

# Generative Medical Event Models Improve with Scale

Shane Waxler<sup>\*1</sup>, Paul Blazek<sup>\*1</sup>, Davis White<sup>\*1</sup>, Daniel Sneider<sup>\*1</sup>, Kevin Chung<sup>1</sup>, Mani Nagarathnam<sup>1</sup>, Patrick Williams<sup>1</sup>, Hank Voeller<sup>1</sup>, Karen Wong<sup>1</sup>, Matthew Swanhorst<sup>1</sup>, Sheng Zhang<sup>2</sup>, Naoto Usuyama<sup>2</sup>, Cliff Wong<sup>2</sup>, Tristan Naumann<sup>2</sup>, Hoifung Poon<sup>2</sup>, Andrew Loza<sup>3</sup>, Daniella Meeker<sup>3, 4</sup>, Seth Hain<sup>1</sup>, and Rahul Shah<sup>†1</sup>

<sup>1</sup>Epic Systems

<sup>2</sup>Microsoft Research

<sup>3</sup>Yale School of Medicine

<sup>4</sup>Cosmos Governing Council

## Abstract

Realizing personalized medicine at scale calls for methods that distill insights from longitudinal patient journeys, which can be viewed as a sequence of medical events. Foundation models pretrained on large-scale medical event data represent a promising direction for scaling real-world evidence generation and generalizing to diverse downstream tasks. Using Epic Cosmos, a dataset with medical events from de-identified longitudinal health records for 16.3 billion encounters over 300 million unique patient records from 310 health systems, we introduce the Cosmos Medical Event Transformer (CoMET) models, a family of decoder-only transformer models pretrained on 118 million patients representing 115 billion discrete medical events (151 billion tokens). We present the largest scaling-law study for medical event data, establishing a methodology for pretraining and revealing power-law scaling relationships for compute, tokens, and model size. Based on this, we pretrained a series of compute-optimal models with up to 1 billion parameters. Conditioned on a patient’s real-world history, CoMET autoregressively generates the next medical event, simulating patient health timelines. We studied 78 real-world tasks, including diagnosis prediction, disease prognosis, and healthcare operations. Remarkably for a foundation model with generic pretraining and simulation-based inference, CoMET generally outperformed or matched task-specific supervised models on these tasks, without requiring task-specific fine-tuning or few-shot examples. CoMET’s predictive power consistently improves as the model and pretraining scale. Our results show that CoMET, a generative medical event foundation model, can effectively capture complex clinical dynamics, providing an extensible and generalizable framework to support clinical decision-making, streamline healthcare operations, and improve patient outcomes.

## 1 Introduction

Safe and effective medical care aims to deliver the right intervention to the right patient at the right time. In pursuit of this goal, patients, clinicians, and health system leaders seek consensus-driven guidelines, integrated data sources, and richer information that captures the full diversity of real-world healthcare. Optimal health outcomes require care that excels across, at minimum, four pillars: accurate diagnosis, reliable prognosis, individualized treatment planning, and optimized clinical workflow [1]. Succeeding across each of these pillars requires understanding a patient’s longitudinal medical history, addressing diagnostic and future uncertainty, incorporating patient values and goals, and adapting reasoning to temporal and clinical contexts.

Real-world data (RWD) and real-world evidence (RWE) offer a scalable path to personalized medical care. RWD-driven insights already inform post-market safety surveillance, regulatory approvals, and guide

---

<sup>\*</sup>Co-first authors

<sup>†</sup>Corresponding author: [rahul@epic.com](mailto:rahul@epic.com)

therapeutic strategies for complex chronic diseases [2–5]. Today, using RWD to generate RWE at scale demands significant analytic expertise and manual curation, constraining its day-to-day impact at the point of care [2, 4]. Unlocking its full potential will require methods that can transform raw data into actionable insights at the point of care in a scalable, generalizable, and personalized way.

Epic Cosmos<sup>1</sup> was created to address these challenges. A collaboration among health systems using Epic that is governed by a peer-elected council of participants, Cosmos aggregates de-identified longitudinal health records for more than 300 million patients and 16.3 billion encounters as of August 2025, de-duplicating each patient’s records across health systems and combining them into a single, integrated longitudinal record. This platform unifies common clinical data—including laboratory results, diagnoses, medications, and procedures—and includes other data relevant to health, such as social drivers of health, cancer staging, genomic variants, and patient-reported outcomes, among many other data types. The de-identified data in Cosmos is intended to support patient care and accelerate scientific discovery. Insights from Cosmos are delivered to clinicians today at the point of care through features in Epic like the Cosmos Median Length of Stay, Look-Alikes, and Best Care Choices for My Patient™. Cosmos data and its downstream applications are only made available to health systems that contribute data to it. Cosmos has also been used to address a wide variety of research priorities [6] such as understanding large trends in healthcare [7, 8], investigating rare diseases [9, 10], and analyzing healthcare utilization [11, 12].

Yet even at the scale of Cosmos, answering a single clinical question requires crafting custom cohort definitions, feature-engineering pipelines, and statistical analyses. To enable personalized medicine and RWE at scale for routine clinical decision-making, we need tools that can learn from the integrated patient record and can answer complex medical inquiries with flexibility, retrieving the right RWE to support clinical decision-making across a wide variety of contexts.

Foundation models pretrained on real-world patient journeys have shown promise in addressing this problem, where a patient journey is formulated as a sequence of medical events. By learning latent representations of complete patient records, generative medical event models can provide patient-specific predictions via simulated health timelines. By simulating multiple probabilistic timelines of a patient’s health, quantitative predictions can be made about the likelihood of events over specific time intervals. A single set of generated trajectories can flexibly address a wide range of clinical queries even without pre-specifying tasks, building task-specific models, or prompting a natural language model with tailored questions. Furthermore, because medical foundation models are trained specifically on medical event tokens, they can be more parameter and token-efficient in their representations. Generative medical event foundation models also offer extensive flexibility for forecasting future events, beyond just binary or quantitative prediction tasks; for example, they can predict the most likely order of events or generate a set of all events within a time frame from among hundreds of thousands of possibilities. Medical foundation models also provide a unique form of interpretability, in that clinicians and researchers can study individual generated trajectories of medical events to better validate and understand the sequence of events which the model predicts may lead up to downstream outcomes.

Previous models, such as CLMBR [13], MOTOR [14], Foresight [15], ETHOS [16], and others (see Section 4.3) have illustrated the feasibility of this approach; however, their scopes are constrained by dataset breadth and depth, leaving the scalability of such approaches largely untested. Moreover, the choices of model size and compute have not been systematically studied and it’s unclear whether they are optimal and how they should scale with available data. This is particularly challenging for RWE studies at the population scale of Cosmos, as suboptimal model size and compute can be extremely costly and wasteful.

To the best of our knowledge, Zhang et al. [17] conducted the first comprehensive study of the scaling laws on structured patient records. They observed power-law relationships among compute, model size, and pretraining data similar to the text domain, albeit with much higher token-to-parameter ratio, which might be attributable to the distinct characteristics of medical events. In this paper, we apply this same methodology to pretraining on Cosmos data, producing to date the largest scaling-law study on real-world patient journeys. The study in Zhang et al. [17] was limited to a de-identified dataset with only 70 thousand patients in emergency medicine. By contrast, our study is not only more than three orders of magnitude larger in terms of the number of patients, but also covers an extremely diverse range of patient populations and health conditions.

---

<sup>1</sup><https://cosmos.epic.com>

We present Cosmos Medical Event Transformer (CoMET), building on advances in generative medical event models by pretraining three decoder-only transformers with up to 1 billion parameters on Cosmos data. These models generate the next medical event—such as a diagnosis, medication order, lab result, the passage of time, or others—and these zero-shot generated sequences of medical events can be used to make clinical predictions. This work makes three contributions:

1. **CoMET models:** We describe the data transformation and training pipeline for medical event data on a dataset of 151 billion tokens from 115 billion medical events across 8.5 billion encounters.
2. **Clinically relevant evaluations:** We show that zero-shot generations with CoMET models demonstrate strong predictive performance on a wide variety of clinical tasks.
3. **Scalability:** We show that scaling up model and dataset size predictably decreases training loss and that minimizing train loss consistently improves downstream evaluation scores.

By learning from the collective experience of care provided by the Cosmos community, CoMET captures patterns from data that are broad, rich, representative, and real. CoMET offers the potential for clinicians, researchers, and health systems to transform that experience into intelligence that uncovers new medical knowledge, enhances healthcare systems, and improves patient outcomes.

## 2 Results

### 2.1 CoMET Training and Inference

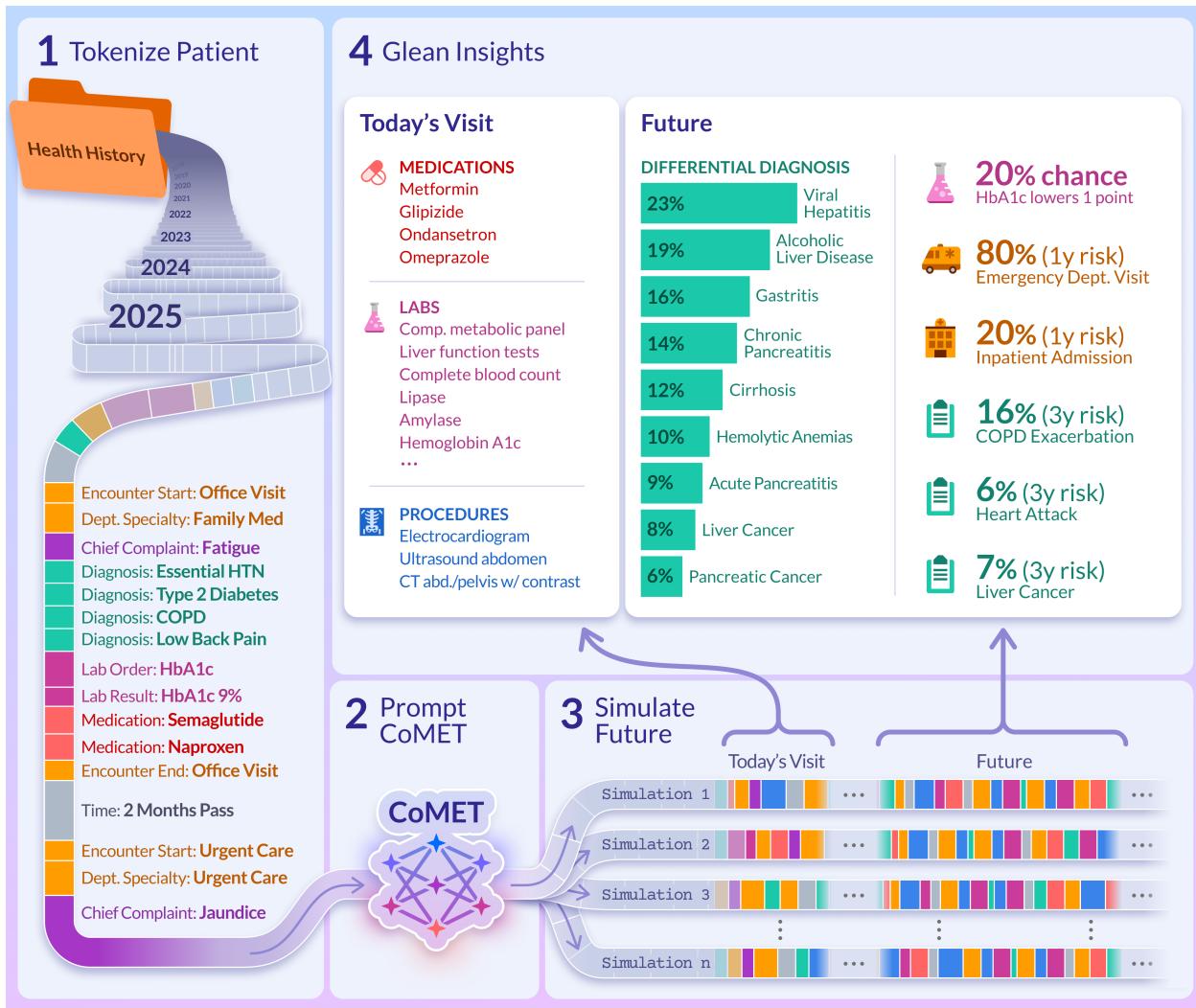
The dataset used for training and evaluating CoMET is a filtered subset of Cosmos that comprises 115 billion medical events from 118 million unique patient records from January 2012 to April 2025 (see [Section 5.1](#) and [Table 4](#)). We transformed each patient’s medical events into a chronological sequence, where events are each represented by compact tokens. Certain tokenization methods were inspired by ETHOS [16] and adapted for the scale and heterogeneity of Cosmos data. [Section 5](#) provides additional information on pre-processing, sequencing, and tokenization of Cosmos data. We trained CoMET using the Qwen2 transformer architecture [18] with random initialization—i.e., without loading any pretrained Qwen2 weights (see [Section 5.3](#) for training details). Three model sizes were trained, as detailed in [Table 1](#). The optimal compute and training tokens used for each were determined by a scaling-law analysis, detailed in [Section 2.6](#).

| Name    | Parameters | Training tokens | Compute (TFLOPs) |
|---------|------------|-----------------|------------------|
| CoMET-S | 62M        | 90B             | 67M              |
| CoMET-M | 119M       | 160B            | 234M             |
| CoMET-L | 1B         | 1.7T            | 14B              |

**Table 1:** Trained compute-optimal CoMET models, with their size in parameters, number of training tokens, and amount of training compute measured in teraFLOPs (floating point operations).

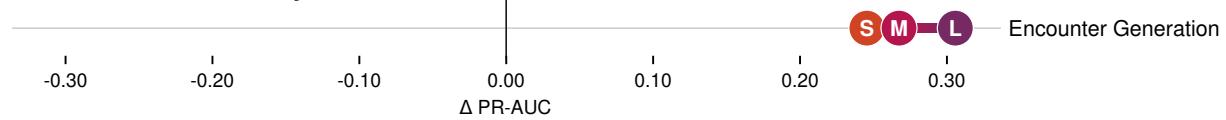
[Figure 1](#) shows an overview of how CoMET inference works, and [Figure 2](#) shows CoMET’s performance on a wide range of clinically relevant evaluations. The model is prompted with a patient’s longitudinal, tokenized record up to a desired time point. CoMET probabilistically generates  $n$  simulations, which are then used to compute all predictions, including event probabilities, distributions, times-to-event, and outcome collections (see [Section 5.5](#) for more details). For all evaluation tasks listed below, models were evaluated on data from the test set (see dataset construction methods in [Section 5.1](#)). Full tables of evaluation results are in [Appendix B](#). [Figure 2](#) provides a high-level snapshot of CoMET’s performance across all evaluation categories, with detailed task-level results and descriptions presented in the subsections that follow.

To contextualize CoMET performance on these downstream tasks, we also trained three classes of supervised task-specific models (see [Section 5.4](#) for details): linear and logistic regression, gradient-boosted decision trees (XGBoost), and supervised transformers trained from a random initialization. Each of these

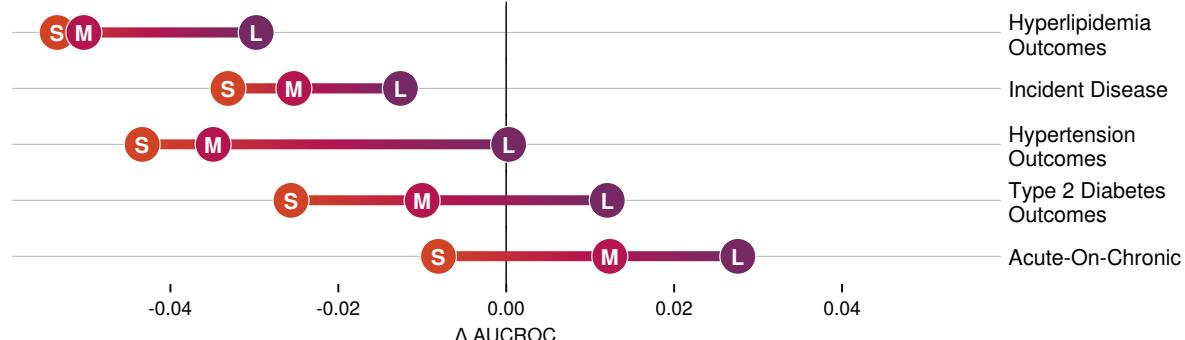


**Figure 1: Overview of CoMET pretraining and inference.** A patient journey is formulated as a sequence of medical events, and CoMET learns by predicting the next medical event. At inference time, CoMET is prompted with a patient's medical event history and simulates potential future trajectories by autoregressively generating the next events. Predictions for any target in CoMET's vocabulary are obtained from these simulated trajectories, enabling broad, out-of-the-box use on downstream tasks without task-specific fine-tuning or few-shot prompts.

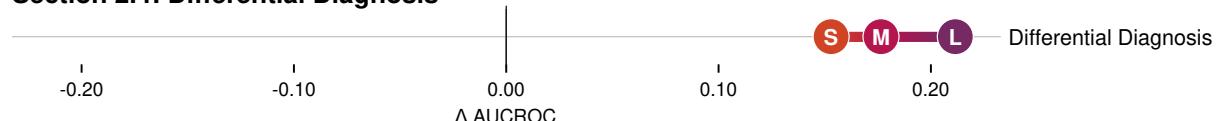
## Section 2.2: Plausibility



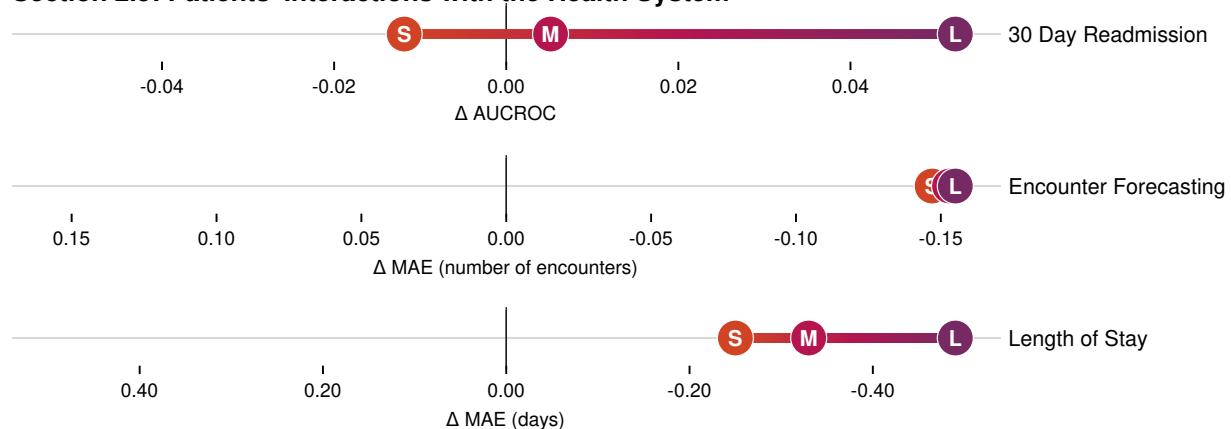
## Section 2.3: Disease Risk



## Section 2.4: Differential Diagnosis



## Section 2.5: Patients' Interactions with the Health System



**Figure 2: Overview of CoMET evaluation performance.** Each point shows the change in median evaluation scores for CoMET-S, CoMET-M, and CoMET-L relative to the best-performing task-specific supervised model in each of the major evaluation categories. For AUCROC and PR-AUC, positive values indicate that CoMET outperforms the task-specific model and negative values indicate under-performance while the opposite is true for MAE. CoMET's performance improved with scale and generally matched or even outperformed the best task-specific supervised methods.

task-specific models was trained on its corresponding downstream task and evaluated using the same datasets and procedures as those applied to CoMET. For simplicity, figures only show the best-performing task-specific model.

## 2.2 CoMET models generate realistic medical event sequences

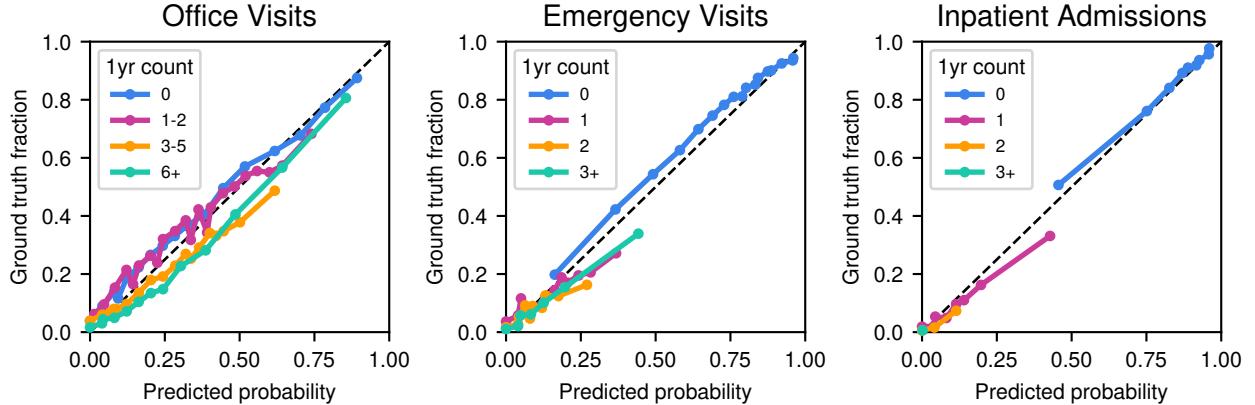
We evaluated alignment between a patient’s ground truth health records and CoMET’s generations in order to validate the plausibility of using these generations for more downstream predictive tasks.

### 2.2.1 Plausibility statistics

We first examined aggregate summary statistics over CoMET generations, prompting CoMET to produce 25 1-year generations for 20,000 patients. We measured CoMET’s ability to generate valid individual medical events that span multiple tokens, including diagnosis codes, medication codes, lab-result events, and department specialties in all encounter headers (see [Appendix C](#) and [Table 6](#)). Generated multi-token events were rarely invalid, and the error rate decreased as model scale increased. Furthermore, the prevalence of individual medical events for diagnoses, medications, labs, and procedures within one year as generated by CoMET all strongly agreed with their corresponding prevalence in the same patients’ 1-year future, as did the 1-year co-occurrence rate of pairs of medical events (see [Appendix D](#) and [Table 6](#)).

### 2.2.2 Encounter types and frequency

In addition to individual medical events, we measured how well CoMET probabilistically generates the number and types of medical encounters a given patient will experience within one year. Using the same model generations as [Section 2.2.1](#), we found each patient’s probability distribution for the number of office visits, emergency visits, and hospital admissions that would occur in the next year. In [Figure 3](#), calibration curves for CoMET-L show good calibration for predicting personalized healthcare needs. Across encounter types and counts, the expected calibration error (ECE) [19] improved with model scale ([Appendix B](#)).



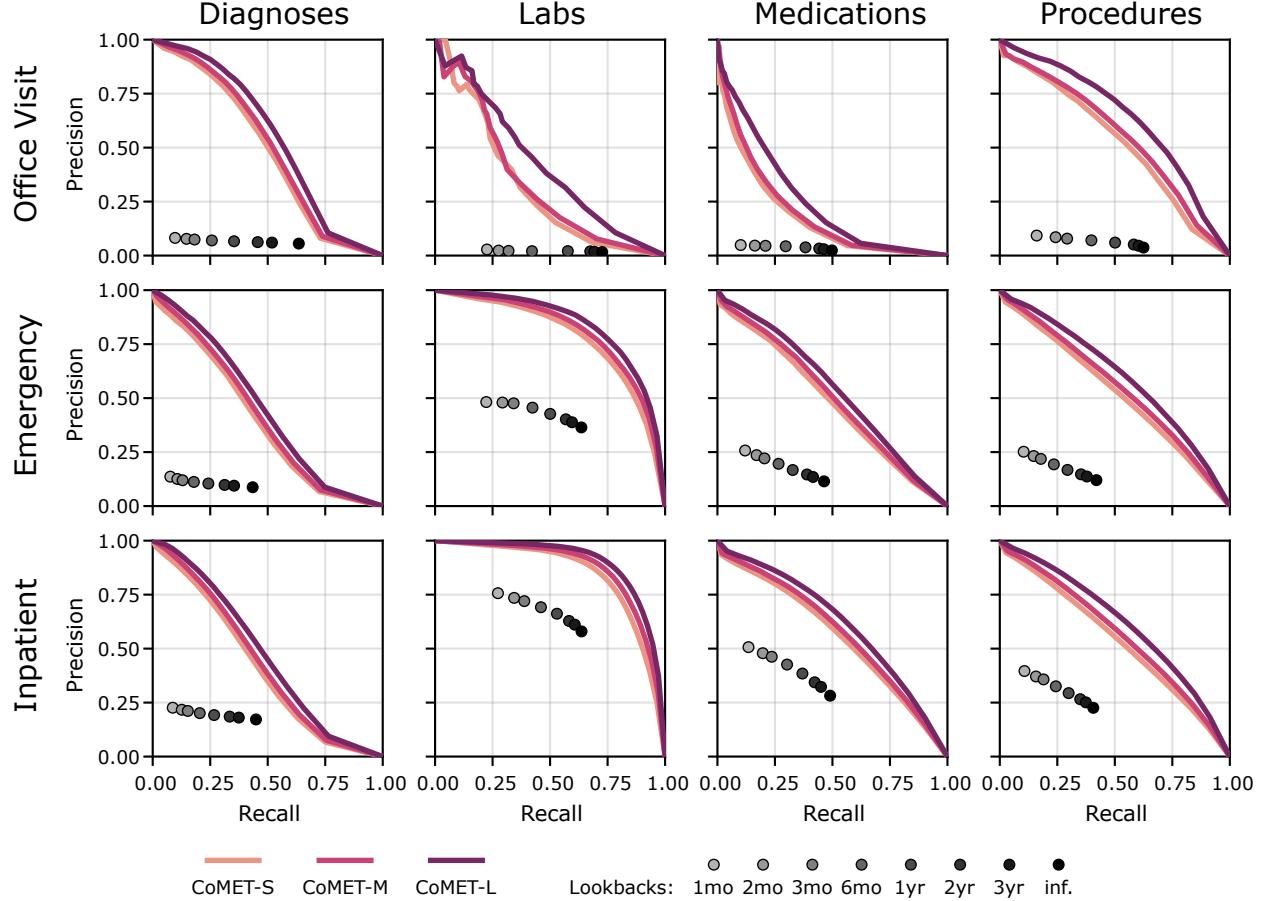
**Figure 3: Calibration plots for encounter frequency.** CoMET-L predicted the probability of how many encounters each patient will have within the next year, for three different encounter types (Office Visit, Emergency, and Inpatient). Each point represents a quantile group containing an equal number of patients with similar predicted probabilities. The horizontal position of each point reflects the group’s average predicted probability and the vertical position reflects the fraction of patients in that group with the specified 1-year count of encounters. Some lines do not span the full horizontal axis because few patients had those predicted probabilities. The diagonal line indicates perfect probability calibration.

