FoMoH: A clinically meaningful foundation model evaluation for structured electronic health records

Chao Pang* **Vincent Jeanselme*** **Young Sang Choi** **Xinzhuo Jiang**

Zilin Jing **Aparajita Kashyap** **Yuta Kobayashi** **Yanwei Li** **Florent Pollet**

Karthik Natarajan **Shalmali Joshi**

Department of Biomedical Informatics
Columbia University

Abstract

Foundation models hold significant promise in healthcare, given their capacity to extract meaningful representations independent of downstream tasks. This property has enabled state-of-the-art performance across several clinical applications trained on structured electronic health record (EHR) data, even in settings with limited labeled data, a prevalent challenge in healthcare. However, there is little consensus on these models' potential for clinical utility due to the lack of desiderata of comprehensive and meaningful tasks and sufficiently diverse evaluations to characterize the benefit over conventional supervised learning. To address this gap, we propose a suite of clinically meaningful tasks spanning patient outcomes, early prediction of acute and chronic conditions, including desiderata for robust evaluations. We evaluate state-of-the-art foundation models on EHR data consisting of 5 million patients from Columbia University Irving Medical Center (CUMC), a large urban academic medical center in New York City, across 14 clinically relevant tasks. We measure overall accuracy, calibration, and subpopulation performance to surface tradeoffs based on the choice of pre-training, tokenization, and data representation strategies. Our study aims to advance the empirical evaluation of structured EHR foundation models and guide the development of future healthcare foundation models.

1 Introduction

Foundation Models [8] have demonstrated state-of-the-art performance across natural language [76], multimodal vision-language [15], multivariate time-series [16], and biological data [20]. Pre-trained on large amounts of unlabelled data, these models learn meaningful representations from simple tasks such as next-token prediction or masked language modeling [17, 26, 59]. These learned representations have proven suitable for diverse downstream tasks, in contrast with the conventional supervised learning paradigm in which models are trained for specific downstream tasks [23, 52, 59, 74]. Increasing availability of digitized structured Electronic Health Records (EHRs), reflective of patients' interactions with the healthcare system, offers an opportunity to leverage foundation models in healthcare [84], where labeled data are often sparse. Multiple foundation models have tackled the

*Equal Contribution

Preprint. Under review.

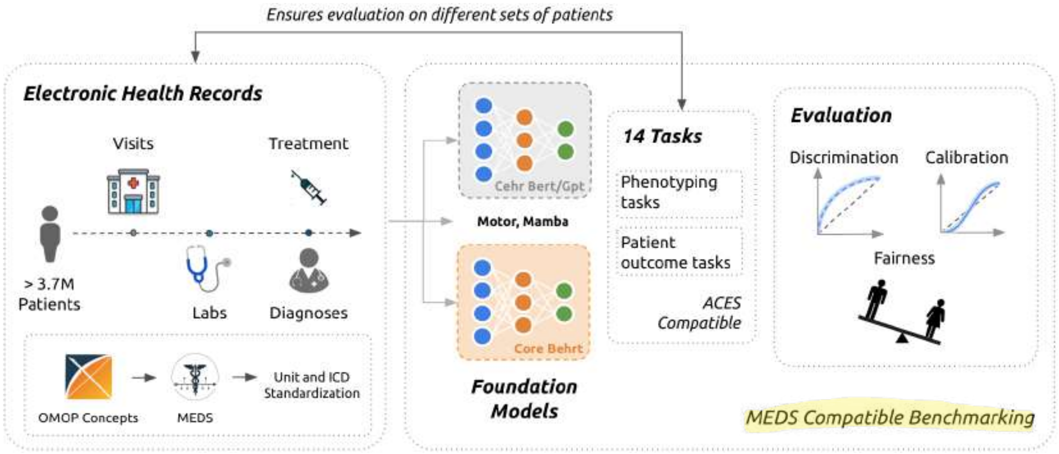

Figure 1. FoMoH is an evaluation of state-of-the-art structured EHR foundation models using (i) clinically meaningful and diverse downstream tasks, (ii) patient cohorts derived to prevent data leakage, (iii) augmented evaluation beyond overall discrimination, such as calibration, and fairness along clinically relevant factors such as healthcare utilization.

unique challenges associated with EHR, such as temporality [56, 67] and long context windows [85], presenting competitive performance across medical tasks [8, 24].

Despite the proliferation of EHR foundation models, there is a critical lack of consensus on optimal design choices and their potential for clinical utility. Clinically-grounded evaluation of structured EHR foundation models is essential to quantify the impact of specific design choices and a standardized measure of progress [8]. Current evaluations have been large-scale but restricted to limited metrics, downstream tasks such as patient mortality, and limited by the lack of data-sharing and interoperability across health institutions [62, 84]. While data sharing is still challenging, several efforts to harmonize EHR data and standardize AI pipelines have removed some barriers to foundation model evaluations in EHRs [3, 28, 39, 63, 87]. Nonetheless, existing benchmarks have focused on performance evaluation along limited design choices such as context-length or model sizes, and evaluating population-level metrics [84]. Evaluating the downstream impact of these choices requires a comprehensive evaluation of population and subpopulation performance, calibration, and discrimination on standardized and diverse clinically meaningful downstream prediction tasks.

We propose a clinically grounded evaluation of structured EHR foundation models on EHR data from Columbia University Irving Medical Center (CUMC), a large academic medical center that caters to a diverse patient population. To ensure reproducibility and transportability, we build on the increasingly robust infrastructure for standardizing EHR data, downstream tasks, and foundation model evaluations [3, 39, 87][2]. Our work's primary contributions, illustrated in Figure 1, are:

1. **Clinically-grounded prediction task desiderata.** The downstream utility of EHR foundation models will be determined by reliable prediction for patient prognosis for healthcare operations (e.g., readmission), patient outcomes, early diagnosis of diverse conditions and pathologies, e.g., chronic conditions, and acute events. We introduce 14 curated and clinically-meaningful downstream prediction tasks consisting of 11 phenotypes and 3 patient prognoses, with ACES-compatible representations [87] — a library for automatic extraction — for reproducibility and transportability across healthcare institutions.

2. **Evaluating foundation models on structured EHR.** We create a MEDS-compatible dataset — a standardized EHR format — to evaluate major foundation models spanning key design choice differences, such as tokenization and pre-training strategies developed at major healthcare institutions, including the assessment of transportability of EHR-based foundation models. The proposed evaluation enables a robust, fair, and clinically meaningful comparison across these design choices, with manual and automated curation to prevent data contamination between pre-training, linear probing,

[2] https://www.ohdsi.org/data-standardization/

and downstream evaluation. Our benchmark, available on Github[3], is intended as a prescriptive framework for future assessments of structured and unstructured EHR foundation models.

3. **Desiderata of evaluation metrics of EHR foundation models.** Motivated by downstream clinical utility, we present a broad set of metrics to capture clinically relevant measures of performance, going beyond discriminative performance, often used as the only evaluation measure in foundation model literature. We measure calibration and fairness of aspects unique to healthcare, such as healthcare utilization. As deployment is a critical bottleneck in healthcare, we evaluate the computational pre-training cost and the transportability of EHR foundation models without local fine-tuning.

2 Related work

There has been a proliferation of unimodal and multimodal foundation models in health and medicine, leveraging structured and unstructured EHR, imaging, genomic, -omics, and other translational medical data, building on the recent success of natural language and vision-based foundation models trained on internet-scale data [18, 20, 23, 29, 46, 52, 75, 84–86]. In this work, we focus on benchmarking EHR-based foundation models, leveraging their structured components (see Table 3 in Appendix A, which describes the datasets, tasks and evaluations of existing EHR foundation models).

Structured EHR Foundation Models. A patient's electronic health record contains clinical events in chronological order, including tokenized occurrences of labs, procedures, prescriptions, and visit types. Consequently, a patient trajectory is a sequence of irregularly sampled events, analogous to natural language. This parallel has led to adapting the language modeling paradigm to EHR. However, EHR-specific challenges require additional innovations in tokenization, data representation, model architectures, and choice of pre-training tasks, for both unstructured and structured data. First, EHRs are marked by irregular sampling, partially informative of the patient's physiology, confounded by the context of patient healthcare utilization patterns, and healthcare processes [1, 5, 6, 11, 32]. Second, EHR tokens may be associated with numerical values from lab tests and other measurements. Third, EHRs cover multiple years of data, requiring long context windows, as longer context windows often reflect sicker patients [85]. Finally, events in EHR often co-occur [89], e.g., multiple conditions may be diagnosed simultaneously following a medical test. Table 3 summarizes the key modeling choices and major structured EHR foundation models, including the pre-training source data. In addition, text-serialization-based representations have been proposed for structured EHR elements [25, 43, 68], which are out of scope for this evaluation.

Data harmonization and representation of structured EHR for foundation modeling. Several pre-processing and harmonization pipelines have been proposed to generate tokenized representations of structured EHR for foundation model training, including FEMR [65] and EventStreamGPT [49]. These frameworks are amenable to AI-friendly standardized representations of structured EHR data, such as the Medical Event Data Standard (MEDS) [3, 39, 66], which are complementary to data harmonization efforts designed to overcome heterogeneity in coding practices across healthcare institutions, to enhance reproducibility and transportability, such as the Observational Medical Outcomes Partnership (OMOP) schema using the OHDSI-Common Data Model [28]. A major challenge in robust evaluation has been the lack of consensus in the downstream task definition, recently addressed by `ACES` [87].

EHR foundation models benchmarking. [86] proposes a comprehensive few-shot evaluation of major structured EHR FMs pre-trained on `Stanford-EHR` data, harmonized in OMOP, on patient outcomes, lab-test results, new diagnoses, and chest radiograph findings. CONTEXT CLUES evaluates the utility of long-context modeling of GPT, LLAMA, MAMBA, and HYENA models, and the impact of irregular sampling, and EHR-specific practices such as copy-forwarding on disease progression tasks [85]. Chen et al. [10] propose MC-BEC, a multimodal benchmark on structured and unstructured emergency care multimodal EHR data for diagnosis and patient deterioration prediction. The majority of EHR benchmarking efforts do not analyze performance beyond AUROC, such as calibration and fairness, especially by variation in healthcare utilization, and propose standardized clinically meaningful downstream tasks, with systematic procedures to account for potential leakage across

[3] `https://github.com/reAIM-Lab/ehr_foundation_model_benchmark/`

variation in major design choices (see Table 1). Our work fills this gap and provides the first evaluation of the transportability of MAMBA models trained on Stanford-EHR on our data[4].

3 FoMoH benchmark

The proposed benchmark consists of four major components. First, we describe standardization and pre-processing from OMOP to a standardized MEDS format, called CUMC-MEDS to enable reproducibility of our AI evaluation. Second, we contribute curated, clinically meaningful downstream prediction tasks, focusing on 3 patient outcomes and 11 phenotypes, spanning chronic disease and acute conditions across multiple clinical specialties. Our phenotypes are designed to avoid data leakage and are ACES-compatible, thus enabling consistent cross-institutional benchmarking on grounded clinical tasks. Third, we benchmark 6 state-of-the-art structured EHR foundation models on CUMC-MEDS transformed data. Finally, we contribute desiderata for enhanced evaluation of downstream EHR foundation model performance, focusing on clinical utility, such as evaluating performance across diverse levels of healthcare utilization and other subpopulation characteristics. Our benchmarking effort introduces a standard for EHR foundation model evaluation. This study was approved under IRB-AAAV2068 at Columbia University.

3.1 Preprocessing

Data. We benchmark performance on EHR data from CUMC, a large academic medical center located in northern Manhattan. CUMC acts as a quaternary care center capable of providing highly specialized care for individuals in the greater New York metropolitan area. The data, therefore, encode a mixture of specialist care (inpatient and outpatient), primary care, and emergency services over a diverse population. While race and ethnicity are sparsely recorded (resp. 61.7% and 75.1% missing), the collected data reflect a diverse population with 7.3% of all patients identifying as non-white and 8.3% as Hispanic or Latino. Further, the database presents 55.7% women and 44.0% men. Additional demographic information can be found in Table 4.

For all patients in the CUMC-EHR, the database encompasses patient demographics, visit details for inpatient and outpatient care, conditions (billing diagnoses and problem lists), medications (outpatient prescriptions and inpatient medication orders and administrations), medical devices, clinical measurements (such as laboratory tests and vital signs), and other clinical observations like symptoms. Out of the more than 6 million patients present in the database, our analysis uses longitudinal health records for ~ 5.3 million patients spanning 1986 to 2023, focusing on prescriptions, procedure events, lab events, patient visits, and setting information. All data have been anonymized and standardized according to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM).

Standardization of measurement units. Similar measurements can be recorded differently in EHRs. We standardize laboratory measurement values and units. We identify the most common unit for each lab and convert, where possible, the unit and associated lab value to this unit.

Standardization of diagnosis codes. We convert ICD-9 codes to ICD-10 using the General Equivalence Mappings (GEM) provided by the Centers for Medicare and Medicaid Services (CMS)[5]. When multiple ICD-10 codes correspond to a given ICD-9 code, we map to the most frequent ICD-10 code. If the mapped ICD-10 code is not present in our dataset, we map to the parent ICD code capturing a more general condition. Any ICD-9 codes absent from the GEM database are dropped.

Handling missingness. For laboratory tests with a concept code but no associated numerical lab value (14.3% of labs) or no unit (5.3% of labs), we keep the laboratory event but indicate that the lab value is missing. Numerical values without associated ICD codes (6.23%) are dropped.

MEDS Conversion. We use MEDS-ETL[6] to convert data from OMOP to MEDS, which represents data as a patient id, time stamp, concept code, and associated value (see Appendix D.1).

[4] Stanford-EHR-pretrained MOTOR could not be directly evaluated on our dataset due to tokenization discrepancies in numerical unit representations in Stanford-EHR and our data.

[5] https://www.cms.gov/medicare/coding/icd10/downloads/icd-10_gem_fact_sheet.pdf

[6] Version 0.3.9 - https://github.com/Medical-Event-Data-Standard/meds_etl/tree/0.3.9

3.2 Downstream task definitions

Downstream evaluations of foundation models should reflect the diversity of downstream applications. In healthcare, these span operational tasks associated with patient outcomes, such as mortality as prevalent in prior benchmarking efforts [10, 39, 50, 85, 86]. Despite their prevalence, definitions of these outcomes and cohorts have varied widely across publications, prohibiting potential comparison and limiting reproducibility and transportability [39, 51]. Our work introduces precise definitions in an ACES-compliant format, leveraging the OHDSI cohort builder to ensure reproducibility. In addition to these tasks, EHR models are frequently used to phenotype both acute and chronic conditions that require timely interventions. We propose 11 phenotypes across diverse clinical settings and reflective of chronic and acute disease courses [7, 45]. Across all tasks, crucially, we maintain any class imbalance present in the true data — rather than sampling negative instances — to ensure downstream clinical utility.

In-hospitalization mortality (`Death`). The cohort includes patients with at least one hospitalization lasting longer than 48 hours, using the admission time as the reference. The prediction is made at 48 hours after admission, referred to as `prediction-time`. Patients who die during the same hospitalization are labeled positive, while those who are discharged alive are labeled as negative. As in other tasks, we require that all patients have at least 2 years of observation prior to the prediction-time (see Figure 3 in Appendix B).

30-day readmission (`Readmission`). We predict a 30-day all-cause readmission following hospital discharge. The `prediction-time` is defined as the discharge time, and any readmissions occurring on the same day are excluded. All patients in the cohort are required to have at least two years of observation prior to the prediction time and must not be censored within 30 days following discharge.

Prolonged length-of-stay (`Long LOS`). We predict whether a hospitalization will last longer than seven days, with the prediction made 48 hours after admission. Patients are required to have at least two years of observation prior to the `prediction-time`.

