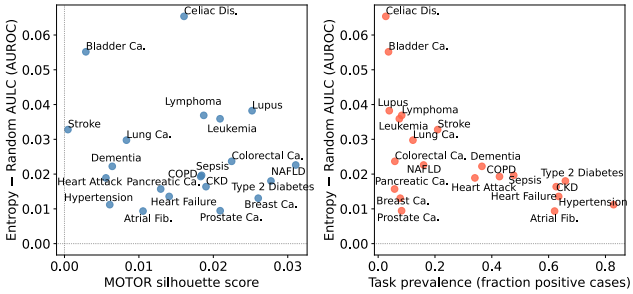

Figure 4. Left: MOTOR silhouette score (class separability) vs. entropy AULC gain over random, per task. No correlation ($r \approx 0$, $p > 0.3$): entropy works regardless of embedding geometry. Right: task prevalence vs. entropy gain; lower prevalence amplifies the advantage.

5.2. Diversity Strategies Under-Enroll Positive Cases

While the representation space is a necessary component, it does not fully explain why diversity-based strategies consistently underperform. We therefore investigate the fraction of positive-class patients selected per round relative to background prevalence (enrichment = 1.0 means proportional) (Figure 3). K-Means, Coreset, and TypiClust maintain enrichment below 1.0 throughout, systematically starving the classifier of rare-event signal. Entropy, however, achieves enrichment at or above 1.0 from round one because it selects boundary samples that are concentrated near the region where positive and negative patients interleave. This effect is strongest for low-prevalence tasks ($\pi < 10\%$), exactly where the AUPRC gap is largest, suggesting that positive-class enrichment is a key mechanism for the entropy advantage.

5.3. Entropy Is Robust to Embedding Separability

Finally, we explore whether class separability in the embedding space modulates strategy performance. For this, we compute the silhouette score of the MOTOR embeddings per task, which ranges from 0.01 (least separable) to 0.11 (most separable).We found no correlation between class separability and entropy AULC gain ($r \approx 0$, $p > 0.3$; Figure 4, left). Instead, gains are driven by prevalence: low-prevalence tasks see the largest advantages (Figure 4, right). This is consistent with the positive-case enrichment mechanism: entropy targets the decision boundary regardless of embedding geometry, while diversity strategies systematically under-enroll positive cases in imbalanced cohorts.

6. Discussion

When labels are not available, entropy sampling with MOTOR embeddings is the clear choice for prospective EHR cohort construction, consistently and significantly outperforming random across all classifiers, metrics, and 21 disease tasks. Diversity methods fail systematically in imbalanced clinical data due to positive-case under-enrollment. This work showcases the critical importance of patient recruitment strategy in biomedical studies, and the value of AL as a decision-support tool for cohort construction.

Impact Statement

This paper advances machine learning for clinical cohort design. Entropy-based active learning enriches rare-disease cases in training data, with potential fairness benefits. We foresee no harmful consequences beyond the general risk of over-reliance on automated patient selection without clinical oversight.

References

Abbas, A., Tirumala, K., Simig, D., Ganguli, S., and Morcos, A. S. Semdedup: Data-efficient learning at web-scale through semantic deduplication. arXiv preprint arXiv:2303.09540, 2023.

Agarwal, I., Killamsetty, K., Popa, L., and Danilevsky, M. DELIFT: Data efficient language model instruction fine-tuning. In International Conference on Learning Representations (ICLR), 2025.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. In International Conference on Learning Representations (ICLR), 2020.

Brier, G. W. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78 (1):1–3, 1950.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. Ethical machine learning in health care. Annual Review of Biomedical Data Science, 4:123–144, 2021.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 785–794, 2016.

Evans, T., Parthasarathy, N., Merzic, H., and Hénaff, O. J. Data curation via joint example selection further accelerates multimodal learning. In Advances in Neural Information Processing Systems (NeurIPS), 2024.

Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data. In International Conference on Machine Learning (ICML), volume 70 of Proceedings of Machine Learning Research, pp. 1183–1192, 2017.