### 2.2.3 Single-encounter generations

We next measured how well CoMET generates the full set of unique diagnosis, medication, lab, and procedure events that will occur during an encounter. For three different encounter types (office visits, emergency visits,

and inpatient admissions), we chose 10,000 random encounters and prompted CoMET with the patient’s history up to and including the target encounter’s header (i.e. encounter type, department specialty, and any chief complaints). We compared the micro-averaged recall and precision of CoMET’s encounter predictions to reference values representing the recall and precision of simply filling the encounter with the patient’s past medical events over various lookback windows.

[Figure 4](#) shows that across encounter types and medical event types, CoMET models demonstrated higher recall and precision than these lookback methods, and that with larger model sizes this performance showed consistent improvements as measured by precision-recall area under the curve (PR-AUC) (see [Table 7](#)). These precision-recall curves indicate CoMET generated medical events the patient had not previously had prior to the current encounter *and* filtered out most past events that were not likely to be repeated in the encounter.



**Figure 4: Medical events predicted for single encounters.** For office visit, emergency visit, and inpatient admissions, 10,000 random encounters of each were selected, and their medical events were compared to the medical events that CoMET predicted over 20 generations. The micro-averaged precisions and recalls are plotted over various thresholds for diagnosis, lab, medication, and procedure medical event types. In order to provide context on CoMET performance, we pooled the patient’s past events over various lookback windows and plotted the precision and recall for each. Higher area under each curve indicates better performance.

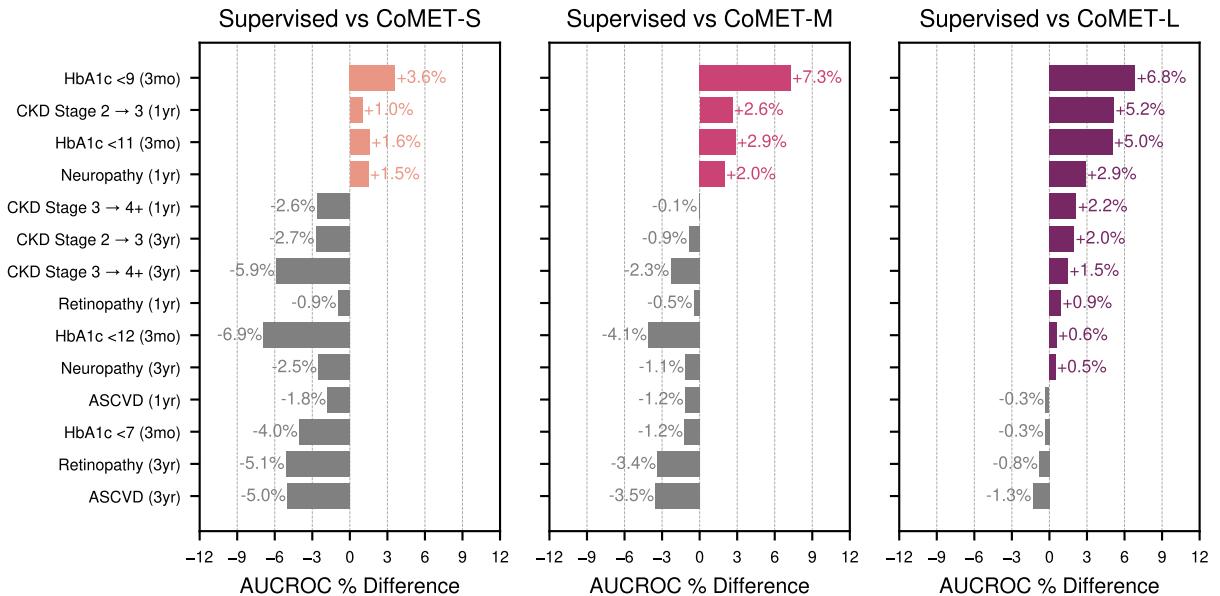
### 2.3 CoMET models can predict personalized future disease risk

We investigated CoMET’s ability to estimate future disease risk across multiple clinical domains and cohorts. Specifically, we categorized our prediction tasks into the following groups: disease-specific outcomes, acute-on-chronic events, and incident disease risk.

### 2.3.1 Disease-Specific Outcome Predictions

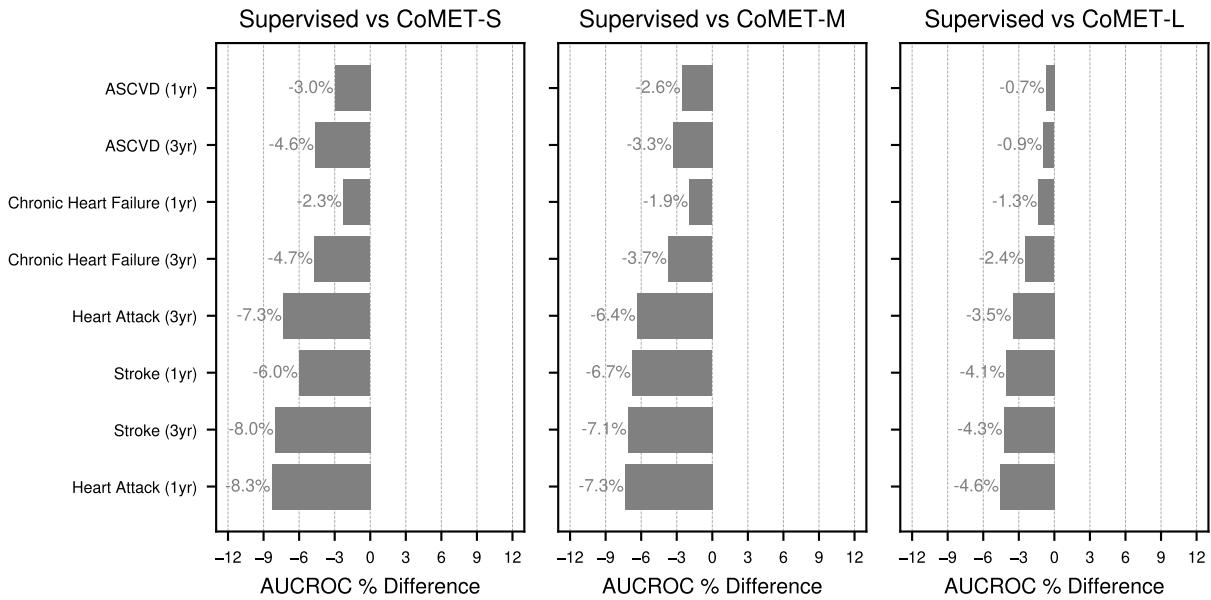
We measured CoMET’s performance on relevant disease-specific outcome prediction tasks, indexed to the time of care decisions. In particular, for type 2 diabetes mellitus (T2DM), hyperlipidemia (HLD), and hypertension (HTN), we examined CoMET’s ability to predict patients’ risk of relevant outcomes at the time of a change in medication management. Adverse outcomes, such as three-year stroke risk, were labeled as binary targets indicating whether the event occurred in the given time frame. Relevant lab results, such as hemoglobin A1c (HgbA1c) or total cholesterol, were labeled as binary targets at certain thresholds (e.g., HgbA1c < 7).

[Figure 5](#) illustrates CoMET’s predictive performance on relevant outcomes for patients initiating a new medication therapy for active T2DM. These include outcomes such as one-year and three-year risk for atherosclerotic cardiovascular disease (ASCVD), chronic kidney disease (CKD) progression from stage 2 through stage 4+, diabetic neuropathy, and diabetic retinopathy, as well as two-to-four-month HgbA1c lab results (the time frame when this lab is recommended to be reassessed after medication changes [20]). CoMET models improved consistently in these discriminative prediction tasks, with CoMET-L outperforming task-specific supervised models on most of these tasks. Scatter plots for predicted lab values can be found in [Figure 24](#).

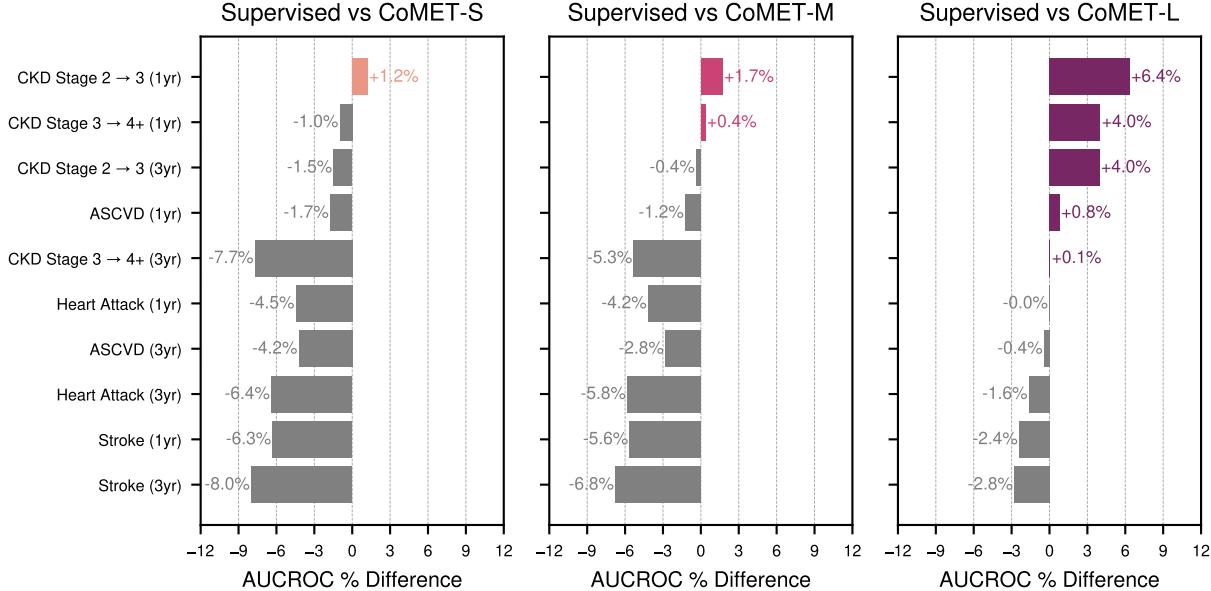


**Figure 5: T2DM-specific outcome predictions.** Percent increase of area under the curve of the receiver operating characteristic curve (AUCROC) from the best-performing task-specific supervised model for each of the three CoMET models on the T2DM-specific outcome prediction tasks.

Likewise, [Figure 6](#) shows results for CoMET predictions on tasks relevant to patients receiving treatment for hyperlipidemia, including one- and three-year risk of ASCVD, heart attacks, strokes, and chronic heart failure (e.g., only the chronic diagnosis codes related to chronic heart failure). CoMET-L achieves an AUCROC of 0.93 for predicting chronic heart failure diagnosis events within a year of changing hyperlipidemia medical management. While CoMET-L performance did not exceed the task-specific supervised models’ performance, CoMET models showed measurable improvements as they scaled up, and absolute AUCROC values were robust. Notably, CoMET did not outperform the task-specific models on the hyperlipidemia outcomes; the factors underlying this underperformance remain unclear and are left to future work.



**Figure 6: Hyperlipidemia-specific outcome predictions.** Percent increase of AUCROC from the best-performing task-specific supervised model for each of the three CoMET models on the hyperlipidemia-specific outcome prediction tasks. CoMET models consistently scored better with scale, yet CoMET-L scored lower than the supervised models on each diagnosis task.



**Figure 7: Hypertension-specific outcome predictions.** Percent increase of AUCROC from the best-performing task-specific supervised model for each of the three CoMET models on the hypertension-specific outcome prediction tasks. CoMET-L scores matched or exceeded the supervised models on 6 out of 10 tasks.

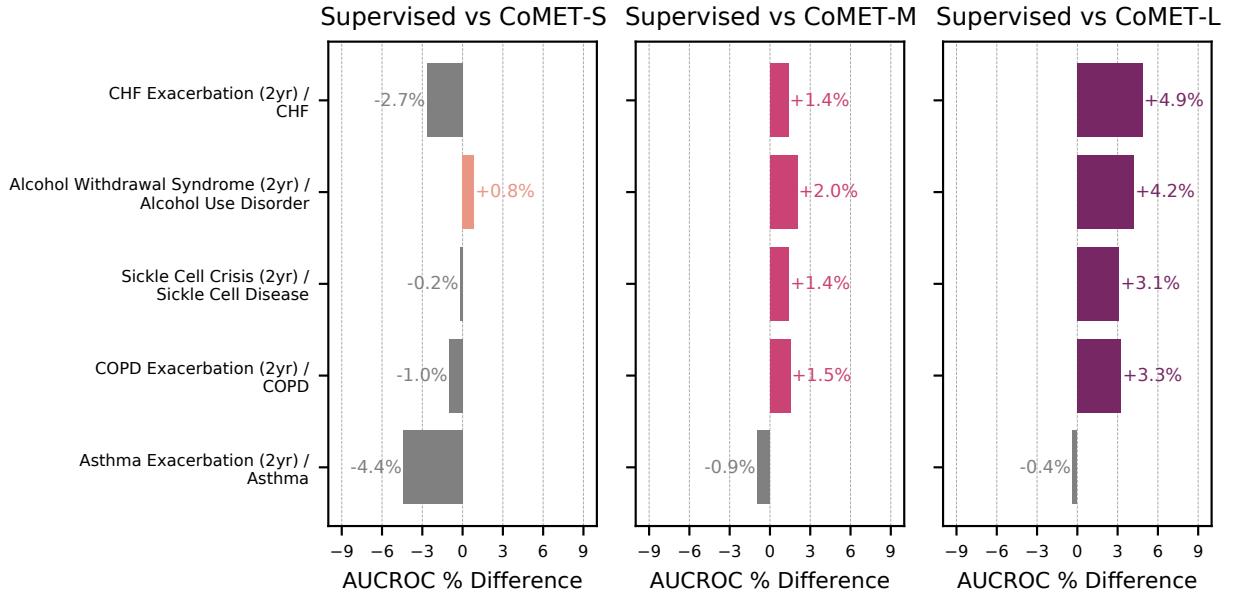
For hypertension-related outcomes, we assessed CoMET performances to make predictions about one- and three-year risk of ASCVD events, heart attack, stroke, and CKD progression from stage 2 to stage 3 and from stage 3 to stage 4<sup>+</sup>. As with the two above cases, CoMET models improved as they increased in scale,

with CoMET-L achieving higher AUCROC scores than the task-specific models on half of these tasks. A full list of evaluation scores across disease-specific outcome tasks can be found in [Table 8](#).

### 2.3.2 Acute-on-Chronic Incidence Prediction

We next evaluated CoMET’s ability to predict the two-year risk of acute-on-chronic clinical events, such as asthma exacerbation or sickle cell crisis. Patients are included in each acute-on-chronic cohort if they demonstrate a medical history of the relevant chronic disease (e.g., sickle cell disease for sickle cell crisis).

We formulated each acute-on-chronic evaluation as a binary classification task: for congestive heart failure (CHF) exacerbations for patients with chronic CHF, asthma attacks for patients with asthma, sickle-cell crises for patients with sickle cell disease, alcohol withdrawal syndrome for patients with alcohol use disorder, and chronic obstructive pulmonary disease (COPD) exacerbations for patients with COPD. Detailed phenotype definitions, prediction date selection, and distinctions between chronic and acute event coding are provided in [Section 5.6.5](#). Approximately 5,000 unique patients are included for each group, with optional upsampling to ensure at least 500 patients experienced the acute event within two years of their prediction date. Dataset characteristics for each task can be found in [Table 19](#).



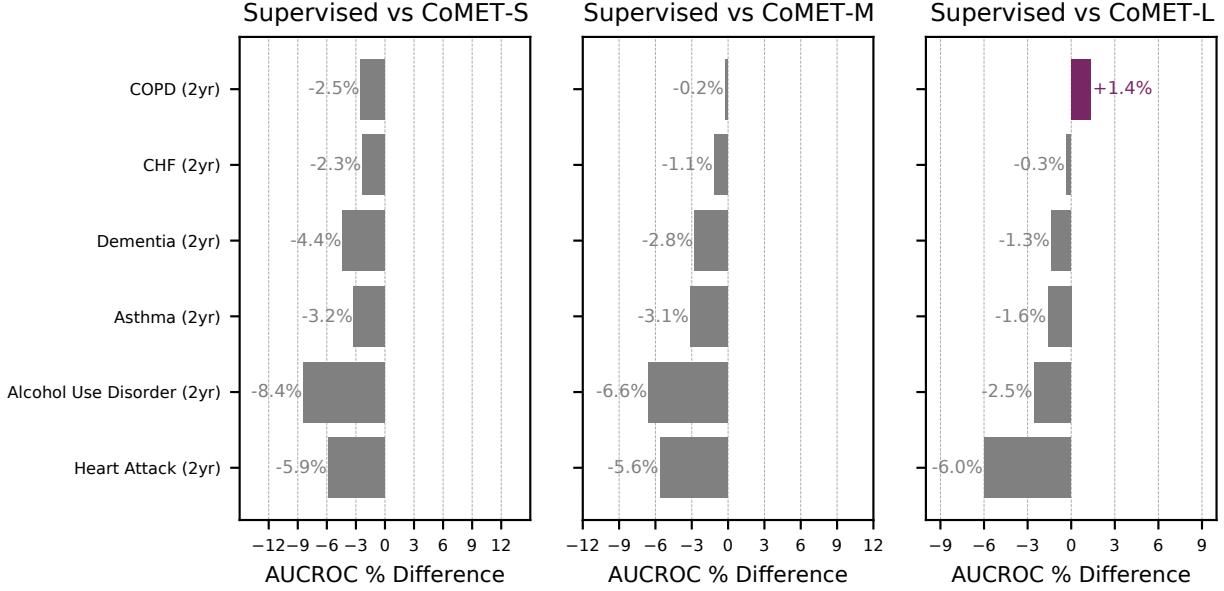
**Figure 8: Acute-On-Chronic Tasks.** Percent increase of AUCROC for each of the three CoMET models compared with the best-performing task-specific supervised model on the acute-on-chronic outcome prediction tasks. CoMET-M and CoMET-L scored higher than the baselines on 4 out of 5 tasks.

CoMET-L achieved a higher AUCROC than the task-specific models across all but one of these acute-on-chronic tasks, as shown in [Figure 8](#), and in all tasks measured by PR-AUC. All tasks showed consistently increasing AUCROC and PR-AUC with larger CoMET model size. Tabulated results can be found in [Table 9](#).

### 2.3.3 Incident Disease Risk Prediction

Finally, we tested CoMET’s performance on predicting the first instance of a variety of disease states within a two-year period in the general population. CoMET generated 60 simulated timelines for approximately 5,000 patients, with each generation spanning two years of tokens. To avoid cohort biases and priors, broad inclusion and minimal exclusion criteria were used on these task cohorts (see [Section 5.6.4](#) for details). Notably, this leads to class imbalance where all targets have a positive prevalence <1.5%. This is quite different from the two preceding disease outcomes tasks, where the inclusion criteria for each task meant that those patients were naturally at increased risk for the prediction targets.

[Figure 9](#) compares CoMET to supervised models on incident disease tasks. AUCROC scores are reported in the main figure for consistency with the other disease risk tasks. CoMET demonstrated higher AUCROC scores than the supervised models on one out of six tasks and shows improvement across most tasks with model scale. However, in tasks like incident disease prediction with highly imbalanced class labels, PR-AUC is more commonly used to judge performance [21, 22]. CoMET-L achieved higher PR-AUC scores than the supervised models on all six tasks. Results for CoMET and supervised models (including PR-AUC) are reported in [Table 10](#). One possible reason that CoMET-L did not outperform most task-specific models in AUCROC is because these tasks had extremely low prevalence and more generations per patient may be necessary. Preliminary evidence for this can be seen in [Section 2.7.2](#).



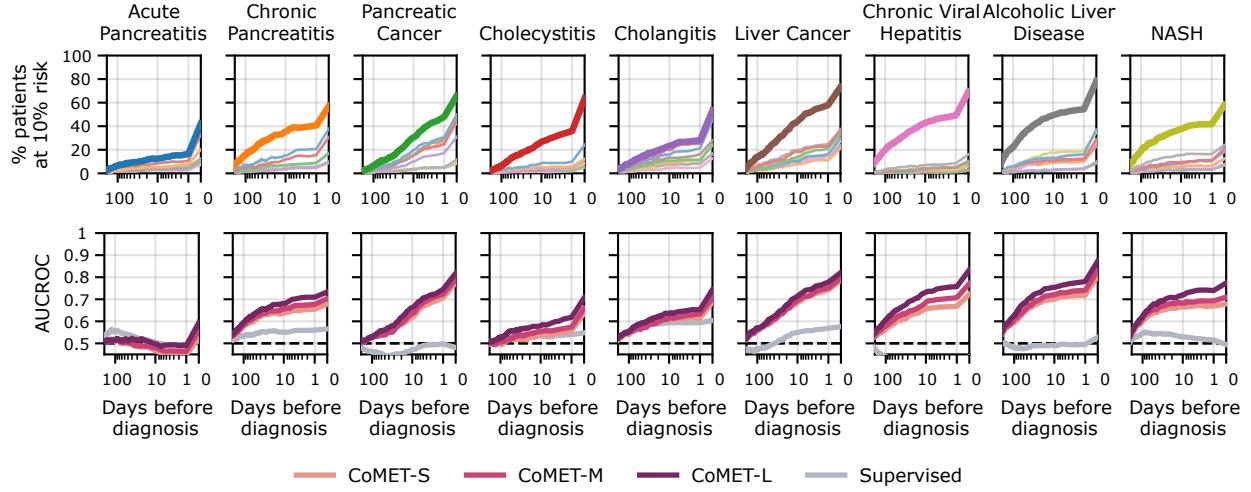
**Figure 9: Incident Disease Risk Predictive Performance.** Percent increase of AUCROC from the best-performing task-specific supervised model for each of the three CoMET models on the six incident-disease prediction tasks. CoMET-L scores higher than the baselines on one out of six tasks.

