HEALTHSAW: SCALING HEALTH WORLD MODEL WITH A FORECAST-DRIVEN BOTTLENECK

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal patient healthcare records are widely considered a key source of clinical intelligence, but incorporating them into foundation models through next-token prediction has achieved limited success. This paper introduces HealthSAW, a Health State–Action World Model that represents patient healthcare records in a learned higher-order representation space, using local encoders to abstract tokens about a patient's health observations into health state vectors and treatment tokens into action vectors. These state–action representations enable a new pre-training paradigm in which the world model can autoregressively predict future events, including diagnoses, treatments, and responses to treatment. To make these vectors capture the patient's health rather than trivial token patterns, HealthSAW introduces a forecast-driven bottleneck: local decoders generate the tokens of the next state/action rather than reconstructing the current ones, and access prior patient history only through these higher-order state–action representations. Our experiments on about 19 million patients and 28 billion healthcare record tokens establish scaling laws for HealthSAW and demonstrate its superior performance across a wide range of disease-risk and treatment-outcome prediction tasks, including generalization to health systems unseen during pretraining. The world model architecture also enables new zero-shot patient simulation capabilities for predicting post-treatment laboratory outcomes, while being $1.8\times$ and $3.2\times$ faster than token-level models in pretraining and simulation, respectively. Further studies demonstrate the necessity of abstraction into the higher-order representation space, the benefits of this abstraction for context efficiency, especially with richer patient observations, and the generalization advantage of the forecast-driven bottleneck.

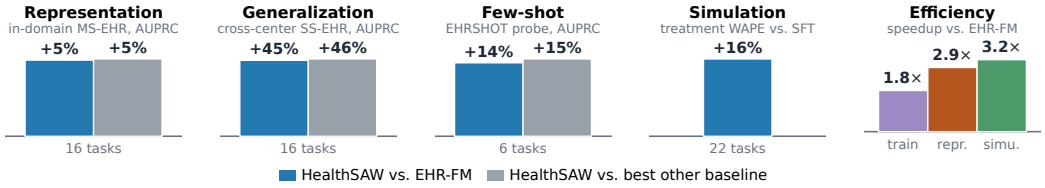


Figure 1: HealthSAW capabilities. The first four panels show relative gains over next-token foundation models and the best competing baselines, including frontier LLMs, medical and EHR foundation models, and classic tree models; the last shows speedups in pretraining, patient representation and disease simulation. MS-EHR (in-domain) contains held-out patients from the multi-site EHR corpus used for pretraining; SS-EHR (cross-center) comes from a single medical center unseen during pretraining. Treatment simulation predicts post-treatment laboratory values, evaluated by weighted absolute percentage error.

1 INTRODUCTION

Patient healthcare records are widely considered a necessary source for foundation models to acquire deeper clinical intelligence: the ability to understand patient health (Guo et al., 2024), forecast disease progression (Kraljevic et al., 2024; Shmatko et al., 2025), identify potential health risks (Rasmy et al., 2021), and simulate patient trajectories (Shmatko et al., 2025). Many of these clinical applications require professional expertise that is difficult for frontier models to acquire solely through training on text from the public web, textbooks, and medical publications: doctors learn not only from books but also through years of clinical practice, seeing patients, treating them, and learning from their outcomes; much of this information resides in electronic healthcare records (EHRs), not on the internet.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The standard next-token prediction approach linearizes a patient's EHR chronologically into a token sequence, represented either as medical codes (e.g., ICD codes) (Li et al., 2020; Steinberg et al., 2021) or as natural-language descriptions of symptoms, diagnoses, and drugs (Ben Shoham & Rappoport, 2024). Pretrained EHR foundation models (EHR-FM) have shown advantages over general LLMs on some clinical tasks (Sun et al., 2026), but have sometimes failed to outperform classic tree models (Chen et al., 2024) and have struggled especially with generalization (Wornow et al., 2023). One challenge is that repetitive patterns in the token space can encourage *trivial learning* (Geirhos et al., 2020; Sun et al., 2026): memorizing long drug names, recurring diagnoses, and common lab-test sequences can lower next-token prediction loss without necessarily capturing the underlying clinical intelligence. Token-by-token generation also *over-complicates* many clinical tasks. Modeling a patient's disease progression does not require stepping through every ICD code in every visit; what matters is how the patient's condition changes from visit to visit (Shmatko et al., 2025).

Building on this view, this paper introduces HealthSAW, a Health State–Action World Model (Ha & Schmidhuber, 2018; Hafner et al., 2020) that represents each visit in a patient's health journey with two vectors in a higher-order representation space: a *patient health state* and a *clinical action*. These state and action vectors abstract the raw tokens: local encoders compress tokens about a patient's health observations within a visit into a local state and tokens about treatments into an action vector; a global Transformer performs autoregressive sequence modeling over these vectors, and its global outputs, which integrate the patient's history, are decoded by local decoders into the raw tokens of the next state or action block. The state–action vectors thus serve as the basic units of sequence modeling: the model steps from visit to visit rather than token by token, reducing token-level over-complication.

To make the state–action representations capture the patient's health rather than trivial token patterns, we learn them through a *forecast-driven bottleneck*: they are trained to forecast the *future* event rather than reconstruct the *current* one, as reconstruction-driven latent world models do (Rombach et al., 2022; NVIDIA et al., 2025). Each state or action vector represents a block of raw tokens, and the entire world model is trained end to end with a next-block prediction objective. Since the decoder can access earlier history only through the global state–action representations, these representations must retain information relevant to how the patient's health evolves, while recurring tokens cannot simply be copied from earlier events.

We conducted large-scale experiments on de-identified healthcare records from about 19 million patients across multiple clinical centers and countries, comprising about 3.5 billion visits and 28 billion raw tokens. We first studied the scaling law of HealthSAW and observed continued performance gains up to three billion parameters. HealthSAW demonstrated strong representation capabilities (Figure 1), outperforming EHR-FM in a controlled comparison with matched parameter counts and identical pretraining and fine-tuning data. Unlike EHR-FM, whose performance is mixed, HealthSAW outperforms state-of-the-art methods, including frontier LLMs, medical and EHR foundation models, and classic tree models, by large margins across a broad set of clinical prediction tasks.

HealthSAW shows its largest gains when generalizing to unseen health systems, with a 45% improvement in AUPRC across 16 risk prediction tasks. With only 128 labeled examples per class, it improves over the best previous method, a classic tree model, by 15% across 6 new-diagnosis tasks on the public EHRSHOT benchmark (Wornow et al., 2023), and leads EHR-FM and all published baselines at every smaller label budget. Further analyses confirm that abstraction from the raw token space into a higher-order representation space is necessary, and that effective pretraining in this space requires the forecast-driven bottleneck: giving the decoder direct access to earlier tokens lowers pretraining loss but degrades performance on every downstream metric due to trivial learning.

The world model architecture naturally enables effective and efficient patient simulation by rolling out future patient trajectories and extracting target "readouts" from them. Without task-specific training, HealthSAW achieves 16% lower treatment-outcome simulation error than EHR-FM across 22 post-treatment laboratory tasks, even though EHR-FM is additionally fine-tuned for each task. The new architecture is also more efficient, running $1.8\times$ faster in pretraining and $3.2\times$ faster in simulation. Abstracting raw tokens into higher-order representations removes token-level redundancy from global sequence modeling and yields substantial context efficiency.

2 Related Work

Advances in general-purpose language models (OpenAI, 2023; Ouyang et al., 2022) have been accompanied by growing interest in medical foundation models, including ClinicalBERT, MEDITRON,

and MedGemma (Huang et al., 2019; Chen et al., 2023; Sellergren et al., 2026), which incorporate domain knowledge through training on medical text and images. A complementary line of work learns longitudinal patient representations directly from structured EHRs: EHRSHOT introduces CLMBR-T-base, pretrained on records from 2.57 million patients, together with a few-shot clinical prediction benchmark (Wornow et al., 2023), while CoMET and EHR-FM report predictive gains from increasing model and training scale (Waxler et al., 2025; Sun et al., 2026). Yet medical specialization does not consistently improve performance: medical LLMs do not reliably outperform their general-domain counterparts on medical QA (Jeong et al., 2024), frontier LLMs can outperform specialized clinical tools in independent evaluations (Vishwanath et al., 2026), and ClinicalBench finds both general-purpose and medical LLMs underperform traditional clinical predictors (Chen et al., 2024).