Phenotyping. We defined 11 phenotypes to assess model performance. Rather than relying solely on ICD codes to define tasks, we established an at-risk cohort and a case cohort for each phenotype using validated phenotyping algorithms. We introduce phenotype-specific rules about temporality to distinguish between the chronic prediction case (e.g., predicting disease onset) and the acute prediction case (e.g., predicting a future adverse event) to ensure clinical utility. Our phenotypes span major health events such as cancer, auto-immune conditions, and chronic conditions such as type-2 diabetes. A detailed definition of these phenotypes is provided in Appendix B.1. Our use of an *at-risk* cohort, rather than a general population cohort, for disease prediction increases task difficulty and clinical utility. All patients with no data during the observation period were excluded to ensure that all models were trained and evaluated on the same set of patients.

3.3 Structured EHR foundation models

We benchmark 6 state-of-the-art foundation models and use linear probing for downstream evaluations. While sophisticated fine-tuning approaches can augment these evaluations, we focus on the pre-training and evaluate paradigm to measure the representation capacity to generalize to clinically meaningful downstream tasks.

We choose foundation models with promising reported performance and covering a diverse range of pre-training objectives, tasks, and tokenization strategies used in the literature, as summarized in Table 1. The following describes the specific design choices made by these models and adjustments enforced to ensure that all models have the same embedding dimensionality (768) with a similar parameter count (120 ± 5% million). Detailed information on how temporal information is integrated is provided in Appendix D.2. Context lengths were chosen as the smallest power of 2 that covers at least 99% of the visits observed in the dataset, as illustrated in Appendix D.3.

CEHR-BERT [56]. CEHR-BERT (Chronological Electronic Health Record BERT) aims to leverage EHR temporal information. To this end, CEHR-BERT introduced Artificial Time Tokens (ATTs), inserted between neighboring events to represent the time intervals to capture fine-grained temporal differences more effectively. CEHR-BERT combines concept embeddings with both absolute and relative time embeddings, which reflect the patient's age at the time of the event. CEHR-BERT focuses

Table 1: EHR foundation models evaluated in this work.

Design Choices	CEHR BERT	CEHR GPT	CORE BEHRT	LLAMA	MAMBA	MOTOR
Architecture	BERT	GPT	BERT	LLAMA	MAMBA	Custom
Context length	2048	2048	2048	8192	8192	8192
Learning objective	MLM	NTP	MLM	NTP	NTP	TTE
Parameter count (in Millions)	114	119	119	124	120	117
Tokenization	Concept	Concept	Concept	Concept & Value	Concept & Value	Concept & Ancestor
Temporal encoding	Age, Timestamp, Inter-event, Position	Age, Inter-event	Age, Timestamp, Position	Position	None	Age, Timestamp
Data Source	Demo., visits, diagnoses, procedures, meds	Demo., visits, diagnoses, procedures meds	Demo., visits, diagnoses meds	All	All	All

MLM: Masked language modeling. NTP: Next token prediction. TTE: Time to event.
Inter-event: time difference between consecutive events that occur on different days.

on structured EHR inputs, specifically using demographic, visit, diagnosis, procedure, and medication data to model temporal clinical trajectories.

CEHR-GPT [57]. CEHR-GPT is a GPT-2 model trained on the patient representation introduced by CEHR-BERT, to enable generating synthetic structured EHR data. CEHR-GPT differs from the previous model through: (i) demographic tokens placed at the start of each patient sequence, (ii) day tokens inserted between neighboring visits to represent time intervals in days, and (iii) day tokens placed within inpatient visits to preserve the duration of hospital stays, to capture patient timelines. CEHR-GPT only operates on a subset of OMOP domains, including demographics, visits, diagnoses, procedures, and medications.

CORE-BEHRT [54].

CORE-BEHRT is a BERT-based model for early diagnosis prediction, which stores patient trajectories as *sentences* with separator tokens to denote the time lapsed between visits. CORE-BEHRT incorporates medication, in addition to diagnosis information used in its precursor model BEHRT [44]. However, it does not use information about procedures or laboratory tests. Contrarily to CEHR-BEHRT, which uses Time2Vec[37] to encode temporality, CORE-BEHRT incorporates temporality through patient age.

LLAMA [21]. LLAMA is a general-purpose foundation model using transformer-based architecture analogous to GPT [59] but differs in the choice of normalization layers. Additionally, LLAMA does not incorporate time but uses rotary positional embeddings (RoPE) in the attention mechanism, to encode the order of medical events. It represents a state-of-art transformer architecture.

MAMBA [22]. MAMBA is a selective state-space model that replaces attention blocks with state-space layers. It can be interpreted as a generalized version of gated Recurrent Neural Networks that capture continuous time-varying inputs. This architecture reduces computational time complexity to linear time compared to transformers, and has been shown to better handle long sequences with temporal dependencies, a common occurrence in EHRs [85].

MOTOR [67]. MOTOR is a Transformer model pre-trained using a piecewise exponential time-to-event (TTE) learning objective. Unlike other architectures that rely on traditional pre-training tasks such as masked language modeling (MLM) or next token prediction (NTP), MOTOR employs TTE predictions to predict over 8, 000 medical codes. This training strategy enables MOTOR to leverage a patient's historical context to predict future events across varying time intervals.

3.4 Baselines

We compare the previous methodologies with two featurization strategies, followed by traditional supervised training. These baselines comprise tabular EHR benchmarks.

FEMR [64] includes normalized age and event counts between the patient's first recorded event and the prediction time. These features are then used to train Logistic Regression and LightGBM [38].

MEDS-TAB [55] extends the previous set of features using customizable time windows with multiple aggregation functions, such as count, value, min, and max. The extracted features are then used to train a Logistic Regression and an XGBoost model [13].

4 Empirical setting

We adopt a local pre-training approach for all foundation models, followed by linear probing. To avoid data leakage, our evaluation relies on a fixed patient split, in which 60% patients are used for training, 10% for hyperparameter tuning during pre-training and linear probing, and the rest for the proposed evaluation. As our evaluation tasks range from $5,000$ to over 2 million records, we cap the training set used for linear probing at $100,000$ and the evaluation set at $50,000$ records for each task.

4.1 Pre-training

Local. We pretrain all models *from scratch* on our EHR data to generate patient embeddings. To this end, we use the same pre-processing, tokenization, hyperparameters, and training objectives as in the original works. To ensure the same convergence levels for all models despite model-specific losses and learning rates, we use a 0.01% change in relative loss as a stopping criterion.

External. We use open weights from a MAMBA model pretrained on `Stanford-EHR`, which we refer to as **MAMBA-TRANSPORT**, on downstream tasks on our EHR data without local fine-tuning. Although MOTOR also provides open weights, we could not evaluate its transportability due to insufficient information on lab measurement units in the original pre-training datasets.

4.2 Linear probing

Linear probing consists of fitting a logistic regression model using representations extracted from frozen foundation models [48]. This common approach is (i) less computationally expensive and data-hungry than further fine-tuning (ii) enables evaluating pretraining design choices on downstream performance. To ensure optimal linear probing performance, we employ a 5-fold cross-validation to select the ℓ_2-penalty λ on a logarithmic scale between 10^{-4} and 10^4.

4.3 Evaluation

For all metrics, confidence intervals were obtained through test set bootstrapping using 100 iterations.

Performance metrics. As both discrimination and calibration are critical to medical applications, our benchmark quantifies discrimination using Area Under the Receiver Operator Curve (AUROC) and calibration using Brier Score (Brier).

Fairness metrics. As medical datasets can reflect historical socio-medical biases [12], our work studies the risk of reinforcing such inequities by quantifying group fairness [4]. Specifically, we measure the previous metrics stratified by reported sex, race, and healthcare utilization[7], and report the associated maximal absolute difference across groups, denoted as $\Delta := \max_g |d_g - d_{\neg g}|$, where d is a performance metric and g is a given group.

Inference computational cost. Due to resource constraints and the response time requirement associated with clinical deployment, inference time and costs are critical for usability. We report inference FLOPs as a proxy to cost (see Appendix E for compute resources used in this evaluation).

[7] We represent healthcare utilization as empirical tertiles of the number of medical encounters a patient has in a year, defined as any day with a recorded visit, condition, procedure, observation, or laboratory test.

5 Results

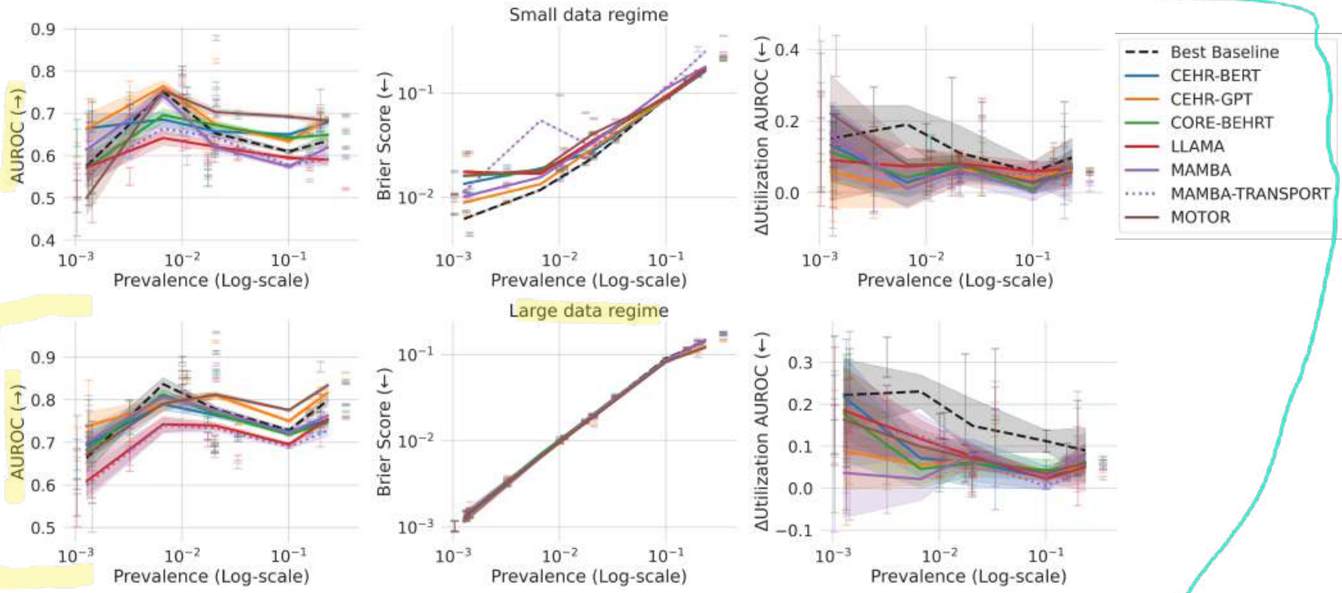

Figure 2. Average linear probing results as a function of the prevalence of positive labels for each task. The leftmost column shows discriminative performance, the middle column shows model calibration, and the right column shows fairness by healthcare utilization. The shaded area represents the bootstrapped standard deviation. The top row reflects the smallest amount of data used for training the linear probing, i.e., 100 points, while the bottom row reflects the largest sample available for each task.

Figure 2 presents the average discriminative, calibration, and fairness performances with respect to healthcare utilization for the considered models across the 14 considered tasks, given the prevalence of each outcome. We further distinguish low and large data regimes corresponding to the smallest and largest amounts of data used for training the linear probing. All task-specific results are deferred to Appendix F. In addition to assessing foundation model fairness according to traditional demographic subgroups (e.g., race, gender), we examine performance based on healthcare utilization as a proxy for access to and trust in the healthcare system. Stratifying model performance based on healthcare utilization also provides insight into how each model performs given different densities of input data and can guide how much data an individual should have for the model to be useful (see Appendix C for the distribution of healthcare utilization in `CUMC-EHR`).

Are foundation model embeddings predictive of clinically meaningful outcomes? While the best discriminative performances on readmission — a common benchmarking task, are achieved by foundation models, performances on tasks with lower prevalence suggest a more nuanced conclusion, with traditional modeling strategies often outperforming all foundation models. For instance, Motor achieves the best performance for readmission but one of the worst for CLL. Further, when limited training data is available with low prevalence, Best (supervised) Baseline is often ranked best on calibration and presents competitive discriminative performance (see Table 2). Low performance on rare conditions suggests that, without fine-tuning, foundation models may not capture the characteristics of rare conditions, as these tokens may be underrepresented during pre-training.

As labeled data are often a critical bottleneck in healthcare, the proposed analysis further quantifies the impact of sample size on performance when probing foundation models. The comparison of small and large-data regimes shows that larger amounts of data lead, on average, to improved discrimination and calibration.

Which data should be used to train foundation models? An important distinction between the considered models is the use of different data sources for the different foundation models. Particularly, Cehr-Bert and Core-Behrt present similar performance despite Cehr-Bert using the additional procedure data. Similarly, Cehr-Gpt performs similarly to Motor despite ignoring laboratory tests (we show the same patterns when comparing two versions of Llama in Appendix F.5). In our analysis of available data sources (Appendix C), we find that lab tests are, on average, more

Table 2. Average performance rank on the proposed dataset stratified per metrics of interest, sorted by average ranking across all metrics.

Foundation	Large data regime			Small data regime		
Models	Calibration	Discrimination	Fairness	Calibration	Discrimination	Fairness
CEHR-GPT	3.07	**2.07**	4.19	3.43	**2.50**	3.86
MOTOR	**2.93**	2.57	3.83	4.71	3.21	4.55
MAMBA	3.57	3.79	4.26	5.07	5.14	4.02
Best Baseline	3.07	3.43	7.10	**1.79**	4.64	6.52
CEHR-BERT	5.93	4.86	4.29	4.29	3.71	5.00
CORE-BEHRT	6.00	4.86	**3.52**	5.43	4.29	4.36
LLAMA	5.29	6.79	4.74	5.29	6.43	4.40
MAMBA-TRANSPORT	6.14	7.64	4.07	6.00	6.07	**3.29**

prevalent. This observation suggests that not all sources of information are valuable for extracting informative embeddings and are tied to the choice of pre-training loss, data sparsity, beyond the inherent informativeness of the data source.

Representing and modeling temporal irregularity. Pre-training losses that explicitly model temporality, particularly irregular sampling, play a central role in foundation model performance. For example, MOTOR which uses TTE loss, and CEHR-GPT, which uses next-token prediction but penalizes prediction of time tokens, i.e. learn to predict inter-event time, demonstrate improved downstream performance, particularly for longer-horizon tasks such as readmission and long-horizon predictions (see Figure F.2). While BERT-based models benefit from long contexts, they do not encode temporality reliably.

Does pre-training improve downstream fairness? In both high-volume and low-volume data regimes, the baseline models result in the largest fairness gaps stratified by healthcare utilization. This pattern holds when examining performance differences based on race and sex (see Appendix F). While foundation models still present differential group performances, the pre-training on large and diverse populations improves fairness, with the average fairness gap close to constant across prevalence and training sizes.

Are EHR foundation models transportable across healthcare institutions? Directly applying the open-weight MAMBA-TRANSPORT model to our dataset leads to degraded performance compared to using the same architecture pre-trained on our data. On some phenotyping tasks, such as chronic lymphocytic leukemia and osteoporosis, its performance falls below the baseline models. This drop may stem from differences in disease prevalence and patient populations across institutions, including lower healthcare utilization and higher density of features (discussed in Appendix C). Additionally, the model's tokenizers rely on data-specific code distributions, inducing a severe lack of overlap and failing to recognize some clinical events in our dataset, highlighting that institution-specific coding practices can hinder model transportability.

6 Discussion

Major Findings. Our results, summarized in Table 2, highlight the current shortcomings of embeddings extracted from state-of-the-art EHR foundation models to surface crucial directions for future advances. First, current foundation models extract representations that do not fully transport between hospitals and do not perform well on rare conditions, due to distributional differences across environments and conditions' characteristics. Since the foundation model paradigm holds greater potential when limited labeled data is available, this highlights the potential for advancing pre-training methods to overcome challenges unique to EHR data. Our experiments echo the literature on the importance of modeling irregularity of EHR data [85]. Second, we highlight the complex relationship between the choice of input data sources, combined with architecture and pre-training choices. For example, laboratory tests appear to be less informative than temporality in EHRs. By measuring performance beyond overall discrimination, we demonstrate that foundation models do not provide conclusive improvement over supervised baselines for low-data and low-prevalence tasks.