## 2.4 CoMET models generate early, quantitative differential diagnoses

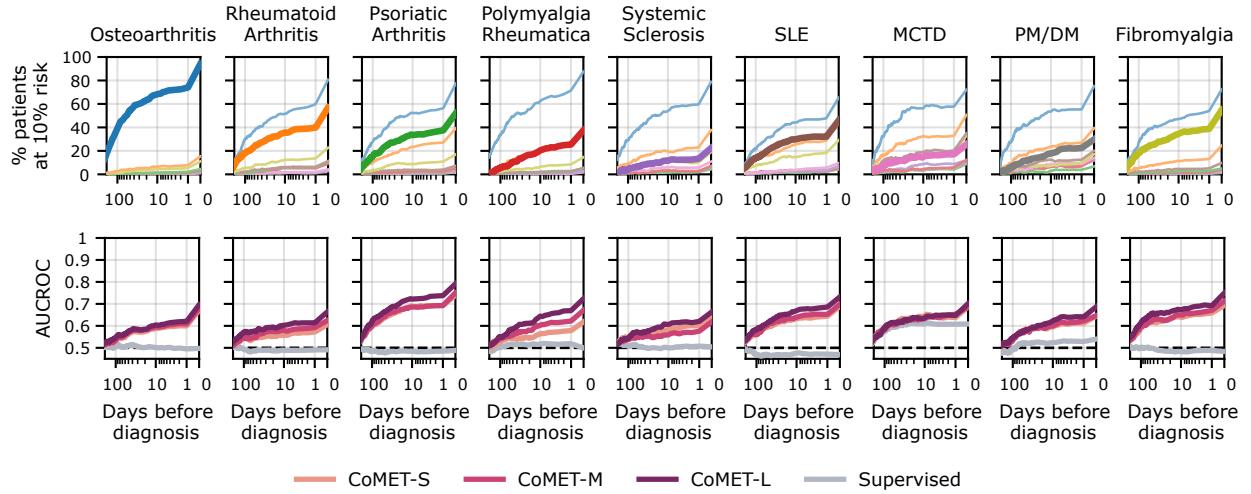
We next assessed whether CoMET can provide ranked, quantitative differential diagnoses for individual patients. We selected two clusters of diagnoses, hepatopancreatobiliary (HPB) diseases and rheumatic diseases, because they span a variety of clinical presentations, can take multiple encounters to diagnose definitively, and have challenges with delayed diagnosis or misdiagnosis [23–30]. Diseases within each cluster often present clinically with overlapping signs, symptoms, and laboratory findings, and as a result they often appear together on differential diagnoses. We selected nine diagnoses for both HPB and rheumatic diseases, and for each we selected a cohort of patients who received that diagnosis (see [Section 5.6.6](#) for diagnoses and their code-based definitions, and [Table 20](#) for sample sizes). For each patient, we chose several points in their history in the one-year span before the first occurrence of their target diagnosis and had CoMET predict their risk for all the cluster’s diagnoses. In this way, CoMET produced a ranked, quantitative differential diagnosis at multiple time points for each patient and flagged patients at risk of receiving these diagnoses. We did not prompt CoMET with any information beyond the start of the encounter at which the patient first received their target diagnosis, so CoMET never sees any diagnostic workup or documentation from this encounter.

Each plot in the top row of [Figure 10](#) and [Figure 11](#) shows how many patients were flagged by CoMET-L as having at least 10% risk for the target diagnosis (bold) and each of the off-target diagnoses within the cluster (lighter, thinner lines). Each plot therefore represents the average differential diagnosis for the cohort over time. For the majority of diagnoses, CoMET-L correctly flagged more than 50% of patients for their

target diagnosis by their final prediction time. For most diagnoses, CoMET-L also flagged more than 25% of patients weeks ahead of time.



**Figure 10: Hepatopancreatobiliary differential diagnosis.** CoMET-L was evaluated on predicting 1.5-year risk of receiving one of the indicated HPB diagnoses. Each column represents a cohort of patients who were ultimately diagnosed with the indicated diagnosis. Each line in the first row represents the percentage of that cohort that was flagged by CoMET-L as having at least 10% risk of a diagnosis. The correct diagnosis is shown in bold, the off-target diagnoses are faint; each line color represents the same diagnosis across the row. The second row shows the AUCROC over time for all three CoMET models and the task-specific supervised model for predicting the indicated diagnosis.



**Figure 11: Rheumatic differential diagnosis.** CoMET-L was evaluated on predicting 1.5-year risk of receiving one of the indicated rheumatic diagnoses. Each column represents a cohort of patients who were ultimately diagnosed with the indicated diagnosis. Each line in the first row represents the percentage of that cohort that was flagged by CoMET-L as having at least 10% risk of a diagnosis. The correct diagnosis is shown in bold, the off-target diagnoses are faint; each line color represents the same diagnosis across the row. The second row shows the AUCROC over time for all three CoMET models and the task-specific supervised model for predicting the indicated diagnosis.

For comparison, we trained task-specific supervised models to predict each of these diagnoses in a wider, more general pool of patients (see [Section 5.4](#) for details on training). These task-specific models were generally not able to differentiate well among patients within these narrower HPB or rheumatic diagnosis

clusters, with AUCROC metrics not increasing appreciably over time (see the bottom row of both [Figure 10](#) and [Figure 11](#)), despite achieving modest sensitivity. In contrast, AUCROC metrics from CoMET generally increase over time and as model size increases. This suggests that CoMET models can effectively distinguish patients with similar presentations but different eventual diagnoses, and they can produce realistic differential diagnoses that become more sensitive and specific as the patient’s clinical presentation and diagnostic workup evolve. AUCROC at the final prediction time is reported for all diagnoses and models in [Table 11](#). Of note, these analyses do not address whether early diagnosis flagging reflects the diagnostic workup being pursued by the patient’s medical providers *versus* the ability to preemptively flag diagnoses before significant clinical suspicion.

[Figure 10](#) shows that, among all nine HPB diseases, the correct diagnosis was the one most commonly flagged by CoMET-L, with the gap between the first- and second-ranked diagnoses generally increasing with time. For acute pancreatitis, AUCROC is hardly better than chance until the target diagnosis date. For diseases with generally more insidious onset (e.g., cancer, chronic viral hepatitis, and alcoholic liver disease), CoMET-L was able to flag many patients much earlier ([Figure 10](#)).

For rheumatic disorders ([Figure 11](#)), most patients were flagged at the 10% risk levels for osteoarthritis at some point, regardless of their eventual diagnosis. This is likely both because osteoarthritis is a common diagnosis for patients with undifferentiated joint pain and because of early inaccurate or imprecise diagnosis and documentation. Among the remaining eight target diagnoses, the correct diagnosis was ranked second for five cohorts, third for two, and fourth for one. AUCROC scores for CoMET models ranged from 0.66-0.79 across rheumatic diagnoses at the final prediction time ([Table 11](#)).

## 2.5 CoMET models forecast patients’ interactions with the health system

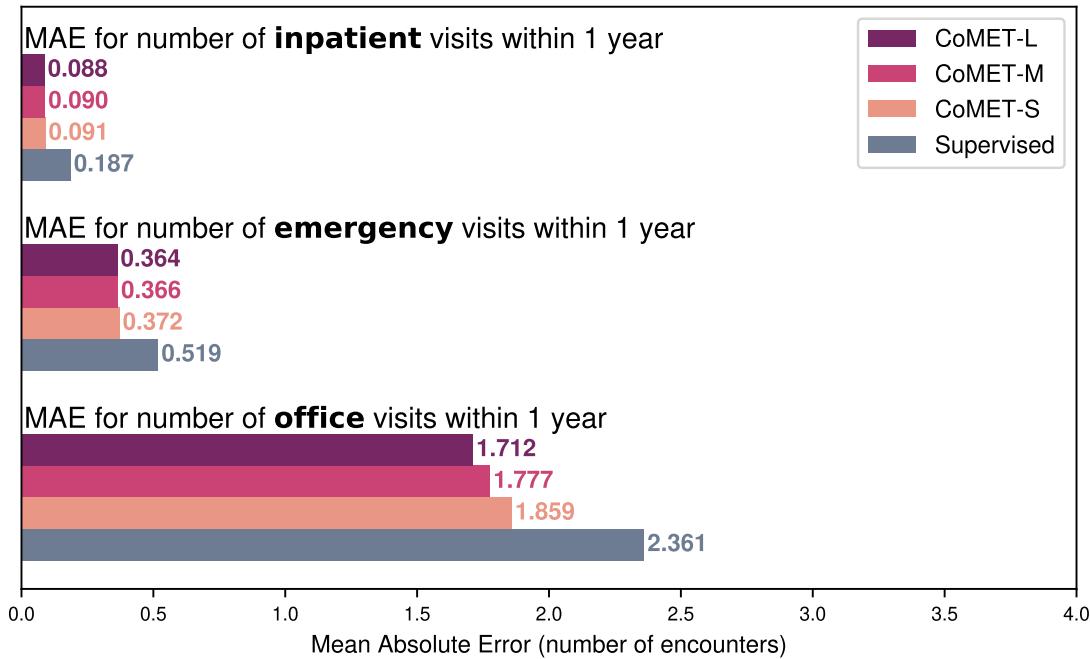
Reliably forecasting health system interactions enables clinicians and health systems to plan for the needs of their patients. We assessed CoMET’s generated patient timelines for their ability to make accurate predictions about a patient’s interactions with the healthcare system.

Having earlier assessed the probability calibration of predicting the number of encounters over a year in [Figure 3](#), we next asked how close these predictions were to the ground truth number of encounters. For inpatient, outpatient, and emergency encounters, all three CoMET models demonstrated lower mean absolute error (MAE) than supervised task-specific regression models for predicting future encounter counts. Results are shown in [Figure 12](#).

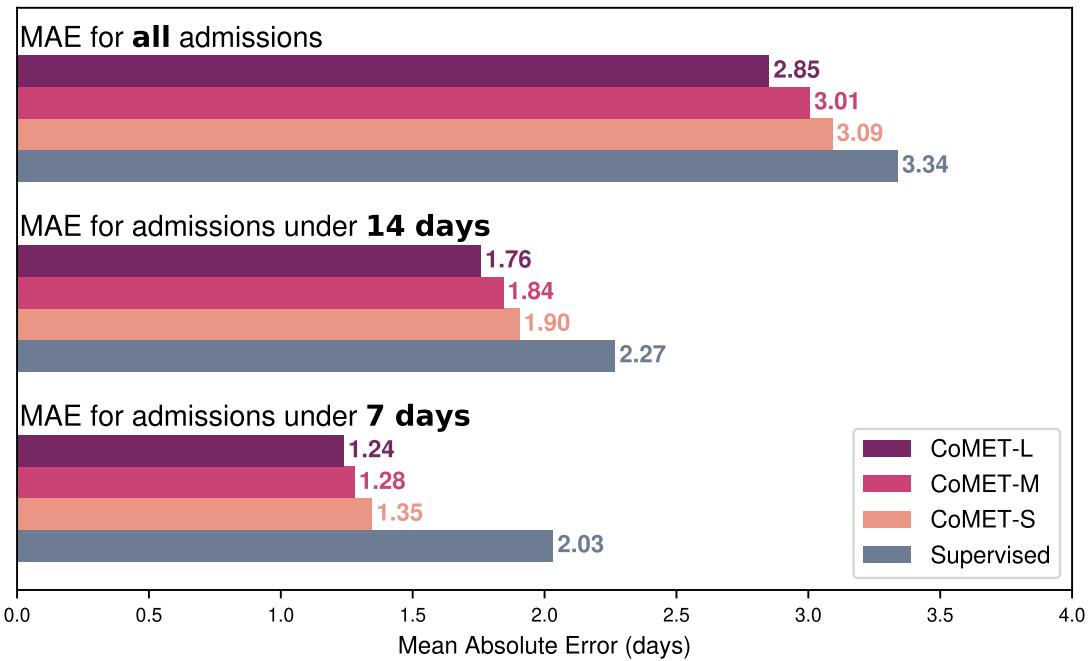
### 2.5.1 Hospital length of stay and 30-day readmission

Accurate hospital length of stay (LOS) prediction helps health systems manage beds and plan patient care to limit discharge delays [31]. We evaluated CoMET’s ability to predict LOS on 10,000 randomly selected hospital admissions (see [Table 21](#) for evaluation set statistics). All models received the patient’s history through the admission encounter’s header (i.e., encounter type, department specialty, and any chief complaints), in addition to the documented primary diagnosis.

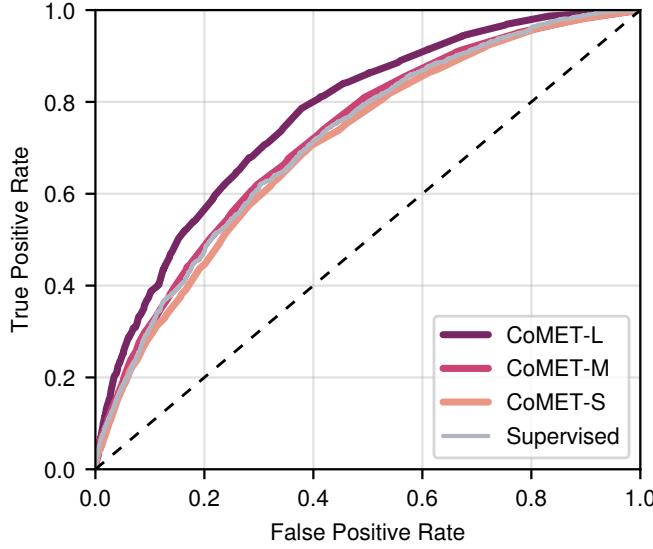
Another metric for assessing healthcare utilization is the 30-day hospital readmission risk, an important and well-studied operational consideration for care transitions, discharge planning, and outpatient follow-up [32]. We randomly selected 10,000 patients being discharged from the hospital and had models predict the probability that a new hospitalization would occur within 30 days (patients who in real life were readmitted within a day were excluded from this analysis because they may often reflect hospital transfers or clerically erroneous discharges [33]). CoMET models saw increases in AUCROC on this task, with CoMET-L and CoMET-M demonstrating higher scores than task-specific supervised models (see [Figure 14](#) and [Appendix B](#)).



**Figure 12: One-year encounter frequency forecasting.** CoMET models were compared with the best-performing supervised task-specific regression model for predicting the number of inpatient, emergency, and outpatient encounters that will occur within a year's time for 18,400 patients. Mean absolute error (MAE) was used to measure the error in predicting these encounter counts; smaller the better.



**Figure 13: Hospital length of stay prediction.** Each of the CoMET models and the best-performing task-specific model evaluated on LOS prediction on a set of 10,000 randomly selected samples. Mean absolute error (MAE) of the length of stay in days was used to compare model performance; smaller the better.

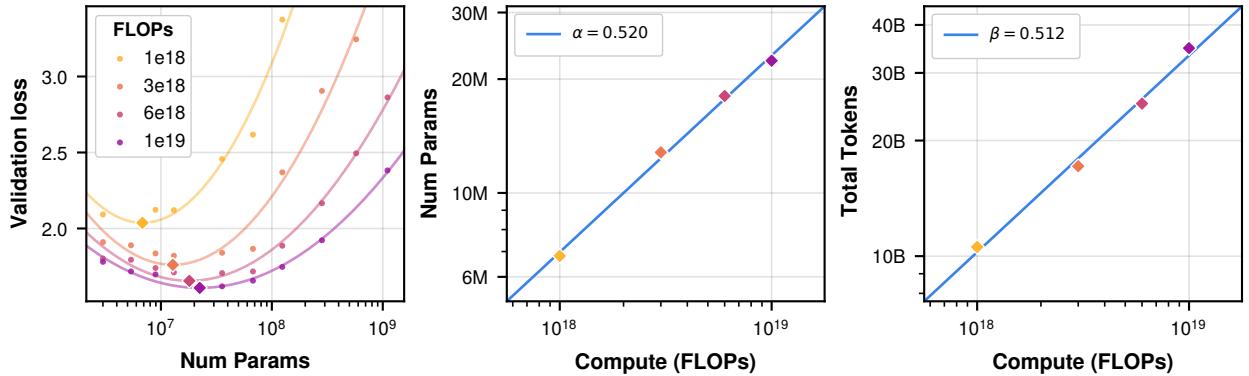


**Figure 14: 30-day readmission prediction.** ROC curves for predicting hospital readmission within 30 days of discharge across 10,000 hospital encounters. CoMET-L achieved an AUCROC of 0.770, with the best-performing supervised task-specific model achieving 0.717.

## 2.6 Training medical event models follows scaling laws

Before training CoMET-S, CoMET-M, and CoMET-L, we first trained many smaller models to derive scaling laws [34, 35] to predict the optimal model size and number of training tokens for a given compute budget, measured in tera floating point operations (TFLOPs). This step was important not only for understanding the best parameters for training models, but also to understand how to optimally scale medical event foundation models on a sufficiently large dataset. Building on recent work demonstrating power-law scaling for generative medical event prediction on the MIMIC-IV dataset [17], we applied the same approach to Cosmos by training a sweep of 10 model sizes ranging from two million to one billion parameters on our dataset of over 136 billion training tokens.

As in Hoffmann et al. [34], we ran a grid search over varying amounts of training TFLOPs to find the optimal scaling of model size and training tokens for the CoMET medical event dataset. After performing smaller training runs at four fixed compute budgets, we fit parabolas to the isoFLOP curves (Figure 15).



**Figure 15: Optimal training of CoMET models follows power laws.** The minimum loss achieved by training runs with fixed compute and varied number of model parameters (left). A log-scale parabola is fit to each isoFLOP curve, and the minimum point of each is marked by a diamond. Middle and right show how the isoFLOP parabolas' minima  $N_{\text{opt}}$  and  $D_{\text{opt}}$ , respectively, vary with the isoFLOP compute on a log-log scale with power-law fits.

The minima of these curves were used to fit power-law equations for the compute-optimal scaling of parameter count  $N_{\text{opt}}$  and training tokens  $D_{\text{opt}}$  with respect to the amount of compute  $C$  used during training

(Figure 15). Our experiments showed that for power laws of the form

$$\begin{aligned} N_{\text{opt}} &= AC^{\alpha} \\ D_{\text{opt}} &= BC^{\beta} \end{aligned} \quad (1)$$

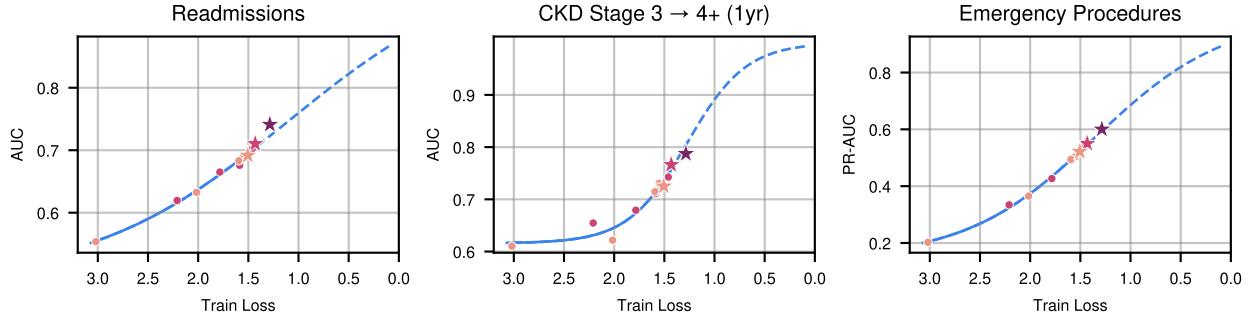
with fit parameters  $A$ ,  $B$ ,  $\alpha$ , and  $\beta$ , we obtained best-fit parameters of  $\alpha = 0.520$  and  $\beta = 0.512$ , closely mirroring coefficients reported for natural-language corpora ( $\alpha = 0.49$  and  $\beta = 0.51$ ) [34]. The near-equality of  $\alpha$  and  $\beta$  implies that, on the compute-optimal frontier, model size and training-token count should be scaled proportionally as total compute increases. We used these best-fit lines to derive the training parameters to train our 3 compute-optimal CoMET models, listed in Table 1.

## 2.7 Scaling medical event model performance

### 2.7.1 Performance vs. loss

Our scaling analysis demonstrated that increasing parameters and training tokens predictably decreases model loss. This raises the question of how minimized loss corresponds with downstream, clinically relevant evaluations. To measure this relationship, we evaluated the CoMET models and several training checkpoints at different train loss on the single-encounter generation, T2DM-specific outcomes, and 30-day readmissions tasks.

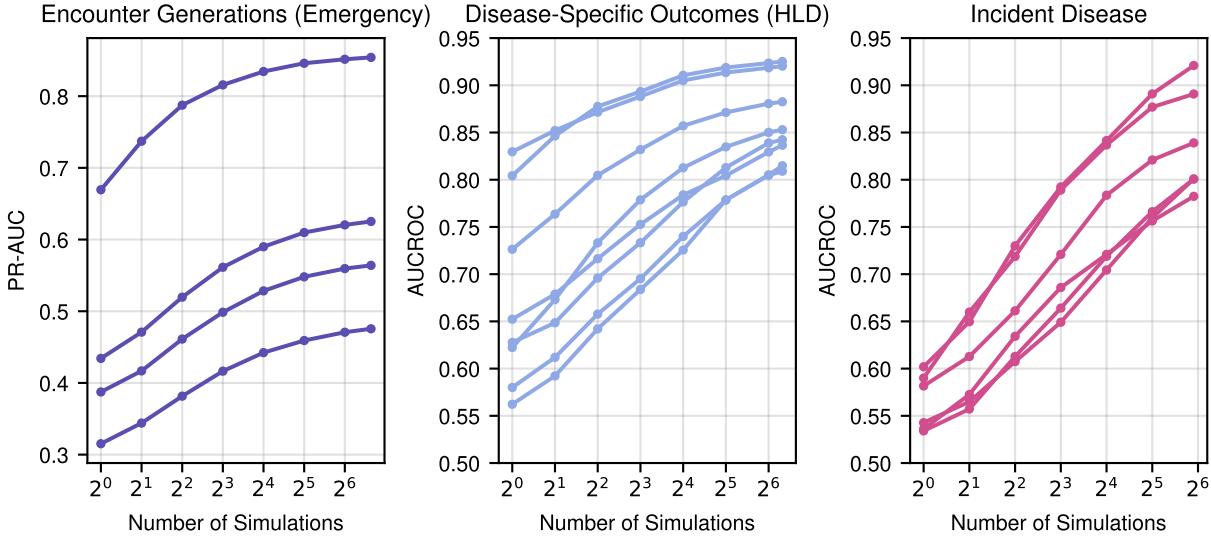
Figure 16 shows a smooth, sigmoidal relationship between the training loss and downstream clinical evaluation scores (the full set of plots can be found in Figure 21, and Figure 22). This relationship holds over many different model sizes and compute budgets. The empirically fitted sigmoid functions show different inflection points and relative slopes, indicating that different clinical tasks require different levels of training to show improvements.



**Figure 16: Downstream performance improves as train loss decreases.** We evaluated each of the CoMET models, as well as earlier checkpoints from the CoMET-S and CoMET-M training runs, on a variety of our downstream evaluations. We fit a sigmoid curve to all points except for CoMET-L to assess the sigmoid curve’s predictive utility. We evaluated all of these models using a more conservative  $n = 20$  simulations.

### 2.7.2 Scaling test-time compute

At inference time, one of the uniquely relevant parameters for CoMET is the number of generated patient timelines,  $n$ . Because probabilities are calculated as an aggregation over  $n$  generations (Section 5.5), increasing  $n$  scales inference costs linearly while reducing the Monte Carlo variance of downstream predictions and increasing the resolution of output probabilities. Unlike in language models, where test-time compute navigates a fully expressive space, in our setting, increasing  $n$  also expands the method’s expressiveness, raising the performance ceiling by reducing the quantization of predictions. To investigate this tradeoff, we varied  $n$  for three representative clinical tasks: hyperlipidemia-specific outcomes (an evaluation with relatively high-prevalence positive outcomes), incident disease prediction (an evaluation with the lowest-prevalence positive outcomes), and single-encounter emergency generations (which is highly multi-target and uses PR-AUC).



**Figure 17: Effect of test-time compute on performance.** For CoMET-L we evaluated model performance against the number of simulations generated, focusing on single-encounter generations for emergency visits (left), HLD-specific outcomes (middle), and incident disease tasks (right). Each line represents a specific task (e.g., one-year ASCVD risk, two-year COPD, etc.) from the titled category. For readability the legend is not included.

All three task groups in Figure 17 show a steady increase in performance as  $n$  increases, with various degrees of plateauing at  $n \gtrsim 64$ . The incident disease evaluations are particularly sensitive to  $n$ , possibly due to most patients having a low probability of having these events. Similar to language models [36], increasing train and test-time compute provide orthogonal directions to improve performance on downstream tasks.

### 3 Discussion

#### 3.1 Key findings

Here we demonstrate that large-scale, medical event foundation models can learn from longitudinal patient records to produce realistic and useful clinical predictions. Trained on a subset of the 16.3B encounters across 300M patient records as of August 2025 from Cosmos, CoMET is a family of decoder-only transformer models up to 1B parameters in size that probabilistically generate medical event sequences which not only preserve event- and encounter-level realism but also demonstrate a broad range of short- and long-term predictive capabilities.

Previous medical event foundation models have been constrained by training on smaller datasets or a limited number of data types [37]. CoMET models were trained directly on sequences of structured medical events spanning core clinical data types (i.e., demographics, encounters, diagnoses, chief complaints, labs, medications, procedures, and time). Rather than relying on natural language inputs and outputs, CoMET directly receives discrete medical events from a patient’s medical record as input, learning event-level structure, long-range temporal dependencies, and the underlying probability distribution of medical events. As of this writing, CoMET represents the largest medical-event foundation model by number of medical events used for training.

Evaluation of medical event foundation models has typically been performed on relatively few evaluation tasks [37]. By assessing CoMET’s performance across a broad and diverse suite of clinical tasks spanning single-encounter generation, disease-specific outcome prediction, acute-on-chronic risk estimation, incident disease detection, differential diagnosis, and operational forecasting, we explored its potential as a flexible and generalizable tool for a range of clinical contexts. CoMET-L, our largest model, matched or outperformed

supervised task-specific models on most of these tasks. Performance gains were evident across multiple clinical domains, supporting CoMET’s generalizability and utility for diverse downstream applications.