Hacohen, G. and Weinshall, D. How to select which active learning strategy is best suited for your specific problem and budget. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

Hacohen, G., Dekel, A., and Weinshall, D. Active learning on a budget: Opposite strategies suit high and low budgets. In International Conference on Machine Learning (ICML), volume 162 of Proceedings of Machine Learning Research, pp. 8175–8195, 2022.

Jiang, L. Y., Liu, X. C., Nejatian, N. P., et al. Health system-scale language models are all-purpose prediction engines. Nature, 619:357–362, 2023.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-W. H., Celi, L. A., and Mark, R. G. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10:1, 2023.

Kirsch, A., van Amersfoort, J., and Gal, Y. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

Lin, Z., Gou, Z., Gong, Y., Liu, X., Shen, Y., Xu, R., Lin, C., Yang, Y., Jiao, J., Duan, N., and Chen, W. Rho-1: Not all tokens are what you need. In Advances in Neural Information Processing Systems (NeurIPS), 2024.

Lu, P.-Y., Cheng, Y.-J., Li, C.-L., and Lin, H.-T. An expanded benchmark that rediscovers and affirms the edge of uncertainty sampling for active learning in tabular datasets. Transactions on Machine Learning Research (TMLR), 2025.

Mindermann, S., Brauner, J. M., Razzak, M. T., Sharma, M., Kirsch, A., Xu, W., Höltgen, B., Gomez, A. N., Morisot, A., Farquhar, S., and Gal, Y. Prioritized training on points that are learnable, worth learning, and not yet learnt. In International Conference on Machine Learning (ICML), volume 162 of Proceedings of Machine Learning Research, pp. 15630–15649, 2022.

Mirzasoleiman, B., Bilmes, J., and Leskovec, J. Coresets for data-efficient neural network training. In International Conference on Machine Learning (ICML), 2020.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., and Rajpurkar, P. Foundation models for generalist medical artificial intelligence. Nature, 616:259–265, 2023.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366:447–453, 2019.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine, 4:86, 2021.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In International Conference on Learning Representations (ICLR), 2018.

Settles, B. Active learning literature survey. Computer Science Technical Reports, 1648, 2009.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. S. Beyond neural scaling laws: Beating power law scaling via data pruning. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Steinberg, E., Jung, K., Shah, N. H., and Fries, J. A. Language models are an effective representation learning technique for electronic health record data. Journal of Biomedical Informatics, 113:103637, 2021.

Tamkin, A., Nguyen, D., Kadakia, S., Lindsey, J., Liang, P., Goodman, N., and Steinhardt, J. Active learning helps pretrained models learn the intended task. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Vysogorets, A. and Gopal, A. Towards efficient active learning in NLP via pretrained representations. arXiv preprint arXiv:2402.15613, 2024.

Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M. A., Fries, J., and Shah, N. H. The shaky foundations of large language models and foundation models for electronic health records. npj Digital Medicine, 6:135, 2023.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. Data-centric artificial intelligence: A survey. arXiv preprint arXiv:2303.10158, 2023.

Zhdanov, F. Diverse mini-batch active learning. arXiv preprint arXiv:1901.05954, 2019.

A. Disease Tasks and Prevalences

Table 2. MIMIC-IV disease prediction tasks and mean cohort prevalence.

Disease task	Prevalence
Coeliac disease	0.027
Bladder cancer	0.036
Lupus	0.039
Pancreatic cancer	0.058
Colorectal cancer	0.059
Leukemia	0.074
Breast cancer	0.078
Prostate cancer	0.083
Lymphoma	0.083
Lung cancer	0.122
NAFLD	0.160
Stroke	0.210
Heart attack	0.340
Dementia	0.365
COPD	0.427
Sepsis	0.477
Atrial fibrillation	0.621
Chronic kidney disease	0.627
Heart failure	0.635
Type 2 diabetes	0.659
Hypertension	0.829

B. Implementation Details

Logistic regression uses L2 regularization with $C = 1.0$ (sklearn default). The MLP has two hidden layers of width 128 with ReLU activations, trained with Adam ($\text{lr} = 10^{-3}$, max 200 epochs, early stopping on validation loss with patience 10). XGBoost uses 100 trees, max depth 4, learning rate 0.1, subsample 0.8. All classifiers are retrained from scratch on the updated labeled set at every AL round; no warm-starting is used.