Parallel work explores clinical world models that capture patient evolution under interventions using state–action abstractions (Qazi et al., 2025; Yang et al., 2026; Mu et al., 2026; Xu et al., 2026). These models predict future patient states under recorded, supplied, or separately selected actions, while temporal progression is fixed or externally specified. Their reported simulations also remain relatively short, covering up to 48 hours, a single hospital admission, or 10–15 imagined ICU steps. More broadly, a common world-model design first maps raw observations into a latent space and then learns dynamics in that space. Latent diffusion and latent world models such as NVIDIA Cosmos first train an encoder–decoder tokenizer to reconstruct the input, then freeze it while a generative model learns in the resulting latent space (Rombach et al., 2022; NVIDIA et al., 2025). Another approach predicts future latent states directly: JEPA-style methods match predicted representations to those produced by a target encoder (Assran et al., 2023), while SMB-Structure combines this objective with next-token supervision on patient records (Adam et al., 2026).

In contrast, HealthSAW uses neither a reconstruction tokenizer nor a target encoder: its local encoders, global Transformer, and local decoders are jointly trained by next-block prediction to learn higher-order state–action representations. Unlike prior clinical world models, it also generates clinical actions and elapsed time during rollout, supporting treatment simulation up to 180 days and disease simulation over 2- or 4-year windows. This forecast-driven bottleneck matters especially for repetitive EHRs, where an objective that can be lowered by reproducing recurring documentation need not capture the underlying patient state (Wornow et al., 2024).

3 Health State–Action World Modeling

HealthSAW aims to learn clinical intelligence from large-scale patient healthcare records more effectively than next-token prediction, the de facto pretraining paradigm for foundation models. It does so by abstracting clinical data into a higher-order representation space (§3.1) and pretraining through a forecast-driven bottleneck (§3.2). The pretrained model then supports a wide range of clinical applications (§3.3).

3.1 Abstraction into State–Action

Patient record formulation. A patient's record is a sequence of visits, each containing the tokens recorded that day: medical codes and their values for diagnoses, measurements, medications, and procedures. Within a visit, these tokens often form an unordered set, so a token-level model serializes them in an arbitrary order into a single stream, with visits separated by time tokens:

$$[\text{visit}_1 \text{ tokens}]\ [\Delta t]\ [\text{visit}_2 \text{ tokens}]\ [\Delta t]\ \cdots$$

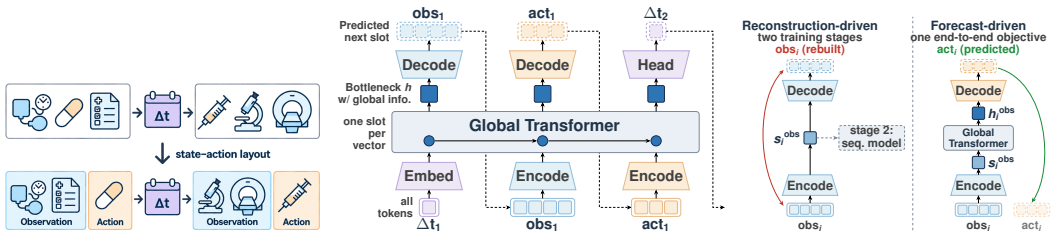
Each token therefore occupies its own position and generation step, while tokens that reveal the patient's condition are interleaved with those that record the care given in response. These two types of tokens play different roles, and what matters is how the patient's condition changes from visit to visit.

HealthSAW regroups the tokens by their role within each visit (Figure 2a). All tokens recorded for a patient on the same day form a visit *event*. Those that reveal the patient's condition (diagnoses, laboratory tests, vital signs, imaging, and diagnostic procedures) form the observation block obs_i, while those that record care (medications, therapeutic procedures, surgeries, referrals, and supplies) form the action block act_i. Following the clinical workflow, the observation block precedes the action block.

With the elapsed time Δt_i since the previous event, event i is $e_i = (\Delta t_i;\ \text{obs}_i;\ \text{act}_i)$, and the record becomes a sequence of slots, three per event: one for the elapsed time and one for each block:

$$[\Delta t_1]\ [\text{obs}_1]\ [\text{act}_1]\ [\Delta t_2]\ [\text{obs}_2]\ [\text{act}_2]\ \cdots$$

HealthSAW's architecture builds on this structure by abstracting each slot into a single vector.



(a) State–action abstraction (b) Overall architecture (c) Forecast-driven bottleneck

Figure 2: HealthSAW overview. (a) Each visit becomes an elapsed time, an observation block and an action block. (b) Local encoders compress blocks into slots for an autoregressive global Transformer, whose outputs predict the next slot. (c) Block representations are learned by forecasting the next block rather than reconstructing the current one.

Architecture. HealthSAW consists of local encoders, an autoregressive global Transformer, and local decoders (Figure 2b). The observation and action encoders use bidirectional attention within each block to produce a local state vector and an action vector, respectively, while the discretized elapsed time is embedded separately:

$$s_i^t = \mathrm{Emb}(\Delta t_i), \qquad s_i^{\mathrm{obs}} = \mathrm{Enc}_{\mathrm{obs}}(\mathrm{obs}_i), \qquad s_i^{\mathrm{act}} = \mathrm{Enc}_{\mathrm{act}}(\mathrm{act}_i).$$

To integrate patient history across events, the global Transformer processes these vectors in slot order and returns a sequence of global outputs:

$$\big(h_1^t, h_1^{\mathrm{obs}}, h_1^{\mathrm{act}}, \ldots, h_i^{\mathrm{act}}\big) = \mathrm{Global}_{\mathrm{AR}}\big(s_1^t, s_1^{\mathrm{obs}}, s_1^{\mathrm{act}}, \ldots, s_i^{\mathrm{act}}\big).$$

Because the Transformer is autoregressive, each output summarizes only the current and preceding slots. Each global output predicts the content of the next slot:

$$\mathrm{obs}_i \sim p(\cdot \mid h_i^t), \qquad \mathrm{act}_i \sim p(\cdot \mid h_i^{\mathrm{obs}}), \qquad \Delta t_{i+1} \sim p(\cdot \mid h_i^{\mathrm{act}}).$$

The observation and action decoders generate each block token by token from the corresponding global output and the tokens already generated within the block, while a time head predicts the next elapsed time. Each local decoder accesses earlier blocks only through its global output.

Thus, the global Transformer represents each event with three slots regardless of how many tokens the blocks contain, whereas a token-level model uses one position and generation step per token. It models how the patient's condition evolves from event to event under recorded care, while the local encoders and decoders handle the tokens within each block. Consequently, the global context length and number of generation steps grow with the number of events rather than raw tokens, making HealthSAW $1.8\times$ faster in pretraining and $3.2\times$ faster in simulation than EHR-FM (§4.4).

3.2 Pretraining with a Forecast-Driven Bottleneck

To make the state–action representations retain information predictive of future events rather than recurring token patterns, the global outputs h must be shaped by the forecasting objective. In two-stage latent world models (§2), latent representations are learned separately from sequence forecasting, typically through reconstruction, and are therefore optimized to preserve the current input rather than specifically the information predictive of the future.

For HealthSAW, we instead train all modules end to end so that each global output predicts the content of the next slot (Figure 2c). Under teacher forcing, we minimize the next-block prediction loss:

$$\mathcal{L} = -\sum\nolimits_i \Big[\log p(\mathrm{obs}_i \mid h_i^t) + \log p(\mathrm{act}_i \mid h_i^{\mathrm{obs}}) + \log p(\Delta t_{i+1} \mid h_i^{\mathrm{act}})\Big].$$

The local decoders compute the observation and action likelihoods autoregressively, while a time head predicts the elapsed time. A local decoder can access preceding tokens within its own block, but reaches earlier blocks only through a single global output, such as $h_i^t = \mathrm{Global}_{\mathrm{AR}}(s_1^t, s_1^{\mathrm{obs}}, s_1^{\mathrm{act}}, \ldots, s_i^t)$ for obs_i. Each global output is therefore an information bottleneck over the patient's history: all information from earlier blocks available to the decoder must pass through it. The forecasting objective determines what is retained by jointly training the local encoders to extract predictive information within blocks and the global Transformer to integrate it across events. This creates the forecast-driven bottleneck: the higher-order representations are optimized to predict future blocks rather than reconstruct their inputs.