Our findings highlight the need to evaluate EHR foundation models on diverse clinically relevant tasks (by population and condition characteristics, clinical settings, EHR-data collection practices, healthcare utilization patterns, etc.) that capture varying outcome prevalence, aligning with the need for a clearer evaluation desiderata [8, 47]. By evaluating linear-probing results beyond overall discrimination, using calibration, fairness measures, and inference computational costs, we disambiguate design choices that lead to robust and meaningful improvements.

6.1 Limitations and future directions.

The following highlights some of the limitations of our work and describes avenues for future work.

Considered metrics. While our benchmark considers a broad range of metrics covering critical evaluation desiderata in healthcare, this work does not quantify the impact on downstream clinical decisions [84]. Collaboration with domain experts and trials remain crucial to ensure additional validation and safe deployment [35, 62]. Moreover, our benchmark does not address the impact of censoring, a pervasive issue in medical datasets [33]. Future iterations will incorporate censoring-adjusted time-to-event metrics to provide a more comprehensive evaluation.

Considered models. To ensure a fair comparison between models, our analysis is limited to models with $\sim$ 120 million parameters and a fixed embedding dimensionality. As model size is often linked to performance and generalization capacity [36], future work should evaluate the impact of these choices on model performance. Further, the proposed benchmark compares state-of-the-art models but does not isolate the impact of architectural inductive biases, tokenizations, and pre-training losses on performance. We will address this in future work to refine our recommendations. Finally, our work focuses on linear probing strategies; fine-tuning strategies may result in conclusions different from those presented in this work. For example, architectures trained with masked loss may benefit from further fine-tuning, as these models are not pre-trained to predict future events at long horizons.

Considered data. While reflective of the robustness of foundation models applied to a large medical center, the conclusions of our work rely on a single center whose characteristics may differ from others. Our focus on reproducibility in our work aims to encourage practitioners to validate these conclusions beyond our benchmarking data, building on our task definitions and evaluation.

6.2 Broader Impact

We contribute a standardized and clinically diverse evaluation desiderata for EHR foundation models, comparing state-of-the-art models on a large and diverse EHR dataset. Our work surfaces future directions for advancing EHR foundation models.

Acknowledgments

VJ, YSC, ZJ, and SJ would like to acknowledge partial support from NIH R01MH137679. AK would like to acknowledge the support of NIH T15LM007079. YK and SJ acknowledge partial support from the SPIRIT Precision Pilot Award at Columbia. ZJ and SJ would like to acknowledge partial support from NIH R01LM014380. SJ would like to acknowledge partial support from the SNF Center for Precision Psychiatry and Mental Health, Columbia. Any opinions, findings, conclusions, or recommendations in this manuscript are those of the authors and do not reflect the views, policies, endorsements, expressed or implied, of any aforementioned funding agencies/institutions.

References

[1] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

[2] American Cancer Society. Risk factors for chronic lymphocytic leukemia. `https://www.cancer.org/cancer/types/chronic-lymphocytic-leukemia/causes-risks-prevention/risk-factors.html`, 2024. Accessed: 2025-05-05.

[3] Bert Arnrich, Edward Choi, Jason Alan Fries, Matthew B.A. McDermott, Jungwoo Oh, Tom Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, and Robin van de Water. Medical event data standard (MEDS): Facilitating machine learning for health. In *ICLR 2024 Workshop on Learning from Time Series For Health*, 2024. URL `https://openreview.net/forum?id=IsHy2ebjIG`.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[5] Brett K Beaulieu-Jones, Daniel R Lavage, John W Snyder, Jason H Moore, Sarah A Pendergrass, and Christopher R Bauer. Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR medical informatics*, 6(1):e8960, 2018.

[6] Brett K Beaulieu-Jones, William Yuan, Gabriel A Brat, Andrew L Beam, Griffin Weber, Marshall Ruffin, and Isaac S Kohane. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ digital medicine*, 4(1):62, 2021.

[7] Saul Blecker, Simon A Jones, Christopher M Petrilli, Andrew J Admon, Himali Weerahandi, Fritz Francois, and Leora I Horwitz. Hospitalizations for chronic disease and acute conditions in the time of covid-19. *JAMA internal medicine*, 181(2):269–271, 2021.

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[9] Centers for Disease Control and Prevention. Risk factors for stroke. `https://www.cdc.gov/stroke/risk-factors/index.html`, 2024. Accessed: 2025-05-05.

[10] Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Reisler, David E Kim, and Pranav Rajpurkar. Multimodal clinical benchmark for emergency care (mc-bec): A comprehensive benchmark for evaluating foundation models in emergency medicine. *Advances in Neural Information Processing Systems*, 36:45794–45811, 2023.

[11] Irene Y* Chen, Shalmali Joshi*, Marzyeh Ghassemi, and Rajesh Ranganath. Probabilistic machine learning for healthcare. *Annual Review of Biomedical Data Science*, 4, 2020.

[12] Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. ***Annual Review of Biomedical Data Science***, 2021. doi: 10.1146/annurev-biodatasci-092820-114757. URL `https://doi.org/10.1146/annurev-biodatasci-092820-114757`.

[13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL `http://dx.doi.org/10.1145/2939672.2939785`.

[14] Goh Eun Chung, Su Jong Yu, Jeong-Ju Yoo, Yuri Cho, Kyu na Lee, Dong Wook Shin, Yoon Jun Kim, Jung-Hwan Yoon, Kyungdo Han, and Eun Ju Cho. Metabolic dysfunction-associated steatotic liver disease increases cardiovascular disease risk in young adults. *Scientific Reports*, 15:5777, 2025. doi: 10.1038/s41598-025-89293-6. URL `https://www.nature.com/articles/s41598-025-89293-6`.

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=vvoWPYqZJA`.

[16] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[18] Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. *arXiv preprint arXiv:2405.14567*, 2024.

[19] Molly T Finnerty, Atif Khan, Kai You, Rui Wang, Gyojeong Gu, Deborah Layman, Qingxian Chen, Noémie Elhadad, Shalmali Joshi, Paul S Appelbaum, et al. Prevalence and incidence measures for schizophrenia among commercial health insurance and medicaid enrollees. *Schizophrenia*, 10(1):68, 2024.

[20] Xi Fu, Shentong Mo, Alejandro Buendia, Anouchka P Laurent, Anqi Shao, Maria del Mar Alvarez-Torres, Tianji Yu, Jimin Tan, Jiayu Su, Romella Sagatelian, et al. A foundation model of transcription across human cell types. *Nature*, pages 1–9, 2025.

[21] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[22] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[23] Fei Guo, Renchu Guan, Yaohang Li, Qi Liu, Xiaowo Wang, Can Yang, and Jianxin Wang. Foundation models in bioinformatics. *National Science Review*, page nwaf028, 2025.

[24] Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, 2024.

[25] Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful ehr encoders. *arXiv preprint arXiv:2502.17403*, 2025.

[26] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79, 2023.

[27] George Hripcsak, RuiJun Chen, and Thomas Falconer. Feasibility of large-scale observational cancer research using the ohdsi network—aim 2 findings. `https://www.ohdsi.org/wp-content/uploads/2015/04/NCI-FinalPresentation-OHDSI.pdf`. Accessed: 2025-05-06.

[28] George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. In *MEDINFO 2015: eHealth-enabled Health*, pages 574–578. IOS Press, 2015.

[29] Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N Nadkarni, Benjamin S Glicksberg, Nils Gehlenborg, and Marinka Zitnik. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613, 2024.

[30] Youn Huh, Yoon Jeong Cho, and Ga Eun Nam. Recent epidemiology and risk factors of nonalcoholic fatty liver disease. *Journal of Obesity & Metabolic Syndrome*, 31(1):17–27, 2022. doi: 10.7570/jomes22021. URL `https://www.jomes.org/journal/view.html?doi=10.7570/jomes22021`.

[31] Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, Min Jae Lee, Eunbyeol Cho, Seong-Eun Moon, Young-Hak Kim, Louis Atallah, and Edward Choi. Genhpf: General healthcare predictive framework for multi-task multi-source learning. *IEEE Journal of Biomedical and Health Informatics*, 28(1):502–513, 2023.

[32] Vincent Jeanselme. *Clinical Presence: Impact on Predictive Modelling and Algorithmic Fairness*. PhD thesis, University of Cambridge, 2024.

[33] Vincent Jeanselme, Nikita Agarwal, and Chen Wang. Review of language models for survival analysis. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

[34] Hyewon Jeong, Nassim Oufattole, Matthew Mcdermott, Aparna Balagopalan, Bryan Jangeesingh, Marzyeh Ghassemi, and Collin Stultz. Event-based contrastive learning for medical time series. *arXiv preprint arXiv:2312.10308*, 2023.

[35] Shalmali Joshi, Iñigo Urteaga, Wouter AC van Amsterdam, George Hripcsak, Pierre Elias, Benjamin Recht, Noémie Elhadad, James Fackler, Mark P Sendak, Jenna Wiens, et al. Ai as an intervention: improving clinical outcomes relies on a causal approach to ai development and validation. *Journal of the American Medical Informatics Association*, page ocae301, 2025.

[36] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[37] Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2Vec: Learning a Vector Representation of Time, July 2019. URL `http://arxiv.org/abs/1907.05321`. arXiv:1907.05321 [cs].

[38] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[39] Aleksia Kolo, Chao Pang, Edward Choi, Ethan Steinberg, Hyewon Jeong, Jack Gallifant, Jason A Fries, Jeffrey N Chiang, Jungwoo Oh, Justin Xu, et al. Meds decentralized, extensible validation (meds-dev) benchmark: Establishing reproducibility and comparability in ml for health. 2024.

[40] Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. Medgpt: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*, 2021.

[41] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.

[42] Junghwan Lee, Cong Liu, Jae Hyun Kim, Alex Butler, Ning Shang, Chao Pang, Karthik Natarajan, Patrick B Ryan, Casey Ta, and Chunhua Weng. Comparative effectiveness of medical concept embedding for feature engineering in phenotyping. *JAMIA Open*, 4:2, 2021.

[43] Simon A Lee, Sujay Jain, Alex Chen, Kyoka Ono, Jennifer Fang, Akos Rudas, and Jeffrey N Chiang. Emergency department decision support using clinical pseudo-notes. *arXiv preprint arXiv:2402.00160*, 2024.

[44] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

[45] Michael Linden, Ulrike Linden, David Goretzko, and Jochen Gensichen. Prevalence and pattern of acute and chronic multimorbidity across all body systems and age groups in primary health care. *Scientific Reports*, 12(1):272, 2022.

[46] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

[47] Faisal Mahmood. A benchmarking crisis in biomedical machine learning. *Nature Medicine*, pages 1–1, 2025.

[48] Markus Marks, Manuel Knott, Neehar Kondapaneni, Elijah Cole, Thijs Defraeye, Fernando Perez-Cruz, and Pietro Perona. A closer look at benchmarking self-supervised pre-training with image classification. *International Journal of Computer Vision*, pages 1–13, 2025.

[49] Matthew McDermott, Bret Nestor, Peniel Argaw, and Isaac S Kohane. Event stream gpt: a data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. *Advances in Neural Information Processing Systems*, 36: 24322–24334, 2023.

[50] Matthew McDermott, Bret Nestor, Peniel Argaw, and Isaac S Kohane. Event stream gpt: a data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. *Advances in Neural Information Processing Systems*, 36, 2024.

[51] Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586):eabb1655, 2021.

[52] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[53] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Risk factors for type 2 diabetes. `https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes`, 2025. Accessed: 2025-05-05.

[54] Mikkel Odgaard, Kiril Vadimovic Klein, Sanne Møller Thysen, Espen Jimenez-Solem, Martin Sillesen, and Mads Nielsen. Core-behrt: A carefully optimized and rigorously evaluated behrt. *arXiv preprint arXiv:2404.15201*, 2024.

[55] Nassim Oufattole, Teya Bergamaschi, Aleksia Kolo, Hyewon Jeong, Hanna Gaggin, Collin M. Stultz, and Matthew B. A. McDermott. Meds-tab: Automated tabularization and baseline methods for meds datasets, 2024. URL `https://arxiv.org/abs/2411.00200`.

[56] Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pages 239–260. PMLR, 2021.

[57] Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S Kalluri, Elise L Minto, Jason Patterson, Linying Zhang, George Hripcsak, Gamze Gürsoy, Noémie Elhadad, et al. Cehr-gpt: Generating electronic health records with chronological patient timelines. *arXiv preprint arXiv:2402.04400*, 2024.

[58] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162/`.

[59] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[60] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

[61] Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *npj Digital Medicine*, 7(1):256, 2024.

[62] Nigam H Shah, David Entwistle, and Michael A Pfeffer. Creation and adoption of large language models in medicine. *Jama*, 330(9):866–869, 2023.

[63] Paul E Stang, Patrick B Ryan, Judith A Racoosin, J Marc Overhage, Abraham G Hartzema, Christian Reich, Emily Welebob, Thomas Scarnecchia, and Janet Woodcock. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Annals of internal medicine*, 153(9):600–606, 2010.

[64] Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective patient representation learning technique for electronic health record data, 2020. URL `https://arxiv.org/abs/2001.05295`.

[65] Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113:103637, 2021.

[66] Ethan Steinberg, Michael Wornow, Suhana Bedi, Jason Alan Fries, Matthew McDermott, and Nigam H Shah. meds_reader: A fast and efficient ehr processing library. *arXiv preprint arXiv:2409.09095*, 2024.

[67] Ethan Steinberg, Yizhe Xu, Jason Alan Fries, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=NialiwI2V6`.

[68] Xiaorui Su, Shvat Messica, Yepeng Huang, Ruth Johnson, Lukas Fesser, Shanghua Gao, Faryad Sahneh, and Marinka Zitnik. Multimodal medical code tokenizer. *arXiv preprint arXiv:2502.04397*, 2025.

[69] Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seng Chan You, RuiJun Chen, and Nicole Pratt. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *The Lancet*, 394(10211):1816–1826, 2019. doi: 10.1016/S0140-6736(19)32317-7. URL `https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)32317-7/abstract`.

[70] Marc A. Suchard, Martijn J. Schuemie, Harlan M. Krumholz, Seng Chan You, RuiJun Chen, Nicole Pratt, Christian G. Reich, Jon Duke, David Madigan, George Hripcsak, and Patrick B. Ryan. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *The Lancet*, 394(10211):1816–1826, 2019. doi: 10.1016/S0140-6736(19)32317-7. URL `https://doi.org/10.1016/S0140-6736(19)32317-7`.

[71] Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seoyoung C You, Ren Chen, Nicole Pratt, Christian Reich, Patrick Ryan, and George Hripcsak. Large-scale evidence generation and evaluation across a network of databases for type 2 diabetes mellitus (legend-t2dm): a protocol for a series of multinational, real-world comparative cardiovascular effectiveness and safety studies. *BMJ Open*, 11(1):e043247, 2021. doi: 10.1136/bmjopen-2020-043247. URL `https://bmjopen.bmj.com/content/11/1/e043247`.

[72] Joel N Swerdel, George Hripcsak, and Patrick B Ryan. Phevaluator: Development and evaluation of a phenotype algorithm evaluator. *Journal of Biomedical Informatics*, 97:103258, 2019.

[73] Joel N. Swerdel, Darmendra Ramcharran, and Jill Hardin. Using a data-driven approach for the development and evaluation of phenotype algorithms for systemic lupus erythematosus. *PLOS ONE*, 18(2):e0281929, 2023. ISSN 1932-6203. doi: 10.1371/journal.pone.0281929.

[74] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[75] Mohan Timilsina, Samuele Buosi, Muhammad Asif Razzaq, Rafiqul Haque, Conor Judge, and Edward Curry. Harmonizing foundation models in healthcare: A comprehensive survey of their roles, relationships, and impact in artificial intelligence's advancing terrain. *Computers in Biology and Medicine*, 189:109925, 2025.