The seed set of 20 patients is drawn with stratified sampling (at least 1 positive, at least 1 negative per fold) to ensure a well-defined binary classifier from round 0. All five folds share the same pool split but different seed draws.

BoW features count per-patient occurrences of condition, drug, measurement, and procedure OMOP concepts appearing strictly before the prediction time t^*. Concepts appearing in fewer than 5 patients in the training split are discarded; the resulting sparse count matrix is ℓ_2-normalized per patient. Feature dimensionality ranges from ∼3,000 to ∼15,000 depending on the task.

C. Per-Disease AULC Results

Table 3 reports AULC (AUROC) for each of the 21 disease tasks, averaged over classifiers and folds.

Table 3. AULC (AUROC) per disease task (MOTOR embeddings), averaged over 3 classifiers and 5 folds. Bold = best strategy per task.

Task	Random	Entropy	Coreset	K-Means	TypiClust
Atrial Fib.	0.687	0.697	0.685	0.648	0.673
Bladder Cancer	0.703	0.759	0.720	0.678	0.688
Breast Cancer	0.830	0.843	0.821	0.789	0.798
Celiac Disease	0.743	0.809	0.781	0.753	0.724
Chr. Kidney Dis.	0.740	0.756	0.738	0.708	0.725
Colorectal Cancer	0.813	0.837	0.811	0.766	0.796
COPD	0.782	0.801	0.780	0.766	0.767
Dementia	0.722	0.745	0.717	0.717	0.721
Heart Attack	0.703	0.722	0.702	0.661	0.693
Heart Failure	0.705	0.718	0.704	0.685	0.693
Hypertension	0.702	0.713	0.710	0.672	0.683
Leukemia	0.776	0.812	0.764	0.753	0.765
Lung Cancer	0.785	0.815	0.771	0.741	0.776
Lupus	0.882	0.920	0.883	0.847	0.861
Lymphoma	0.744	0.781	0.745	0.724	0.736
NAFLD	0.816	0.839	0.818	0.789	0.812
Pancreatic Cancer	0.796	0.812	0.769	0.759	0.785
Prostate Cancer	0.815	0.824	0.816	0.801	0.811
Sepsis	0.758	0.778	0.753	0.735	0.738
Stroke	0.639	0.672	0.657	0.621	0.638
Type 2 Diabetes	0.824	0.842	0.819	0.792	0.815

D. Supplementary Figures

D.1. Terminal AUPRC per Task

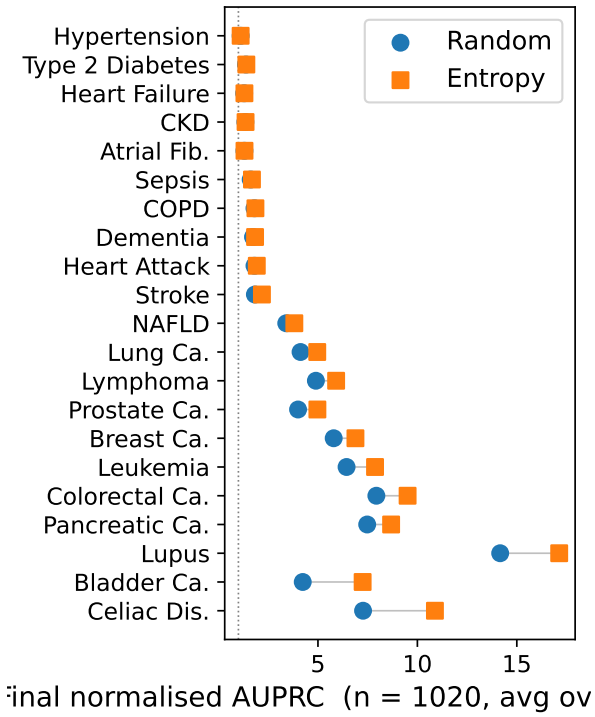

Figure 5. Terminal normalized AUPRC per task, sorted by prevalence (rarest top). Entropy (squares) leads random (circles) on 20/21 tasks.