We also conducted the largest scaling law analysis for medical event data to date—more than  $300\times$  larger than the largest previous work [17]. To investigate medical event scaling laws, we employ the same methodology from [34], exploring isoFLOPs across four fixed compute points. Our isoFLOPs analysis yields similar exponents to those shown in the natural language (NL) domain ( $\alpha = 0.520$  and  $\beta = 0.512$ ), suggesting we should scale training tokens and model parameters proportionally to train compute-optimal models [34]. In contrast to work in the NL domain, we observe an optimal token-to-parameter ratio of 1,000:1, closely matching results from prior work on medical event scaling laws [17].

Lastly, we found that loss minimization predictably translates into better downstream performance on our suite of evaluations. Nearly every clinical task we evaluated exhibited sigmoidal improvements in their performance metrics as we minimized training loss. These trends were consistent across task families and clinical domains, and most tasks had not yet plateaued, indicating substantial remaining headroom. This result highlights pretraining loss as a useful proxy for downstream capability, and suggests that continuing to scale will yield better downstream utility. In addition to training-time scaling, we also observe benefits from increased inference-time compute: generating more patient simulations per prediction consistently improves downstream performance metrics across many clinical tasks.

## 3.2 Limitations

Even with these strengths, CoMET has clear limitations.

**Real-world data.** CoMET is trained on real-world healthcare data in Cosmos, and it is subject to imprecision and errors in documentation and clinical practice [38]. Several automated data quality control processes in Cosmos aim to improve data completeness and identify irregularities, and these periodic analyses are fed back to the contributing organizations [6]. Cosmos aggregates data across 310 health systems with extensive linking and deduplication efforts, but individuals may receive care at other health systems that is not recorded in Cosmos. This leads to data gaps before, within, and after observed medical event sequences. Certain prediction targets may be missing or occur earlier than the first record in Cosmos. We mitigate this by training and evaluating on patients who meet predefined criteria for observable healthcare utilization (see Section 5.1) and by the use of time tokens, which allow the model to read and generate patterns of absent data.

**Tokenization.** While the tokenizer for CoMET can be used to encode a wide range of clinical data, the discretization of continuous values like laboratory results and time masks smaller changes in these values that may be important for some prediction tasks. Future work can investigate tradeoffs in performance and model complexity between a pure categorical approach and a hybrid approach that handles numeric values separately [39–43]. In addition, we did not conduct an evaluation for training a tokenizer or optimizing vocabulary size, presenting a rich avenue for future work. Tokenizers trade vocabulary size for sequence length which can impact performance in non-trivial ways with model scale [44]. As shown in Figure 18, many patients’ entire tokenized history exceeded the model’s current context window of 8,192, presenting an opportunity to improve performance with longer context models. Finally, while CoMET tokenized and was trained on a core set of structured clinical data types Section 5.2, many other data types remain to be added.

**Evaluation.** Disease phenotype definitions were based on International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) codes, categorical encounter types, demographics, event timings, and other similar features. While these features form reasonable phenotypes, they do not always correspond directly to how events transpired in the real world. This presents an opportunity for future work involving prospective validation of CoMET’s outputs. Finally, evaluations focused on aggregate performance across individuals in the test set and did not investigate performance and calibration within specific subpopulations, which will be an additional area for future work.

## 3.3 Future work

The above constraints outline the frontier of CoMET’s capabilities. However, each constraint naturally suggests consequent experiments and improvements.

**New event types.** First, including additional structured data from Cosmos, such as genomic variants, social drivers of health, and cancer staging data, would give CoMET a more robust representation of patient health timelines. The addition of specialty-focused data types would further improve CoMET’s ability to both contextualize and make certain predictions for a wide range of medical specialties, professions, and use cases. Pediatric data, including mother-baby links, could enable CoMET to capture developmental physiology, age-specific drug dosing, vaccine schedules, and congenital condition timelines. Lastly, multimodal data such as waveform vitals, images, and clinical notes would close information gaps that structured medical event data cannot capture.

**Counterfactual reasoning.** Future work will incorporate decision-effect estimation objectives and reinforce counterfactual consistency; explicit what-if analyses (e.g., “What if metformin were initiated today?”) will allow us to move beyond risk analysis and towards actionable risk mitigation.

**Time to event analysis.** Our evaluation tasks focused primarily on binary predictive outcomes, and censoring was not explicitly accounted for in our analysis of real and synthetic patient timelines. This does not take advantage of the full time-to-event predictive capacity that CoMET has to offer.

**Fine-tuning.** While CoMET achieves strong zero-shot performance, it can also be fine-tuned to improve task-specific performance and extend its capabilities to out-of-vocabulary tasks. Fine tuning offers an alternative to increasing test-time compute, eliminating sampling variance and enabling multiple precise calibrated predictions to be made with a single compute-efficient forward pass.

**Prospective studies.** Prospective studies of CoMET could help assess how well CoMET outputs align with clinical practice, which would inform downstream applications and model retraining schedules.

**Human factors research and governance.** CoMET models are generalizable to many clinical tasks, and deeply understanding best practices for introducing and scaling downstream applications within clinical workflows is key to interpretability, transparency, and usability. As with other healthcare technology, strong evaluation frameworks and thoughtful governance strategies are important to make CoMET applications sustainable, responsible, and fair.

Together, these findings position CoMET as a general-purpose engine for real-world evidence: it can screen populations for incident disease risk, forecast individual health timelines, surface differential diagnoses, and anticipate health system interactions—all from a single set of autoregressive generations.

## 4 Related Work

In this section, we highlight prior work that informs and contextualizes our study, focusing on three key areas: (1) large-scale, real-world medical event databases such as Epic Cosmos, (2) scaling laws for large language models, and (3) medical foundation models trained on structured and unstructured health data. This is not intended as an exhaustive survey of all related literature; rather, we summarize representative work that captures the state of the art, key challenges, and motivations for our approach.

### 4.1 Epic Cosmos

As described above, Cosmos is a collaboration for integrating and linking electronic health record (EHR) data between 310 healthcare organizations<sup>2</sup> in the United States, Canada, Lebanon, and Saudi Arabia as of August 2025 [6]. The Cosmos population from healthcare organizations in the United States is largely representative of census data. While some variation from national statistics has been noted, Cosmos has been shown to accurately reflect information and trends in other national databases (e.g. transplant registry, comorbidity patterns) [45, 46]. At the time of writing, most of the 121 published studies using Cosmos focus on describing the epidemiology or outcomes of specific conditions or services [47–49].

The current work is distinct insofar as most of our objectives are predictive rather than descriptive or associative—we predict diagnoses, acute exacerbation of chronic disease, and utilization outcomes. With that in mind, some of the conditions and outcomes we explored in this evaluation have been studied at a smaller

---

<sup>2</sup><https://cosmos.epic.com/community>

scale in previous research using Cosmos data. These studies highlight challenges and opportunities. Varghese et al. [50] used Cosmos to demonstrate that the type of subtle differences CoMET captures impact risk of acute events among patients with T2DM. Patel et al. [51] demonstrated that Cosmos rates of systemic lupus erythematosus (SLE) based on structured data in diagnosis codes were lower than expected, and paradoxical mortality findings in [52] highlight the challenges of working with EHR data. Relatedly, Cosmos has been used for developing and validating risk scores and strata in other conditions, including cancer-associated thrombosis [53], vision impairment in multiple sclerosis [54], and perinatal cardiovascular events among patients with rare congenital diseases [55]. Several Cosmos publications have studied healthcare utilization outcomes in distinct subpopulations [56–58]. Section 2.6 demonstrates how forecasting these types of metrics can be scaled with CoMET. While the present evaluation did not include the detailed phenotype modeling and validation included in some of these studies, CoMET data can readily be analyzed and fine tuned with greater specificity and logical complexity employed in other Cosmos studies. In this way, CoMET may inform and be informed by more traditional Cosmos research.

## 4.2 Scaling laws for LLMs

Research into the scaling behavior of machine learning models has fundamentally shaped the trajectory of large language models (LLMs). A pivotal study by Kaplan et al. [35] showed that as we increase a model’s size, the amount of training data, and the compute budget, the performance improves following a smooth power-law trend—a predictable pattern that governs how model performance grows with more resources. Henighan et al. [59] extended this analysis beyond text to other domains like images, video, and multi-modal tasks, finding that larger transformer models consistently yield better predictive power across modalities as compute and model size grow, again following power-law improvement curves. A further demonstration of scale came from Brown et al. [60], who presented 175B parameter GPT-3 and found that simply making the model extremely large unlocked emergent capabilities; notably the ability to perform new tasks in a few-shot setting without any task-specific training. Subsequently, Hoffmann et al. [34] observed that many earlier large models were trained with too little data. They trained a 70B Chinchilla with four times more data than GPT-3, using the same compute budget, and found it outperformed much larger models. The frontier of scale was later pushed further by Chowdhery et al. [61] and OpenAI [62, 63], where the performance continued to improve log-linearly with scale. Studies of LLM scaling laws provide a guiding principle for our work, which asks whether similar predictable gains occur when we scale up models on more patient data.

## 4.3 Medical foundation models

Inspired by the success of foundation models in general domains, researchers have started developing large-scale models tailored to electronic health records. BEHRT [64] introduced a BERT-like transformer model trained on the longitudinal medical histories of 1.6 million patients. This approach yielded substantial improvements for disease prediction compared to prior state-of-the-art models. Med-BERT [65] was developed from 28 million patients, which further validated the power of large-scale pretraining on structured EHR data. CLMBR [13] introduced an autoregressive next-code predictor trained on 3.4 million patient records; the representations it learned boosted AUCROC across multiple downstream tasks, especially in low-data settings. MOTOR [14] trained a transformer-based model for time-to-event (TTE) prediction using 55 million patient records, demonstrating the transfer learning ability for the TTE foundation model.

Moving beyond encoder-only models like BERT, recent research has explored generative and autoregressive transformers that can explicitly model the sequence of events in a patient’s timeline. CEHR-GPT [66] was one of the first attempts to train GPT models on structured EHR data. It showed the synthetic data generated by the model effectively captures the intricate patterns present in EHR data. Introduced by Renc et al. [16], ETHOS is trained on tokenized event streams of patient health timelines and tasked with predicting the next set of events in a patient’s record. Interestingly, it does so in a zero-shot fashion, effectively learning a simulator of possible futures for a patient. In a similar vein, Event Stream GPT [39] provides tooling to convert complex, irregular medical event sequences into a format that a transformer can ingest, and methods to handle the causally ordered generation of events. TransformEHR [67] is an encoder-decoder transformer pretrained with a “visit masking” strategy: it masks out all the medical codes in some future visit and trains the model to reconstruct them from the preceding history.

Other work has focused on incorporating unstructured data into medical foundation models. Foresight [15] integrated unstructured text with structured EHR data, where important details from doctors’ notes were converted into standardized medical concepts and combined with coded data as input to a GPT-based model. Foresight demonstrated the feasibility of training one large model to handle many prediction tasks across different institutions.

While these EHR-focused foundation models have shown encouraging results, they also highlight challenges. A recent comprehensive review by Wornow et al. [37] examined over 80 such models and found that many were trained on relatively narrow datasets (e.g., a single hospital’s records) and evaluated on surrogate tasks that may not translate to real clinical impact. EHRSHOT [68] and FoMoH [69] introduced new benchmark suites designed around patient timelines that extended beyond intensive care settings. These benchmarks emphasize robust, fair, and clinically meaningful evaluations. Instead of zero-shot approaches, both studies focused on smaller-scale models ( $\sim 100M$  parameters) trained using a pretrain-then-finetune paradigm. Recently, Zhang et al. [17] conducted the first study on scaling laws for EHR foundation models. They investigated how model performance scales with size and data volume on smaller scales using MIMIC-IV [70], identifying consistent patterns such as power-law relationships between compute resources, model parameters, and clinical utility.

## 5 Methods

### 5.1 Data Pre-Processing

To ensure that Cosmos supplies the foundation model with records of sufficient depth and quality, we apply a three-stage filter before tokenization.

**1. Patient selection.** We retain only adults who have meaningful longitudinal follow-up within Cosmos:

- Age between 18 and 120 years on January 1, 2012.
- At least two successive face-to-face encounters within a two-year period between January 1, 2012 and April 17, 2025.

The encounter-frequency requirement screens out patients whose primary care occurs outside Cosmos-contributing sites, while the 2012 index date avoids the sparse documentation that characterizes earlier years and simplifies downstream temporal alignment. We exclude pediatric patients because of significantly different care patterns (e.g., routine well-child visits, growth-chart measurements, age-specific dosing).

**2. Encounter selection.** From the eligible patients we include encounters based on these criteria:

- Encounter belongs to an eligible patient.
- The encounter start date on or after January 1, 2012 and before April 17, 2025.
- Encounter type is associated with direct clinical care (e.g., outpatient visit, emergency-department stay, inpatient admission, telehealth, and many others). We discard canceled appointments, test records, and administrative placeholders that rarely carry coded clinical data.

Encounter types span core clinical areas such as office visits (17.1%), emergency (2.1%), surgery (1.0%), and inpatient (0.76%), but also include a broad range of other encounter types from telemedicine (0.88%) and home care (1.0%) to anesthesia (0.97%) and prenatal visits (0.39%).

**3. Post-filter cleanup.** Patients left with zero qualifying encounters after the above steps are removed to prevent empty timelines. This results in 118M unique patient records in our full dataset.

**4. Train/Test Split.** 90% of the patients were randomly allocated for training, with the remaining 10% reserved for all evaluations. Experiments using temporal train/test splits to assess generalization across time are an important consideration for future work.

This pipeline yields a cohort whose records are dense enough to train a sequence model while minimizing noise from sporadic documentation and non-clinical artifacts. We applied very minimal filtering of patient records or input data to reflect the diversity of patients and the realities of real world healthcare documentation. For a full breakdown of our dataset, see [Table 4](#).

## 5.2 Tokenization Details

Our tokenization method adapts a few key techniques from Renc et al. [16]. In general, medical events are placed in chronological order according to the instant at which they were documented, with some type-specific nuances noted below. The token vocabulary is defined *a priori* based on the individual events that are possible rather than frequency-based methods like byte-pair encoding. When possible we tokenized using ontologies with codes that directly capture hierarchical and categorical information, such as ICD-10-CM and Anatomical Therapeutic Chemical (ATC). All CoMET models have a vocabulary size of 7,105, which is summarized in [Table 2](#).

| Event Type                      | Tokens/event | Number of Unique Tokens |
|---------------------------------|--------------|-------------------------|
| Sex                             | 1            | 6                       |
| Race                            | 1            | 7                       |
| Age and Years since 1970        | 1            | 24                      |
| Beginning of Sequence           | 1            | 1                       |
| Encounter Starts & Ends         | 1            | 226 (113 types)         |
| Department Specialties          | 1            | 299                     |
| Chief Complaint (Name)          | 1            | 1231                    |
| Chief Complaint (Body Location) | 1            | 67                      |
| Diagnoses                       | 1-3          | 2429                    |
| Lab Results                     | 1            | 1000                    |
| Lab Quantiles                   | 1            | 10                      |
| Medication Orders               | 1-3          | 289                     |
| Procedures                      | 1            | 1500                    |
| Time                            | $\geq 1$     | 13                      |

**Table 2:** Medical events included in CoMET’s vocabulary. The tokenizer also includes separation, padding, and unknown tokens.

### 5.2.1 Demographics

Patient history sequences begin with a set of demographics tokens, which represent patient attributes that are not tied to a single encounter. We included demographics tokens for sex, race, and age at first encounter, along with the number of years from 1970 to the start of the patient’s medical history (both in 5-year buckets). Additionally, we added a “Beginning of Sequence” token to denote the start of the patient timeline after the demographics section. Sex is bucketed into “Male”, “Female”, “Unknown”, “Masked”, “Other”, and “Unspecified”.

### 5.2.2 Encounters

Encounters are book-ended with start- and end-encounter tokens—each with a denoted encounter type (e.g., “Emergency\_Start”, “Emergency\_End”). Each encounter header contains the start token, a department specialty token, and possible chief complaint tokens. Chief complaints in Cosmos consist of a name and an optional body location, which we separated into two unique consecutive tokens per chief complaint. When

multiple encounters overlap, their encounter header and end-encounter tokens each appear at the proper instant in the timeline. In general, we do not enforce rules for overlapping or nested encounter-start or -end tokens; their placement in the patient sequence corresponds only to the time at which the event happened, with the tokens for a given encounter header always appearing consecutively.

### 5.2.3 Diagnoses

Diagnosis events are represented by the associated ICD-10-CM code. ICD-10-CM codes were split up into three tokens by category and sub-category (first 3 characters), specific details (characters 4-5), and additional details and extensions (characters 6-7). Diagnoses are represented in CoMET patient trajectories by 1-3 tokens depending on the specificity of the documentation in the patient’s chart. Because diagnosis events in Cosmos only have documentation resolution at the day level, they are placed at the very beginning of the encounter after the encounter header or at midnight of their documented date, whichever comes later. Diagnoses with the same date are sequenced in random order.

### 5.2.4 Labs

Our dataset contains the 1,000 most common numeric lab tests from Cosmos (representing >99% of all numeric lab test results). Each lab result is represented by two consecutive tokens which identify the test performed and the quantile of the numeric result, following ETHOS [16]. Lab components are represented by Logical Observation Identifiers Names and Codes (LOINC®)<sup>3</sup> codes, and numeric results are stratified by LOINC and measurement unit. For each LOINC-unit pair we partitioned values into 10 equal-frequency bins and mapped them to generic tokens. Compared to uniform-width binning, quantiles better accommodate non-linear mappings between heterogeneous units and better balance token frequencies, though they can compress rare but clinically meaningful extremes. The lab-result token pairs are placed in the sequence at the instance of collection. This ensures the sequence mirrors the patient’s evolving clinical state but calls for careful consideration to ensure that only information available at the evaluation time is used.

### 5.2.5 Medications

Medication orders are represented by an ATC code without other data. ATC codes were split up into three sets of tokens by anatomical group and therapeutic subgroup (characters 1-3), pharmacological and chemical subgroup (characters 4-5), and chemical substance (characters 6-7). Medication orders are represented in CoMET patient trajectories using three consecutive tokens to represent the full ATC code. They are placed in the sequence at the instant of the order.

### 5.2.6 Procedures

We extracted all billed procedures with Current Procedure Terminology (CPT®)<sup>4</sup> codes in Cosmos into our dataset. Our codes include such events as conventional procedures and imaging tests, as well as CPT codes associated with other billed codes like lab panel orders and level of service. We only tokenized the 1,500 most common procedures in our dataset to avoid sparse procedure tokens in our vocabulary (this represented 97.3% of all procedure events). They are placed in the sequence at procedure start instant.

### 5.2.7 Time

The passage of time was represented in the medical event sequence by one or more tokens that represent a time interval. The token is selected from a set of time ranges following Renc et al. [16], ranging from “1-5 minutes” to “3-6 months”. Events that occur within a shorter time span than the smallest time range are not separated using any time tokens. Events that occur more than 6 months apart have one or more “6 month” tokens, rounded to the nearest integer.

<sup>3</sup>LOINC® is a registered trademark of Regenstrief Institute, Inc.

<sup>4</sup>CPT® is a registered trademark of the American Medical Association.

### 5.2.8 Excluded Data

CoMET is limited to structured data and does not include data such as clinical notes, images, or free-text results from diagnostic procedures. Additional structured data, such as vitals, allergies, and medication administrations, were not included at this time.

## 5.3 Model Training Details

All CoMET models are built on the Qwen2 architecture. An overview of the architectural hyperparameters is shown in [Table 3](#). All variants are trained with a context window of 8,192 tokens.

| Model   | Params | Layers | Dimension | Heads | MLP dimension |
|---------|--------|--------|-----------|-------|---------------|
| CoMET-S | 62M    | 6      | 768       | 12    | 3072          |
| CoMET-M | 119M   | 12     | 768       | 12    | 3072          |
| CoMET-L | 1B     | 16     | 2048      | 32    | 8192          |

**Table 3:** Summary of CoMET model configurations, including Small (S), Medium (M), and Large (L) model sizes. Abbreviation: MLP = multi-layer perceptron

### 5.3.1 CoMET training and scaling laws

CoMET models are built on the Qwen2 architecture [18]. Qwen2 incorporates pre-layer normalization, SwiGLU activations, rotary positional embeddings, and grouped-query attention, all of which improve training time and stability. CoMET models were trained with cross-entropy loss and standard hyperparameters, slightly adjusting them as determined by experimentation to account for possible differences in training transformers on medical events rather than natural language. Batch size was fixed at 512 sequences, and input sequences were densely packed in order to fully use the context window during training, with only a separation token marking different patients since our experiments demonstrated this was sufficient for training. Following Hoffmann et al. [34], we employed a 10x learning rate decay with cosine schedule and used AdamW as our optimizer.

We estimated the compute  $C$  used in training a model by the number of TFLOPs required in the forward and backward passes, using PaLM’s methodology [61], where  $N$  is the number of parameters in the model and  $D$  is the number of training tokens used. To obtain our power-law fit for optimal parameter count  $N_{\text{opt}}$  and number of tokens  $D_{\text{opt}}$  for a fixed  $C$ , we fixed  $C$  and varied  $N$  and  $D$  to obtain a set of isoFLOP experiments ([Figure 15\[A\]](#)). Each isoFLOP’s loss-versus-log  $N$  and loss-versus-log  $D$  points were fit to parabolas, and the minimum point of the parabola served as  $N_{\text{opt}}$  and  $D_{\text{opt}}$  for that value of  $C$ . We then plotted log  $N_{\text{opt}}$  vs. log  $C$  ([Figure 15\[B\]](#)) and log  $D_{\text{opt}}$  vs. log  $C$  ([Figure 15\[C\]](#)) and fit a power law, using the first-degree  $\alpha$  and  $\beta$  terms, respectively, to measure the relative power-law scaling of optimal model size versus optimal training tokens.

## 5.4 Benchmarking with Task-Specific Supervised Models

To contextualize the performance of CoMET, we implemented three baseline approaches representing distinct modeling paradigms. These included (1) linear and logistic regression models, representing classical linear approaches; (2) gradient-boosted decision trees, a strong non-linear method widely used for structured data; and (3) supervised transformers trained from a random initialization, a flexible deep learning architecture that retains the temporal information of each patient record. Each baseline model was trained independently for its corresponding downstream task and evaluated using the same datasets and procedures as those applied to the CoMET foundational models.

**Sample construction.** For each downstream task, we applied the evaluation’s labeling logic to the CoMET training set to construct supervised examples. We used all supervised examples available in the CoMET training dataset up to a maximum of 5 million patient histories per task to fit the baselines. These samples

were grouped by patient id and stratified by the task label where applicable. We reserved an additional 625,000 stratified rows when available or 10% of the available training rows as a validation set for early stopping and hyper-parameter selection, leaving the designated CoMET development and test splits untouched for final evaluation.

**Input representation** With the exception of the supervised transformer baselines, input prompts were converted into a bag-of-words (BoW) count vector over the full CoMET vocabulary ([Table 2](#)). Columns were normalized such that the maximum absolute value of each feature in the training set was 1.0 before fitting the linear and logistic models; XGBoost consumed raw counts. Preliminary experiments with inverse-frequency class-weighting showed no material AUCROC gain on these large samples, so we report unweighted results for simplicity.

**Linear and logistic regression.** We trained linear models with mini-batch stochastic gradient descent for up to 1,000 epochs over the five million row dataset with early stopping after 5 epochs without improvement. Classification used a logistic loss; regression used a squared loss with an L1 regularization penalty.

**Gradient-boosted decision trees.** We used a gradient-boosted decision trees classifier or regressor with `n_estimators=10,000`, `max_depth=6`, `learning_rate=0.1`, and `subsample=0.8`. Training employed early stopping with a patience of 100 boosting rounds. Other hyper-parameters followed XGBoost 1.5.2 library defaults unless specified.

**Supervised Transformer.** We trained the closest task-specific supervised counterpart to CoMET-M by using the same architecture as CoMET-M ([Table 3](#)) and the same tokenized input prompts. Unlike the bag-of-words inputs, this baseline preserves model architecture and the temporal sequence information in the original prompts. We attached a task-specific classification or regression head to the model and optimized with AdamW (weight decay 0.01), using a linear-decay schedule with warmup (500 warmup steps) and a peak learning rate of  $1 \times 10^{-4}$ . Training ran for 10,000 update steps (global batch size 128; 1.28M samples). We applied gradient-norm clipping at 1.0 and dropout 0.0.

We evaluated the supervised transformer baseline on a representative subset of tasks. The results are summarized in [Figure 25](#) and [Table 15](#). Across the evaluated tasks, XGBoost consistently outperformed logistic regression and was generally stronger than the transformer baseline, despite requiring less compute. Based on these findings, we used XGBoost as the primary task-specific supervised comparator but always report the strongest baseline result obtained. Estimating the value added by pretraining and the transfer-learning it enables would require a deeper analysis of performance saturation under an unbounded task-specific compute budget and is left to future work.