The bottleneck also blocks a key route to *trivial learning* (Geirhos et al., 2020). A token-level decoder can lower its loss on a chronic diagnosis recorded at every event simply by copying its earlier occurrence. HealthSAW's decoders cannot access that occurrence directly, so the loss can decrease only if information about the underlying condition is retained in the global outputs, which are also used by downstream tasks. In §4.5, a variant whose decoder directly accesses earlier tokens achieves a lower pretraining loss but performs worse downstream, while reconstruction-based two-stage variants also underperform HealthSAW.

3.3 Application of the Learned World Model

The pretrained HealthSAW supports a range of clinical tasks involving patient modeling, either through scenario-specific fine-tuning or directly through zero-shot inference. Here, we use it to form patient representations for supervised prediction and to simulate future trajectories.

Patient representation. The three global outputs from a patient's last event i, h_i^t, h_i^{obs}, and h_i^{act}, summarize the history up to that event; their concatenation forms the patient representation. This representation serves as input to a wide range of clinical prediction tasks, such as disease-risk prediction. For each task, a linear head maps the representation to the task output,

$$\hat{y} = W[h_i^t;\, h_i^{\text{obs}};\, h_i^{\text{act}}].$$

The head is trained on task labels either jointly with the local encoders and global Transformer (supervised fine-tuning) or alone on the frozen representation; the latter setting is suitable for tasks with few labels.

Patient simulation. The world model naturally supports patient simulation by rolling out future trajectories from the recorded history, event by event. After the time head samples $\Delta t_{i+1} \sim p(\cdot \mid h_i^{\text{act}})$, the model encodes it, updates the global output, and decodes the next observation block:

$$s_{i+1}^t = \text{Emb}(\Delta t_{i+1}), \quad h_{i+1}^t = \text{Global}_{\text{AR}}\big(s_1^t, s_1^{\text{obs}}, s_1^{\text{act}}, \ldots, s_{i+1}^t\big), \quad \text{obs}_{i+1} \sim p(\cdot \mid h_{i+1}^t).$$

The generated observation block and then the action block pass through the same encode, update, and decode process, yielding act_{i+1} from h_{i+1}^{obs} and the next elapsed time from h_{i+1}^{act}. Repeating this until the specified time horizon or a generated death yields a simulated future record containing both care and the patient's response to it.

For an outcome of interest, such as a patient's response to a drug, we generate multiple trajectories from the same history and read the outcome within the target time window from each. Aggregating across trajectories yields diagnosis probabilities or laboratory-value predictions that are less sensitive to any single sampled future, without task-specific training. Readout and aggregation details are in Appendices F.2 and G.1.

4 Experiments

We first describe the data and pretraining under data-limited scaling, then evaluate HealthSAW on disease-risk prediction within the pretraining domain and on unseen health systems (§4.2), training-free simulation of treatment outcomes (§4.3), and ablations of the state–action abstraction and forecast-driven bottleneck (§4.4 and 4.5).

4.1 Data and Pretraining

Data sources. Our pretraining data comprise two sources. *CLAIM* is a large-scale, dense longitudinal claims dataset from a universal healthcare system. *Multi-Site EHR* (MS-EHR) is a sparse EHR dataset drawn from connected hospital chains across the United States.

For downstream evaluation, we use three additional sources. *Single-Site EHR* (SS-EHR) contains dense records from a leading medical center within the same national healthcare system as CLAIM. Its patients do not appear in CLAIM, and its records overlap CLAIM's time span but extend about five years beyond it. *PanCan* is a curated cancer cohort comprising sparse, single-site records from a top-10 U.S. hospital. *EHRSHOT* (Wornow et al., 2023) is a public few-shot clinical prediction benchmark on which we evaluate 6 new-diagnosis tasks. Corpus statistics and benchmark details are in Appendices A.4 and E.

We obtained the de-identified patient records through an academic collaboration for CLAIM, a commercial provider under a data-use agreement for MS-EHR, institutional agreements for SS-EHR and PanCan, and the published access terms for EHRSHOT. Appendix A details data licensing, privacy safeguards, and applicable IRB review.

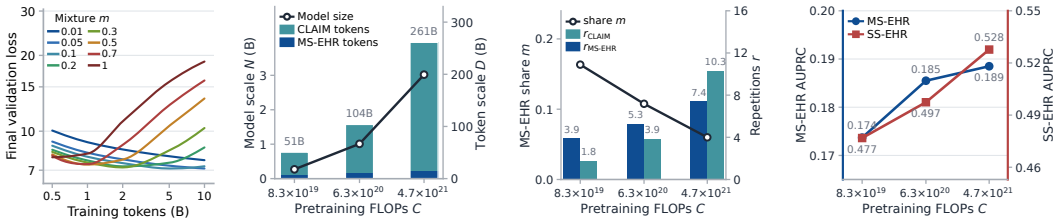


Figure 3: Data-limited scaling, left to right. (a) Mixture-dependent validation loss with only 100M MS-EHR tokens. (b) Data split into MS-EHR and CLAIM tokens at MS-EHR shares m of 16.3%, 10.8%, and 6.0%. (c) Optimal MS-EHR share m and corresponding repetitions r over each corpus. (d) Downstream AUPRC with scaling, each model pretrained at its optimal MS-EHR share.

Data-limited scaling of pretraining. Before scaling up, we study whether larger models and more compute improve downstream clinical prediction. Because healthcare data are scarce relative to compute, pretraining quickly enters a data-limited regime.

We pretrain HealthSAW from scratch on CLAIM and 80% of MS-EHR patients, reserving the rest for validation and testing. Adapting Sedova et al. (2026) to health data, we fit a repetition-aware mixture law that predicts loss from model size, training budget, and the MS-EHR share m, while accounting for diminishing returns from repeated exposure to each corpus (Figure 3a). Minimizing the predicted loss over m yields the optimal mixture for each model size and budget (Figures 3b and 3c; Appendices B and C). To the best of our knowledge, this is the first repetition-aware scaling law for health world models. We pretrain three model sizes at their optimal shares and fine-tune each on the 16 disease-risk tasks of §4.2. Scaling from 0.27B to 3.02B parameters increases macro AUPRC by 8.5% on MS-EHR and 10.7% on SS-EHR (Figure 3d), showing that larger-scale pretraining improves patient representations even in this data-limited regime.

4.2 Patient Representations and Transfer

We evaluate HealthSAW on disease prediction within the pretraining domain and on generalization to unseen health systems, comparing it with language models, tree models, and EHR foundation models.

Task construction. On MS-EHR, we construct 16 disease-risk tasks: seven cancers, five cardiovascular or cerebrovascular conditions, and four neurodegenerative conditions. Inputs end at the prediction anchor; a task-specific *washout period* precedes the outcome window, in which a target diagnosis defines a positive. Appendix D.1 details the prior-history and washout exclusions used to prevent label leakage. Among the three datasets held out from pretraining, SS-EHR follows the same task construction as MS-EHR; PanCan contains one pancreatic cancer risk task constructed analogously to the MS-EHR cancer tasks; and EHRSHOT provides six new-diagnosis tasks with k labeled examples per class. On MS-EHR and SS-EHR, all methods use identical prediction anchors and patient-level 80/10/10 training, validation, and test splits.

Evaluation metrics. We use macro AUPRC, the standard metric for rare positives (Saito & Rehmsmeier, 2015), and report macro positive predictive value (PPV) and sensitivity at 99% specificity following previous work (Sun et al., 2026).

Baselines. We compare HealthSAW with three token-level EHR models that use the same records without state–action abstraction: EHR-FM (Sun et al., 2026), trained with next-token prediction; SMB (JEPA), our implementation of SMB-Structure (Adam et al., 2026), combining JEPA-style prediction with next-token training; and Apollo (Zhang et al., 2026), trained with masked-token prediction. We also evaluate Qwen 3.5-0.8B (Qwen Team, 2026) and MedGemma 1.5-4B (Sellergren et al., 2026), fine-tuned on text-rendered histories; LightGBM (Ke et al., 2017) and XGBoost (Chen & Guestrin, 2016), trained on history-token counts; and zero-shot Claude 4.5 Haiku and GPT-5.6 Luna on the same text histories. On EHRSHOT (Wornow et al., 2023), we report the benchmark's published baselines.