D.2. BoW vs. MOTOR Representation

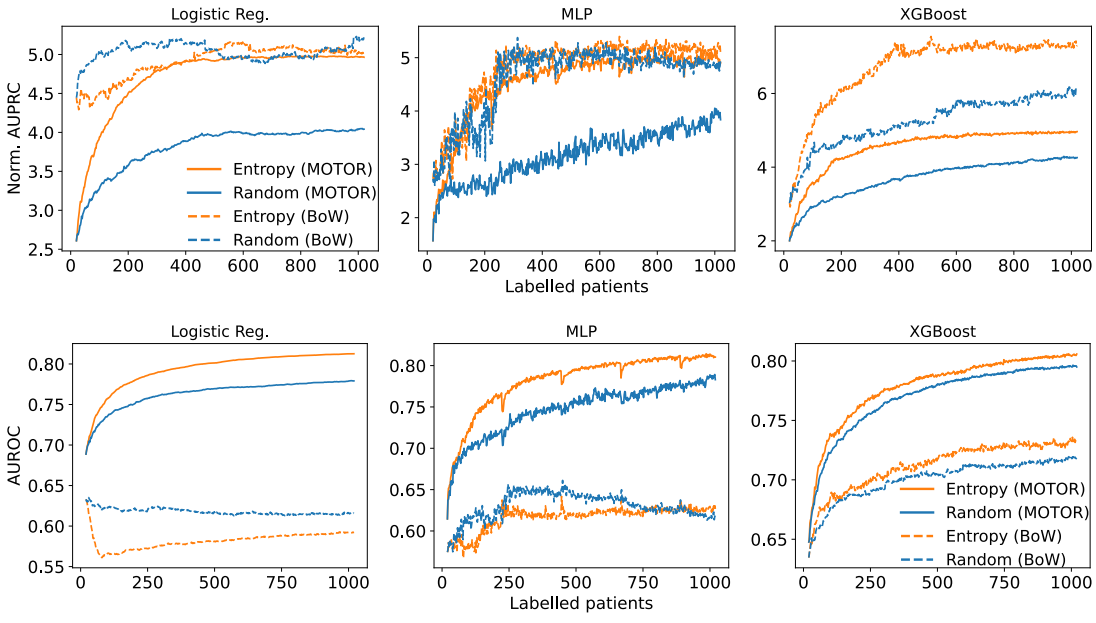

Figure 6. Entropy vs. random on MOTOR (solid) and BoW (dashed). Top: normalized AUPRC. Bottom: AUROC. On BoW, entropy loses for logistic regression and MLP on both metrics; XGBoost retains an advantage.