[76] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[77] UC San Diego Health. Osteoporosis risk factors. `https://health.ucsd.edu/care/endocrinology-diabetes/osteoporosis/risk-factors/`, 2024. Accessed: 2025-05-05.

[78] UpToDate. Diagnosis of acute myocardial infarction. `https://www.uptodate.com/contents/diagnosis-of-acute-myocardial-infarction?search=AMI&source=search_result&selectedTitle=2~150&usage_type=default&display_rank=2`, . Accessed: 2025-05-06.

[79] UpToDate. Diagnosis of celiac disease in adults. `https://www.uptodate.com/contents/diagnosis-of-celiac-disease-in-adults`, . Accessed: 2025-05-06.

[80] UpToDate. Overview of hypertension in adults. `https://www.uptodate.com/contents/overview-of-hypertension-in-adults?search=prevalence%20of%20hypertension&source=search_result&selectedTitle=2%7E150&usage_type=default&display_rank=2#H8`, . Accessed: 2025-05-06.

[81] UpToDate. Hypertension in adults: Epidemiology. `https://www.uptodate.com/contents/the-prevalence-and-control-of-hypertension-in-adults?search=hypertension&topicRef=3852&source=see_link#H1773280799`, . Accessed: 2025-05-06.

[82] UpToDate. Clinical manifestations, diagnosis, and staging of exocrine pancreatic cancer. `https://www.uptodate.com/contents/clinical-manifestations-diagnosis-and-staging-of-exocrine-pancreatic-cancer?search=pancreatic%20cancer&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1`, . Accessed: 2025-05-06.

[83] UpToDate. Systemic lupus erythematosus in adults: Clinical manifestations and diagnosis. `https://www.uptodate.com/contents/systemic-lupus-erythematosus-in-adults-clinical-manifestations-and-diagnosis?search=lupus&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1`, . Accessed: 2025-05-06.

[84] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1): 135, 2023.

[85] Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Re, Sanmi Koyejo, and Nigam Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehr data. In *The Thirteenth International Conference on Learning Representations*, 2024.

[86] Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36, 2024.

[87] Justin Xu, Jack Gallifant, Alistair Johnson, and Matthew B.A. McDermott. ACES: Automatic cohort extraction system for event-stream datasets. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=P4XmKjXTrM`.

[88] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

[89] Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature communications*, 14(1):7857, 2023.

A Existing EHR foundation models

Table 3 describes existing EHR foundation models that have been proposed in the literature, their associated characteristics, and evaluation metrics.

Table 3. Structured EHR foundation models, pre-training/training datasets, tokenization strategy, and evaluation measures (grouped by training data, chronologically ordered).

Foundation	Dataset	Tokenization			Pretrain Loss	Evaluation		
Model		Time	Concept	Value		Disc.	Cal.	Fair.
BEHRT [44]	Clinical Practice Research Datalink	PE & Age-Embedding	Code-based	None	MLM	✓		
Med-BERT [60]	Cerner Health Facts	PE & Age-Embedding	Code-based	None	MLM	✓		
CLMBR [65]	EHRSHOT	RoPE	Code-based	None	NTP	✓		
MedGPT [40]	London Hospital + MIMIC-III	RoPE	Text & Code based	None	NTP	✓		
CEHR-BERT [56]	CUMC-NYP	Time2vec+ATT	Code-based	None	MLM	✓	✓	
GatorTron [88]	UF Health, Pubmed, Wikipedia + MIMIC III	None	Text-based	None	MLM & SOP	✓		
GenHPF [31]	MIMIC-III, IV + eICU	ALiBi	Text-based	Digit Place Embedding	SSL	✓		
Foresight [41]	London Hospital + MIMIC-III	Sinusoidal	Text & Code based	None	NTP	✓		
MOTOR [67]	STANFORD-EHR + MERATIVE	RoPE	Code-based	None	TTE	✓	✓	✓
TransformEHR [89]	VA Bedford	Sinusoidal	Code-based	None	Visit MLM	✓		✓
EBCL [34]	MIMIC-III	Custom	Code-based	CLIP	EBCL	✓		
CEHR-GPT [57]	CUMC-NYP	ATT	Code-based	None	NTP			
CORE-BEHRT [54]	Capital Region of Denmark	RoPE	Code-based	None	MLM	✓		
EHRMAMBA [18]	MIMIC-IV	Time2vec	Code-based	None	NTP	✓		
ETHOS [61]	MIMIC-IV	Age quantiles + ATT	Code-based	Quantiles	NTP	✓		
Long-context LLAMA,MAMBA [85]	TRANSPORT	RoPE	Code-based	None	NTP	✓		

Notes: Disc.: Discrimination; Cal.: Calibration. Fair.:Fairness. Checkmarks (✓) indicate the aspect was evaluated in the paper. PE: Positional embedding. ATT: Artificial Time Tokens. MLM: Masked language modeling. NTP: Next token prediction. SOP: Sentence order prediction. SSL: Self-supervised learning. TTE: Time to event. EBCL: Event-based contrastive learning

B Phenotype Definitions

For most phenotypes, we used the OHDSI Phenotype Library to generate the at-risk and case cohorts. The phenotype library is a publicly accessible and version-controlled catalog of phenotypes generated by members of the OHDSI community. We evaluated our phenotype algorithms using OHDSI PheValuator [72]. All phenotypes are available in the GitHub repository as JSON files and are fully reproducible on any dataset that employs the OMOP-CDM.

The at-risk cohorts were defined primarily through three methods: (i) identifying risk factors for the condition through UpToDate[8], an evidence-based clinical resource that provides point-of-care medical information summaries about a wide variety of medical topics; (ii) identifying non-descendant diagnostic codes linguistically related to the condition of interest through GloVe embeddings [42, 58]; and (iii) leveraging the ICD hierarchy to identify subchapters of codes related to the condition of interest.

Each phenotype has three major temporal components: cohort entry date, index date, and cohort exit date (Figure 3). The observation period is defined as the time between the cohort entry date and the index date. The model uses data in the observation period to predict the outcome. The time from the end of the index date to the cohort exit date comprises the prediction period, where the goal of the model is to predict whether or not the target diagnosis occurs in this period.

For all disease phenotypes, the observation start date corresponds with the patient's first entry in the database. The patient enters the at-risk cohort when the inclusion criteria for being at-risk are met, and the entrance event to the at-risk cohort is named "at-risk cohort inclusion event". Index event, i.e., the patient encounter when the model predicts whether the patient belongs in the case or control group for the phenotype at that time, will always be defined after the patient enters the at-risk cohort, i.e., after the at-risk cohort inclusion event. The prediction window is 1 year for all phenotypes. The exact specifications of all phenotypes, including specific concept codes, are stored in JSON files in the GitHub repository.

B.1 Celiac Disease

Celiac is a chronic disease. The inclusion event for the at-risk cohort can be a family history of celiac [79], or a related code as identified by GloVE embeddings. Inclusion in the case cohort is based on a celiac disease phenotype created as part of the HERA characterization study and requires a diagnosis code of celiac disease (SNOMED code 396331005, ICD-10 K90.0, or ICD-9 579.0).

Patients who satisfy the inclusion criteria for the case cohort prior to the at-risk cohort are not included in the phenotyping task.

B.2 Acute Myocardial Infarction (AMI)

AMI, also known as a heart attack, is an acute condition. The inclusion event for the at-risk cohort is one of the AMI risk factors from [78], a related code identified by GloVe embeddings, or a diagnosis from the "Ischemic heart diseases" chapter (ICD10: I20-I25). The inclusion event for the AMI case cohort is an appropriate diagnostic code (SNOMED: 22298006) during an emergency room or inpatient visit. We specifically omit the SNOMED code 1755008 (prior myocardial infarction) from the inclusion event. Additionally, we require a 180-day washout period between AMI diagnoses to ensure that each diagnosis refers to a separate event. This phenotype was also used in [69].

Patients who satisfy the inclusion criteria for the case cohort less than 7 days prior to entry of the at-risk cohort are not included in the phenotyping task.

B.3 Systemic lupus erythematosus (SLE)

SLE, or lupus, is a chronic disease. We derive the phenotype for SLE from the OHDSI-generated phenotype, which has been previously validated via chart review [73]. The entry event for the at-risk cohort includes: differential diagnoses for SLE [83], related codes generated by GloVe embeddings, and any code within the "Systemic connective tissue disorders" subchapter of ICD10 (M30-M36). Additionally, the validated phenotype for SLE includes a set of concepts titled "Signs and symptoms

[8] `https://www.uptodate.com/contents/search`

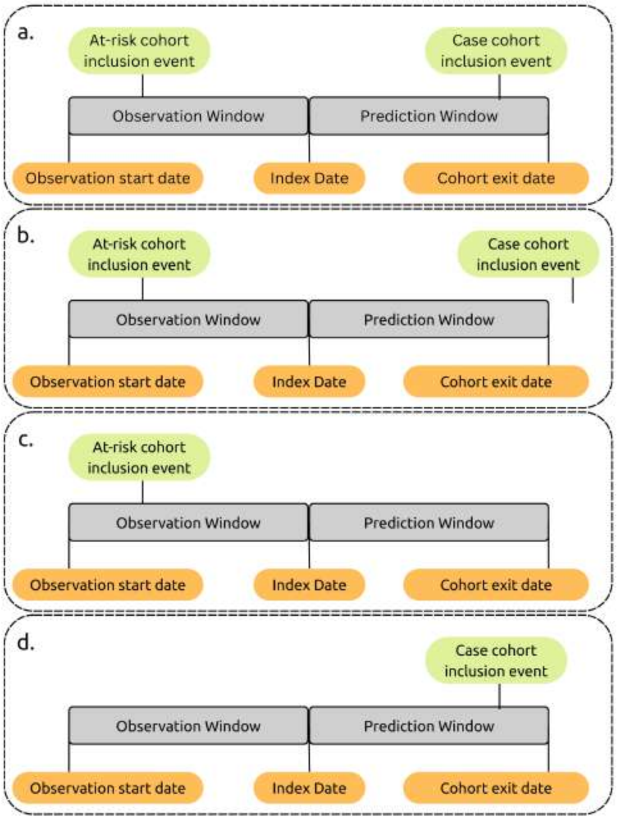

Figure 3. The proposed phenotypes are structured using consistent temporal components: (observation entry, index, and cohort exit) and clear inclusion criteria for the at-risk and case cohorts. (a) Example of an individual included in the case cohort, as their case-cohort inclusion event occurs during the prediction window. (b-c) Examples of individuals included in the at-risk, but not the case cohort, since they have an at-risk inclusion event but do not have a case cohort inclusion event during the prediction window. (d) Example of an individual not included in the at-risk or case-cohorts, as they are missing an at-risk inclusion event.

suggestive of SLE"; any of these condition occurrences can also constitute an entry event for the at-risk cohort. The entry event for the case cohort starts with at least one condition occurrence from the predefined concept set "Signs and symptoms suggestive of SLE" or a drug exposure from the concept set “SLE treatments”. In both cases, this initial event must be followed by a diagnosis from the SLE concept set within 90 days. Due to its chronic nature, only the first diagnosis of SLE is considered during phenotyping.

Patients who satisfy the inclusion criteria for the case cohort prior to the at-risk cohort are not included in the phenotyping task.

B.4 Pancreatic Cancer

Pancreatic cancer is also a chronic disease. The at-risk cohort entry event is defined as a common differential diagnosis and symptoms for pancreatic cancer [82], a related diagnosis code derived from GloVe embeddings, or a diagnosis code from the "Malignant neoplasms of digestive organs" subchapter (ICD10: C15-C26). The inclusion event for the case cohort is a diagnosis of "Neoplasm of pancreas" (SNOMED: 126859007) with no diagnoses of "Benign neoplasm of pancreas" (SNOMED: 92264007) and "Benign tumor of exocrine pancreas" (SNOMED: 271956003). We defined our pancreatic cancer cohort following the "Phenotyping and Validation of Cancer Diagnoses" [27].

Patients who satisfy the inclusion criteria for the case cohort prior to the at-risk cohort will not be included in the phenotyping task.

B.5 Hypertension (HTN)

Hypertension is also a chronic disease. The entry event for the at-risk cohort includes risk factors identified from UpToDate [80, 81], and condition codes from the "Hypertensive diseases" ICD10 subchapter (I10-I1A). No GloVe embeddings were found for hypertensive disorder. The inclusion event for the case cohort is the first diagnosis of hypertensive disorder (SNOMED: 38341003) or its descendants.

Patients who satisfy the inclusion criteria for the case cohort prior to the at-risk cohort will not be included in the phenotyping task.

B.6 Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD)

MASLD is a chronic disease. The at-risk event is an occurrence of one of the MASLD risk factors: insulin resistance or Type 2 diabetes, metabolic syndrome, obesity, dyslipidemia, sarcopenia, hypertension, hypertriglyceridemia, polycystic ovarian syndrome, chronic kidney disease, hypobetalipoproteinemia, lysosomal acid lipase deficiency, defects in mitochondrial fatty acid oxidation [14, 30]. Inclusion into the case cohort is based on the PheKB phenotype algorithm[9], which consists of ICD9 (571.5, 571.8, 571.9) and ICD10 codes (K75.81, K76.0, K76.9). The diagnosis of hepatic steatosis from imaging, medications, and clinical notes is not considered for simplicity.

B.7 Ischemic Stroke

Ischemic stroke is an acute condition. The inclusion event for the at-risk cohort includes the following risk factors: previous stroke, transient ischemic attack, hypertension, high cholesterol, heart disease, diabetes, obesity, and sickle cell disease [9]. The inclusion event for the case cohort is an ischemic stroke code (ICD10: $I63^*$) that occurs during an inpatient or emergency room visit [70].

B.8 Osteoporosis

Osteoporosis is a chronic disease. The at-risk cohort entry event includes any of the following risk criteria: women aged 65 and older, men aged 70 and older, anyone 50 or older who has had a fracture or with the following risk factors, such as White or Asian women, smokers and heavy drinkers, who weigh less than 125 pounds, who have undergone bariatric surgery, with kidney failure, inflammatory bowel disease, rheumatoid arthritis, liver disease or an eating disorder or who take oral corticosteroids daily or other high-risk medications (e.g., thyroid hormone replacement, immunosuppressant drugs, warfarin) [77]. The inclusion event for the case cohort is a diagnosis code for "Osteoporosis" (SNOMED: 80502) or any of its descendants.

B.9 Chronic Lymphoid Leukemia (CLL)

Inclusion events for the CLL at-risk cohort consist of any of the following risk factors: age greater than or equal to 50 with a CLL symptom, or with a family history of CLL. CLL symptoms include weight loss, splenomegaly, shortness of breath, loss of appetite, localized enlarged lymph nodes, and fatigue [2]. The inclusion event for the case cohort was generated using the "Chronic lymphoid leukemia" concept set, excluding "Acute lymphoblastic leukemia". The phenotype definition was created as part of the HERA characterization study[10].

B.10 Type 2 Diabetes Mellitus (T2DM)

T2DM is a chronic disease. The inclusion event for the at-risk cohort is one of the following: over 35 years old and overweight or obese, with prediabetes and a family history of diabetes, or with a history of gestational diabetes [53]. The case cohort was generated from the OHDSI Legend study [71]. The cohort includes patients with a T2DM diagnosis on two separate days, or with one T2DM diagnosis and one antidiabetes drug, or one T2DM diagnosis and two high HbA1C lab tests (test results between 6.5 and 30) at least 7 days before the diagnosis date.

[9] `https://phekb.org/phenotype/non-alcoholic-fatty-liver-disease-nalfd-alcoholic-fatty-liver-disease-ald`

[10] `https://data.ohdsi.org/HERACharacterization/`

B.11 Schizophrenia

Schizophrenia is a chronic disease. Inclusion into the at-risk cohort requires a non-schizophrenia spectrum psychotic disorder. The case cohort requires a subsequent diagnosis of schizophrenia or schizoaffective disorder. Individuals with schizophreniform disorder were excluded from both cohorts, as their ambiguous diagnostic status could constitute data leakage. We further require that the individual be between 10 and 35 years old at the prediction time, as this is the typical age range of schizophrenia onset. All individuals had at least 3 years of observation, with at least one diagnosis or prescription in the last 3 years of observation. The schizophrenia phenotype, including the 3-year observation lookback, was previously validated in [19].