Implementation details. For a controlled comparison, HealthSAW, EHR-FM, SMB, and Apollo are pretrained from scratch under matched conditions: the same corpus, a token-proportional mixture of the two sources, the same training schedule, and a scale of about 1B parameters. All four are then adapted to each task as in §3.3, using the same data and settings. Supervised fine-tuning (SFT) trains the model and a linear classifier on task labels; on EHRSHOT, few-shot transfer fits only the classifier on frozen HealthSAW and EHR-FM representations. More details are in Appendices D and E.

Table 1: Disease risk prediction on MS-EHR (in-domain) and SS-EHR (cross-dataset), with 16 tasks each. We report macro AUPRC, PPV, and sensitivity at 99% specificity by disease group and overall (task counts in parentheses). * Zero-shot API LLMs are evaluated on a 4,000-row test subsample per task. HealthSAW-3B is the 3B model from the scaling study (§4.1). Best per dataset, group, and metric in bold; second best underlined. AUROC and per-task results are in Appendix D.4.

	Cancer (7)			Cardiovascular (5)			Neurodegenerative (4)			All (16)		
Model	AUPRC	PPV	Sens.	AUPRC	PPV	Sens.	AUPRC	PPV	Sens.	AUPRC	PPV	Sens.
MS-EHR (in-domain)												
Claude 4.5 Haiku*	0.044	0.062	0.090	0.138	0.248	0.070	0.053	0.071	0.067	0.075	0.123	0.078
GPT-5.6 Luna*	0.076	0.071	0.116	0.181	0.311	0.078	0.101	0.116	0.138	0.115	0.157	0.109
LightGBM	0.063	0.095	0.146	0.236	0.366	0.139	0.088	0.130	0.148	0.123	0.188	0.144
XGBoost	0.073	0.101	0.162	0.221	0.356	0.139	0.092	0.136	0.162	0.124	0.189	0.155
Qwen 3.5	0.065	0.093	0.157	0.226	0.344	0.121	0.088	0.126	0.139	0.121	0.180	0.141
MedGemma 1.5	0.067	0.097	0.158	0.234	0.350	0.124	0.086	0.118	0.123	0.124	0.181	0.139
Apollo	0.086	0.110	0.181	0.264	0.392	0.158	0.093	0.135	0.156	0.144	0.204	0.168
EHR-FM	0.113	0.134	0.229	0.303	0.432	0.199	0.118	0.167	0.210	0.173	0.235	0.215
SMB (JEPA)	0.111	0.133	0.226	0.306	0.435	0.203	0.118	0.162	0.203	0.173	0.235	0.213
HealthSAW	0.117	0.140	0.241	0.321	0.445	0.213	**0.125**	0.168	**0.213**	0.183	0.242	0.226
HealthSAW-3B	**0.129**	**0.144**	**0.251**	**0.324**	**0.447**	**0.215**	0.122	**0.169**	**0.213**	**0.189**	**0.244**	**0.230**
SS-EHR (cross-dataset generalization)												
Claude 4.5 Haiku*	0.037	0.058	0.031	0.133	0.203	0.022	0.104	0.180	0.139	0.084	0.134	0.055
GPT-5.6 Luna*	0.051	0.081	0.061	0.175	0.309	0.069	0.105	0.156	0.132	0.103	0.171	0.081
LightGBM	0.311	0.316	0.402	0.372	0.484	0.232	0.328	0.310	0.339	0.334	0.367	0.333
XGBoost	0.222	0.241	0.277	0.324	0.451	0.185	0.218	0.247	0.233	0.253	0.308	0.237
Qwen 3.5	0.075	0.112	0.101	0.234	0.371	0.089	0.107	0.149	0.107	0.133	0.202	0.099
MedGemma 1.5	0.061	0.094	0.069	0.225	0.362	0.086	0.117	0.171	0.137	0.126	0.197	0.091
Apollo	0.174	0.232	0.235	0.377	0.529	0.223	0.144	0.200	0.166	0.230	0.317	0.214
EHR-FM	0.303	0.306	0.346	0.482	0.588	0.315	0.239	0.290	0.284	0.343	0.390	0.321
SMB (JEPA)	0.303	0.303	0.347	0.468	0.580	0.304	0.248	0.272	0.253	0.341	0.382	0.310
HealthSAW	0.470	0.378	0.492	0.630	0.668	0.487	0.374	0.336	0.352	0.496	0.458	0.455
HealthSAW-3B	**0.512**	**0.390**	**0.522**	**0.664**	**0.681**	**0.527**	**0.386**	**0.354**	**0.383**	**0.528**	**0.472**	**0.489**

Table 2: Transfer to PanCan: pancreatic cancer risk. AUPRC, PPV and sensitivity at 99% specificity; best per column in bold.

Model	AUPRC	PPV	Sens.
LightGBM	0.4608	0.6667	0.1034
XGBoost	0.4545	0.6591	0.1000
Qwen 3.5	0.3305	0.5833	0.0724
MedGemma 1.5	0.4124	0.6739	0.1069
Apollo	0.4333	0.6875	0.1138
EHR-FM	0.4648	0.7059	0.1241
SMB (JEPA)	0.4840	0.7115	0.1276
HealthSAW	**0.5174**	**0.7887**	**0.1931**

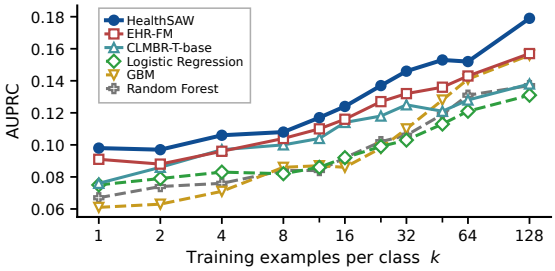


Figure 4: Few-shot transfer to EHRSHOT: mean AUPRC across 6 new-diagnosis tasks with k labeled training examples per class, from 1 to 128.

In-domain evaluation. On MS-EHR patients held out from pretraining, HealthSAW leads every baseline on all three macro metrics, both overall and within each disease group; HealthSAW-3B improves further overall (Table 1). These results show that HealthSAW's generative pretraining yields patient representations that better capture disease risk from clinical histories.

Cross-dataset generalization. On SS-EHR, HealthSAW outperforms every baseline across all three metrics, with its largest gains over token-level models occurring on this dataset: +45% and +46% AUPRC over EHR-FM and SMB, both of which trail LightGBM in the cancer and neurodegenerative groups. All EHR and tree-based models perform substantially better on SS-EHR than on MS-EHR, which contains much less information per event. Consistently, retaining more content from each SS-EHR event raises AUPRC for both HealthSAW and EHR-FM, with larger gains for HealthSAW (§4.4). These results suggest that richer records amplify the benefit of HealthSAW's abstraction.

On PanCan, HealthSAW outperforms the best baseline, SMB (JEPA), with relative gains of 6.9% in AUPRC, 10.9% in PPV, and 51.3% in sensitivity (Table 2). On EHRSHOT's six new-diagnosis tasks, HealthSAW's frozen representations outperform EHR-FM and published baselines in mean AUPRC across all evaluated k-shot settings (Figure 4); with 128 examples per class, EHR-FM matches the tree baselines while HealthSAW remains 15% ahead of them.

Together, these results show strong generalization across healthcare systems, with the largest gains on denser records and consistent few-shot gains from frozen representations. Analyses in §4.4 and §4.5 examine the contributions of the state–action abstraction and forecast-driven bottleneck.

4.3 Patient Simulation

Predicting a patient's response to treatment helps clinicians select effective therapies and trial designers anticipate drug effects. The many combinations of treatments, outcomes, and time horizons make training a separate predictor for each impractical, especially when labeled data are scarce. A world model offers a natural alternative by simulating future patient trajectories to estimate outcomes across these settings (§3.3). Here, we evaluate HealthSAW on treatment-outcome prediction through patient simulation.

Table 3: Treatment outcome simulation on MS-EHR at 20 trajectories per patient. SFT fine-tunes EHR-FM on each task's labels; the simulation columns read outcomes from trajectories generated by one model per method. Best per metric in bold.