D.3. AULC Distributions and Learning Curves

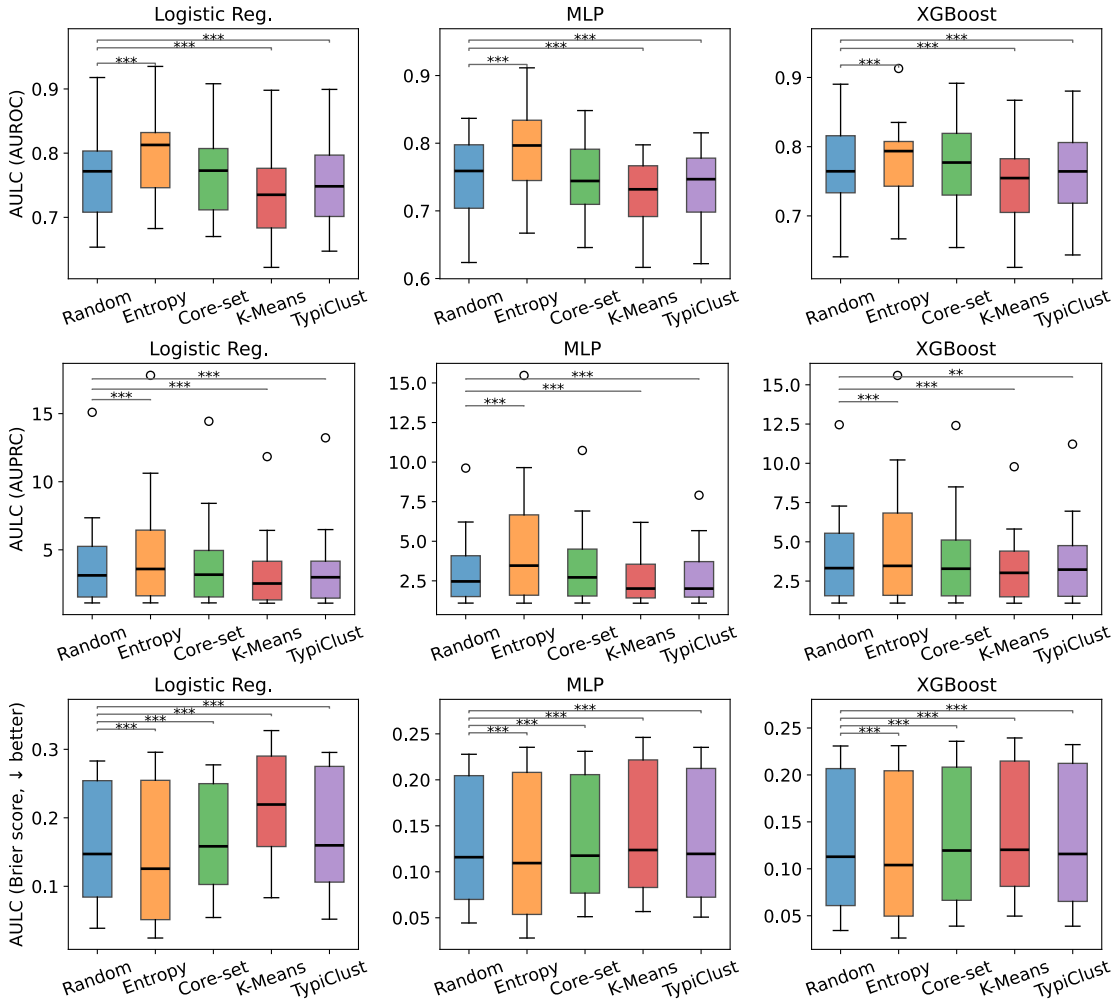

Figure 7. AULC distribution across 21 MIMIC-IV disease tasks (MOTOR embeddings), for AUROC (top), $AUPRC_n$ (middle), and Brier (bottom). Significance brackets: Wilcoxon tests vs. Random, Bonferroni-corrected.

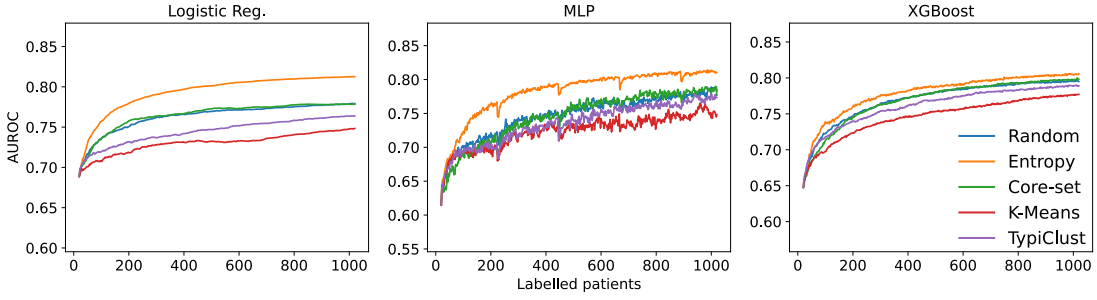

Figure 8. Mean AUROC learning curves across 21 tasks. Shaded bands: ±1 std.