## 5.5 Inference Details

Inference with CoMET was performed by generating future medical event sequences via Monte Carlo sampling. For a given inference case, a patient history up to a specific moment was given as context to the model, and the model generated a specified number  $d$  of output tokens  $n$  times at a temperature of 1, where  $d$  and  $n$  depended on the needs of the individual task. If the amount of patient history exceeded the size of the model’s context window (8,192 tokens for all CoMET models), the history is left-truncated as necessary. When this is necessary, the model loses the ability to see demographics and older information, but this aligns with how the model sees truncated sequences during pretraining.

All of our evaluations ask questions about events over a specified time range, not number of tokens. The model’s  $i$ th generation is a series of tokens  $(y_{i1}, y_{i2}, \dots, y_{id})$ , where  $y_{ij} \in V$  are tokens in the vocabulary  $V$ . Times are assigned by incrementing the current time at generated time tokens based on the geometric midpoint of the interval. For each token  $y \in V$  there is an associated time bucket  $(t_{\min}^{(y)}, t_{\max}^{(y)})$ , which is  $(0, 0)$  for non-time tokens, and the geometric midpoint is given by  $\Delta t_y = \sqrt{t_{\min}^{(y)} t_{\max}^{(y)}}$ . Therefore, for generated tokens we create a sequence of event tuples  $T_i = ((y_{i1}, t_{i1}), (y_{i2}, t_{i2}), \dots, (y_{id}, t_{id}))$ , where  $t_{ij} = t_{i(j-1)} + \Delta t_{y_{i(j-1)}}$  and  $(y_{i0}, t_{i0})$  is the final event-time tuple in the model’s context window. Because generated trajectories

might not reach the full time length when generating by number of tokens, generated trajectories must be right-censored, and so the probability that a token in the target set  $S$  occurs within time  $\tau$  when using  $n$  generations is given by:

$$P_\tau(S) = \frac{\sum_{i=1}^n \mathbb{1}[\exists j : y_{ij} \in S \text{ and } t_{ij} \leq \tau]}{\sum_{i=1}^n \max(\mathbb{1}[t_{id} > \tau], \mathbb{1}[\exists j : y_{ij} \in S \text{ and } t_{ij} \leq \tau])} \quad (2)$$

If the denominator is zero, we exclude the patient from the evaluation. Similar logic is used for getting the probability of counts of events in generated sequences, as well as the distribution of time-to-event outcomes. Future work is needed to improve model generations to always reach  $\tau$  regardless of output length.

## 5.6 Evaluation Details

All evaluations were performed using patients from the held-out test set of patients.

### 5.6.1 Plausibility statistics

We selected 20,000 random patients and performed generations starting from the end of the patient’s last encounter that ended prior to 2022. To select for patients with some minimal degree of activity within health systems contributing data to Cosmos, we required that these patients have at least one prior encounter as well as one encounter beyond a year’s time from the generation start. We used  $n = 25$  generations for each patient, using up to 2000 tokens to reach one year’s time, and discarding those generations that did not reach time. We used these generations to measure the percent of multi-token events that were invalid ([Appendix C](#)), the prevalence and co-occurrence rates ([Appendix D](#)), and the probability calibrations of the number and types of encounters ([Section 2.2.2](#)). For prevalence and co-occurrence, we measured the overall agreement between ground truth and predicted rates using root-mean squared log error (RMSLE):

$$\text{RMSLE} = \sqrt{\sum_{i=1}^N [\log_{10}(x_i + \epsilon) - \log_{10}(\hat{x}_i + \epsilon)]^2} \quad (3)$$

where  $N$  is the number of event concepts, and  $(x_i, \hat{x}_i)$  are the prevalence or co-occurrence rate of individual concepts in ground truth and generated sequences, making RMSLE essentially the log-transformed root-mean square error (RMSE). For the probability calibration plots, we pooled patients into cohorts using equal quantile bin edges, and in each probability cohort measured the average predicted probability and the ground truth positive fraction. We quantified overall calibration for each count bucket using expected calibration error (ECE), which is just the MAE of the calibration curve with respect to the diagonal (i.e., perfect calibration).

### 5.6.2 Single-encounter generations

Encounters were selected at random from across the test set without filtering for patients with a certain amount of previous history. This was done to mimic real clinical practice. We evaluated encounters that were either office visits, emergency visits, or inpatient admissions from the emergency department. For each encounter type, 10,000 encounters were selected, and all the medical events that occurred between encounter start and encounter end were tabulated, specifically for diagnosis, medication, lab, and procedure data types. Prompts terminated at the end of the encounter header (i.e., encounter type, department specialty, and any chief complaints), and generations stopped once the model generated the appropriate encounter end token. Twenty generations were used per sample, with up to 2,000 tokens allowed per generation. In order to be considered a true positive, the exact same code had to appear in both the patient simulation and the ground truth encounter, even for multi-token diagnosis and medication events. Micro-averaged precision and recall were computed for each encounter type and data type using different thresholds, and PR-AUC was determined from these precision-recall curves.

For reference points, we pooled together the medical events from the patient’s history prior to the encounter and measured the precision and recall of this method. We did this for different lookback windows, indicated by the gray dots in [Figure 4](#).

### 5.6.3 Disease-Specific Outcomes

Each sample provided to CoMET includes a patient’s health timeline from the first token up to and including the order of a medication indicated for one of the following conditions: T2DM, hypertension, and hyperlipidemia. When a sample exceeded the available context or left too little room for generation, we applied left truncation, removing tokens from the beginning of the sample. Given a sample, CoMET generated up to 2,000 new tokens from which predictions were made programmatically.

Binary predictions, such as determining whether an adverse event occurred within a determined time  $t$ , were made by scanning the generation up to time  $t$  for a set of ICD-10-CM codes that describe an adverse outcome. Continuous predictions of lab values were made by uniformly sampling from within the range of values encapsulated by the relevant lab’s quantile bucket. If multiple labs of the desired type were produced in a given patient simulation, the average of the uniform samples was used. For all prediction types, generations that did not reach the time threshold  $t$  required by outcome definitions were discarded. CoMET was evaluated on 30,000 samples per condition and predictions extracted from the model’s generations were micro-averaged over 80 generations per sample prior to calculating metrics.

### 5.6.4 Incident Disease Risk Prediction

All tasks here were formulated as binary prediction of a patient disease state within a 2 year window. Each prediction point in this cohort is unique to one patient and was chosen to be the last instant of a randomly selected outpatient encounter between 2020 and 2022. All patients were required to have at least 2 encounters of either outpatient or “emergent” encounter types in the 2 years prior to the prediction date. “Emergent” encounters are inpatient admission, emergency department, and urgent care visits. Additionally, all patients were required to be at least 18 and less than 120 years old at the prediction date to be included in these cohorts. For each incident disease prediction tasks, patients are excluded if they have had any of the diagnosis codes of any type from the phenotype definition before the prediction point.

Each target’s phenotype was classified as “chronic” or “emergent” and defined by a list of ICD-10-CM codes within certain encounter types (see [Table 24](#)). A chronic target was marked positive if a diagnosis appeared in at least two outpatient encounters or in a single emergent encounter in the patient’s history. An emergent target was marked positive only if a diagnosis occurred during an emergent encounter. Only encounter diagnoses and billing diagnoses were used for this target gathering. Patients in the “Dementia (2yr)” cohort were also required to be 60 years or older at the prediction date to make the task more difficult.

Roughly 5K patients and prediction points were sampled from the test set for each task (see [Table 19](#) for exact counts). Since there were broad inclusion and minimal exclusion criteria, the dataset was naturally very imbalanced with all tasks having <1.5% positive prevalence. Random upsampling on the minority class was performed for each task cohort so at minimum 500 positive samples are present. During the calculation of performance metrics, samples were weighted by the inverse of the resampling factor so that the class proportions in the analysis matched those of in the full test set:

$$w_+ = \frac{P_+^{\text{original}}}{P_+^{\text{resample}}}, \quad (4)$$

$$w_- = \frac{P_-^{\text{original}}}{P_-^{\text{resample}}}, \quad (5)$$

where  $w_+$  and  $w_-$  are the weights assigned to positive and negative samples, and  $P_{\pm}^{\text{original}}$  and  $P_{\pm}^{\text{resample}}$  denote the number of positive/negative cases before and after up-sampling, respectively. These weights restore the population-level prevalence, ensuring that prevalence-dependent metrics remain unbiased despite the synthetic inflation of positives.

Inference on these tasks followed our typical strategy with the following modifications. We generated  $n = 60$  simulations for each patient initially with 2,000 tokens, and we retried up to 2 times to try to get the patient generations to the 2-year prediction time duration. When evaluating patient timelines generated by CoMET, the timeline would be evaluated to have the target if the timeline has any occurrence of any ICD-10-CM codes from the target phenotype within the corresponding encounter types for that phenotype.

### 5.6.5 Acute on Chronic Event Prediction

The non-task specific inclusion criteria and the prediction date selection criteria used for incident disease risk prediction were also used as the base criteria for acute-on-chronic event prediction, with additional task-specific inclusion criteria applied for each condition. Each acute-on-chronic event prediction task consists of a chronic and an acute (“emergent”) phenotype, such as sickle cell disease and sickle cell crisis respectively. Find the full list of phenotypes in [Table 24](#).

To be included in the prediction cohort, a patient must meet both the base inclusion criteria and have the chronic phenotype before the prediction date. This means they must have at least two outpatient encounters or one emergent encounter with a clinical or billed diagnosis code from the phenotype before the prediction date.

To meet the acute phenotype criteria, a patient must have an emergent encounter with an encounter or billing diagnosis code within the prediction window in their ground truth data. Patients who have already had this acute event prior to the prediction date are still included within the cohort. A patient timeline generated by CoMET would be evaluated to have the target if the timeline has any occurrence of any ICD-10-CM codes from the acute phenotype within an emergent encounter. The same sampling strategy, sample weighting for performance metrics, and sampling parameters as incident disease risk prediction are used in this evaluation. [Table 19](#) has more details on dataset size and positive prevalence.

### 5.6.6 Differential diagnosis

For the HPB and rheumatic diagnosis clusters, we selected 9 conditions from each to represent a range of prevalence and disease types. Specific definitions of each are in [Table 22](#). Patients were selected for evaluation based on the presence of relevant diagnosis codes in their record. We used retrospective inclusion based on the diagnosis outcomes in order to focus this evaluation on patients who have one of the target diagnoses. In order to be included in a disease cohort, the patient had to have at least two occurrences of one of the eligible ICD-10-CM codes at separate encounters (this was done to mitigate diagnoses documented as part of rule-out diagnostic tests but for which the patient did not receive additional care). Time was measured relative to the encounter containing the first occurrence of an eligible ICD-10-CM code, which was then used as the index time  $t = 0$ . Additionally, to be included, patients were required to both have an encounter *prior to* one year before the index diagnosis date as well as one encounter *within* one year before the index diagnosis date. We then selected patients who received one of these diagnoses. Once a patient was selected, the first occurrence of each off-target diagnosis was identified for that patient. Selecting at most 9 random encounters from the year *prior* to  $t = 0$ , we had CoMET generate from the end of these encounters using  $n = 40$  generations and up to 2,000 output tokens to predict a patient’s 18-month risk of having that diagnosis. Additionally, we also generated from the encounter header for the encounter at  $t = 0$  to measure CoMET’s final predictions. We chose not to generate from within this encounter due to an increased chance of information leakage and to focus the evaluation on *early* diagnosis detection and differential diagnosis.

For the training of the task-specific supervised models, we selected many additional patients. For a proper comparison to CoMET, we chose to include patients who were being worked up for, or were at risk of developing, a much wider set of diagnoses than HPB or rheumatic diagnoses. This makes the comparison to CoMET more appropriate and also better reflects the diverse clinical scenarios a deployed system might encounter when flagging patients at risk of being diagnosed with specific diseases. For training, we expanded our pool of eligible patients by broadening the pool of possible inclusion diagnoses from those in our HPB and rheumatic lists to all 3-character ICD-10-CM codes, and selected multiple prediction dates for each as described above. We then trained 18 task-specific binary classifiers, one for each of the 18 HPB and rheumatic diagnoses.

In order to visualize and contextualize the results, a patient’s predicted risk over the subsequent 18 months was used as their diagnosis risk at a moment in time, and that risk remained the same until their next encounter for which we had model predictions. In [Figure 10](#) and [Figure 11](#), the predictions are plotted for each day from 6 months out up to  $t = 0$ . We first assessed CoMET’s ability to flag at-risk patients, choosing a threshold probability of 10%. Each individual plot only contains patients we had selected as having the indicated diagnosis at  $t = 0$ . For each of the off-target lines in the plot, we also excluded patients whose ground truth record ever contained that off-target diagnosis (in the past or in the future), so that the off-target diagnosis lines represent the percent of patients flagged with an “incorrect” diagnosis.

### 5.6.7 Forecasting patients’ interactions with the health system

We used the same patients and CoMET generations described above in [Section 5.6.1](#). For each encounter type indicated, we computed the mean number of encounters predicted by CoMET in the following year, as well as that predicted by the task-specific regression model.

### 5.6.8 Length of stay

To evaluate CoMET’s ability to predict inpatient length of stay, we randomly selected 10,000 Hospital Admissions from our test set and labeled each with the total length of stay from admission to discharge in seconds. For each hospital admission, we prompt CoMET with the patient’s history up until the point of admission (including the department specialty and any associated chief complaint and flagged primary encounter diagnosis events). We then generate up to 2,000 new tokens with  $n = 20$  inpatient trajectories until the encounter’s stop token appears. We take the median time-to-encounter-end as the model’s prediction.

### 5.6.9 30-day readmission

We constructed prediction targets by selecting 10,000 patients discharged between January 1, 2020, and the present date, subsequently determining whether these patients experienced a hospital readmission within 30 days. Readmissions occurring within 24 hours of discharge were excluded from this analysis because they may often reflect hospital transfers or clerically erroneous discharges [33]. A 30-day readmission in a CoMET generation was defined as the beginning of a new inpatient encounter within 30 days of the discharge time.

## 6 Acknowledgments

We thank Zach Galvin and Zhuowen Nie for managing the high-performance computing environment; Andrea Noel, Kersten Bartelt, and Jackie Gerhart for carefully reviewing our evaluations to ensure clinical accuracy; Brian Olson and Amy Kim for creating and advising on the graphic designs; Samson Race Dorfman for his help reviewing code and data quality; Phil Lindemann, Matthew Lungren, Jonathan Carlson, and Joe Petro for project guidance; the Epic Cosmos R&D team for building and maintaining the essential infrastructure; and the Cosmos Community for the courage demonstrated in creating Cosmos, the dataset that made this work possible.

## References

- [1] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019. doi: 10.1056/NEJMra1814259.
- [2] Rachel E. Sherman, Steven A. Anderson, Gerald J. Dal Pan, Gerry W. Gray, Thomas Gross, Nina L. Hunter, Lisa LaVange, Danica Marinac-Dabic, Peter W. Marks, Melissa A. Robb, Jeffrey Shuren, Robert Temple, Janet Woodcock, Lilly Q. Yue, and Robert M. Califf. Real-World Evidence — What Is It and What Can It Tell Us? *New England Journal of Medicine*, 375(23):2293–2297, 2016. doi: 10.1056/NEJMsb1609216.
- [3] U.S. Food and Drug Administration. Use of Electronic Health Record Data in Clinical Investigations: Guidance for Industry. [FDA Guidance Website](#), Jul 2018. Guidance Document, docket number FDA-2016-D-1224. Accessed 2025-07-28.
- [4] John Concato and Jacqueline Corrigan-Curay. Real-World Evidence—Where Are We Now? *The New England Journal of Medicine*, 386(18):1680–1682, Apr 2022. doi: 10.1056/NEJMp2200089.
- [5] U.S. Congress. H.R. 34—21st Century Cures Act. [Congress.gov - Bill H.R. 34](#), 2016. Public Law 114–255; accessed 2024-07-10.

- [6] Yasir Tarabichi, Adam Frees, Steven Honeywell, Courtney Huang, Adam M Naidech, Jason H. Moore, and David C. Kaelber. The Cosmos Collaborative: A Vendor-Facilitated Electronic Health Record Data Aggregation Platform. *ACI Open*, 5(1):36–46, 2021. doi: 10.1055/s-0041-1731004.
- [7] Sara W. Kelly, Dale L. Smith, Zachary Davis, Robin Mermelstein, and Niranjan S. Karnik. Waves of Change: A Nationwide Analysis of Acute Pediatric Healthcare Encounters Attributed to Substance Use Before, During, and Following the COVID-19 Pandemic. *Journal of the American Academy of Child & Adolescent Psychiatry*, 2025. ISSN 0890-8567. doi: 10.1016/j.jaac.2025.06.019.
- [8] Rashmi Lamsal, Cameron G. Estrich, Danessa Sandmann, Kersten Bartelt, and Ruth D. Lipman. Declining US dental amalgam restorations in US Food and Drug Administration-identified populations: 2017–2023. *Journal of the American Dental Association*, 155(10):816–824, Oct 2024. ISSN 0002-8177. doi: 10.1016/j.adaj.2024.07.015.
- [9] Leon S. Moskatel, Oyindamola Ogunlaja, and Niushen Zhang. Prevalence, demographics, comorbidities, and treatment patterns of patients with the trigeminal autonomic cephalgias: A retrospective analysis of United States electronic health records. *BMC Neurology*, 25(299), Jul 2025. ISSN 1471-2377. doi: 10.1186/s12883-025-04314-1.
- [10] Allison Kranyak, Jillian Rork, Joshua Levy, and Timothy E. Burdick. Alopecia areata and thyroid screening in down syndrome: Leveraging Epic Cosmos data set. *Journal of the American Academy of Dermatology*, 89(2):360–361, 2023. ISSN 0190-9622. doi: 10.1016/j.jaad.2023.03.026.
- [11] Nahiyah Bin Noor, George Pro, Mahip Acharya, Hari Eswaran, and Corey J. Hayes. Association between different modalities of opioid use disorder-related care delivery and opioid use disorder-related patient outcomes: A retrospective cohort study. *Addictive Behaviors Reports*, 21:100588, 2025. ISSN 2352-8532. doi: <https://doi.org/10.1016/j.abrep.2025.100588>.
- [12] Robert W. Turer, Samuel A. McDonald, Robin T. Higashi, Bhaskar Thakur, Andrew P. Bain, Ann Marie Navar, and Bryan D. Steitz. Who uses patient portals? a national cross-sectional study using epic cosmos. *Journal of General Internal Medicine*, May 2025. ISSN 1525-1497. doi: 10.1007/s11606-025-09585-5.
- [13] Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfahl, and Nigam H. Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021. ISSN 1532-0464. doi: 10.1016/j.jbi.2020.103637.
- [14] Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. MOTOR: A Time-To-Event Foundation Model For Structured Medical Records. <https://arxiv.org/abs/2301.03150>, 2023.
- [15] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, James T. Teo, and Richard J. B. Dobson. 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, Apr 2024. ISSN 2589-7500. doi: 10.1016/S2589-7500(24)00025-6.
- [16] 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), Sep 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01235-0.
- [17] Sheng Zhang, Qin Liu, Naoto Usuyama, Cliff Wong, Tristan Naumann, and Hoifung Poon. Exploring Scaling Laws for EHR Foundation Models. <https://arxiv.org/abs/2505.22964>, 2025.
- [18] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu

Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. <https://arxiv.org/abs/2407.10671>, 2024.

- [19] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb 2015. doi: 10.1609/aaai.v29i1.9602.
- [20] American Diabetes Association Professional Practice Committee. 6. glycemic goals and hypoglycemia: Standards of care in diabetes—2025. *Diabetes Care*, 48(Supplement 1):S128–S145, Dec 2024. ISSN 0149-5992. doi: 10.2337/dc25-S006.
- [21] Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, 10(3):1–21, Mar 2015. doi: 10.1371/journal.pone.0118432.
- [22] Brice Ozenne, Fabien Subtil, and Delphine Maucort-Boulch. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, 68(8):855–859, 2015. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2015.02.010.
- [23] George G.A Pujalte and Shirley A. Albano-Aluquin. Differential Diagnosis of Polyarticular Arthritis. *American Family Physician*, 92(1):35–41, Jul 2015.
- [24] Melanie Sloan, Rupert Harwood, Stephen Sutton, David D’Cruz, Paul Howard, Chris Wincup, James Brimicombe, and Caroline Gordon. Medically explained symptoms: a mixed methods study of diagnostic, symptom and support experiences of patients with lupus and related systemic autoimmune diseases. *Rheumatology Advances in Practice*, 4(1):rkaa006, Feb 2020.
- [25] H S El-Gabalawy, P Duray, and R Goldbach-Mansky. Evaluating patients with arthritis of recent onset: studies in pathogenesis and prognosis. *JAMA*, 284(18):2368–2373, Nov 2000.
- [26] Paul Y Kwo, Stanley M Cohen, and Joseph K Lim. ACG Clinical Guideline: Evaluation of Abnormal Liver Chemistries. *American Journal of Gastroenterology*, 112(1):18–35, Jan 2017.
- [27] Nicole E Rich and Amit G Singal. Overdiagnosis of hepatocellular carcinoma: Prevented by guidelines? *Hepatology*, 75(3):740–753, Mar 2022.
- [28] Douglas S Swords, Mary C Mone, Chong Zhang, Angela P Presson, Sean J Mulvihill, and Courtney L Scaife. Initial Misdiagnosis of Proximal Pancreatic Adenocarcinoma Is Associated with Delay in Diagnosis and Advanced Stage at Presentation. *Journal of Gastrointestinal Surgery*, 19(10):1813–1821, Oct 2015.
- [29] James M. Halle-Smith, David Bartlett, Nikolaos Chatzizacharias, Bobby VM. Dasari, Ravi Marudanayagam, Robert P. Sutcliffe, Rachel M. Brown, and Keith J. Roberts. Investigating misdiagnosis of suspected cancer among patients undergoing pancreaticoduodenectomy: three decades of experience. *HPB*, 26:344–351, Mar 2024.
- [30] Kush M. Patel, Jingwen Zhang, Justin Marsden, Chloe Bays, Patrick D. Mauldin, and Andrew D. Schreiner. Missed and Delayed Diagnoses of Chronic Liver Disease in Primary Care Patients with Cirrhosis. *Digestive Diseases and Sciences*, 69, Oct 2024.
- [31] Lalit Garg, Sally I. McClean, Maria Barton, Brian J. Meenan, and Ken Fullerton. Intelligent Patient Management and Resource Planning for Complex, Heterogeneous, and Stochastic Healthcare Systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 42(6):1332–1345, 2012. doi: 10.1109/TSMCA.2012.2210211.
- [32] Stephen F. Jencks, Mark V. Williams, and Eric A. Coleman. Rehospitalizations among Patients in the Medicare Fee-for-Service Program. *New England Journal of Medicine*, 360(14):1418–1428, 2009. doi: 10.1056/NEJMsa0803563.

- [33] Yale New Haven Health Services Corporation Center for Outcomes Research and Evaluation (YN-HHSC/CORE). 2024 Hospital-Wide Readmission Measure Updates and Specifications Report (Version 13.0). Technical report, Centers for Medicare & Medicaid Services, Apr 2024.
- [34] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. <https://arxiv.org/abs/2203.15556>, 2022.
- [35] 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. <https://arxiv.org/abs/2001.08361>, 2020.
- [36] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. <https://arxiv.org/abs/2408.03314>, 2024.
- [37] 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(135), 2023. doi: 10.1038/s41746-023-00879-8.
- [38] Fang Liu and Demosthenes Panagiotakos. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*, 22(287), 2022. doi: 10.1186/s12874-022-01768-6.
- [39] Matthew B. A. McDermott, Bret Nestor, Peniel Argaw, and Isaac Kohane. Event Stream GPT: A Data Pre-processing and Modeling Library for Generative, Pre-trained Transformers over Continuous-time Sequences of Complex Events. <https://arxiv.org/abs/2306.11547>, 2023.
- [40] Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanusse, Michael McCabe, Ruben Ohana, Liam Parker, Bruno Régaldo-Saint Blancard, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. xVal: A Continuous Numerical Tokenization for Scientific Language Models. <https://arxiv.org/abs/2310.02989>, 2024.
- [41] Andrew J. Loza, Jun Yup Kim, Shangzheng Song, Yihang Liu, Joseph J. Y. Sung, R Andrew Taylor, and Dennis L. Shung. multivariateGPT: a decoder-only transformer for multivariate categorical and numeric data. <https://arxiv.org/abs/2505.21680>, 2025.
- [42] Yury Gorishniy, Ivan Rubachev, and Artem Babenko. On Embeddings for Numerical Features in Tabular Deep Learning. <https://arxiv.org/abs/2203.05556>, 2023.
- [43] Inkit Padhi, Yair Schiff, Igor Melnyk, Mattia Rigotti, Youssef Mroueh, Pierre Dognin, Jerret Ross, Ravi Nair, and Erik Altman. Tabular Transformers for Modeling Multivariate Time Series. <https://arxiv.org/abs/2011.01843>, 2021.
- [44] Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling Laws with Vocabulary: Larger Models Deserve Larger Vocabularies. <https://arxiv.org/abs/2407.13623>, 2024.
- [45] Michal A. Mankowski, Sunjae Bae, Alexandra T. Strauss, Bonnie E. Lanze, Babak J. Orandi, Darren Stewart, Allan B. Massie, Mara A. McAdams-DeMarco, Eric K Oermann, Marlena Habal, Eduardo Iturrate, Sommer E. Gentry, Dorry L. Segev, and David Axelrod. Generalizability of kidney transplant data in electronic health records — The Epic Cosmos database vs the Scientific Registry of Transplant Recipients. *American Journal of Transplantation*, 25(4):744–755, 2025. doi: 10.1016/j.ajt.2024.11.008.
- [46] Eric G.R. Kim and David C. Kaelber. Phenotypic prevalence of obesity and metabolic syndrome among an underdiagnosed and underscreened population of over 50 million children and adults. *Frontiers in Genetics*, Volume 13 - 2022, 2022. ISSN 1664-8021. doi: 10.3389/fgene.2022.961116.