	SFT	Simulation	
Group (tasks)	EHR-FM	EHR-FM	HealthSAW
Metric (%)	WAPE↓	WAPE↓ (Cov.↑)	WAPE↓ (Cov.↑)
Blood/coag. (7)	16.31	17.91 (83.37%)	**15.34** (**99.72**%)
Renal/electr. (6)	22.66	18.76 (89.67%)	**16.31** (**98.29**%)
Hepatobiliary (3)	61.69	61.83 (80.32%)	**54.89** (**95.01**%)
Metabolic (3)	33.27	30.12 (79.76%)	**27.93** (**96.25**%)
Inflammatory (3)	60.60	54.15 (88.00%)	**48.83** (**93.00**%)
Overall (22)	32.58	30.74 (86.27%)	**27.28** (**98.48**%)

Task construction. We construct 22 treatment-outcome tasks on MS-EHR across five outcome groups, each pairing a treatment with a laboratory test that assesses treatment response or potential harm. Examples include HbA1c two to three months after an SGLT2 inhibitor to assess glycemic control, and potassium two to six weeks after spironolactone to monitor hyperkalemia risk. Predictions are made at the recorded treatment event and target the first laboratory measurement within the specified follow-up window. Task details are in Appendix F.

Evaluation metrics. We evaluate weighted absolute percentage error (WAPE), defined as each task's mean absolute error (MAE) relative to its mean absolute target. Simulation coverage (Cov.) is the share of samples with at least one valid readout before fallback.

Baselines. HealthSAW receives no task-specific training. Given the record through each treatment event, it rolls out 20 trajectories that preserve the recorded treatment and generate subsequent care and measurements, then reads the target laboratory value from them. EHR-FM is simulated in the same way, token by token. We also compare with per-task EHR-FM SFT, which fine-tunes EHR-FM with a regression head on each task's labels, yielding 22 separate predictors.

Results. Without task-specific training, a single HealthSAW model achieves the lowest WAPE and highest simulation coverage across all outcome groups, with an overall WAPE of 27.28%, 16.3% lower than the 22 supervised predictors, and coverage of 98.48% (Table 3). EHR-FM simulation also outperforms its per-task SFT baseline, suggesting that modeling broader patient trajectories can benefit individual treatment-outcome prediction. HealthSAW's event-level generation also enables faster simulation than EHR-FM (§4.4). A single pretrained model therefore addresses treatment questions that would otherwise require separate labeled predictors.

4.4 Necessity of Abstraction via State–Action Representations

To isolate the effect of state–action organization, we pretrain a variant that preserves HealthSAW's architecture, objective, corpus, and training schedule but randomly shuffles the tokens within each event before splitting them into two blocks. The two blocks therefore contain the same event content but no longer preserve their state–action semantics. Under the downstream SFT protocol of §4.2, HealthSAW outperforms this variant by 2.8% in macro AUPRC on MS-EHR and 23.0% on SS-EHR, where each event contains nearly twice as much information as in MS-EHR (Table 4; Appendix H).

Because the degree of abstraction depends on event density, we further evaluate HealthSAW and EHR-FM on SS-EHR using 25%, 50%, 75%, and 100% of each event's original content. As event density

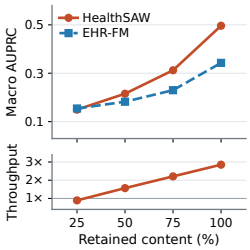


Figure 5: Retained event content on SS-EHR: macro AUPRC (top) and HealthSAW's throughput over EHR-FM (bottom).

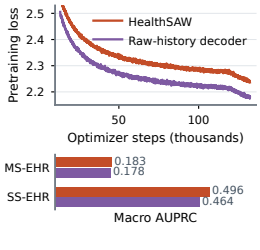


Figure 6: Direct history access: pretraining loss (top) and macro AUPRC (bottom) of HealthSAW and the raw-history decoder.

Table 4: Randomized block content and reconstruction-driven two-stage training against HealthSAW. 16-task macro AUPRC per dataset; sensitivity at 99% specificity, best in bold.

	MS-EHR		SS-EHR	
Model	AUPRC	Sens.	AUPRC	Sens.
Random blocks	0.1777	0.2208	0.4034	0.3718
Cont. two-stage	0.1574	0.1915	0.4787	0.4453
VQ two-stage	0.1241	0.1428	0.4449	0.4141
HealthSAW	**0.1827**	**0.2255**	**0.4963**	**0.4554**

increases, HealthSAW's advantage over EHR-FM grows steadily, reaching a 45% improvement in macro AUPRC at full content (Figure 5; Appendix H). Efficiency shows a similar trend: HealthSAW's patient-representation throughput starts below EHR-FM's at 25% content and ends at nearly three times EHR-FM's throughput, because EHR-FM's sequence length grows with the number of tokens per event while HealthSAW continues to use three slots per event. Thus, the benefits of state–action abstraction grow with event richness, in both accuracy and efficiency.

Abstracting raw tokens into higher-order representations also improves context and computational efficiency: patient histories occupy far fewer global sequence positions, while HealthSAW achieves $1.8\times$ EHR-FM's pretraining token throughput, computes patient representations $2.9\times$ faster, and runs long-horizon disease simulation $3.2\times$ faster. Measurement protocols and complete results are in Appendix I.

4.5 Benefits of Forecast-Driven Bottleneck

HealthSAW's local decoders access earlier blocks only through the global Transformer's representations h, creating an information bottleneck. To test its role, we pretrain a variant whose decoder directly accesses raw tokens from earlier blocks. Despite achieving a lower pretraining loss, this variant performs worse downstream (Figure 6). Because downstream prediction reads h rather than decoder outputs, this gap reflects what is encoded in the representation: direct token access lets the decoder exploit recurring patterns without encoding the relevant information in h. The lower loss therefore reflects trivial learning (Geirhos et al., 2020) rather than better patient representations.

We next test whether compression can be learned separately from dynamics. Following the two-stage approach of §3.2, we train a tokenizer to reconstruct state–action blocks with either continuous vectors or vector-quantized codes, freeze its encoders, and then learn to predict the next block's representation. HealthSAW, which learns compression and forecasting jointly, outperforms both variants on both datasets (Table 4), suggesting that reconstruction may not preserve the information most useful for forecasting. Full configurations and results are in Appendix H.

5 Conclusion

This paper introduces HealthSAW, a patient world model for learning clinical intelligence from large-scale patient healthcare records. To address the trivial learning and token-level over-complication of next-token prediction on EHRs, HealthSAW abstracts raw visit tokens into a higher-order space of health states and clinical actions and trains the model end to end with a forecast-driven bottleneck and next-block prediction. Our experiments across large patient populations and multiple health systems demonstrate the benefits of this architecture and pretraining paradigm on a wide range of tasks. On unseen health systems in particular, HealthSAW outperforms next-token foundation models matched in data, token budget, and parameter count by up to 45%, while the latter sometimes fail to outperform classic tree models. The higher-order representations also improve context and computational efficiency, yielding 1.8–$3.2\times$ gains on structured EHR data.

The world model also naturally supports patient simulation beyond supervised prediction. We demonstrate treatment-outcome simulation as a first step, while future work can extend HealthSAW to applications such as clinical decision support, clinical development, and population health management. Our scaling study across model size, compute, and data scale shows continued gains with scale, while the stronger results on denser records suggest a complementary direction: enriching patient representations with more complete health observations.

AI USE STATEMENT

In this work, we used generative AI tools to draft parts of the manuscript and edit it for readability, to create and modify figures, and to identify relevant literature. We did not use them for any task that requires disclosure under the ICLR 2027 AI policy, such as proposing hypotheses, designing experiments, implementing methods or interpreting results. We have reviewed all AI-assisted work and take responsibility for the final content of the paper, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only retrospective, de-identified patient records; no data were linked to any external information, no attempt was made to re-identify any individual, and the paper reports aggregate results only. Access to the CLAIM data was obtained through an academic collaboration; the study completed the required institutional review board (IRB) process, including submission of the data schema, the research protocol, the data-access procedures and the list of authorized users, and was granted an IRB exemption. The MS-EHR patients were sampled from de-identified longitudinal EHR data provided by a commercial healthcare data provider that works with multiple U.S. EHR repositories and aggregates records from established healthcare systems across several U.S. states, and that authorizes use of these data for non-profit academic research under an executed data-use agreement. Under that agreement, all records are de-identified in accordance with HIPAA guidelines and the provider's certified de-identification workflows, including removal of direct identifiers and date shifting or other transformations where required; access is restricted to authorized researchers; re-identification, redistribution and publication of the underlying patient-level dataset are prohibited; and data-security, confidentiality and data-deletion requirements apply, all of which this study follows. SS-EHR records were accessed under an active IRB-approved protocol and PanCan records under an institutional data-use agreement, each analyzed within the environment its agreement requires, and EHRSHOT was used under its research data-use agreement; no patient-level data from any cohort are released with this paper. The zero-shot LLM baselines received only de-identified test histories rendered as text, through zero-data-retention API endpoints. HealthSAW is a research model; its risk scores and simulated trajectories are not validated for clinical use and may reflect the care patterns of the health systems in its training data. Appendix A details the authorization and governance of each cohort.