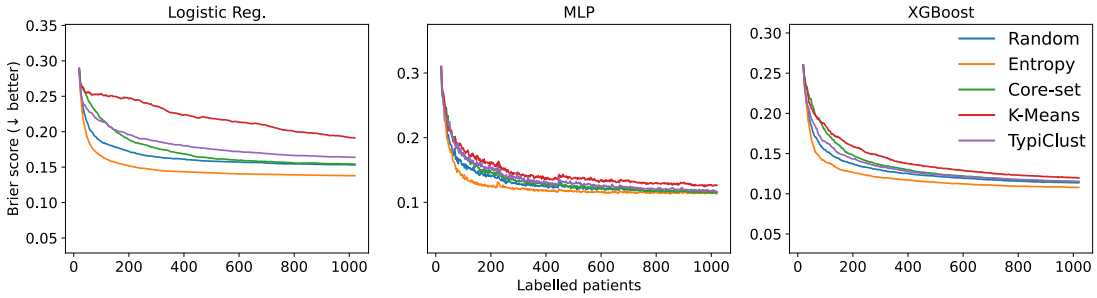

Figure 9. Mean Brier score learning curves (lower = better). Shaded bands: ±1 std.

D.4. Cohort Size Comparison

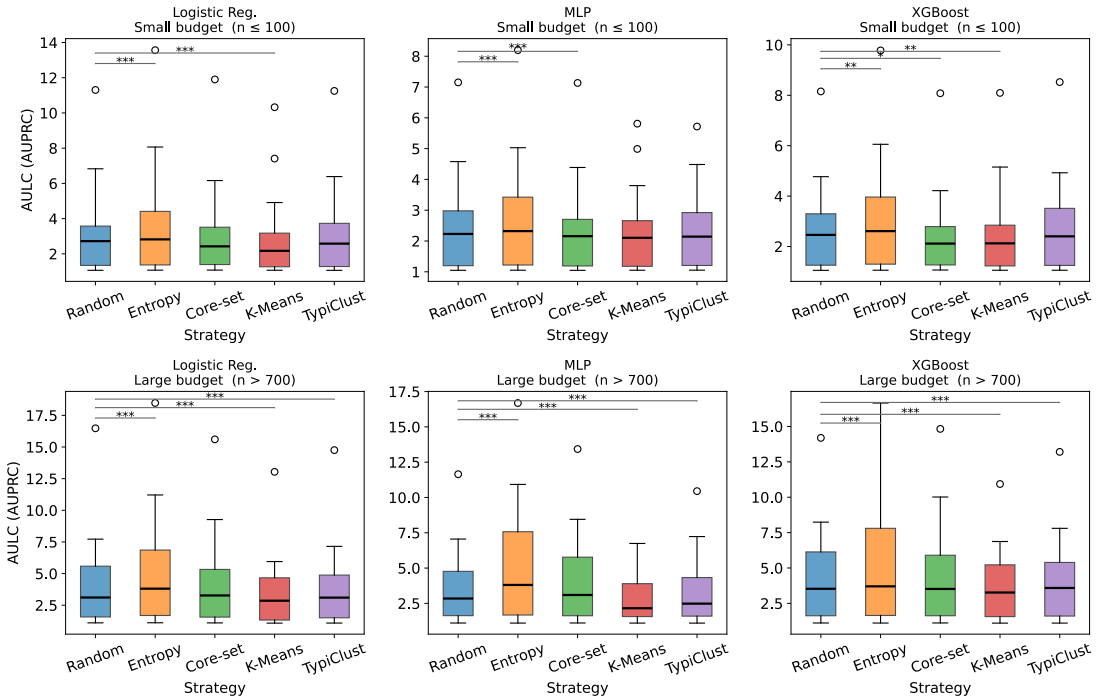

Figure 10. AULC (AUPRC) at small ($n \leq 100$) and large ($n \geq 700$) cohort sizes (MOTOR embeddings, 21 tasks). Entropy leads random at both budgets; model-free strategies do not recover at larger cohorts.