- [47] Tyler B Nofzinger, Timothy T Huang, Christopher Eduard R Lingat, Gaurang M Amonkar, Emily E Edwards, Albert Yu, Alexander D Smith, Nasser Gayed, and Heidi L Gaddey. Vaccine fatigue and influenza vaccination trends across Pre-, Peri-, and Post-COVID-19 periods in the United States using epic's cosmos database. *PLoS One*, 20(6):e0326098, 2025.
- [48] JM Chowdhury, CS King, M Desai, A Kasarabada, M Patel, and MG Kashiouris. Epidemiology and Outcomes of Veno-arterial ECMO in Acute Pulmonary Embolism: A Retrospective Cohort Study Using Nationwide EHR Data. *American Journal of Respiratory and Critical Care Medicine*, 211:A3737–A3737, 2025.
- [49] Moeun Son, Kieran Gallagher, Justin Y Lo, Eric Lindgren, Heather H Burris, Kevin Dysart, Jay Greenspan, Jennifer F Culhane, and Sara C Handley. Coronavirus disease 2019 (COVID-19) Pandemic and Pregnancy Outcomes in a U.S. Population. *Obstetrics & Gynecology*, 138(4):542–551, 2021.
- [50] Jithin Sam Varghese, Zhongyu Li, Aamna Soniwala, and Mohammed Ali. Abstract 076: Risk of Atherosclerotic Cardiovascular Disease and Heart Failure by Type 2 Diabetes Subtypes. *Circulation*, 151(Suppl\_1), 2025. doi: 10.1161/cir.151.suppl\\_\\_1.076.
- [51] Jay Patel, Lixia Yao, Ernest Vina, David Fleece, Jayatilleke Arundathi, Roberto Caricchio, and Huanmei Wu. Phenotype Systemic Lupus Erythematosus Patients from Epic Cosmos. *Studies in Health Technology and Informatics*, 310:159–163, 2024. ISSN 1879-8365. doi: 10.3233/SHTI230947.
- [52] Craig C. Hofmeister, Mingyuan Zhang, Vikas A. Gupta, Nisha S. Joseph, Jonathan L. Kaufman, Ajay K. Nooka, Madhav V. Dhodapkar, Sagar Lonial, and Douglas W. Sborov. Survival of Multiple Myeloma Patients in Epic Cosmos Confounded By Death Reporting. *Blood*, 144:5152, 2024. doi: 10.1182/blood-2024-202100.
- [53] Ang Li, Rockbum Kim, Omid Jafari, Shengling Ma, Jun Yang Jiang, Nathanael R Fillmore, Jennifer La, and Christopher I Amos. External Validation of EHR-CAT Risk Assessment Model for Cancer Associated Thrombosis in 150 Healthcare Organizations. *Blood*, 144:812, 2024. doi: 10.1182/blood-2024-202183.
- [54] Brandon Buxton, Amr Hassan, Nevin Shalaby, John Lindsey, John Lincoln, Elmer Bernstam, Wagida Anwar, Degui Zhi, and Laila Rasmy. Vision Impairment prediction for patients diagnosed with Multiple Sclerosis: Cosmos based machine learning model training and evaluation. *medRxiv*, 2023. doi: 10.1101/2023.11.10.23298366.
- [55] Gianna L Wilkie, Rahi Patel, Lara Kovell, and Anna Whelan. Maternal Cardiac and Perinatal Outcomes Among Pregnant Individuals with Ebstein's Anomaly. *The American Journal of Cardiology*, 2025.
- [56] Michael Gottlieb, Eric Moyer, Kyle Bernard, and Kevin G Buell. Delays in Diagnosis of Diabetes Mellitus Among Emergency Department Patients with Hyperglycemia Using the Epic Cosmos Database. *Journal of General Internal Medicine*, pages 1–4, 2025.
- [57] Christopher S Evans, Robert W Turer, John J Hanna, Emilie Pendley, and Richard J Medford. Emergency Department Return Visits After Having Left Without Being Seen. *JAMA*, 333(9):806–808, 2025.
- [58] E.E. Moin, N.J. Seewald, and S.D. Halpern. Characteristics and Prognosis of Patients Experiencing Recurrent ICU Admissions. *American Journal of Respiratory and Critical Care Medicine*, 211(Abstracts), 2025.
- [59] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling Laws for Autoregressive Generative Modeling. <https://arxiv.org/abs/2010.14701>, 2020.
- [60] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack

Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. <https://arxiv.org/abs/2005.14165>, 2020.

- [61] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways. <https://arxiv.org/abs/2204.02311>, 2022.
- [62] OpenAI. GPT-4 Technical Report. <https://arxiv.org/abs/2303.08774>, 2024.
- [63] OpenAI. OpenAI GPT-4.5 System Card. <https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf>, 2025. Accessed: 2025-02-27.
- [64] Yikuan Li, Shishir Rao, José Solares Roberto Ayala, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhuand, Kazem Rahimiand, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10:7155, 2020.
- [65] 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*, 4:86, 2021.
- [66] Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Elise L. Minto, Jason Patterson, Lining Zhang, George Hripcsak, Gamze Gürsoy, Noémie Elhadad, and Karthik Natarajan. CEHR-GPT: Generating Electronic Health Records with Chronological Patient Timelines. <https://arxiv.org/abs/2402.04400>, 2024.
- [67] 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(7857), 2023.
- [68] Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A. Fries, and Nigam H. Shah. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. <https://arxiv.org/abs/2307.02028>, 2023.
- [69] Chao Pang, Vincent Jeanselme, Young Sang Choi, Xinzhuo Jiang, Zilin Jing, Aparajita Kashyap, Yuta Kobayashi, Yanwei Li, Florent Pollet, Karthik Natarajan, and Shalmali Joshi. FoMoH: A clinically meaningful foundation model evaluation for structured electronic health records. <https://arxiv.org/abs/2505.16941>, 2025.
- [70] Alistair E.W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023. URL <https://www.nature.com/articles/s41597-023-02136-9>.
- [71] Chronic Conditions. [Chronic Conditions Data Warehouse Website](#), 2025. Centers for Medicare & Medicaid Services.
- [72] Tessa L. Steel, Theresa E. Matson, Kevin A. Hallgren, Malia Oliver, Helen E. Jack, Douglas Berger, and Katharine A. Bradley. Incidence of Hospitalizations Involving Alcohol Withdrawal Syndrome in a Primary Care Population. *JAMA*, 7(10):e2438128, Oct 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.38128.

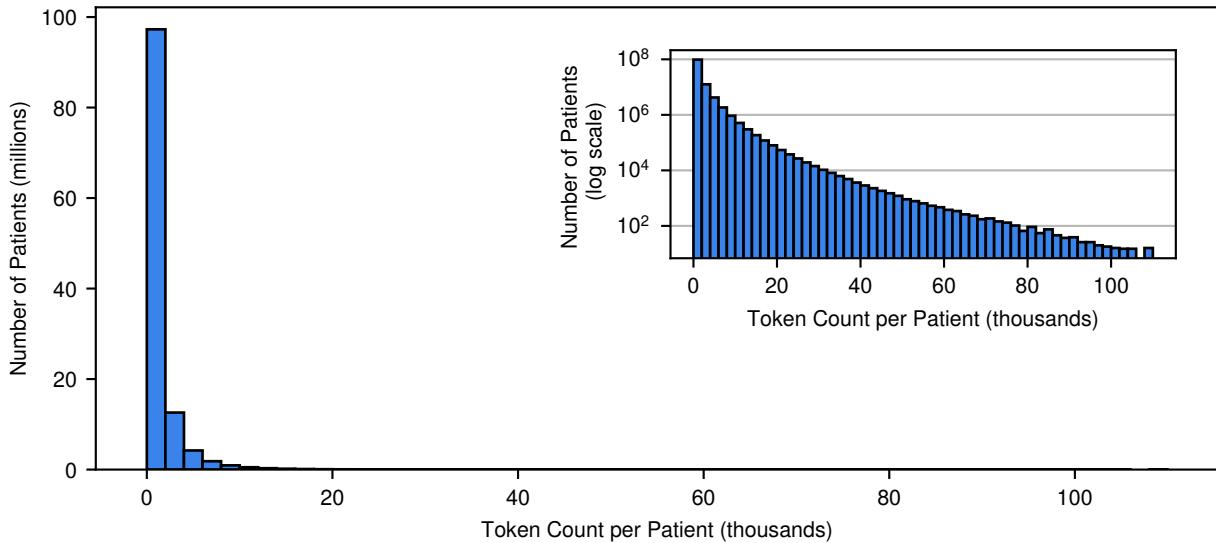
- [73] Jove Graham, Andy Iverson, Joao Monteiro, Katherine Weiner, Kara Southall, Katherine Schiller, Mudit Gupta, and Edgar P. Simard. Applying computable phenotypes within a common data model to identify heart failure patients for an implantable cardiac device registry. *IJC Heart and Vasculature*, 39:100974, 2022. ISSN 2352-9067. doi: 10.1016/j.ijcha.2022.100974.
- [74] Scott D. Grosse, Nancy S. Green, and Sarah L. Reeves. Administrative data identify sickle cell disease: A critical review of approaches in U.S. health services research. *Pediatric Blood & Cancer*, 67(12):e28703, 2020. doi: 10.1002/pbc.28703.

## Appendix A: CoMET dataset statistics

After data pre-processing as described in [Section 5.1](#), data from Epic Cosmos was transformed into a subset used for the CoMET models, which was split into a train and test set at the patient level. Summary statistics of this dataset are shown in [Table 4](#). Cosmos maintains and reports on over 1,000 metrics to assess data quality, focused on completeness, conformance, and plausibility. Some of these data quality metrics as measured on the CoMET dataset subset are displayed in [Table 5](#).

| Characteristic                 | Group                  | Train  | Test   | Total  |
|--------------------------------|------------------------|--------|--------|--------|
| Total Counts                   | Patients               | 106 M  | 11.8 M | 118 M  |
|                                | Events                 | 104 B  | 11.5 B | 115 B  |
|                                | Encounters             | 7.65 B | 850 M  | 8.50 B |
|                                | Diagnoses              | 15.3 B | 1.70 B | 17.0 B |
|                                | Labs                   | 15.8 B | 1.76 B | 17.6 B |
|                                | Medications            | 7.35 B | 817 M  | 8.17 B |
|                                | Procedures             | 9.98 B | 1.11 B | 11.1 B |
|                                | Tokens                 | 136 B  | 15.1 B | 151 B  |
|                                | Tokens/Patient         | 1278.3 | 1277.7 | 1278.3 |
| Age                            | Tokens/Encounter       | 17.7   | 17.7   | 17.7   |
|                                | 18 - 39                | 43.6 M | 4.84 M | 48.4 M |
|                                | 40 - 59                | 38.8 M | 4.31 M | 43.1 M |
|                                | 60 - 79                | 21.2 M | 2.35 M | 23.5 M |
| Race                           | 80+                    | 2.90 M | 323 K  | 3.22 M |
|                                | White                  | 74.9 M | 8.33 M | 83.3 M |
|                                | BoAA                   | 14.6 M | 1.62 M | 16.2 M |
|                                | Asian                  | 4.67 M | 519 K  | 5.19 M |
|                                | AI/AN                  | 1.00 M | 112 K  | 1.12 M |
|                                | NHOPI                  | 386 K  | 42.7 K | 429 K  |
| Ethnicity                      | Other                  | 4.96 M | 551 K  | 5.51 M |
|                                | Not Hispanic or Latino | 85.8 M | 9.54 M | 95.3 M |
|                                | Hispanic or Latino     | 11.3 M | 1.26 M | 12.6 M |
| Sex                            | Unspecified            | 9.34 M | 1.04 M | 10.4 M |
|                                | Female                 | 58.4 M | 6.49 M | 64.9 M |
|                                | Male                   | 48.0 M | 5.33 M | 53.3 M |
|                                | Unknown                | 36.3 K | 4.10 K | 40.4 K |
|                                | Masked                 | 4.97 K | 572    | 5.54 K |
|                                | Other                  | 984    | 119    | 1.10 K |
| Number of source organizations | Unspecified            | 877    | 91     | 968    |
|                                | 1                      | 61.9 M | 6.89 M | 68.8 M |
|                                | 2                      | 31.0 M | 3.44 M | 34.5 M |
|                                | 3                      | 10.1 M | 1.12 M | 11.2 M |
|                                | 4+                     | 3.38 M | 376 K  | 3.76 M |

**Table 4:** CoMET dataset counts in the train, test, and full data sets, organized by various data types, demographics, and the number of source organizations contributing to each patient record. Abbreviations: BoAA = Black or African American, AI/AN = American Indian or Alaska Native, NHOPI = Native Hawaiian or Other Pacific Islander. Only breakdowns for First Race are shown.



**Figure 18: Distribution of token counts per patient.** Histogram of all patient token counts in the train and test sets, with a bin size of 2,000 tokens. Note: The inset plot has a log scaled y-axis to effectively visualize the long tail, and bins with counts less than or equal to 10 have been omitted.

| Data Type   | Data Quality Metric                    | Percentage |
|-------------|--|------------|
| Patients    | Has first race                         | 94.4%      |
|             | Has ethnicity                          | 91.2%      |
|             | Has legal sex                          | 100.0%     |
|             | Has birth date                         | 100.0%     |
|             | Birth date after death                 | 0.0%       |
|             | Has diagnosis                          | 98.4%      |
|             | Has lab results                        | 87.5%      |
|             | Has medications                        | 92.5%      |
|             | Has problem list diagnosis             | 78.8%      |
| Encounters  | Has start date                         | 100.0%     |
|             | Has end instant                        | 100.0%     |
|             | Has department specialty               | 85.8%      |
|             | Has chief complaint                    | 32.0%      |
|             | Has admission instant (inpatient only) | 100.0%     |
|             | Has discharge instant (inpatient only) | 100.0%     |
|             | Has encounter type                     | 100.0%     |
|             | Start instant outside lifetime         | 0.1%       |
| Medications | Has medication code                    | 100.0%     |
|             | Has frequency                          | 88.2%      |
|             | Has dose                               | 79.4%      |
|             | Min dose < 0                           | 0.0%       |
| Labs        | Has LOINC code                         | 100.0%     |
|             | Has collected instant                  | 99.8%      |
|             | Has resulted instant                   | 99.5%      |
|             | Collected after resulted               | 0.2%       |
|             | Resulted > 30 days after encounter end | 0.2%       |

**Table 5:** CoMET dataset data quality metrics by data type

## Appendix B: Tabulated results

| Evaluation                 | Task                  | Metric  | CoMET-S | CoMET-M      | CoMET-L        |
|----------------------------|-----------------------|---------|---------|--------------|----------------|
| Invalid multi-token events | Diagnoses             | Percent | 0.057%  | 0.029%       | <b>0.011%</b>  |
|                            | Medications           | Percent | 0.037%  | 0.023%       | <b>0.0087%</b> |
|                            | Lab results           | Percent | 0.0016% | 0.0009%      | <b>0.0002%</b> |
|                            | Encounter headers     | Percent | 0.012%  | 0.011%       | <b>0.006%</b>  |
| Event prevalence           | Diagnoses             | RMSLE   | 0.311   | 0.292        | <b>0.281</b>   |
|                            | Labs                  | RMSLE   | 0.291   | 0.292        | <b>0.278</b>   |
|                            | Medications           | RMSLE   | 0.240   | 0.233        | <b>0.215</b>   |
|                            | Procedures            | RMSLE   | 0.311   | 0.289        | <b>0.264</b>   |
| Event co-occurrence        | Diagnosis-Diagnosis   | RMSLE   | 0.117   | 0.107        | <b>0.102</b>   |
|                            | Diagnosis-Lab         | RMSLE   | 0.178   | 0.174        | <b>0.170</b>   |
|                            | Diagnosis-Medication  | RMSLE   | 0.104   | 0.101        | <b>0.097</b>   |
|                            | Diagnosis-Procedure   | RMSLE   | 0.194   | 0.188        | <b>0.182</b>   |
|                            | Lab-Lab               | RMSLE   | 0.142   | 0.138        | <b>0.135</b>   |
|                            | Lab-Medication        | RMSLE   | 0.118   | 0.118        | <b>0.116</b>   |
|                            | Lab-Procedure         | RMSLE   | 0.216   | 0.215        | <b>0.210</b>   |
|                            | Medication-Medication | RMSLE   | 0.049   | 0.049        | <b>0.047</b>   |
|                            | Medication-Procedure  | RMSLE   | 0.127   | 0.126        | <b>0.122</b>   |
|                            | Procedure-Procedure   | RMSLE   | 0.170   | 0.168        | <b>0.164</b>   |
| Encounter frequency        | Office visit, 0       | ECE     | 0.065   | 0.050        | <b>0.043</b>   |
|                            | Office visit, 1-2     | ECE     | 0.058   | 0.057        | <b>0.051</b>   |
|                            | Office visit, 3-5     | ECE     | 0.056   | <b>0.049</b> | 0.050          |
|                            | Office visit, 6+      | ECE     | 0.068   | 0.055        | <b>0.041</b>   |
|                            | Emergency, 0          | ECE     | 0.035   | <b>0.028</b> | 0.030          |
|                            | Emergency, 1          | ECE     | 0.031   | 0.029        | <b>0.029</b>   |
|                            | Emergency, 2          | ECE     | 0.028   | 0.025        | <b>0.022</b>   |
|                            | Emergency, 3+         | ECE     | 0.024   | 0.020        | <b>0.018</b>   |
|                            | Inpatient, 0          | ECE     | 0.027   | 0.022        | <b>0.017</b>   |
|                            | Inpatient, 1          | ECE     | 0.029   | 0.027        | <b>0.024</b>   |
|                            | Inpatient, 2          | ECE     | 0.011   | 0.011        | <b>0.009</b>   |
|                            | Inpatient, 3+         | ECE     | 0.005   | 0.004        | <b>0.004</b>   |

**Table 6:** Percent, RMSLE, and ECE score comparisons across plausibility and encounter frequency tasks (Section 2.2.1 and Section 2.2.2) for each CoMET model.

| Evaluation                  | Task                      | Reference | CoMET-S | CoMET-M | CoMET-L      |
|-----------------------------|---------------------------|-----------|---------|---------|--------------|
| Single-encounter generation | Office visit, Diagnoses   | 0.098     | 0.480   | 0.502   | <b>0.548</b> |
|                             | Office visit, Labs        | 0.129     | 0.312   | 0.331   | <b>0.421</b> |
|                             | Office visit, Medications | 0.075     | 0.176   | 0.195   | <b>0.251</b> |
|                             | Office visit, Procedures  | 0.127     | 0.520   | 0.552   | <b>0.638</b> |
|                             | Emergency, Diagnoses      | 0.106     | 0.390   | 0.412   | <b>0.450</b> |
|                             | Emergency, Labs           | 0.415     | 0.782   | 0.808   | <b>0.840</b> |
|                             | Emergency, Medications    | 0.170     | 0.475   | 0.497   | <b>0.536</b> |
|                             | Emergency, Procedures     | 0.158     | 0.522   | 0.550   | <b>0.600</b> |
|                             | Inpatient, Diagnoses      | 0.171     | 0.408   | 0.431   | <b>0.469</b> |
|                             | Inpatient, Labs           | 0.595     | 0.846   | 0.871   | <b>0.899</b> |
|                             | Inpatient, Medications    | 0.320     | 0.552   | 0.574   | <b>0.616</b> |
|                             | Inpatient, Procedures     | 0.237     | 0.537   | 0.565   | <b>0.619</b> |

**Table 7:** PR-AUC score comparisons for single-encounter generation tasks (Section 2.2.3) for each CoMET model and various baseline lookback windows.

| Condition       | Task (time horizon)                      | Supervised   | CoMET-S | CoMET-M | CoMET-L      |
|-----------------|--|--------------|---------|---------|--------------|
| Type 2 Diabetes | ASCVD (1yr)                              | <b>0.878</b> | 0.861   | 0.867   | 0.875        |
|                 | ASCVD (3yr)                              | <b>0.907</b> | 0.857   | 0.871   | 0.894        |
|                 | CKD Prog. Stage 2 → 3 (1yr)              | 0.710        | 0.720   | 0.736   | <b>0.762</b> |
|                 | CKD Prog. Stage 2 → 3 (3yr)              | 0.766        | 0.739   | 0.757   | <b>0.785</b> |
|                 | CKD Prog. Stage 3 → 4 <sup>+</sup> (1yr) | 0.774        | 0.748   | 0.773   | <b>0.796</b> |
|                 | CKD Prog. Stage 3 → 4 <sup>+</sup> (3yr) | 0.799        | 0.740   | 0.775   | <b>0.813</b> |
|                 | Diabetic Neuropathy (1yr)                | 0.885        | 0.900   | 0.905   | <b>0.914</b> |
|                 | Diabetic Neuropathy (3yr)                | 0.906        | 0.881   | 0.895   | <b>0.911</b> |
|                 | Diabetic Retinopathy (1yr)               | 0.899        | 0.889   | 0.894   | <b>0.908</b> |
|                 | Diabetic Retinopathy (3yr)               | <b>0.910</b> | 0.859   | 0.876   | 0.902        |
|                 | HgbA1c < 7 (60-120 days)                 | <b>0.764</b> | 0.724   | 0.752   | 0.761        |
|                 | HgbA1c < 9 (60-120 days)                 | 0.687        | 0.724   | 0.760   | <b>0.756</b> |
| Hypertension    | HgbA1c < 11 (60-120 days)                | 0.710        | 0.726   | 0.738   | <b>0.760</b> |
|                 | HgbA1c < 12 (60-120 days)                | 0.730        | 0.661   | 0.689   | <b>0.736</b> |
|                 | ASCVD (1yr)                              | 0.854        | 0.837   | 0.842   | <b>0.862</b> |
|                 | ASCVD (3yr)                              | <b>0.893</b> | 0.851   | 0.865   | 0.889        |
|                 | CKD Prog. Stage 2 → 3 (1yr)              | 0.707        | 0.719   | 0.724   | <b>0.771</b> |
|                 | CKD Prog. Stage 2 → 3 (3yr)              | 0.772        | 0.757   | 0.768   | <b>0.811</b> |
|                 | CKD Prog. Stage 3 → 4 <sup>+</sup> (1yr) | 0.748        | 0.738   | 0.751   | <b>0.788</b> |
|                 | CKD Prog. Stage 3 → 4 <sup>+</sup> (3yr) | 0.813        | 0.736   | 0.760   | <b>0.813</b> |
|                 | Heart Attack (1yr)                       | <b>0.828</b> | 0.784   | 0.786   | 0.828        |
|                 | Heart Attack (3yr)                       | <b>0.870</b> | 0.806   | 0.811   | 0.853        |
| Hyperlipidemia  | Stroke (1yr)                             | <b>0.836</b> | 0.773   | 0.780   | 0.812        |
|                 | Stroke (3yr)                             | <b>0.867</b> | 0.787   | 0.780   | 0.840        |
|                 | ASCVD (1yr)                              | <b>0.860</b> | 0.830   | 0.834   | 0.853        |
|                 | ASCVD (3yr)                              | <b>0.892</b> | 0.845   | 0.859   | 0.883        |
|                 | Heart Attack (1yr)                       | <b>0.855</b> | 0.772   | 0.781   | 0.809        |
|                 | Heart Attack (3yr)                       | <b>0.878</b> | 0.804   | 0.814   | 0.842        |
|                 | Chronic Heart Failure (1yr)              | <b>0.938</b> | 0.916   | 0.919   | 0.925        |
|                 | Chronic Heart Failure (3yr)              | <b>0.945</b> | 0.898   | 0.908   | 0.920        |
|                 | Stroke (1yr)                             | <b>0.855</b> | 0.796   | 0.788   | 0.815        |
|                 | Stroke (3yr)                             | <b>0.879</b> | 0.799   | 0.808   | 0.837        |

**Table 8:** AUCROC on disease-specific outcome tasks (Section 2.3.1) for each CoMET model and the best-performing supervised task-specific model.