REPRODUCIBILITY STATEMENT

The model is specified in §3 (data organization and architecture in §3.1, the next-block objective in §3.2, and simulation and readouts in §3.3), and the experimental setup in §4.1. The appendix details the data, models and experiments: corpus statistics and patient partitions (Appendix A); model configurations and pretraining for HealthSAW and every baseline (Appendix B); the derivation and fitted parameters of the scaling law (Appendix C); the construction of the 16 disease-risk tasks and the per-task SFT settings (Appendix D); the EHRSHOT probing protocol and the PanCan setup (Appendix E); the construction of the 22 treatment-outcome tasks and the simulation readout rules (Appendix F); supplementary disease simulation, including temporal merging and sampling (Appendix G); the ablation configurations and the NLL normalization (Appendix H.1); and the timing protocols (Appendices G.4 and I). Complete per-task results accompany each section. EHRSHOT is available under its research data-use agreement; CLAIM, MS-EHR, SS-EHR and PanCan records are available only under their respective data-use agreements, and the paper reports aggregate results only.

REFERENCES

Irsyad Adam, Zekai Chen, David Laprade, Shaun Porwal, David Laub, Erik Reinertsen, Arda Pekis, and Kevin Brown. The Patient is not a Moving Document: A World Model Training Paradigm for Longitudinal EHR, 2026. URL https://arxiv.org/abs/2601.22128.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding

predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023. URL https://arxiv.org/abs/2301.08243.

Ofir Ben Shoham and Nadav Rappoport. CPLLM: Clinical prediction with large language models. *PLOS Digital Health*, 3(12):e0000680, 2024. doi: 10.1371/journal.pdig.0000680.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Shuang Zhou, Yuan Luo, Rui Zhang, Danielle Bitterman, Fei Wang, and Kai Shu. ClinicalBench: Can LLMs beat traditional ML models in clinical prediction? *arXiv preprint arXiv:2411.06469*, 2024. URL https://arxiv.org/abs/2411.06469.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. MEDITRON-70B: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023. URL https://arxiv.org/abs/2311.16079.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam Shah, and Lillian Sung. A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*, 7:171, 2024. doi: 10.1038/s41746-024-01166-w. URL https://www.nature.com/articles/s41746-024-01166-w.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

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, 2022. URL https://arxiv.org/abs/2203.15556.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Daniel P. Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. Medical adaptation of large language and vision-language models: Are we making progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12143–12170. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.677.

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, 2020. URL https://arxiv.org/abs/2001.08361.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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, 2024. doi: 10.1016/S2589-7500(24)00025-6.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(1):7155, 2020. doi: 10.1038/s41598-020-62922-y.

Linjie Mu, Zhongzhen Huang, Yannian Gu, Shengqian Qin, Shaoting Zhang, and Xiaofan Zhang. EHRWorld: A Patient-Centric Medical World Model for Long-Horizon Clinical Trajectories, 2026. URL https://arxiv.org/abs/2602.03569.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *Journal of Machine Learning Research*, 26(53):1–66, 2025.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Yen-Chen Lin, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.02155.

Mohammad Areeb Qazi, Maryam Nadeem, and Mohammad Yaqub. Beyond generative AI: World models for clinical prediction, counterfactuals, and planning, 2025. URL https://arxiv.org/abs/2511.16333.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86, 2021. doi: 10.1038/s41746-021-00455-y.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

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):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Anastasiia Sedova, Skyler Seto, Natalie Schluter, and Pierre Ablin. Scaling laws for mixture pretraining under data constraints, 2026. URL https://arxiv.org/abs/2605.12715.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, Harshad Purandare, Abhishek Bijay Mishra, Sam Schmidgall, Tao Tu, Anil Palepu, Chunjong Park, Tim Strother, Rahul Thapa, Yong Cheng, Preeti Singh, Kat Black, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Joelle Barral, Tris Warkentin, Shravya Shetty, Dale Webster, Sunny Virmani, David F. Steiner, Can Kirmizibayrak, and Daniel Golden. MedGemma 1.5 technical report, 2026. URL https://arxiv.org/abs/2604.05081.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025. doi: 10.1038/s41586-025-09529-3.

Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021. doi: 10.1016/j.jbi.2020.103637.

Liwen Sun, Hao-Ren Yao, Ophir Frieder, Xiang Qian, and Chenyan Xiong. Scaling electronic health record foundation models for population health management, 2026. URL https://arxiv.org/abs/2506.00209.

Krithik Vishwanath, Anton Alyakin, Mrigayu Ghosh, Ali Hage, Sean N. Neifert, Cordelia Orillac, Nataniel J. Mandelberg, Hammad A. Khan, Jin Vivian Lee, Jie J. Yao, William Robert Small, Aakaash Varma, D. Brock Hewitt, Yindalon Aphinyanaphongs, Daniel Alexander Alber, and Eric Karl Oermann. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nature Medicine*, 32(7):2405–2409, 2026. doi: 10.1038/s41591-026-04431-5.

Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative Medical Event Models Improve with Scale, 2025. URL https://arxiv.org/abs/2508.12104.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A. Fries, and Nigam H. Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 36, pp. 67125–67137, 2023. doi: 10.52202/075280-2933.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H. Shah. Context clues: Evaluating long context models for clinical prediction tasks on EHRs. *arXiv preprint arXiv:2412.16178*, 2024. URL https://arxiv.org/abs/2412.16178.

Qianyi Xu, Gousia Habib, Feng Wu, Dilruk Perera, and Mengling Feng. medDreamer: Model-based reinforcement learning with latent imagination on complex EHRs for clinical decision support. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 1693–1704. ACM, 2026. doi: 10.1145/3770854.3780292.

Yixuan Yang, Mehak Arora, Ryan Zhang, Baraa Abed, Junseob Kim, Tilendra Choudhary, Md Hassanuzzaman, Kevin Zhu, Ayman Ali, Chengkun Yang, Alasdair Edward Gent, Victor Moas, and Rishikesan Kamaleswaran. Clin-JEPA: A Multi-Phase Co-Training Framework for Joint-Embedding Predictive Pretraining on EHR Patient Trajectories, 2026. URL https://arxiv.org/abs/2605.10840.

Andrew Zhang, Tong Ding, Sophia J. Wagner, Caiwei Tian, Ming Y. Lu, Rowland Pettit, Joshua E. Lewis, Alexandre Misrahi, Dandan Mo, Long Phi Le, and Faisal Mahmood. A multimodal and temporal foundation model for virtual patient representations at healthcare system scale, 2026. URL https://arxiv.org/abs/2604.18570.

Sheng Zhang, Qin Liu, Naoto Usuyama, Cliff Wong, Tristan Naumann, and Hoifung Poon. Exploring scaling laws for ehr foundation models, 2025. URL https://arxiv.org/abs/2505.22964.

A DATA

A.1 IRB, DATA AUTHORIZATION, AND COMPLIANCE

We obtain the datasets through authorized academic collaborations or a data-use agreement with a commercial provider. The CLAIM corpus was accessed under formal institutional agreements and data-use authorizations, with an IRB exemption determination. The SS-EHR corpus was accessed under an active IRB-approved research protocol and an approved institutional data compliance letter. The CLAIM and SS-EHR corpora were fully de-identified prior to our access. The MS-EHR corpus was obtained from a commercial data provider under a data-use agreement for academic research, as fully de-identified data supported by an independent third-party HIPAA compliance assessment. The PanCan cohort was evaluated through an authorized academic collaboration under a formal licensing agreement.