As with all chronic diseases, the observation start date is the start of the patient's observed data in the database. The cohort exit date is either a schizophrenia diagnosis or the end of observed time (if the patient does not have schizophrenia. The index date is 90 days prior to the cohort exit date, leaving a 90-day censor time to avoid temporal data leakage. We use the full patient history (limited only by the model's context length) to predict the diagnostic transition from psychosis to schizophrenia because of the relatively small cohort size: 636 patients with diagnostic transition from psychosis to schizophrenia and 2,634 with psychosis only.

C Overview of CUMC-EHR and Stanford-EHR datasets

The following plots describe the distributions of the different feature groups in CUMC-EHRand Stanford-EHR. While all models were tested on our dataset, MAMBA-TRANSPORT was trained on Stanford-EHR and then tested on CUMC-EHR to assess the model's transportability. Figures 4, 5 and 6 describe different characteristics of both our dataset and Stanford-EHR (EHRSHOT), a publicly available subset of the original dataset (also summarized numerically in Table 5)) to understand the origin of the lack of inter-hospital transportability observed in our results.

Table 4: Demographic distribution of CUMC-EHR and Stanford-EHR (EHRSHOT) data.

		CUMC-EHR % (n)	Stanford-EHR (EHRSHOT) % (n)
Sex	**Female**	55.7 (3,744,185)	51.1 (3,441)
	Male	44.0 (2,955,940)	48.9 (3,298)
	Missing	0.3 (16,814)	0 (0)
Ethnicity	**Hispanic or Latino**	8.3 (557,173)	15.4 (1,038)
	Not Hispanic or Latino	16.6 (1,118,232)	77.7 (5,236)
	Missing	75.1 (5,041,534)	6.9 (465)
Race	**Asian**	1.2 (83,697)	15.5 (1,043)
	Black or African American	5.9 (396,677)	4.4 (298)
	Native Hawaiian or Other Pacific Islander	0.1 (6,534)	1.1 (74)
	Native American or Alaskan Native	0.1 (7,443)	0.4 (25)
	White	16.9 (1,133,364)	55.4 (3,736)
	Other	14.1 (944,342)	0 (0)
	Missing	61.7 (4,144,882)	23.2 (1,563)

Table 5. Means (standard deviation) number of events differentiated by data types available in CUMC-EHR and Stanford-EHR (EHRSHOT) data.

	CUMC-EHR	Stanford-EHR (EHRSHOT)
Healthcare utilization (distinct events per year)	118.4 (162.1)	28.6 (59.1)
Conditions (occurrences per year)	98.9 (392.9)	57.4 (243.8)
Medications (occurrences per year)	84.3 (508.6)	133.2 (583.5)
Procedures (occurrences per year)	45.8 (222.7)	85.6 (398.6)
Lab Tests (occurrences per year)	817.0 (4823.0)	1,717.5 (7,142.5)

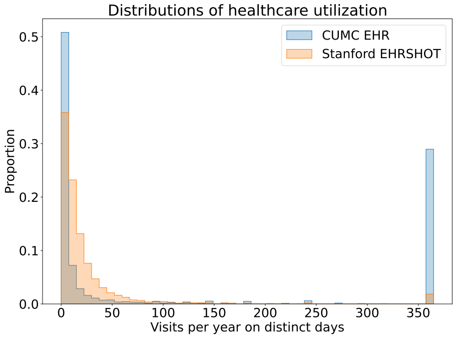

Figure 4. We plot the distributions of healthcare utilization, measured as visits per year for both datasets. Healthcare utilization was measured as the number of distinct interactions with the healthcare system per year. We could any event (condition, drug, procedure, measurement, observation, or lab test) as part of a "distinct interaction", but count a maximum of one "interaction" per day (e.g. if a person receives two diagnoses and a prescription on the same day, we assume this all comes from the same visit). We divide the number of distinct interactions by the number of years the person is observed within the dataset. We find that, relative to `Stanford-EHR` (EHRSHOT), the `CUMC-EHR` dataset has more patients with either extremely low or extremely high (e.g., every day) healthcare utilization.

D Models-specific preprocessing and pre-training

The following describes any model-specific preprocessing and parameters used for the models.

D.1 MEDS-processing

For each individual, medical events are extracted from all relevant OMOP domain tables, including `person`, `visit_occurrence`, `condition_occurrence`, `drug_exposure`, `procedure_occurrence`, `measurement`, `device_exposure`, `observation`, and `death`. Demographic information is placed at the beginning of each patient sequence, followed by clinical events ordered chronologically. For non-numeric records, the `subject_id`, `concept_code`, and `timestamp` are extracted into the MEDS event format. For numeric or categorical events—such as measurement records—associated values and units are stored in the `numeric_value`, `text_value`, and `unit` fields, respectively.

By design, MEDS-ETL prioritizes source concept codes over standard concept codes during conversion. Standardized concept codes come from vocabularies defined as standard by the OHDSI community. As a result, OMOP uses standard concept IDs (e.g., SNOMED for conditions), whereas MEDS uses source concept IDs (e.g., ICD-9/10). We use the MEDS format for MOTOR, FEMR, MEDS-TAB, and `Context Clues`, as these models were originally developed and extensively evaluated using the MEDS. Although CEHR-BERT and CEHR-GPT support the MEDS data format, the reliance on source concept codes may result in sub-optimal model performance due to the many-to-many mappings between terminologies. Moreover, the original authors of these models explicitly stated that they used standard concept codes extracted from OMOP, not source codes. For fairness, we use OMOP when training and evaluating CEHR-BERT and CEHR-GPT. Patient sequences used for pre-training and evaluation were extracted from OMOP using the respective original code bases[11] [56, 57]. EHR preprocessing for CORE-BEHRT is also conducted without the MEDS conversion using the custom processing and tokenization pipeline [54].

D.2 Encoding temporal information

The considered foundation models integrate temporal information using different encoding mechanisms. Table 6 details, for each temporal information, the associated encoding strategy.

MOTOR incorporates two timestamp features — absolute time and its square — by replacing the last two dimensions of the hidden states. It integrates sinusoidal age embeddings at each layer of the

[11] `https://github.com/knatarajan-lab/cehrbert_data`

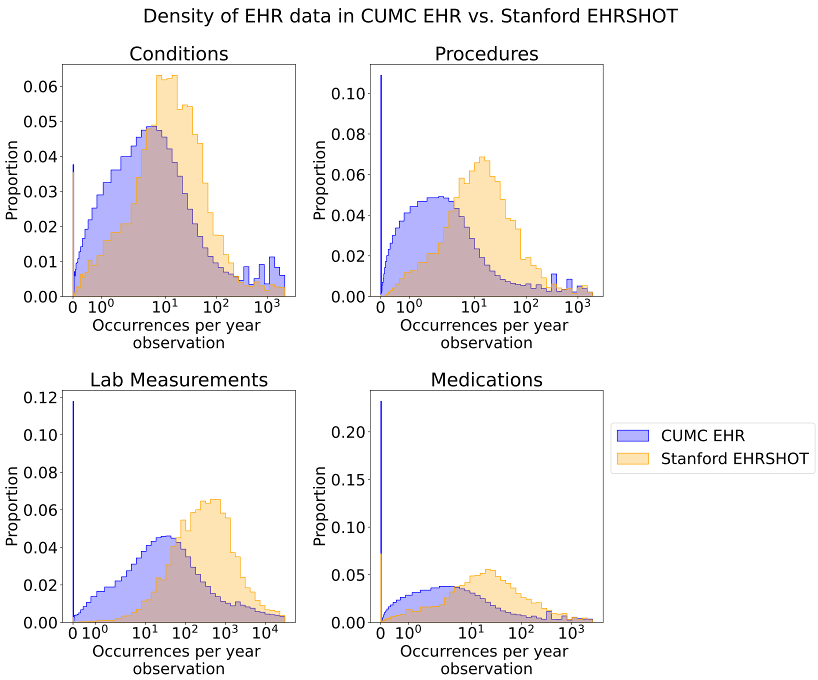

Figure 5. In order to compare the density of data at each institution, we measure the number of occurrences of each data type (a. conditions, b. medications, c. procedures, d. lab tests) per year of observation for each patient. Unlike with the healthcare utilization metric, we count multiple instances of a data type separately (e.g. all diagnoses made on the same day will contribute to the "density" measurement). We find that `Stanford-EHR` (EHRSHOT) data is more dense than `CUMC-EHR` data across all measured data types.

Table 6: Temporal Information Input and Encoding Mechanisms Across Models

Model	Age	Time Stamp	Inter-event	Position
MOTOR	Sinusoidal	Time Features	None	None
CEHR-BERT	Time2Vec	Time2Vec	Time token	Sinusoidal
CEHR-GPT	Age Token	None	Time token	None
CORE-BEHRT	Time2Vec	Time2Vec	None	Sinusoidal
MAMBA	None	None	None	None
LLAMA	None	None	None	Sinusoidal

Inter-event: time difference between consecutive events that occur on different days.

FEMRdecoder using RoPE, allowing age and time information to influence the model throughout its depth.

CEHR-BERT integrates timestamp, age, and visit positional embeddings through a feedforward network before passing them to the encoder layers. Additionally, it inserts time tokens between consecutive visits and between inpatient events occurring on different days to capture the temporal irregularity of EHR sequences.

CEHR-GPT encodes temporal information by augmenting patient sequences with explicit time-related tokens. It introduces a start-year token and an age token at the beginning of each trajectory, and inserts fine-grained day tokens between visits and between inpatient events on different days. These temporal embeddings—corresponding to year, age, and day tokens—are learned directly through the model's next-token prediction objective.

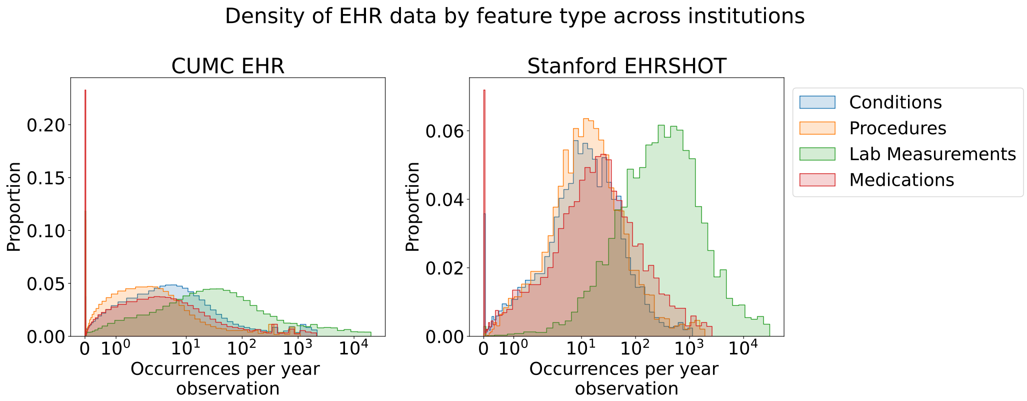

Figure 6. Using the same measure of density described in Figure 5, we compare the density of each feature type in a. [Institution] data and b. `Stanford-EHR`-sub data. At both institutions, the density of labs is much higher than any of the other feature types, with the difference in distributions being more pronounced at [Institution], as there are a larger number of patients who are missing certain feature types altogether (medications, in particular).

Core-Behrt adds time embeddings, age embeddings, and visit positional embeddings to the concept embeddings before passing them to the Bert encoder layers. Additionally, it applies RoPE positional embeddings to the hidden states at every encoder layer.

Llama does not incorporate age, absolute time, or inter-event temporal information, relying solely on the default RoPE positional embeddings. Mamba similarly does not encode any explicit temporal signals.

D.3 Context length selection

As tokenization choices result in different lengths of embedded data, we selected each model's context lengths based on the percentage of visits covered by a given choice. The following illustrates the frequency of visits for each number of tokens. Using these figures, we selected the context lengths presented in Table 1 to cover at least 99% of the visits.

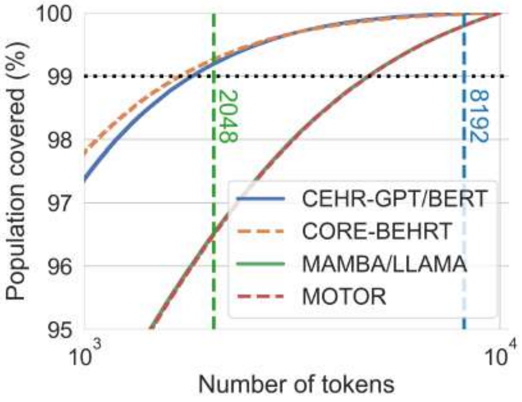

Figure 7: Percentage of the population covered given different context windows.

D.4 Other hyperparameters

All other pre-training hyperparameters are described in Table 7.

Table 7: Considered foundation models for electronic health records.

Hyperparameter	CEHR BERT	CEHR GPT	CORE BEHRT	LLAMA	MAMBA	MOTOR
Architecture	BERT	GPT	BERT	LLAMA	MAMBA	Custom
Embedding size	768	768	768	768	768	768
Learning rate	0.0002	0.0002	0.00005	0.0002	0.0002	0.00001
Weight decay	0.01	0.1	0.01	0.1	0.1	0.1
Optimizer	`Adam` $\lambda_1 = 0.9$ $\lambda_2 = 0.999$	`Adam` $\lambda_1 = 0.9$ $\lambda_2 = 0.95$	`AdamW` $\lambda_1 = 0.9$ $\lambda_2 = 0.999$	`AdamW` $\lambda_1 = 0.9$ $\lambda_2 = 0.95$	`AdamW` $\lambda_1 = 0.9$ $\lambda_2 = 0.95$	`AdamW` $\lambda_1 = 0.9$ $\lambda_2 = 0.95$

D.5 Baselines

FEMR-LightGBM Hyperparameter search for LightGBM is presented in Table 8, based on the initial code of FEMR[12].

Table 8: Hyperparameter Search Space for FEMR LightGBM

Parameter	Type	Range	Scale
`lambda_l1`	Float	$[10^{-8}, 10.0]$	Log
`lambda_l2`	Float	$[10^{-8}, 10.0]$	Log
`num_leaves`	Integer	$[2, 256]$	Linear
`feature_fraction`	Float	$[0.4, 1.0]$	Linear
`bagging_fraction`	Float	$[0.4, 1.0]$	Linear
`bagging_freq`	Integer	$[1, 7]$	Linear
`min_child_samples`	Integer	$[5, 100]$	Linear

MEDS-TAB- XGBoost Featurization, using MEDS-TAB 0.1, was performed on each task. Extracted features consist of counts, sums, squared sums, minimum and maximum values over multiple time windows of 1, 7, 30, 365 days, and the full length of stay prior to prediction times. Then, an XGBoost model was trained with hyperparameters selected over 100 draws from a search over parameters described in Table 9.

FEMR-Logistic Regression and MEDS-TAB-Logistic Regression A standard 10-fold cross-validation was employed to select the optimal inverse of the $l2$ penalty (λ), using values chosen on a logarithmic scale between 10^{-4} and 10^4.

E Compute resources

All experiments were run on a server with 4 NVIDIA H100 NVL GPUs, 2 Intel(R) Xeon(R) Platinum 8480+ CPUs (56 cores each) with 2Tb of memory. Table 10 describes the estimated training time for the final run. As these final numbers do not include prior iterations of the models, our total estimates are about $1,400$ compute hours.

Additionally, as computing resources may be a limitation for the application of foundation models in healthcare, Table 11 reports the number of FLOPs for inference corresponding to the computational requirement at deployment.