D.5. Per-Task Heatmaps

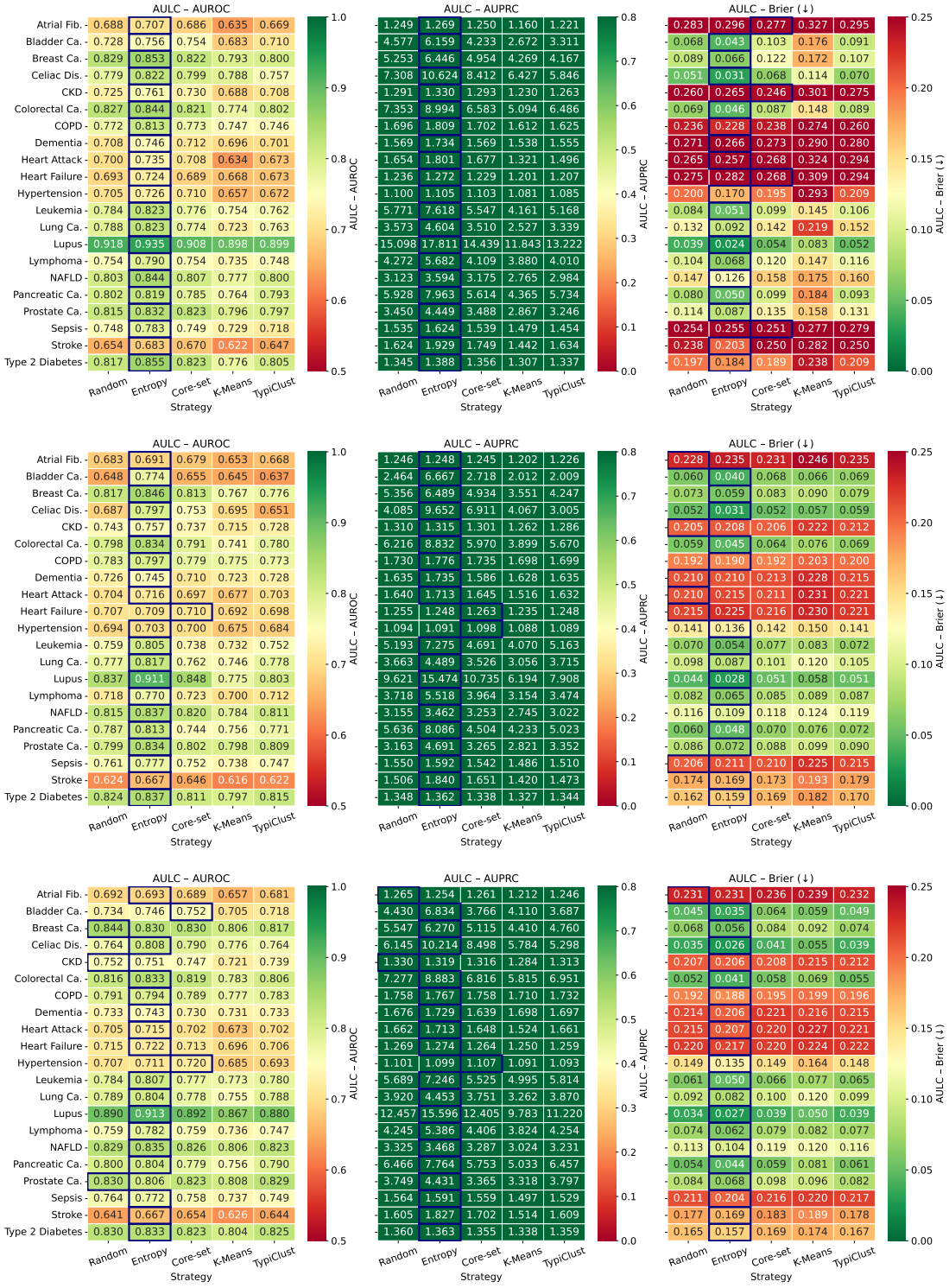

Figure 11. Per-task AULC (AUROC, AUPRC, Brier) heatmaps. Navy border: best strategy per task. Top: logistic regression. Middle: MLP. Bottom: XGBoost.