A.2 DATA ACCESS AND GOVERNANCE

The CLAIM and SS-EHR corpora were accessed within their respective authorized academic and institutional environments. No data were transferred to public or unauthorized systems; the zero-shot LLM baselines received only de-identified test histories through zero-data-retention API endpoints. The MS-EHR corpus was delivered through secure transfer after full de-identification, supported by a HIPAA compliance assessment on the data vendor's side. No dataset was linked to external identifying information. Only authorized users can access and analyze the data. The PanCan cohort was evaluated exclusively within our collaborating institution's secure environment. It is strictly required that no data leave their system.

A.3 DATA CARD

Table 5 summarizes the four non-public datasets.

Table 5: Characteristics of the four non-public datasets: source, modality, setting, role, access and compliance basis, de-identification, linkage and raw-data availability.

Attribute	CLAIM	SS-EHR	MS-EHR	PanCan
Data source	Health system claims	Single medical center	Connected hospital systems	Partner institution
Data modality	Claims	Longitudinal EHR	Sparse longitudinal EHR	Clinical cohort
Setting	Large-scale claims	Single-site	Multi-site	Partner-site evaluation
Access basis	Academic agreement	Institutional collaboration	Commercial provider, academic DUA	Institutional license
Compliance basis	IRB exempt + data authorization	Active IRB + institutional authorization	HIPAA de-identification assessment	Institutional data-use agreement
Fully de-identified	Yes	Yes	Yes	Yes
Cross-dataset linkage	None	None	None	None
Role in study	Claims evaluation	Single-site evaluation	Multi-site evaluation	External validation
Raw data availability	Restricted	Institution-restricted	Licensed	Institution-only

A.4 CORPUS STATISTICS

Table 6 reports statistics for the CLAIM corpus and the MS-EHR corpus, which together form the joint pretraining mixture of §4.1, and for SS-EHR, which is held out from that pretraining. The

Table 6: Corpus statistics before task selection, partitioning, temporal merging or runtime truncation. An event is one patient-day; total tokens include one prefix token per event, and per-event token counts exclude it but include value-bin tokens, so they count tokenized content rather than distinct concepts. CLAIM is a claims corpus; joint pretraining uses CLAIM with the 80% training partition of MS-EHR.

				Events per patient		Tokens per event		
Dataset	Patients	Events	Total tokens	Mean	Median	Obs.	Act.	Median total
CLAIM	16,223,983	3,197,897,936	23,870,087,133	197.11	152	2.59	3.87	5
MS-EHR	2,000,000	209,680,719	2,654,553,581	104.84	48	10.04	1.62	2
SS-EHR	717,013	81,002,606	1,617,358,633	112.97	61	15.55	3.42	9

additional transfer datasets, PanCan and EHRSHOT, are described in Appendix E. Observation and action tokens are counted separately, since the two blocks carry different amounts of content per event.

The separate counts show that the claims corpus and the two EHR corpora are organized differently. CLAIM records more action than observation tokens per event, 3.87 against 2.59, whereas both EHR corpora are observation-heavy, 10.04 against 1.62 in MS-EHR and 15.55 against 3.42 in SS-EHR. CLAIM also has the longest histories, 197.11 events per patient, and the shortest events.

A.5 Patient Partitions

For MS-EHR and SS-EHR, patients are split into non-overlapping training, validation, and test partitions of approximately 80%, 10%, and 10% before task construction, and all anchors from a patient remain in one partition across tasks. MS-EHR training patients belong to the joint-pretraining pool, while validation and test patients are excluded from pretraining. SS-EHR is excluded from pretraining of the models evaluated in §4.2. The treatment and disease simulation benchmarks (§4.3, Appendix G) use MS-EHR.

B Model and Pretraining

B.1 Model

Parameter accounting. Table 7 lists HealthSAW's configuration. Trainable pretraining parameters are 1.007B for HealthSAW, 0.994B for EHR-FM and Apollo and 1.003B for SMB, giving the EHR-pretrained models comparable total scale. EHR-FM, SMB and Apollo are token-level Transformers of 22 layers at width 1,536 and embed each token at that width, holding a 120,365-entry table twice, once as input and once as the untied output projection.

Architecture. HealthSAW adds learned sex and age embeddings to its event prefix; EHR-FM, SMB and Apollo encode sex and age as tokens. The three slots of a HealthSAW event share one positional encoding, given by the event's time. HealthSAW predicts elapsed time as a coarse bin refined by one of 64 conditional sub-bins; the sub-bin head receives h_i^{act} and, during training, the embedding of the true next coarse bin.

B.2 Pretraining

Schedule and optimization. The four-epoch pretraining schedule used for the primary representation comparison is configured for 132,164 optimization steps in `bf16` mixed precision, with a source-token budget of 3,072 per sample. The EHR models share AdamW pretraining with peak learning rate 2×10^{-4}, batch size 256, betas $(0.9, 0.95)$, weight decay 0.1 and gradient clipping at 1.0. Learning rates use 10% linear warmup, 80% constant rate and 10% exponential decay to half the peak.

Objectives. HealthSAW is trained with the next-block prediction loss of §3.2: its time term is the sum of the coarse-bin and sub-bin cross-entropies, and the observation, action, coarse-time and

Table 7: HealthSAW architecture. Counts describe one stack; there is one local encoder and one local decoder per block type (observation, action). Observation and action tokens use separate embedding tables over the shared vocabulary, with untied output projections. The EHR-pretrained baselines keep their token-level designs at the same total parameter budget (Appendix B.1).

Component	HealthSAW
Global stack: layers / width / heads	24 / 1536 / 24
Local encoder: layers / width / heads	6 / 768 / 12
Local decoder: layers / width / heads	3 / 768 / 12
MLP width: global / local	4096 / 2048
Clinical vocabulary	120,365
Code embedding width	768
Trainable pretraining parameters	1B

sub-bin terms are weighted equally. EHR-FM uses next-token cross-entropy. SMB adds masked latent MSE with weight 0.2 to next-token cross-entropy and uses EMA momentum 0.996. Apollo replaces next-token prediction with masked-token prediction: attention is bidirectional, 30% of input positions have their embedding replaced by a learned mask vector, and cross-entropy is scored only at those positions; it otherwise shares EHR-FM's architecture, token sequences and schedule.

Dataset alignment. Because CLAIM records dominate the pretraining mixture, the simulation experiments first align HealthSAW and EHR-FM to MS-EHR with one further epoch of pretraining on its training split under each model's original objective, and use the final checkpoints. For treatment simulation (Appendix F), one checkpoint per model is shared across all 22 tasks; both models use 2,688 updates, peak learning rate 2×10^{-4} and global batch 256, with 268 warmup, 2,152 constant-rate and 268 exponential-decay updates and a terminal-to-peak learning-rate ratio of 0.1. For disease simulation (Appendix G), one checkpoint per model and temporal resolution is shared across the 16 tasks: original-resolution alignment uses learning rate 2×10^{-5} and batch 256, merged alignment 10^{-5} and batch 64, with the AdamW settings and schedule proportions above; HealthSAW's time-loss weights are 0.1 each, with observation/action weights 1. The per-task alignment used only for temporal merging is described in Appendix G.3.

C Data-Limited Scaling with a Repetition-Aware Mixture

This appendix details the scaling analysis summarized in §4.1.

HealthSAW pretrains on two corpora of very different size: the MS-EHR training partition supplies roughly 2.11 billion unique tokens and CLAIM 23.87 billion (Table 6). Both fall far short of a compute-optimal single pass for the three model sizes (rungs) in Table 8, so the question is how to divide a fixed budget between a small in-domain corpus and a large out-of-domain one. We fit a repetition-aware mixture law with the MS-EHR share m of the token budget as its free parameter. Extending Sedova et al. (2026), the law discounts *each* corpus by a saturating function of its own repetition count before combining them into an effective token count D_{eff}. We write m^* for the loss-minimizing share, and $r_{\text{MS-EHR}}$, r_{CLAIM} for the passes a given share makes over each corpus. Fitted on a 200-run grid and evaluated on held-out MS-EHR validation loss, the law reads

$$L(N, D, m) \;=\; 3.760 \;+\; \frac{7.424 \times 10^{22}}{N^{2.935}} \;+\; \frac{231.9}{D_{\text{eff}}^{0.1793}} \;+\; 0.900\, m, \tag{1}$$

with N in parameters and D_{eff} in tokens. Appendix C.1 derives D_{eff}, lists all ten fitted parameters and describes the grid.