| Task (time horizon)                                      | Supervised   |        | CoMET-S |        | CoMET-M |        | CoMET-L      |              |
|--|--------------|--------|---------|--------|---------|--------|--------------|--------------|
|  | AUCROC       | PR-AUC | AUCROC  | PR-AUC | AUCROC  | PR-AUC | AUCROC       | PR-AUC       |
| CHF Exacerbation (2yr) / CHF                             | 0.788        | 0.475  | 0.767   | 0.405  | 0.799   | 0.475  | <b>0.827</b> | <b>0.511</b> |
| Alcohol Withdrawal Syndrome (2yr) / Alcohol Use Disorder | 0.823        | 0.434  | 0.830   | 0.441  | 0.839   | 0.445  | <b>0.857</b> | <b>0.490</b> |
| Sickle Cell Crisis (2yr) / Sickle Cell Disease           | 0.886        | 0.801  | 0.884   | 0.790  | 0.898   | 0.814  | <b>0.913</b> | <b>0.844</b> |
| COPD Exacerbation (2yr) / COPD                           | 0.820        | 0.447  | 0.812   | 0.397  | 0.832   | 0.448  | <b>0.847</b> | <b>0.479</b> |
| Asthma Exacerbation (2yr) / Asthma                       | <b>0.835</b> | 0.285  | 0.798   | 0.230  | 0.827   | 0.293  | 0.832        | <b>0.300</b> |

**Table 9:** AUCROC and PR-AUC on acute-on-chronic tasks (Section 2.3.2) for each CoMET model and the best-performing supervised task-specific model.

| Task (time horizon)        | Supervised   |        | CoMET-S |        | CoMET-M |        | CoMET-L      |              |
|----------------------------|--------------|--------|---------|--------|---------|--------|--------------|--------------|
|                            | AUCROC       | PR-AUC | AUCROC  | PR-AUC | AUCROC  | PR-AUC | AUCROC       | PR-AUC       |
| COPD (2yr)                 | 0.828        | 0.084  | 0.807   | 0.048  | 0.826   | 0.110  | <b>0.839</b> | <b>0.114</b> |
| CHF (2yr)                  | <b>0.894</b> | 0.125  | 0.873   | 0.091  | 0.883   | 0.143  | 0.891        | <b>0.184</b> |
| Dementia (2yr)             | <b>0.933</b> | 0.106  | 0.893   | 0.063  | 0.908   | 0.144  | 0.921        | <b>0.152</b> |
| Asthma (2yr)               | <b>0.795</b> | 0.058  | 0.770   | 0.030  | 0.770   | 0.083  | 0.782        | <b>0.065</b> |
| Alcohol Use Disorder (2yr) | <b>0.822</b> | 0.027  | 0.753   | 0.015  | 0.768   | 0.026  | 0.801        | <b>0.072</b> |
| Heart Attack (2yr)         | <b>0.852</b> | 0.063  | 0.802   | 0.054  | 0.804   | 0.069  | 0.801        | <b>0.092</b> |

**Table 10:** AUCROC and PR-AUC on incident disease risk prediction (Section 2.3.3) for each CoMET model and the best-performing supervised task-specific model.

| Evaluation              | Task                            | Supervised | CoMET-S | CoMET-M | CoMET-L      |
|-------------------------|---------------------------------|------------|---------|---------|--------------|
| Hepatopancreato-biliary | Acute Pancreatitis              | 0.512      | 0.526   | 0.557   | <b>0.594</b> |
|                         | Chronic Pancreatitis            | 0.565      | 0.677   | 0.703   | <b>0.733</b> |
|                         | Pancreatic Cancer               | 0.481      | 0.774   | 0.789   | <b>0.818</b> |
|                         | Cholecystitis                   | 0.548      | 0.637   | 0.659   | <b>0.706</b> |
|                         | Cholangitis                     | 0.603      | 0.686   | 0.706   | <b>0.744</b> |
|                         | Liver Cancer                    | 0.576      | 0.787   | 0.797   | <b>0.821</b> |
|                         | Chronic Viral Hepatitis         | 0.438      | 0.723   | 0.771   | <b>0.833</b> |
|                         | Alcoholic Liver Disease         | 0.528      | 0.806   | 0.830   | <b>0.873</b> |
|                         | Non-Alcoholic Steatohepatitis   | 0.495      | 0.679   | 0.710   | <b>0.774</b> |
| Rheumatic               | Osteoarthritis                  | 0.497      | 0.671   | 0.679   | <b>0.696</b> |
|                         | Rheumatoid Arthritis            | 0.491      | 0.605   | 0.624   | <b>0.664</b> |
|                         | Psoriatic Arthritis             | 0.488      | 0.740   | 0.750   | <b>0.791</b> |
|                         | Polymyalgia Rheumatica          | 0.500      | 0.618   | 0.671   | <b>0.724</b> |
|                         | Systemic Sclerosis              | 0.504      | 0.636   | 0.618   | <b>0.663</b> |
|                         | Systemic Lupus Erythematosus    | 0.469      | 0.689   | 0.699   | <b>0.732</b> |
|                         | Mixed Connective Tissue Disease | 0.609      | 0.687   | 0.687   | <b>0.701</b> |
|                         | Polymyositis/Dermatomyositis    | 0.540      | 0.644   | 0.647   | <b>0.687</b> |
|                         | Fibromyalgia                    | 0.484      | 0.693   | 0.712   | <b>0.750</b> |

**Table 11:** AUCROC for differential diagnosis prediction (Section 2.4) at  $t = 0$  (i.e., the start of the encounter of the target diagnosis event) for each CoMET model and the best-performing supervised task-specific model.

| Task (time horizon)              | Supervised | CoMET-S | CoMET-M | CoMET-L      |
|----------------------------------|------------|---------|---------|--------------|
| Number of inpatient visits (1yr) | 0.187      | 0.091   | 0.090   | <b>0.088</b> |
| Number of emergency visits (1yr) | 0.519      | 0.372   | 0.366   | <b>0.364</b> |
| Number of office visits (1yr)    | 2.361      | 1.859   | 1.777   | <b>1.712</b> |

**Table 12:** MAE for one-year encounter count forecasting (Section 2.5) for each CoMET model and the best-performing supervised task-specific model.

| Task                     | Supervised | CoMET-S | CoMET-M | CoMET-L      |
|--------------------------|------------|---------|---------|--------------|
| Admissions under 7 days  | 2.032      | 1.346   | 1.279   | <b>1.238</b> |
| Admissions under 14 days | 2.267      | 1.904   | 1.843   | <b>1.757</b> |
| All admissions           | 3.339      | 3.091   | 3.006   | <b>2.851</b> |

**Table 13:** MAE on length of stay tasks (Section 2.5.1) measured in days for each CoMET model and the best-performing supervised task-specific model.

| Model      | AUCROC       |
|------------|--------------|
| CoMET-L    | <b>0.770</b> |
| CoMET-M    | 0.723        |
| CoMET-S    | 0.706        |
| Supervised | 0.718        |

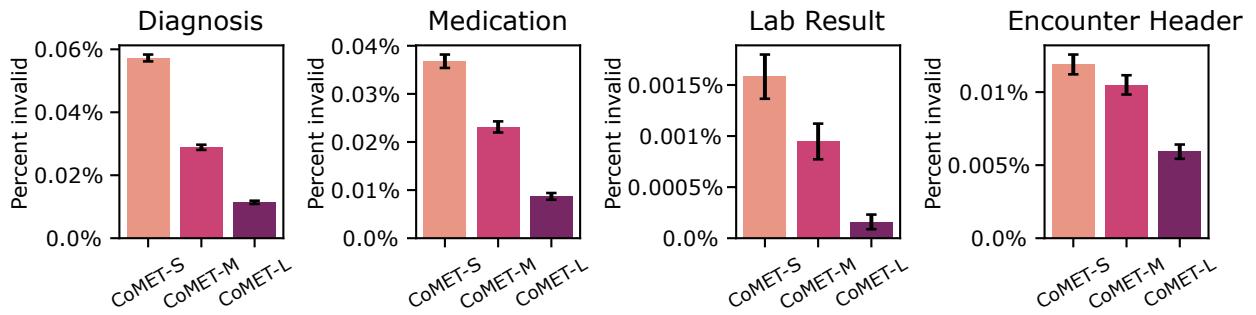
**Table 14:** AUCROC scores on the 30-day readmission task (Section 2.5.1) for the CoMET models and the best-performing task-specific model.

| Evaluation     | Task                     | Transformer | XGBoost | Linear Regression |
|----------------|--------------------------|-------------|---------|-------------------|
| Length of Stay | All Admissions           | <b>3.34</b> | 3.57    | 3.67              |
|                | Admissions under 14 days | <b>2.27</b> | 2.53    | 2.58              |
|                | Admissions under 7 days  | <b>2.03</b> | 2.34    | 2.35              |

**Table 15:** MAE (days) on the length of stay task (Section 2.5.1) across the task-specific supervised models, including a supervised 119M parameter transformer described in Section 5.4.

## Appendix C: Multi-token event validity

Each of the CoMET models was evaluated for the frequency of generating invalid multi-token events. A diagnostic event is considered invalid if there are ICD-10-CM tokens in combination that do not correspond to real ICD-10-CM codes. The same was done for medication orders with ATC codes. A lab result is considered invalid if the lab token is not followed by a lab result quantile token. Encounter headers were considered invalid if an encounter start token was not directly followed by a department specialty token (as noted in Table 5, not all encounters have a specified department specialty, but the encounter headers still have a token with a specialty type of ‘unspecified’). These experiments were done with 20,000 patients, each with 25 1-year generations. The denominator for each metric is the number of tokens that initiated an event.

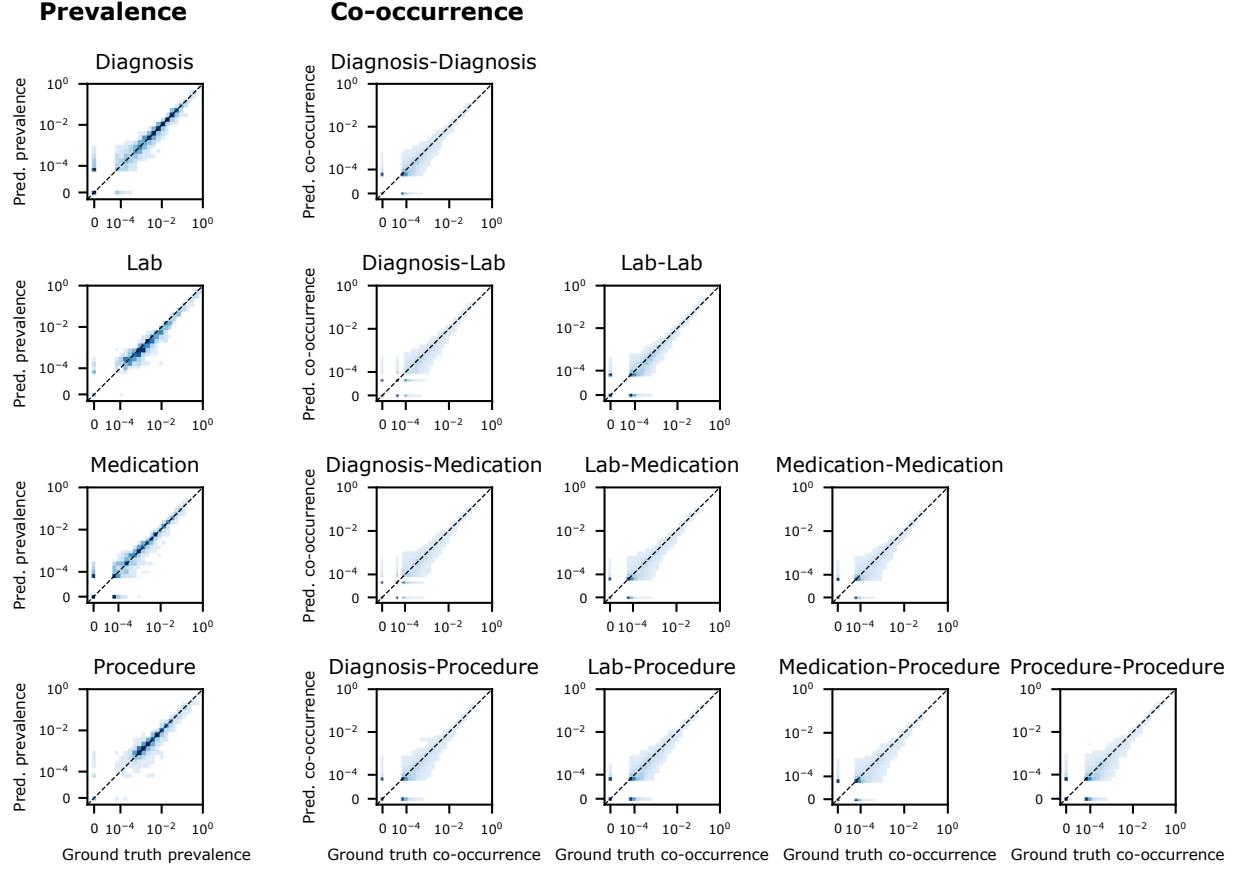


**Figure 19: Generation syntactic validity.** The percent of diagnosis, medication, lab result, and encounter header events that were invalidly generated by each of the CoMET models.

## Appendix D: Event prevalence and co-occurrence

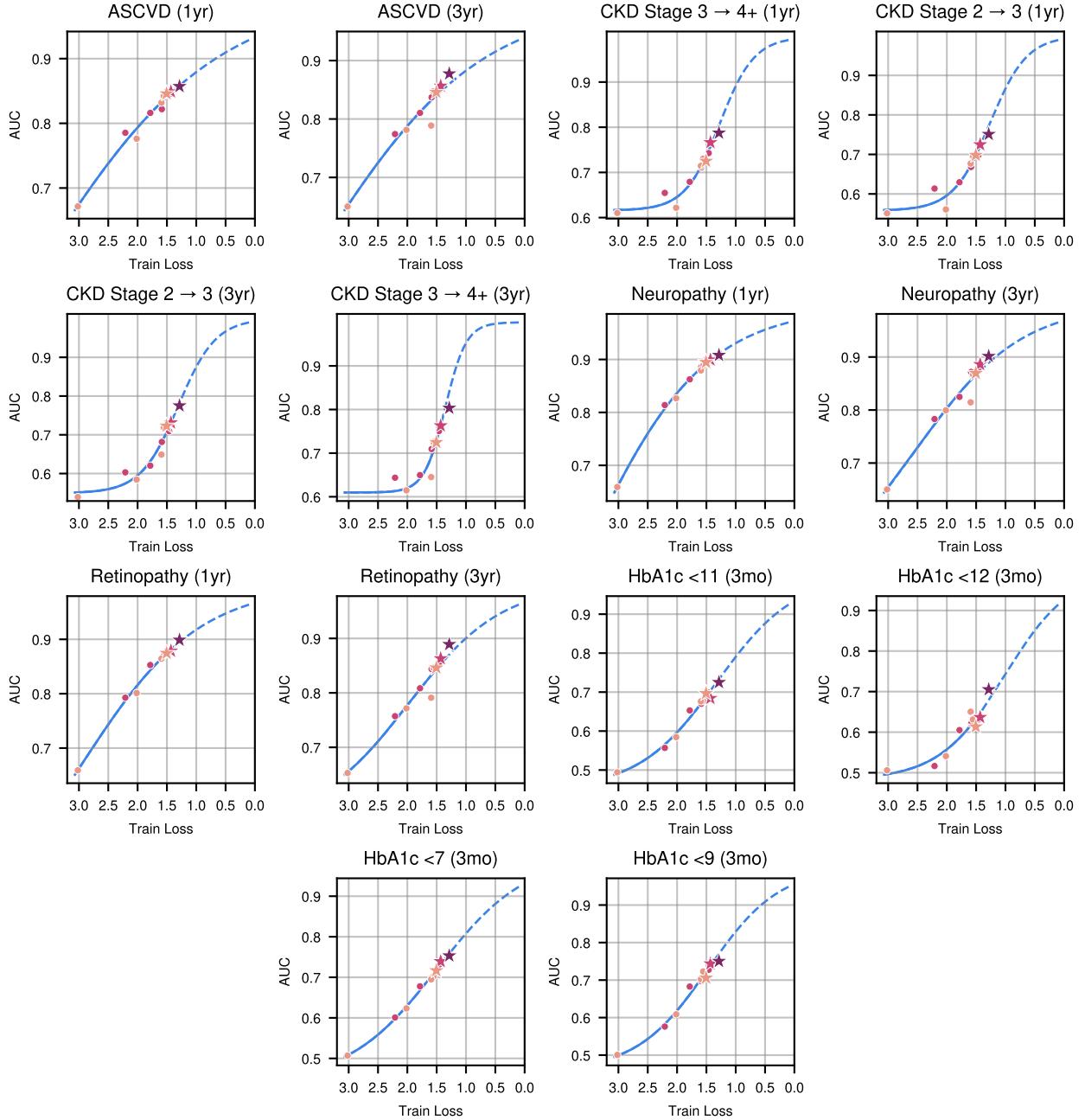
One way to assess the plausibility of CoMET’s generations is to measure the prevalence of generated medical events. Given the model has to learn how to properly generate syntactically valid medical event sequences, it is not necessarily trivial to also learn the frequency at which different events of different types should appear for individual patients. For 20,000 patients, we generated 1 year’s worth of sequences 25 times and measured, on average, how many times different medical events occurred. We measured this for diagnoses (at the 3-character ICD-10-CM code level for simplicity), labs, medications, and procedures, as seen in Figure 20.

In addition to measuring agreement in the prevalence of single events in generated sequences, we next asked whether pairs of events co-occur within a patient’s generated sequence the proper amount. This second-degree question is a quick, aggregate measure that the model understands the relationship between events, both within and across data types. We measured the fraction of patients in which each possible pair of events co-occurred. Results in Figure 20 and Table 6 show that CoMET generally does well at generating events at the same rate of co-occurrence as observed in real patient medical event sequences.

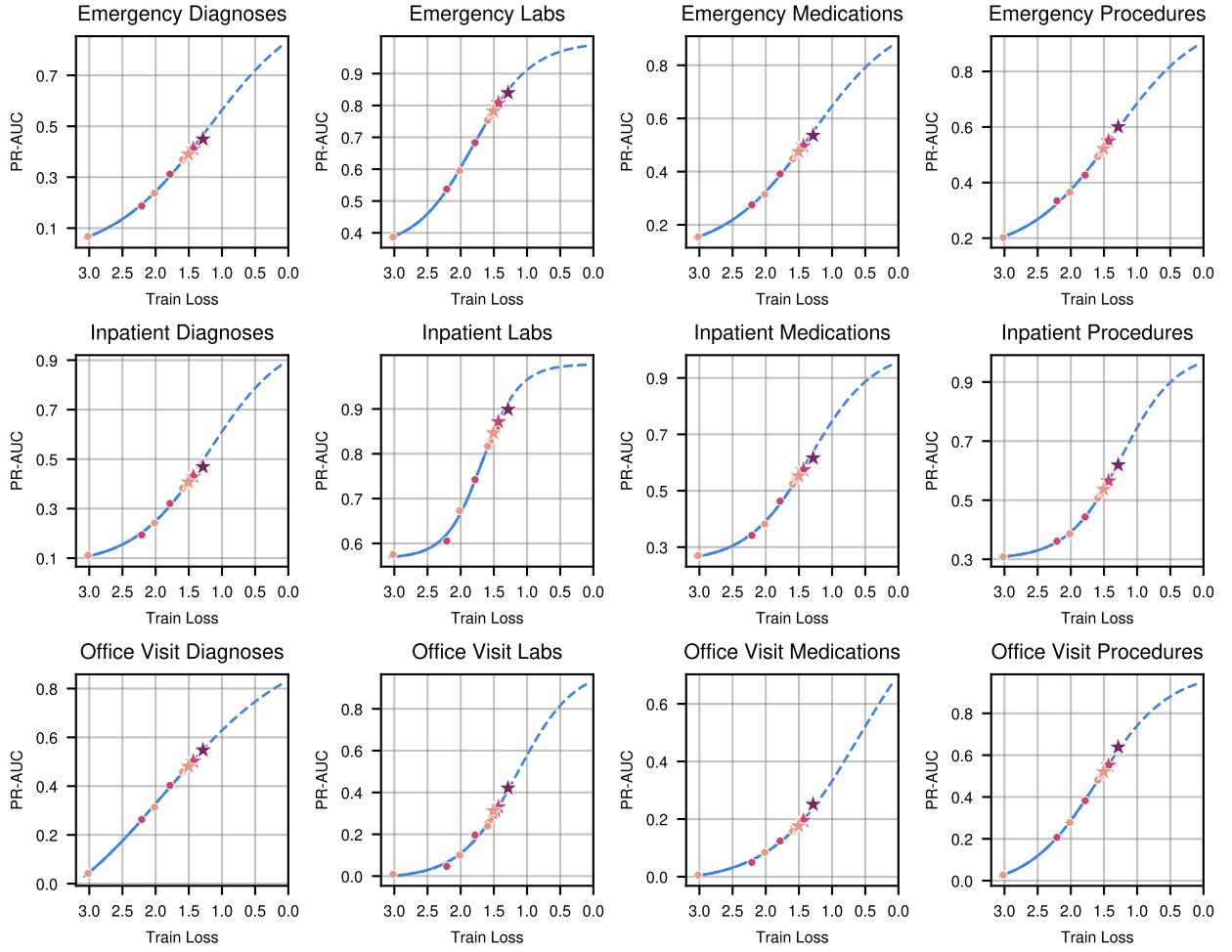


**Figure 20: Plausibility of predicted medical event prevalence and pairwise co-occurrence.** On the left are heatmap scatter plots for the prevalence of different discrete medical events in CoMET-L’s predicted generations versus the ground truth. The plots include diagnoses (first 3 ICD-10-CM characters only), medications, labs, and procedures. On the right are heatmap scatter plots showing the fraction of patients in which pairs of medical events co-occurred within one year, comparing CoMET-L’s predicted generations versus ground truth. Note that because these are all log-log plots, zero prevalence and co-occurrence values are also shown but with an explicit gap.

## Appendix E: Section 2.7.1 results

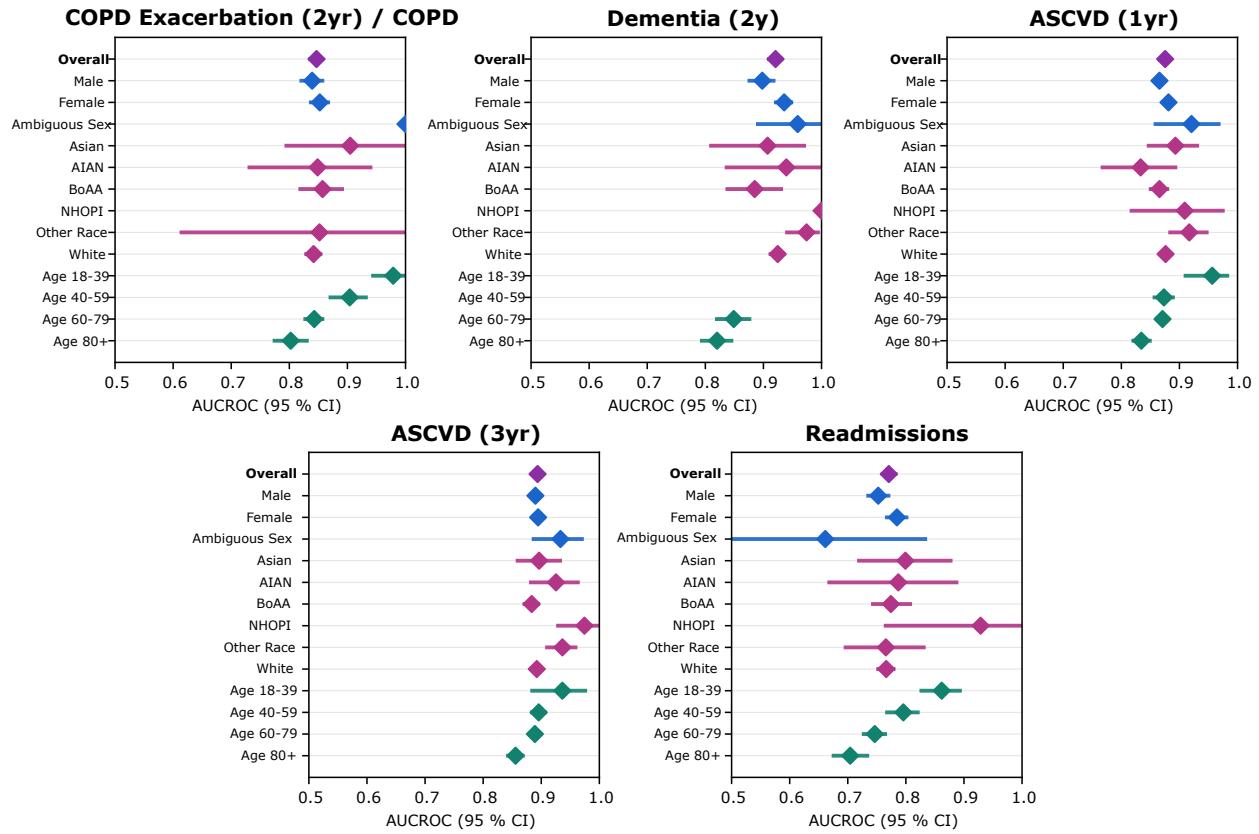


**Figure 21: T2DM-specific outcome performance measured by AUCROC improves as train loss decreases.** We evaluated each CoMET model, along with earlier checkpoints from the CoMET-S and CoMET-M training runs, on all T2DM tasks. We fit a sigmoid curve to all points except for CoMET-L to assess the sigmoid curve's predictive utility. We evaluated each model using a more conservative  $n = 20$  simulations. Stars indicate compute-optimal models.