[12] `https://github.com/ChaoPang/femr/tree/omop_meds_v3_tutorial`

Table 9: Hyperparameter Search Space for Meds-Tab

Parameter	Type	Range	Scale
`model.eta`	Float	$[0.001, 1]$	Log
`model.lambda`	Float	$[0.001, 1]$	Log
`model.alpha`	Float	$[0.001, 1]$	Log
`model.subsample`	Float	$[0.5, 1]$	Linear
`model.min_child_weight`	Float	$[10^{-2}, 100]$	Linear
`model.max_depth`	Integer	$[2, 16]$	Linear
`training_params.num_boost_round`	Integer	$[100, 1000]$	Linear
`training_params.early_stopping_rounds`	Integer	$[1, 10]$	Linear
`tabularization.min_code_inclusion_count`	Integer	$[10, 1000000]$	Log

Table 10: Pretraining and linear probing total running time (in days).

Metrics	Baseline	Cehr Bert	Cehr Gpt	Motor	Llama	Mamba	Core Behrt
Final run	7	1	1.5	1.5	1.2	1.9	4
Prior experimentation	14	3	9	9	10	10	3

F Additional results

F.1 Average performance comparison

Section 6 presents the average ranking across tasks. Similarly, Table 12 presents the average performance across tasks, demonstrating the negative impact of a small training set on foundation models' performances.

Table 11: Inference FLOPs.

Metrics	**CEHR BERT**	**CEHR GPT**	**MOTOR**	**LLAMA**	**MAMBA**	**CORE BEHRT**
Inference FLOPs	$2.02e^{16}$	$1.78e^{16}$	$5.1e^{16}$	$7.53e^{16}$	$9.32e^{16}$	$1.52e^{16}$

Table 12. Average performance on the proposed dataset stratified per metrics, sorted by average ranking across all metrics.

Foundation	Large data regime			Small data regime		
Model	Calibration	Discrimination	Fairness	Calibration	Discrimination	Fairness
CEHR-GPT	0.04 (0.05)	0.79 (0.07)	0.06 (0.06)	0.06 (0.07)	0.68 (0.07)	0.06 (0.05)
MOTOR	0.04 (0.05)	0.78 (0.10)	0.08 (0.09)	0.06 (0.06)	0.66 (0.11)	0.08 (0.08)
MAMBA	0.05 (0.06)	0.76 (0.07)	0.06 (0.06)	0.06 (0.07)	0.63 (0.07)	0.06 (0.05)
Best Baseline	0.04 (0.05)	0.76 (0.09)	0.13 (0.10)	0.05 (0.07)	0.64 (0.09)	0.12 (0.09)
CEHR-BERT	0.05 (0.06)	0.75 (0.06)	0.08 (0.07)	0.06 (0.06)	0.67 (0.05)	0.08 (0.05)
CORE-BEHRT	0.05 (0.06)	0.75 (0.07)	0.06 (0.07)	0.06 (0.07)	0.65 (0.07)	0.07 (0.05)
LLAMA	0.05 (0.06)	0.71 (0.09)	0.08 (0.06)	0.06 (0.07)	0.60 (0.06)	0.08 (0.07)
MAMBA-TRANSPORT	0.05 (0.06)	0.70 (0.08)	0.07 (0.06)	0.08 (0.11)	0.61 (0.06)	0.06 (0.06)

F.2 Task-specific performances

The main text focuses on the average performance across tasks. The following results present detailed figures for all outcomes of interest. Specifically, Figures present the discriminative and calibration for all tasks and models given increasing amounts of training data, while Figures present the fairness discriminative gap for all tasks stratified by sex, race[13], and healthcare utilization.

[13] Due to the granularity of available demographic data, we consider two groups of interest: Black and White patients.

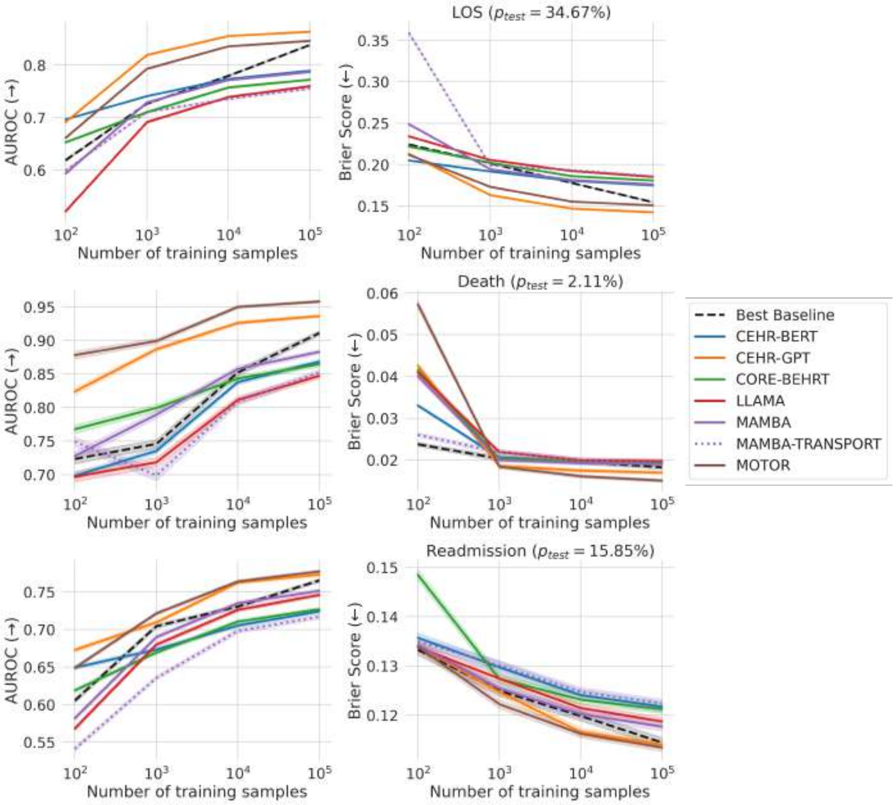

Figure 8. Discriminative performances and calibration results for all outcomes for increasing number of training points (Shaded area represents bootstrapped standard deviation).

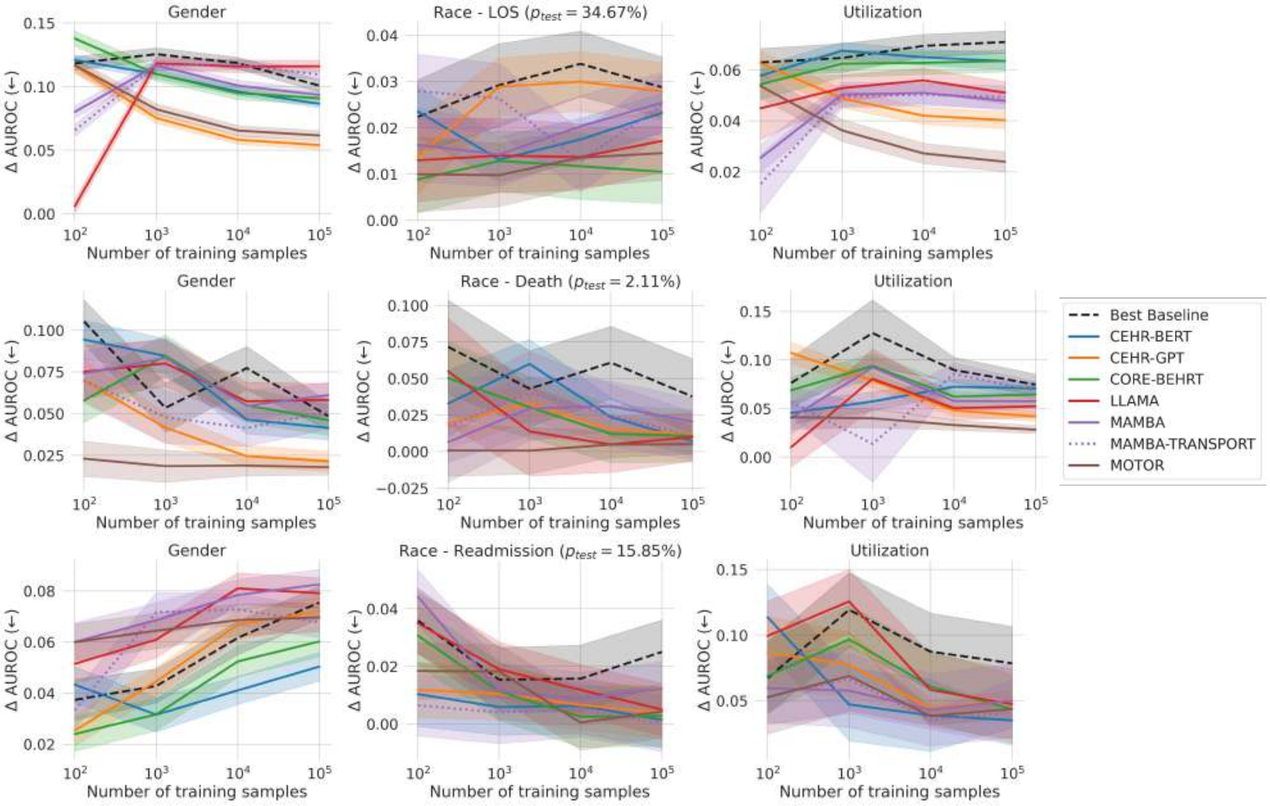

Figure 9. Discriminative performances and calibration results for all outcomes for increasing number of training points (Shaded area represents bootstrapped standard deviation).

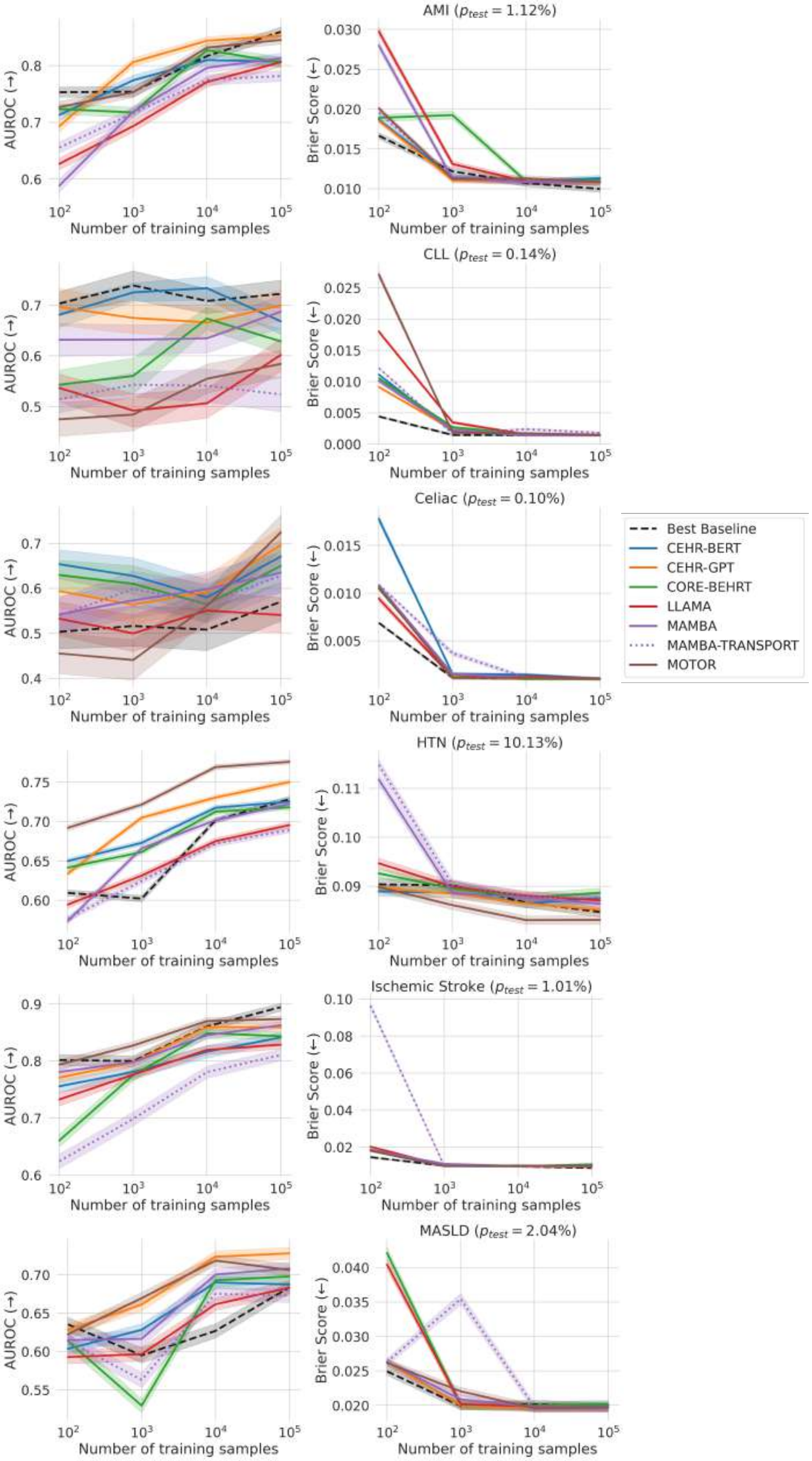

Figure 10. Discriminative performances and calibration results for all phenotypes for increasing number of training points (Shaded area represents bootstrapped standard deviation).

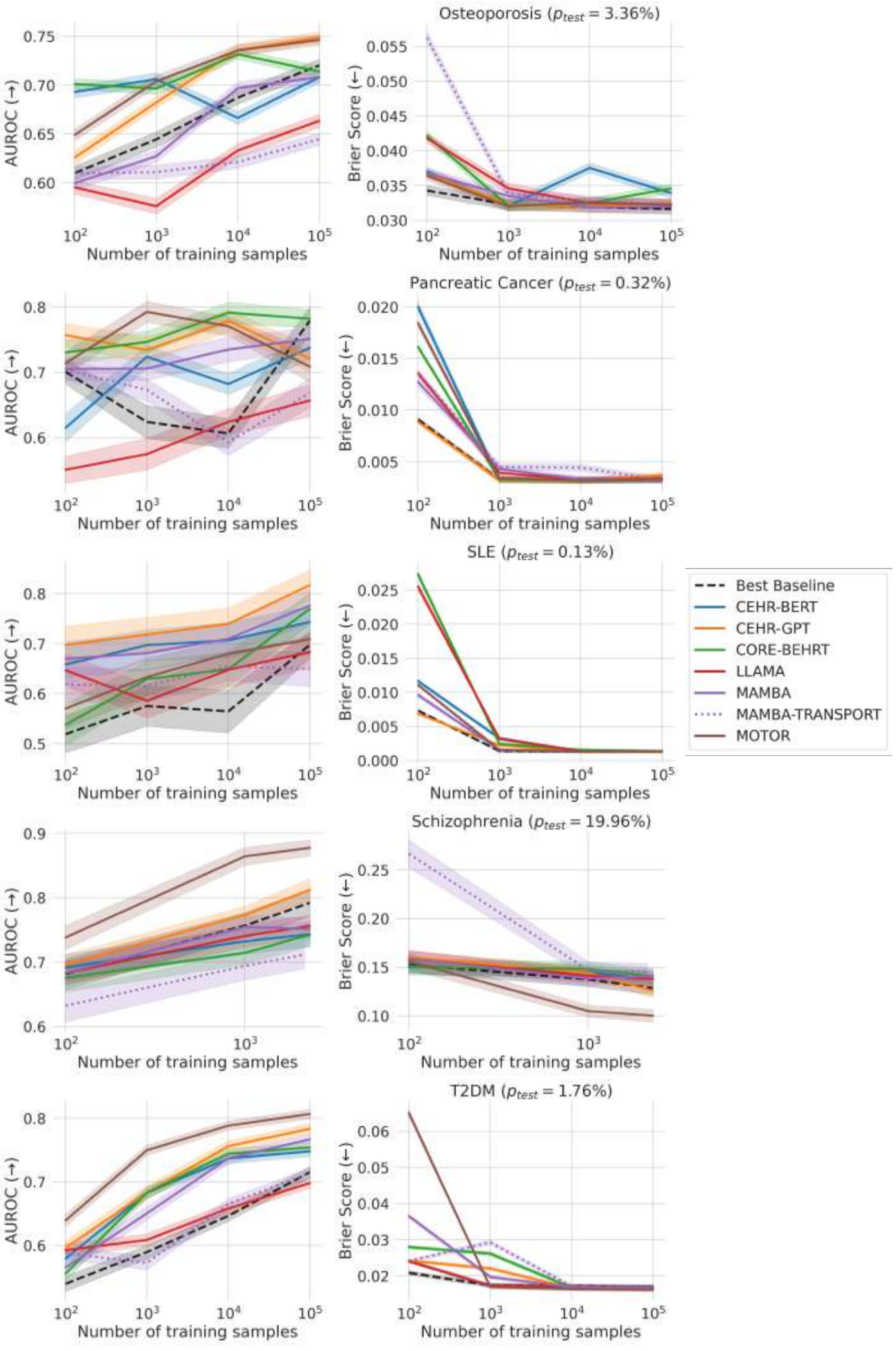

Figure 11. Discriminative performances and calibration results for all phenotypes for increasing number of training points (Shaded area represents bootstrapped standard deviation) - Continued.

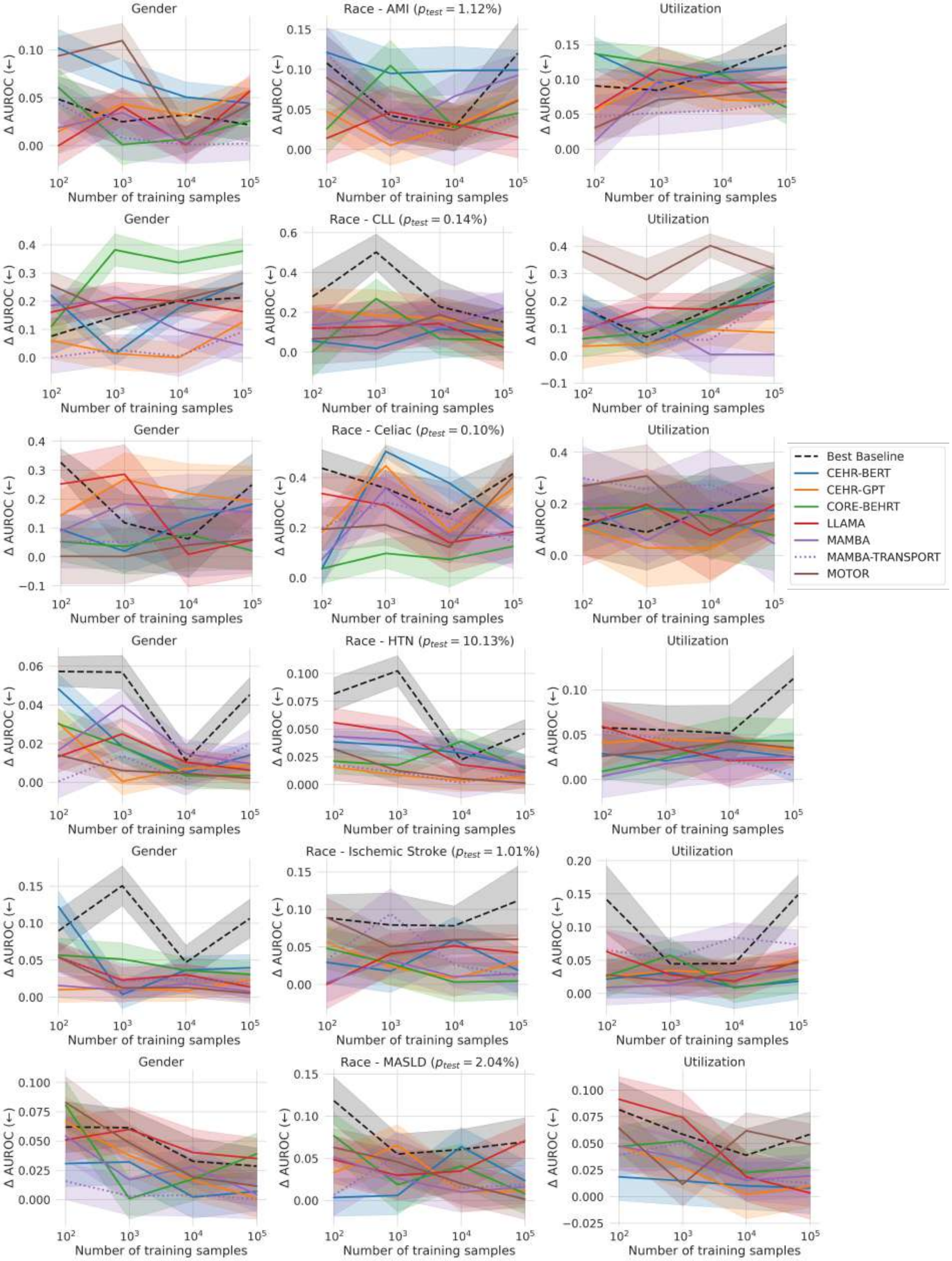

Figure 12. Discriminative performances and calibration results for all phenotypes for increasing number of training points (Shaded area represents bootstrapped standard deviation).

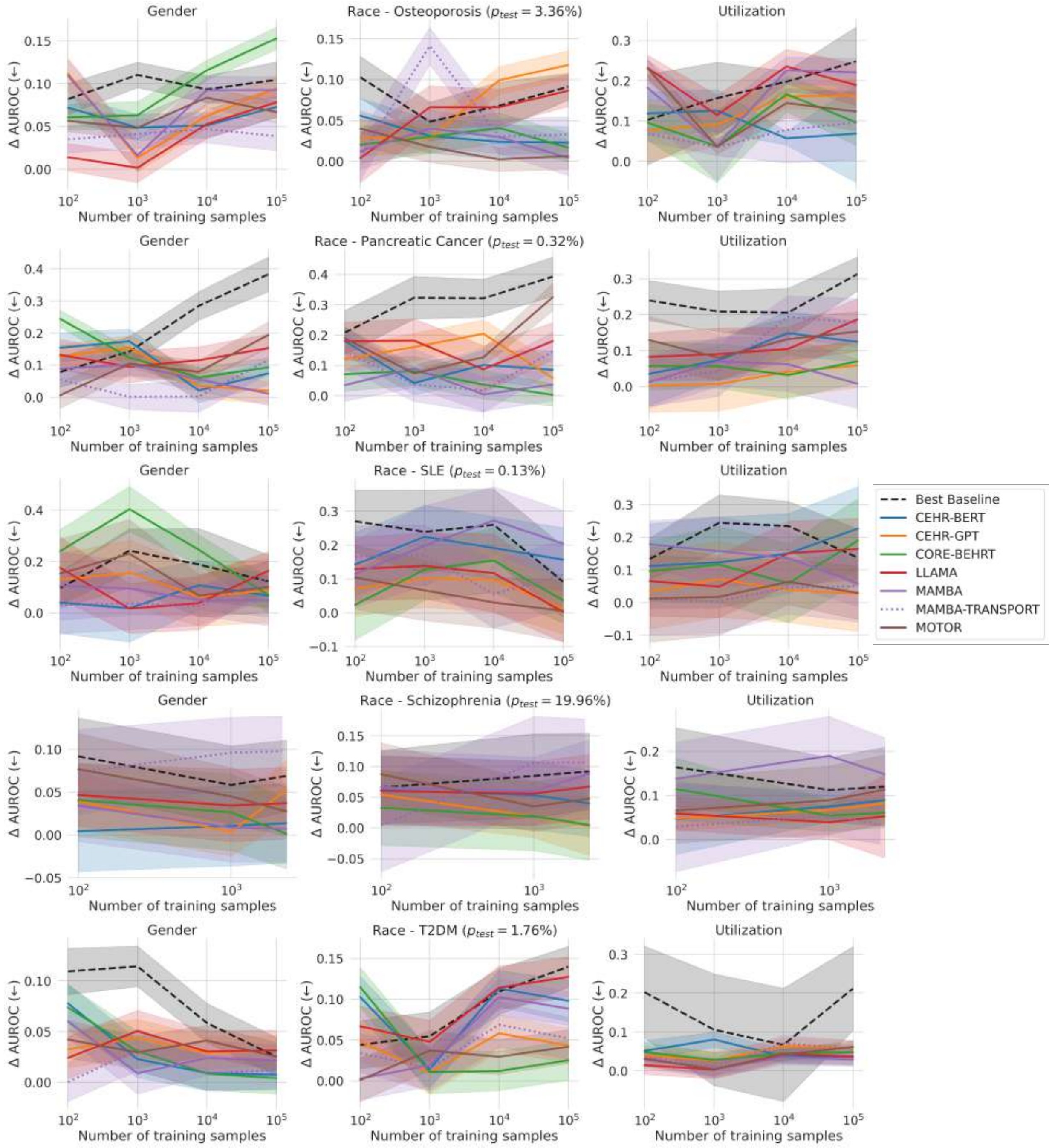

Figure 13. Discriminative performances and calibration results for all phenotypes for increasing number of training points (Shaded area represents bootstrapped standard deviation) - Continued.

F.3 Performances using the largest training set

Tables 13 and 14 present the discriminative tasks over all considered tasks when using all extracted training points. Similarly Tables 15 and 16 present the calibration. Tables 17 and 18, Tables 19 and 20 and Tables 21 and 22 respectively present the maximal discriminative performances for the sex, ethnicity and utilization splits.

Table 13: Outcome discriminative performance results. *Best performances are marked in **bold**.*

	Best Baseline	**CEHR BERT**	**CEHR GPT**	**CORE BEHRT**	**LLAMA**	**MAMBA**	**MAMBA-TRANSPORT**	**MOTOR**
Long LOS	0.838 (0.002)	0.789 (0.001)	**0.864** (0.001)	0.772 (0.002)	0.760 (0.002)	0.787 (0.002)	0.756 (0.002)	0.846 (0.002)
Mortality	0.911 (0.004)	0.868 (0.004)	0.936 (0.003)	0.864 (0.004)	0.847 (0.005)	0.883 (0.004)	0.852 (0.004)	**0.958** (0.002)
Readmission	0.765 (0.003)	0.724 (0.003)	0.773 (0.003)	0.727 (0.003)	0.746 (0.003)	0.752 (0.003)	0.717 (0.003)	**0.777** (0.003)

Table 14: Phenotype discriminative performance results. *Best performances are marked in **bold**.*

Phenotype	**Best Baseline**	**CEHR BERT**	**CEHR GPT**	**CORE BEHRT**	**LLAMA**	**MAMBA**	**MAMBA-TRANSPORT**	**MOTOR**
AMI	**0.859** (0.008)	0.807 (0.008)	0.853 (0.007)	0.805 (0.010)	0.806 (0.008)	0.813 (0.009)	0.782 (0.009)	0.845 (0.007)
CLL	**0.722** (0.027)	0.668 (0.025)	0.699 (0.031)	0.628 (0.025)	0.602 (0.032)	0.687 (0.032)	0.524 (0.034)	0.584 (0.027)
Celiac	0.570 (0.043)	0.671 (0.036)	0.696 (0.039)	0.650 (0.034)	0.541 (0.039)	0.635 (0.045)	0.628 (0.038)	**0.724** (0.038)
HTN	0.728 (0.004)	0.725 (0.004)	0.750 (0.003)	0.718 (0.003)	0.696 (0.004)	0.722 (0.004)	0.689 (0.004)	**0.776** (0.003)
Ischemic Stroke	**0.894** (0.007)	0.841 (0.009)	0.859 (0.007)	0.844 (0.007)	0.829 (0.008)	0.863 (0.006)	0.811 (0.010)	0.873 (0.006)
MASLD	0.684 (0.008)	0.687 (0.009)	**0.728** (0.008)	0.698 (0.008)	0.683 (0.009)	0.709 (0.008)	0.673 (0.009)	0.706 (0.008)
Osteoporosis	0.720 (0.006)	0.708 (0.006)	**0.749** (0.006)	0.714 (0.006)	0.663 (0.007)	0.708 (0.006)	0.645 (0.006)	0.746 (0.005)
Pancreatic Cancer	0.779 (0.021)	0.737 (0.018)	0.721 (0.020)	**0.782** (0.014)	0.657 (0.024)	0.751 (0.017)	0.669 (0.022)	0.709 (0.024)
SLE	0.697 (0.027)	0.743 (0.032)	**0.817** (0.030)	0.769 (0.031)	0.683 (0.031)	0.776 (0.034)	0.649 (0.035)	0.709 (0.030)
Schizophrenia	0.792 (0.021)	0.743 (0.020)	0.812 (0.017)	0.742 (0.017)	0.756 (0.016)	0.751 (0.017)	0.712 (0.022)	**0.877** (0.012)
T2DM	0.715 (0.009)	0.748 (0.008)	0.783 (0.008)	0.754 (0.008)	0.698 (0.008)	0.767 (0.007)	0.714 (0.008)	**0.807** (0.007)

AMI: Acute myocardial infarction. CLL: Chronic lymphocytic leukemia. HTN: Hypertension. MASLD: Metabolic dysfunction-associated steatotic liver disease. SLE: Systemic lupus erythematosus. T2DM: Type 2 diabetes mellitus.

Table 15: Outcome calibration results. *Best performances are marked in **bold**.*

	Best Baseline	CEHR BERT	CEHR GPT	CORE BEHRT	LLAMA	MAMBA	MAMBA-TRANSPORT	MOTOR
Long LOS	0.155 (0.001)	0.175 (0.001)	**0.142** (0.001)	0.181 (0.001)	0.185 (0.001)	0.176 (0.001)	0.186 (0.001)	0.151 (0.001)
Mortality	0.018 (0.001)	0.019 (0.000)	0.017 (0.000)	0.019 (0.001)	0.020 (0.001)	0.019 (0.001)	0.019 (0.001)	**0.015** (0.000)
Readmission	0.114 (0.001)	0.122 (0.001)	0.114 (0.001)	0.121 (0.001)	0.119 (0.001)	0.118 (0.001)	0.122 (0.001)	**0.113** (0.001)

Table 16: Phenotype calibration results. *Best performances are marked in **bold**.*

Phenotype	Best Baseline	CEHR BERT	CEHR GPT	CORE BEHRT	LLAMA	MAMBA	MAMBA TRANSPORT	MOTOR
AMI	**0.010** (0.000)	0.011 (0.000)	0.011 (0.000)	0.011 (0.000)	0.011 (0.000)	0.011 (0.000)	0.011 (0.000)	0.011 (0.000)
CLL	**0.001** (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)	0.002 (0.000)	0.001 (0.000)
Celiac	0.001 (0.000)	0.001 (0.000)	**0.001** (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)
HTN	0.085 (0.001)	0.088 (0.001)	0.085 (0.001)	0.089 (0.001)	0.087 (0.001)	0.086 (0.001)	0.088 (0.001)	**0.083** (0.001)
Ischemic Stroke	**0.009** (0.000)	0.011 (0.000)	0.010 (0.000)	0.011 (0.000)	0.010 (0.000)	0.010 (0.000)	0.010 (0.000)	0.010 (0.000)
MASLD	0.020 (0.001)	0.020 (0.001)	0.020 (0.001)	0.020 (0.001)	0.020 (0.001)	**0.020** (0.001)	0.020 (0.001)	0.020 (0.001)
Osteoporosis	**0.032** (0.001)	0.034 (0.001)	0.032 (0.001)	0.035 (0.001)	0.032 (0.001)	0.032 (0.001)	0.032 (0.001)	0.032 (0.001)
Pancreatic Cancer	0.003 (0.000)	0.003 (0.000)	0.004 (0.000)	0.003 (0.000)	0.003 (0.000)	0.003 (0.000)	**0.003** (0.000)	0.003 (0.000)
SLE	0.001 (0.000)	0.001 (0.000)	**0.001** (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)
Schizophrenia	0.128 (0.007)	0.138 (0.008)	0.126 (0.007)	0.141 (0.007)	0.138 (0.007)	0.141 (0.008)	0.145 (0.009)	**0.100** (0.006)
T2DM	0.017 (0.001)	0.017 (0.000)	0.016 (0.000)	0.017 (0.000)	0.017 (0.001)	0.017 (0.000)	0.017 (0.001)	**0.016** (0.000)

AMI: Acute myocardial infarction. CLL: Chronic lymphocytic leukemia. HTN: Hypertension. MASLD: Metabolic dysfunction-associated steatotic liver disease. SLE: Systemic lupus erythematosus. T2DM: Type 2 diabetes mellitus.