**Figure 22: Single-encounter generations improve as train loss decreases.** We evaluated each CoMET model, along with earlier checkpoints from the CoMET-S and CoMET-M training runs, on all of the single-encounter completion tasks. We fit a sigmoid curve to all points except for CoMET-L to assess the sigmoid curve’s predictive utility. We evaluated each model using  $n = 20$  simulations. Stars indicate compute-optimal models.

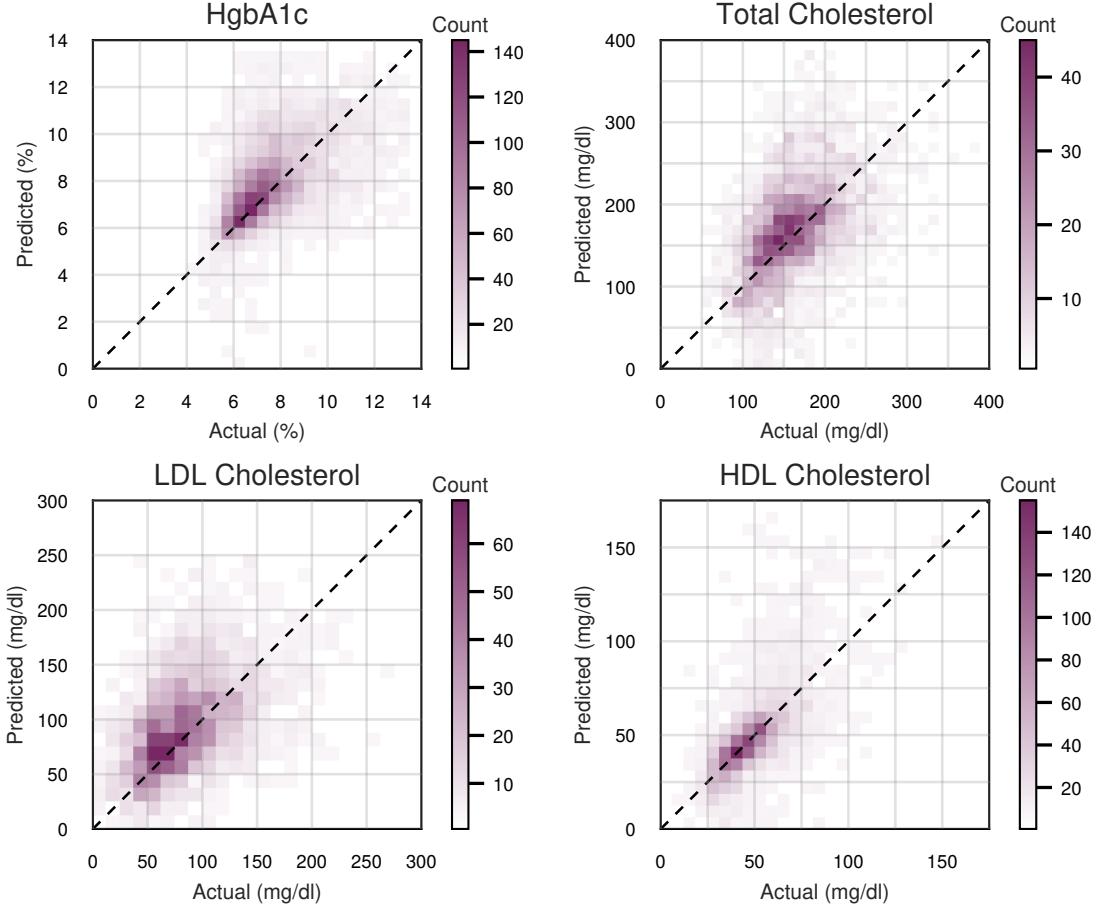
## Appendix F: Bias and Fairness



**Figure 23: Bias Impact via Subcohort Analysis.** CoMET-L's AUCROCs with 95% bootstrapped CI stratified by demographic groups are displayed for five different predictive tasks for brevity. In order of which they appear they are: one from acute-on-chronic event, one from incident disease risk, two from T2DM-specific outcomes, and one from operational outcomes.

## Appendix G: Additional Results for Disease-Specific Outcomes

CoMET predicts lab values that exhibit moderate correlation with actual observed values. CoMET predictions are noisy as a consequence of the limited expressiveness of the model’s lab result value bucketed tokens (see [Section 5.2.4](#) for how labs are tokenized).



**Figure 24:** Heat map scatter plots comparing 3-month CoMET-L predictions of lab values to the patient’s actual results. Plots show HgbA1c for patients with diabetes and show total cholesterol, LDL cholesterol, and HDL cholesterol for patients with hyperlipidemia.

| Model      | HgbA1c      |             | HDL Cholesterol |              | LDL Cholesterol |              | Total Cholesterol |              |
|------------|-------------|-------------|-----------------|--------------|-----------------|--------------|-------------------|--------------|
|            | MAE         | RMSE        | MAE             | RMSE         | MAE             | RMSE         | MAE               | RMSE         |
| CoMET-L    | 1.21        | 1.69        | 11.57           | 18.30        | 31.51           | 41.91        | 43.01             | 58.66        |
| CoMET-M    | 1.24        | 1.73        | 13.15           | 20.36        | 33.41           | 44.75        | 45.75             | 61.98        |
| CoMET-S    | 1.30        | 1.81        | 13.62           | 21.06        | 33.86           | 44.70        | 44.88             | 60.61        |
| Supervised | <b>1.05</b> | <b>1.43</b> | <b>10.07</b>    | <b>13.49</b> | <b>26.58</b>    | <b>34.43</b> | <b>31.44</b>      | <b>42.06</b> |

**Table 16:** MAE and RMSE values for the corresponding subplots in [Figure 24](#) across all CoMET models and the best performing task-specific model.

## Appendix H: Evaluation Dataset Characteristics

| Condition       | Task (time horizon)                      | Test Sample Size | Positive Prevalence |
|-----------------|--|------------------|---------------------|
| Type 2 Diabetes | ASCVD (1yr)                              | 20,653           | 0.137               |
|                 | ASCVD (3yr)                              | 12,331           | 0.353               |
|                 | CKD Prog. Stage 2 → 3 (1yr)              | 1,462            | 0.159               |
|                 | CKD Prog. Stage 2 → 3 (3yr)              | 923              | 0.462               |
|                 | CKD Prog. Stage 3 → 4 <sup>+</sup> (1yr) | 3,795            | 0.107               |
|                 | CKD Prog. Stage 3 → 4 <sup>+</sup> (3yr) | 2,071            | 0.338               |
|                 | Diabetic Neuropathy (1yr)                | 21,879           | 0.337               |
|                 | Diabetic Neuropathy (3yr)                | 15,196           | 0.604               |
|                 | Diabetic Retinopathy (1yr)               | 20,272           | 0.110               |
|                 | Diabetic Retinopathy (3yr)               | 11,293           | 0.266               |
|                 | HgbA1c < 7 (60-120 days)                 | 5,608            | 0.401               |
|                 | HgbA1c < 9 (60-120 days)                 | 5,608            | 0.827               |
|                 | HgbA1c < 11 (60-120 days)                | 5,608            | 0.950               |
|                 | HgbA1c < 12 (60-120 days)                | 5,608            | 0.975               |
| Hypertension    | ASCVD (1yr)                              | 21,344           | 0.180               |
|                 | ASCVD (3yr)                              | 13,687           | 0.419               |
|                 | CKD Prog. Stage 2 → 3 (1yr)              | 1,429            | 0.227               |
|                 | CKD Prog. Stage 2 → 3 (3yr)              | 985              | 0.541               |
|                 | CKD Prog. Stage 3 → 4 <sup>+</sup> (1yr) | 4,157            | 0.120               |
|                 | CKD Prog. Stage 3 → 4 <sup>+</sup> (3yr) | 2,326            | 0.354               |
|                 | Heart Attack (1yr)                       | 20,301           | 0.049               |
|                 | Heart Attack (3yr)                       | 11,139           | 0.158               |
|                 | Stroke (1yr)                             | 20,254           | 0.045               |
|                 | Stroke (3yr)                             | 10,939           | 0.138               |
| Hyperlipidemia  | ASCVD (1yr)                              | 22,300           | 0.140               |
|                 | ASCVD (3yr)                              | 13,996           | 0.352               |
|                 | Heart Attack (1yr)                       | 21,515           | 0.037               |
|                 | Heart Attack (3yr)                       | 11,884           | 0.119               |
|                 | Chronic Heart Failure (1yr)              | 22,845           | 0.200               |
|                 | Chronic Heart Failure (3yr)              | 14,771           | 0.394               |
|                 | Stroke (1yr)                             | 21,540           | 0.036               |
|                 | Stroke (3yr)                             | 11,825           | 0.109               |

**Table 17:** Test-set sample sizes and positive class prevalence for each disease-specific adverse outcome (Section 2.3.1), grouped by clinical condition.

| Task               | Test Sample Size | Positive Prevalence |
|--------------------|------------------|---------------------|
| 30-day Readmission | 10,000           | 0.1074              |

**Table 18:** Test set sample size and positive class prevalence for the 30-day readmission task (Section 2.5.1).

| Task Type             | Task (time horizon)                                      | Patient Count | Resampled Positive Prevalence | Real Positive Prevalence |
|-----------------------|--|---------------|-------------------------------|--------------------------|
| Acute-on-Chronic      | Alcohol Withdrawal Syndrome (2yr) / Alcohol Use Disorder | 5032          | 0.111                         | 0.105                    |
|                       | Asthma Exacerbation (2yr) / Asthma                       | 5199          | 0.097                         | 0.039                    |
|                       | CHF Exacerbation (2yr) / CHF                             | 5000          | 0.184                         | 0.190                    |
|                       | COPD Exacerbation (2yr) / COPD                           | 5004          | 0.147                         | 0.129                    |
|                       | Sickle Cell Crisis (2yr) / Sickle Cell Disease           | 1933          | 0.339                         | 0.339                    |
| Incident Disease Risk | Alcohol Use Disorder (2yr)                               | 5369          | 0.093                         | 0.003                    |
|                       | Asthma (2yr)   | 5116          | 0.098                         | 0.008                    |
|                       | CHF (2yr)  | 5202          | 0.096                         | 0.013                    |
|                       | COPD (2yr)   | 5083          | 0.099                         | 0.011                    |
|                       | Dementia (2yr)   | 5369          | 0.093                         | 0.006                    |
|                       | Heart Attack (2yr)                                       | 5330          | 0.094                         | 0.010                    |

**Table 19:** Summary statistics for acute-on-chronic ([Section 2.3.2](#)) and incident disease risk ([Section 2.3.3](#)). Note the sickle cell crisis task has fewer than 5,000 patients since there were only 1,933 patients who met inclusion criteria in the test set.

| Diagnosis                       | Test Sample Size |
|---------------------------------|------------------|
| Acute Pancreatitis              | 1239             |
| Chronic Pancreatitis            | 1002             |
| Pancreatic Cancer               | 980              |
| Cholecystitis                   | 1118             |
| Cholangitis                     | 967              |
| Liver Cancer                    | 970              |
| Chronic Viral Hepatitis         | 992              |
| Alcoholic Liver Disease         | 1027             |
| Non-Alcoholic Steatohepatitis   | 979              |
|                                 |                  |
| Osteoarthritis                  | 1256             |
| Rheumatoid Arthritis            | 1110             |
| Psoriatic Arthritis             | 951              |
| Polymyalgia Rheumatica          | 937              |
| Systemic Sclerosis              | 723              |
| Systemic Lupus Erythematosus    | 1015             |
| Mixed Connective Tissue Disease | 572              |
| Polymyositis/Dermatomyositis    | 466              |
| Fibromyalgia                    | 1067             |

**Table 20:** Summary statistics for the HPB and rheumatic differential diagnosis evaluation sets ([Section 2.4](#)).

| Category              | N     | Actual Standard Deviation | Actual Mean |
|-----------------------|-------|---------------------------|-------------|
| all admissions        | 10000 | 8.20                      | 5.65        |
| admissions $\leq$ 4d  | 5503  | 0.95                      | 2.28        |
| admissions $\leq$ 7d  | 7762  | 1.64                      | 3.15        |
| admissions $\leq$ 10d | 8734  | 2.26                      | 3.72        |
| admissions $\leq$ 14d | 9305  | 2.93                      | 4.21        |
| admissions $\leq$ 50d | 9968  | 5.55                      | 5.36        |

**Table 21:** Summary statistics for the length of stay evaluation set ([Section 2.5.1](#)).

## Appendix I: Disease Phenotypes

Disease phenotypes for each category of evaluation task are built with a variety of criteria, such as encounter type inclusions, number of occurrences, and code groupers, which are described in [Section 5.6](#). Below are the code sets we used for defining each phenotype.

| Disease Type           | Disease                         | ICD-10-CM Codes |
|------------------------|---------------------------------|-----------------|
| Hepatopancreatobiliary | Acute Pancreatitis              | K85             |
|                        | Chronic Pancreatitis            | K86.0, K86.1    |
|                        | Pancreatic Cancer               | C25             |
|                        | Acute Cholecystitis             | K81             |
|                        | Cholangitis                     | K83.0           |
|                        | Chronic Viral Hepatitis         | B18             |
|                        | Liver Cancer                    | C22             |
|                        | Alcoholic Liver Disease         | K70             |
|                        | Non-Alcoholic Steatohepatitis   | K75.81          |
| Rheumatic              | Osteoarthritis                  | M15-19          |
|                        | Rheumatoid Arthritis            | M05, M06        |
|                        | Psoriatic Arthritis             | L40.5           |
|                        | Systemic Lupus Erythematosus    | M32             |
|                        | Polymyalgia Rheumatica          | M35.3, M31.5    |
|                        | Mixed Connective Tissue Disease | M35.1           |
|                        | Polymyositis/Dermatomyositis    | M33             |
|                        | Systemic Sclerosis              | M34             |
|                        | Fibromyalgia                    | M79.7           |

**Table 22:** Phenotypes and associated ICD-10-CM codes for the HPB and rheumatic differential diagnosis tasks. All subcategories of the ICD-10-CM codes here are also used. We considered a patient to have received the diagnosis if it appears at least twice in the patient’s record on different dates. If an off-target diagnosis appears exactly once in a patient’s record, that patient is excluded from analyses related to that off-target diagnosis.

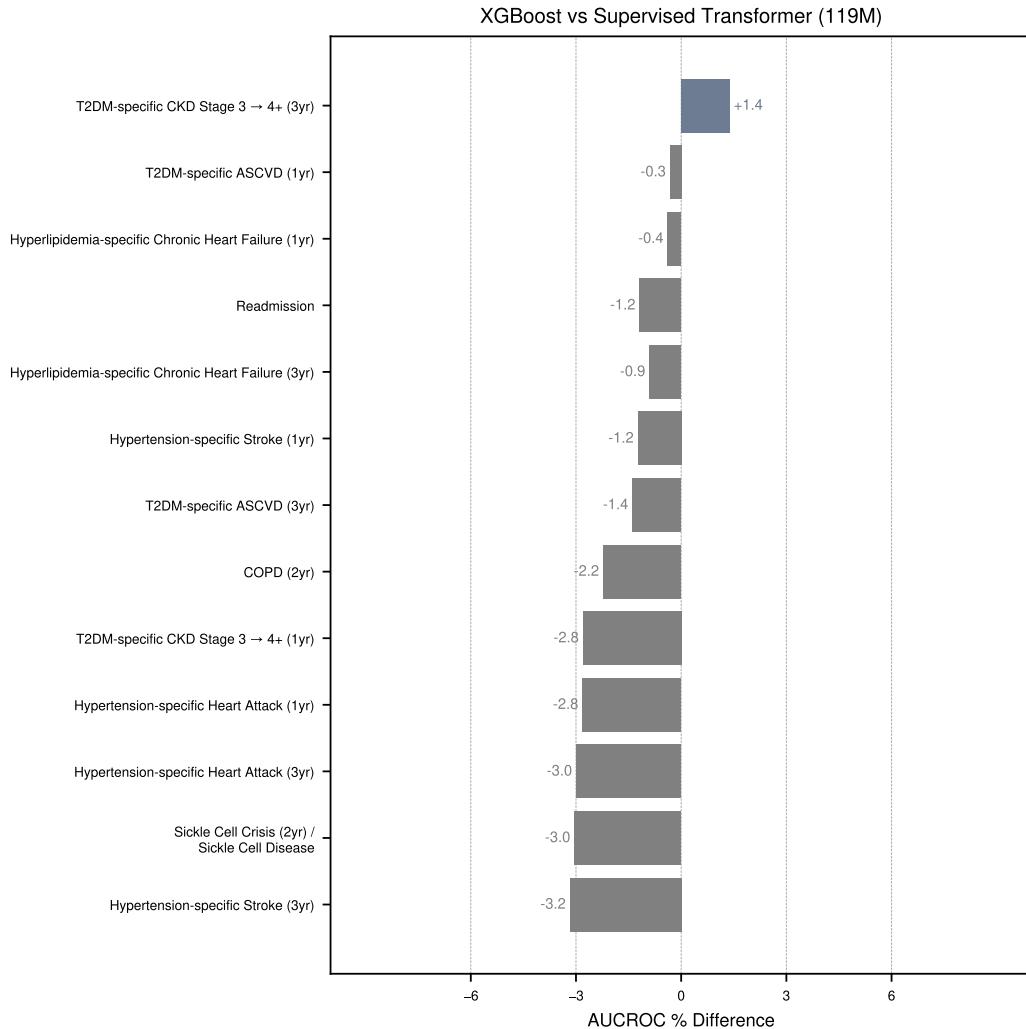
| Disease                     | Medical Codes for Disease-Specific Outcomes Tasks  |
|-----------------------------|--|
| ASCVD                       | G45, I20.0, I20.8, I20.9, I21, I22, I24.9, I25.11, I25.7, I25.810, I25.812, I63, I70.2, I70.3, I70.4, I70.5, I70.6, I70.7, I70.8, I70.9, I73.9, I75.029, I77.6, Q28.8, Q87.89, T82.21, T82.310, T82.311, T82.312, T82.320, T82.321, T82.322, T82.330, T82.331, T82.332, T82.390, T82.391, T82.392, Z95.5 |
| CKD (Stage 2)               | N18.2  |
| CKD (Stage 3)               | N18.3  |
| CKD (Stage 4 <sup>+</sup> ) | I12.0, I13.11, I13.2, N18.4, N18.5, N18.6  |
| Diabetic Neuropathy         | E08.4, E10.4, E11.4, E13.4   |
| Diabetic Retinopathy        | E08.31, E08.32, E08.33, E08.34, E08.35, E10.31, E10.32, E10.33, E10.34, E10.35, E11.31, E11.32, E11.33, E11.34, E11.35, E13.31, E13.32, E13.33, E13.34, E13.35   |
| Heart Attack                | I21, I22   |
| Chronic Heart Failure       | I50  |
| Stroke                      | I60, I61, I63  |
| HgbA1c                      | LOINC Codes: 4548-4, 4549-2, 17855-8, 17856-6  |

**Table 23:** Associated codes for Disease-Specific Outcomes tasks. Subcategories of listed medical codes were included as part of the task definition. All medical codes are ICD-10-CM unless otherwise specified.

| Disease                          | Phenotype | ICD-10-CM Codes   |
|----------------------------------|-----------|---|
| Heart Attack[71]                 | Emergent  | I21.01, I21.02, I21.09, I21.11, I21.19, I21.21, I21.29, I21.3, I21.4, I21.9, I21.A1, I21.A9, I21.B, I22.0, I22.1, I22.2, I22.8, I22.9, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.7, I23.8  |
| Alcohol Use Disorder*[72]        | Chronic   | F10.10, F10.120, F10.121, F10.129, F10.14, F10.150, F10.151, F10.159, F10.180, F10.181, F10.182, F10.188, F10.19, F10.20, F10.220, F10.221, F10.230, F10.231, F10.232, F10.239, F10.24, F10.250, F10.251, F10.259, F10.26, F10.27, F10.280, F10.281, F10.282, F10.288, F10.29   |
| Alcohol Withdrawal Syndrome*[72] | Emergent  | F10.130, F10.131, F10.132, F10.139, F10.229, F10.230, F10.231, F10.232, F10.239, F10.930, F10.931, F10.932, F10.939   |
| Asthma*[71]                      | Chronic   | J45.20, J45.21, J45.22, J45.30, J45.31, J45.32, J45.40, J45.41, J45.42, J45.50, J45.51, J45.52, J45.901, J45.902, J45.909, J45.990, J45.991, J45.998  |
| Asthma Exacerbation*† [71]       | Emergent  | J45.21, J45.22, J45.31, J45.32, J45.41, J45.42, J45.51, J45.52, J45.901, J45.990  |
| COPD*[71]                        | Chronic   | J40, J41.0, J41.1, J41.8, J42, J43.0, J43.1, J43.2, J43.9, J44.0, J44.1, J44.81, J44.89, J44.9, J47.0, J47.1, J47.9, J98.2, J98.3   |
| COPD Exacerbation*† [71]         | Emergent  | J44.0, J44.1, J47.0, J47.1  |
| CHF*[73]                         | Chronic   | I09.81, I11.0, I13.0, I13.2, I50.1, I50.20, I50.21, I50.22, I50.23, I50.30, I50.31, I50.32, I50.33, I50.40, I50.41, I50.42, I50.43, I50.810, I50.811, I50.812, I50.813, I50.814, I50.82, I50.83, I50.84, I50.89, I50.9  |
| CHF Exacerbation*† [73]          | Emergent  | I50.21, I50.23, I50.31, I50.33, I50.41, I50.43, I50.811, I50.813  |
| Dementia[71]                     | Chronic   | F01.50, F01.51, F01.511, F01.518, F01.52, F01.53, F01.54, F01.A0, F01.A11, F01.A18, F01.A2, F01.A3, F01.A4, F01.B0, F01.B11, F01.B18, F01.B2, F01.B3, F01.B4, F01.C0, F01.C11, F01.C18, F01.C2, F01.C3, F01.C4, F02.80, F02.81, F02.811, F02.818, F02.82, F02.83, F02.84, F02.A0, F02.A11, F02.A18, F02.A2, F02.A3, F02.A4, F02.B0, F02.B11, F02.B18, F02.B2, F02.B3, F02.B4, F02.C0, F02.C11, F02.C18, F02.C2, F02.C3, F02.C4, F03.90, F03.91, F03.911, F03.918, F03.92, F03.93, F03.94, F03.A0, F03.A11, F03.A18, F03.A2, F03.A3, F03.A4, F03.B0, F03.B11, F03.B18, F03.B2, F03.B3, F03.B4, F03.C0, F03.C11, F03.C18, F03.C2, F03.C3, F03.C4, F05, G13.8, G31.01, G31.1, G31.2, G31.83, G94, R41.81 |
| Sickle Cell Crisis*[74]          | Emergent  | D57.0, D57.00, D57.01, D57.02, D57.03, D57.09, D57.21, D57.211, D57.212, D57.213, D57.218, D57.219, D57.41, D57.411, D57.412, D57.413, D57.414, D57.418, D57.419, D57.43, D57.431, D57.432, D57.433, D57.434, D57.438, D57.439, D57.445, D57.451, D57.452, D57.453, D57.454, D57.458, D57.459, D57.81, D57.811, D57.812, D57.813, D57.814, D57.818, D57.819   |
| Sickle Cell Disease*† [74]       | Chronic   | D57, D57.0, D57.00, D57.01, D57.02, D57.03, D57.09, D57.1, D57.2, D57.20, D57.21, D57.211, D57.212, D57.213, D57.218, D57.219, D57.4, D57.40, D57.41, D57.411, D57.412, D57.413, D57.414, D57.418, D57.419, D57.42, D57.43, D57.431, D57.432, D57.433, D57.434, D57.438, D57.439, D57.44, D57.45, D57.451, D57.452, D57.453, D57.454, D57.458, D57.459, D57.81, D57.811, D57.812, D57.813, D57.814, D57.818, D57.819  |

**Table 24:** Phenotypes and associated ICD-10-CM codes for acute-on-chronic and incident disease risk prediction tasks. Asterisked phenotypes are used in acute-on-chronic tasks. Phenotypes with daggers are used for the acute definition of an acute-on-chronic task that we curated to be a subset of the cited chronic definition.

## Appendix J: Comparison of task-specific supervised models



**Figure 25: XGBoost vs. Supervised Transformer (119M).** AUCROC comparisons between the supervised transformer and XGBoost models on a representative subset of classification tasks. Differences are shown in terms of percentage points with XGBoost performance as the baseline.