Table 17: Outcome sex gap in discriminative performance. *Best performances are marked in **bold**.*

	Best Baseline	CEHR BERT	CEHR GPT	CORE BEHRT	LLAMA	MAMBA	MAMBA-TRANSPORT	MOTOR
Long LOS	0.101 (0.005)	0.086 (0.003)	**0.054** (0.003)	0.091 (0.005)	0.116 (0.005)	0.093 (0.004)	0.109 (0.004)	0.062 (0.004)
Mortality	0.049 (0.008)	0.041 (0.007)	0.021 (0.006)	0.046 (0.009)	0.058 (0.010)	0.061 (0.008)	0.051 (0.010)	**0.018** (0.005)
Readmission	0.075 (0.006)	**0.050** (0.006)	0.072 (0.006)	0.060 (0.006)	0.079 (0.006)	0.083 (0.006)	0.068 (0.006)	0.070 (0.006)

Table 18: Phenotype sex gap in discriminative performance.*Best performances are marked in **bold**.*

Phenotype	Best Baseline	CEHR BERT	CEHR GPT	CORE BEHRT	LLAMA	MAMBA	MAMBA-TRANSPORT	MOTOR
AMI	0.022 (0.017)	0.044 (0.016)	0.057 (0.013)	0.026 (0.016)	0.057 (0.017)	0.044 (0.015)	**0.002** (0.017)	0.043 (0.014)
CLL	0.213 (0.041)	0.264 (0.043)	0.125 (0.063)	0.378 (0.045)	0.163 (0.060)	**0.045** (0.063)	0.091 (0.069)	0.261 (0.050)
Celiac	0.249 (0.106)	0.182 (0.099)	0.192 (0.122)	**0.021** (0.062)	0.058 (0.124)	0.155 (0.127)	0.092 (0.147)	0.061 (0.106)
HTN	0.045 (0.009)	0.014 (0.007)	0.009 (0.007)	0.004 (0.007)	0.006 (0.007)	0.010 (0.007)	0.020 (0.007)	**0.002** (0.006)
Ischemic Stroke	0.106 (0.026)	0.040 (0.018)	0.023 (0.015)	0.031 (0.017)	0.014 (0.016)	0.008 (0.013)	0.019 (0.018)	**0.005** (0.013)
MASLD	0.029 (0.015)	0.007 (0.017)	0.001 (0.016)	0.039 (0.018)	0.035 (0.017)	0.004 (0.015)	**0.001** (0.017)	0.011 (0.017)
Osteoporosis	0.104 (0.021)	0.073 (0.012)	0.094 (0.015)	0.153 (0.013)	0.078 (0.017)	0.093 (0.016)	**0.039** (0.017)	0.067 (0.012)
Pancreatic Cancer	0.383 (0.054)	0.074 (0.037)	0.020 (0.038)	0.093 (0.030)	0.153 (0.043)	**0.011** (0.037)	0.116 (0.043)	0.194 (0.041)
SLE	0.124 (0.103)	0.067 (0.081)	0.087 (0.125)	0.069 (0.068)	0.163 (0.075)	**0.049** (0.099)	0.134 (0.072)	0.101 (0.099)
Schizophrenia	0.069 (0.041)	0.014 (0.046)	0.052 (0.035)	**0.001** (0.041)	0.037 (0.042)	0.007 (0.039)	0.098 (0.041)	0.028 (0.028)
T2DM	0.025 (0.019)	0.008 (0.015)	0.032 (0.014)	**0.004** (0.015)	0.031 (0.020)	0.021 (0.016)	0.012 (0.018)	0.025 (0.014)

AMI: Acute myocardial infarction. CLL: Chronic lymphocytic leukemia. HTN: Hypertension. MASLD: Metabolic dysfunction-associated steatotic liver disease. SLE: Systemic lupus erythematosus. T2DM: Type 2 diabetes mellitus.

Table 19: Outcome racial gap in discriminative performance. *Best performances are marked in **bold**.*

	Best Baseline	CEHR BERT	CEHR GPT	CORE BEHRT	LLAMA	MAMBA	MAMBA TRANSPORT	MOTOR
Long LOS	0.029 (0.007)	0.023 (0.004)	0.028 (0.006)	**0.010** (0.007)	0.017 (0.007)	0.025 (0.007)	0.025 (0.006)	0.015 (0.006)
Mortality	0.038 (0.026)	0.007 (0.010)	0.012 (0.008)	0.011 (0.018)	0.010 (0.017)	0.021 (0.016)	0.011 (0.016)	**0.005** (0.007)
Readmission	0.025 (0.011)	0.001 (0.010)	0.004 (0.009)	0.003 (0.010)	0.005 (0.009)	0.012 (0.009)	**0.000** (0.010)	0.004 (0.009)

Table 20: Phenotype racial gap in discriminative performance.*Best performances are marked in **bold**.*

Phenotype	Best Baseline	CEHR BERT	CEHR GPT	CORE BEHRT	LLAMA	MAMBA	MAMBA TRANSPORT	MOTOR
AMI	0.120 (0.037)	0.099 (0.025)	0.063 (0.025)	0.046 (0.029)	**0.015** (0.025)	0.093 (0.027)	0.043 (0.030)	0.061 (0.028)
CLL	0.151 (0.083)	0.099 (0.112)	0.111 (0.087)	0.061 (0.108)	**0.025** (0.112)	0.220 (0.086)	0.191 (0.100)	0.083 (0.124)
Celiac	0.416 (0.073)	0.204 (0.044)	0.356 (0.078)	**0.126** (0.058)	0.183 (0.096)	0.172 (0.116)	0.144 (0.106)	0.405 (0.098)
HTN	0.046 (0.012)	0.017 (0.011)	0.007 (0.010)	0.014 (0.012)	0.011 (0.012)	0.016 (0.011)	0.010 (0.012)	**0.001** (0.010)
Ischemic Stroke	0.111 (0.046)	0.019 (0.031)	0.030 (0.022)	**0.005** (0.024)	0.042 (0.023)	0.014 (0.024)	0.011 (0.027)	0.060 (0.020)
MASLD	0.069 (0.029)	0.024 (0.024)	0.011 (0.020)	0.008 (0.024)	0.071 (0.020)	0.016 (0.023)	0.019 (0.022)	**0.001** (0.024)
Osteoporosis	0.091 (0.016)	0.023 (0.019)	0.118 (0.017)	0.017 (0.018)	0.087 (0.019)	**0.004** (0.022)	0.033 (0.021)	0.007 (0.016)
Pancreatic Cancer	0.392 (0.064)	0.086 (0.062)	0.059 (0.054)	**0.003** (0.036)	0.180 (0.061)	0.037 (0.053)	0.146 (0.064)	0.325 (0.045)
SLE	0.092 (0.105)	0.157 (0.094)	0.012 (0.074)	0.037 (0.067)	**0.003** (0.090)	0.204 (0.098)	0.112 (0.102)	0.006 (0.090)
Schizophrenia	0.092 (0.062)	0.040 (0.054)	**0.003** (0.048)	0.005 (0.057)	0.068 (0.052)	0.089 (0.055)	0.107 (0.071)	0.048 (0.040)
T2DM	0.140 (0.025)	0.098 (0.025)	0.043 (0.021)	**0.025** (0.025)	0.127 (0.025)	0.089 (0.020)	0.052 (0.026)	0.042 (0.021)

AMI: Acute myocardial infarction. CLL: Chronic lymphocytic leukemia. HTN: Hypertension. MASLD: Metabolic dysfunction-associated steatotic liver disease. SLE: Systemic lupus erythematosus. T2DM: Type 2 diabetes mellitus.

Table 21. Outcome healthcare utilization gap in discriminative performance. *Best performances are marked in **bold**.*

	Best Baseline	CEHR BERT	CEHR GPT	CORE BEHRT	LLAMA	MAMBA	MAMBA TRANSPORT	MOTOR
Long LOS	0.071 (0.004)	0.063 (0.003)	0.040 (0.003)	0.064 (0.004)	0.051 (0.004)	0.048 (0.004)	0.049 (0.004)	**0.024** (0.004)
Mortality	0.075 (0.011)	0.071 (0.006)	0.042 (0.005)	0.064 (0.010)	0.052 (0.011)	0.057 (0.008)	0.068 (0.009)	**0.028** (0.004)
Readmission	0.078 (0.028)	**0.035** (0.006)	0.050 (0.007)	0.044 (0.026)	0.047 (0.027)	0.050 (0.028)	0.041 (0.025)	0.044 (0.007)

Table 22. Phenotype healthcare utilization gap in discriminative performance.*Best performances are marked in **bold**.*

Phenotype	Best Baseline	CEHR BERT	CEHR GPT	CORE BEHRT	LLAMA	MAMBA	MAMBA TRANSPORT	MOTOR
AMI	0.149 (0.032)	0.117 (0.018)	0.068 (0.017)	**0.060** (0.024)	0.096 (0.022)	0.081 (0.019)	0.068 (0.021)	0.087 (0.019)
CLL	0.267 (0.066)	0.253 (0.054)	0.085 (0.077)	0.267 (0.053)	0.197 (0.064)	**0.005** (0.081)	0.225 (0.070)	0.319 (0.055)
Celiac	0.262 (0.100)	0.175 (0.095)	0.154 (0.079)	0.077 (0.136)	0.195 (0.141)	**0.046** (0.151)	0.169 (0.110)	0.141 (0.058)
HTN	0.112 (0.026)	0.025 (0.028)	0.033 (0.007)	0.043 (0.024)	0.022 (0.008)	0.026 (0.007)	**0.005** (0.007)	0.035 (0.006)
Ischemic Stroke	0.149 (0.029)	**0.018** (0.027)	0.050 (0.020)	0.022 (0.020)	0.048 (0.023)	0.035 (0.022)	0.074 (0.022)	0.046 (0.016)
MASLD	0.058 (0.021)	0.008 (0.024)	0.010 (0.020)	0.027 (0.020)	**0.003** (0.024)	0.019 (0.020)	0.013 (0.022)	0.049 (0.020)
Osteoporosis	0.248 (0.085)	**0.069** (0.120)	0.164 (0.041)	0.097 (0.057)	0.189 (0.052)	0.221 (0.035)	0.097 (0.093)	0.124 (0.048)
Pancreatic Cancer	0.313 (0.048)	0.124 (0.057)	0.059 (0.062)	0.070 (0.060)	0.187 (0.058)	**0.008** (0.068)	0.178 (0.068)	0.153 (0.056)
SLE	0.137 (0.087)	0.227 (0.128)	**0.027** (0.116)	0.190 (0.129)	0.164 (0.094)	0.059 (0.071)	0.051 (0.110)	0.029 (0.081)
Schizophrenia	0.120 (0.089)	0.089 (0.035)	0.082 (0.033)	0.061 (0.034)	0.053 (0.094)	0.148 (0.083)	**0.035** (0.044)	0.113 (0.078)
T2DM	0.212 (0.108)	0.036 (0.022)	0.059 (0.018)	0.048 (0.021)	0.036 (0.019)	**0.029** (0.017)	0.057 (0.021)	0.062 (0.015)

AMI: Acute myocardial infarction. CLL: Chronic lymphocytic leukemia. HTN: Hypertension. MASLD: Metabolic dysfunction-associated steatotic liver disease. SLE: Systemic lupus erythematosus. T2DM: Type 2 diabetes mellitus.

F.4 Detailed baseline results

For completeness, Tables 23 and 24 detail the discriminative performances for each considered baseline model. All previous tables and figures reflect the best performance across these different models.

Table 23: Outcome prediction results for considered baselines.

Task	Femr- GBM	Femr- LR	Meds-Tab- XGB	Meds-Tab- LR
Long LOS	0.808 (0.002)	0.727 (0.003)	**0.838** (0.002)	0.801 (0.002)
Mortality	0.896 (0.004)	0.833 (0.006)	**0.911** (0.004)	0.880 (0.006)
Readmission	0.745 (0.003)	0.682 (0.003)	**0.765** (0.003)	0.721 (0.003)

LR = Logistic Regression, GBM = LightGBM, XGB = XGBoost

Table 24: Phenotype prediction results for considered baselines.

Phenotype	Femr- GBM	Femr- LR	Meds-Tab- XGB	Meds-Tab- LR
AMI	0.775 (0.009)	0.669 (0.014)	**0.859** (0.008)	0.751 (0.011)
CLL	0.489 (0.035)	**0.722** (0.027)	0.666 (0.025)	0.716 (0.037)
Celiac	**0.570** (0.043)	0.529 (0.041)	0.495 (0.035)	0.547 (0.041)
HTN	0.699 (0.004)	0.579 (0.005)	**0.728** (0.004)	0.562 (0.005)
Ischemic Stroke	0.850 (0.008)	0.723 (0.013)	**0.894** (0.007)	0.710 (0.014)
MASLD	0.668 (0.009)	0.538 (0.009)	**0.684** (0.008)	0.564 (0.010)
Osteoporosis	0.673 (0.006)	0.530 (0.009)	**0.720** (0.006)	0.581 (0.008)
Pancreatic Cancer	0.671 (0.025)	0.467 (0.027)	**0.779** (0.021)	0.533 (0.028)
SLE	0.696 (0.032)	0.399 (0.031)	**0.697** (0.027)	0.437 (0.029)
Schizophrenia	0.742 (0.021)	0.745 (0.020)	**0.792** (0.021)	0.730 (0.020)
T2DM	**0.715** (0.009)	0.577 (0.012)	0.705 (0.008)	0.630 (0.011)

LR = Logistic Regression, GBM = LightGBM, XGB = XGBoost

F.5 Importance of input modalities

Our main results highlight that some foundation models appear to benefit from ignoring lab data. To investigate this point further, we propose an additional experiment in which we retrain LLAMA with and without this input, to measure the impact of this modality on the quality of the extracted representations. Tables 25 and 26 present the associated results when considering the largest set of training points for linear probing. While only focusing on one architecture, these results provide further evidence of the limited values of laboratory modality when using linear probing for downstream tasks. This observation may be explained by the fixed embedding dimensions: more input information needs to be included in the same representation. Our future work will investigate this problem further.

Table 25: Outcome discriminative performance results for LLAMAwith and without lab values.

	LLAMA	LLAMA (NO LAB)
Long LOS	0.760 (0.002)	**0.798** (0.002)
Mortality	0.847 (0.005)	**0.900** (0.004)
Readmission	0.746 (0.003)	**0.754** (0.003)

Table 26: Phenotype discriminative performance results for LLAMAwith and without lab values.

	LLAMA	LLAMA (NO LAB)
AMI	**0.806** (0.008)	0.797 (0.008)
CLL	**0.602** (0.032)	0.596 (0.034)
Celiac	0.541 (0.039)	**0.610** (0.038)
HTN	0.696 (0.004)	**0.702** (0.003)
Ischemic Stroke	0.829 (0.008)	**0.840** (0.007)
MASLD	**0.683** (0.009)	0.675 (0.008)
Osteoporosis	0.663 (0.007)	**0.687** (0.007)
Pancreatic Cancer	**0.657** (0.024)	0.613 (0.023)
SLE	**0.683** (0.031)	0.644 (0.034)
Schizophrenia	0.756 (0.016)	**0.761** (0.019)
T2DM	**0.698** (0.008)	0.696 (0.009)