For fixed N and D, the share is chosen by grid search over $[0, 1]$:

$$m^* \;=\; \arg\min_{m \in [0,1]} \; L(N, D, m). \tag{2}$$

Repeated data yields diminishing returns. The discount is exponential in the number of passes and saturates at a fitted ceiling of 3.81 effective passes for MS-EHR and 5.31 for CLAIM: however

Table 8: Repetition-aware mixture scaling across three HealthSAW rungs. N: parameters; D: pretraining tokens; C: pretraining FLOPs; m: the MS-EHR share of the token budget as trained by minimizing the fitted law at N and D, with $r_{\text{MS-EHR}}$, r_{CLAIM} the passes it makes over each corpus. Downstream columns are macro averages over the 16 disease tasks, with PPV and sensitivity at 99% specificity; per-task values in Table 13. †HealthSAW-1B† is pretrained at the MS-EHR share $m^* = 0.1077$, the optimal mixture ratio predicted by the scaling law, with the data selection and update budget of HealthSAW-1B; Figure 3 uses it as the 1B rung.

						MS-EHR			SS-EHR		
N	D	C	m	$r_{\text{MS-EHR}}$	r_{CLAIM}	AUPRC	PPV	Sens.	AUPRC	PPV	Sens.
0.27 B	50.7 B	8.3×10^{19}	0.1634	3.92	1.78	0.1737	0.2339	0.2149	0.4768	0.4544	0.4439
1.01 B	103.9 B	6.3×10^{20}	0.0814	4.00	4.00	0.1827	0.2423	0.2255	0.4963	0.4580	0.4554
1.01 B†	103.9 B	6.3×10^{20}	0.1077	5.29	3.88	0.1855	0.2442	0.2285	0.4974	0.4583	0.4574
3.02 B	261.3 B	4.7×10^{21}	0.0600	7.41	10.29	**0.1885**	**0.2445**	**0.2301**	**0.5277**	**0.4721**	**0.4890**

often a corpus is repeated, it contributes at most that many passes' worth of data. This reproduces, for EHR data, the behavior of data-constrained language models (Muennighoff et al., 2025): after a few epochs another pass over the scarce corpus buys less than spending the same tokens elsewhere.

Mixture training raises repetition tolerance. Because budget spent once one corpus's returns have saturated can be absorbed by the other, the mixture tolerates far more total repetition than a single corpus would: at the 3B rung the fitted optimum takes 7.41 passes over MS-EHR and 10.29 over CLAIM (Table 8) while still improving predicted loss. The optimal share also shifts with scale: m^* falls from 0.1634 at 0.27B to 0.1077 at 1B and 0.0600 at 3B, so larger models draw a smaller fraction of their tokens from the in-domain corpus even though it matches the evaluation distribution more closely. Two fitted quantities explain this: CLAIM tolerates about 39% more effective repetition than MS-EHR, which moves the optimum toward the larger corpus, while a MS-EHR token is worth roughly 55 CLAIM tokens, which keeps m^* well above zero.

Table 8 reports downstream performance across all three trained rungs. From 0.27B through 1B to 3B, macro AUPRC rises from 0.1737 to 0.1855 to 0.1885 on MS-EHR and from 0.4768 to 0.4974 to 0.5277 on SS-EHR. PPV and sensitivity at 99% specificity also improve at each rung on both datasets, and the larger gain from 1B to 3B is on SS-EHR, where AUPRC improves on 14 of 16 tasks (per-task values in Table 13; 3B SFT settings in Appendix D.2).

Actual and fitted mixture shares. m^* is the share minimizing the fitted law at each rung's N and D: 0.1634 at 0.27B, 0.1077 at 1B and 0.0600 at 3B. The basin of shares whose predicted loss is within the fit's error (RMSE 0.190) of the minimum is 0.044–0.461, 0.024–0.362 and 0.011–0.287, respectively. The 0.27B and 3B rungs were trained at m^*, whereas the main 1B model retains the token-proportional four-epoch schedule of §4.1, which is $m = 0.0814$; Table 8 reports the shares as trained. For the 1B budget, the default $m = 0.0814$ and fitted $m^* = 0.1077$ yield predicted losses of 5.7573 and 5.7489, respectively. HealthSAW-1B† repeats the 1B pretraining at the MS-EHR share $m^* = 0.1077$, the optimal mixture ratio predicted by the scaling law, and keeps the data selection and update budget of HealthSAW-1B; it is the 1B rung in Figure 3, and its per-task results are in Table 13.

C.1 Derivation and Fitted Parameters

Setup. Pretraining draws a budget of D tokens from a scarce in-domain *target* corpus of D_{T} unique tokens (the MS-EHR training partition) and an abundant out-of-domain *generic* corpus of D_{G} unique tokens (CLAIM). A single free parameter, the target share $m \in (0, 1]$, splits the budget, so each corpus is traversed

$$r = \frac{mD}{D_{\text{T}}}, \qquad r_g = \frac{(1-m)D}{D_{\text{G}}} \tag{3}$$

times. Both are repetition counts, not epoch caps: $r > 1$ means the target corpus is seen more than once.

Diminishing returns from repetition. A repeated token is worth less than a fresh one. We discount each corpus by a saturating function of its own repetition count,

$$g(x,\rho) = \begin{cases} x, & x \le 1, \\ 1 + \rho\big(1 - e^{-(x-1)/\rho}\big), & x > 1, \end{cases} \tag{4}$$

which is C^1-continuous at $x = 1$, assigns full value to every token in the first pass, and asymptotically approaches $1 + \rho$ as the number of passes increases. The repetition scale ρ controls diminishing returns: beyond the first pass, the marginal value $g'(x) = e^{-(x-1)/\rho}$ decreases by a factor of e every ρ passes. Thus, at $x = 1 + \rho$, a repeated token retains $e^{-1} \approx 0.37$ of a fresh token's value; the ceiling is approached only as $x \to \infty$. We fit ρ separately for each corpus (r_1 for the target and r_2 for the generic corpus), allowing their sensitivity to repetition to differ. Each corpus therefore contributes at most ρ additional effective passes beyond the first, imposing a learned upper bound on the benefit of repeated exposure.

Effective data and loss. The discounted corpora combine into a single effective token count in which one target token is worth τ generic tokens,

$$D_{\text{eff}} = D_{\text{G}}\, g(r_g, r_2) \;+\; \tau D_{\text{T}}\, g(r, r_1), \tag{5}$$

and predicted validation loss follows a Chinchilla-style form with an additional linear mixture penalty,

$$L(N, D, m) = E + \frac{C}{N^{\beta}} + \frac{BN^{\delta}}{D_{\text{eff}}^{\alpha}} + \gamma m. \tag{6}$$

The optimal share m^* minimizes Equation 6 over m at fixed N and D; the reported $r_{\text{MS-EHR}}$ and r_{CLAIM} are r and r_g evaluated at the share each rung was trained with. This extends the mixture scaling law of Sedova et al. (2026) by giving the generic domain its own repetition discount; taking $r_2 \to \infty$ recovers their fresh-data term $(1 - m)D$.

Fitting. The ten parameters are fitted on a grid of 200 pretraining runs spanning *five* configurations of model size and target-pool size, crossed with eight mixture shares $m \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0\}$ and five budgets $D \in \{0.5, 1, 2, 5, 10\}$ billion tokens, for $5 \times 8 \times 5 = 200$ runs. The five configurations are not a full cross product: three target pools at the base model size ($D_{\text{T}} \in \{40\text{M}, 100\text{M}, 1\text{B}\}$ tokens at $N = 274$M), plus two larger models on the 100M pool ($N \in \{587\text{M}, 1.01\text{B}\}$); every configuration uses $D_{\text{G}} = 10D_{\text{T}}$. Each run contributes its *best* loss on a held-out MS-EHR validation set, so all grid points and every predicted L^* are scored on the same in-domain data. Figure 7 plots the *final* validation loss of the same 200 runs instead: the two coincide wherever a run is still improving at the end of training, but diverge under heavy repetition, where late-stage overfitting lifts the final loss well above the best one reached (on the 40M pool at $D = 10$B and $m = 1$, 29.06 final against 9.23 best). The figure therefore shows a sharper upturn than the points the law is fitted on, while the empirical r^* of its fourth panel is unchanged by the choice. Fitting minimizes a Huber objective ($\delta = 0.05$) by Adam followed by L-BFGS with strong-Wolfe line search, over 200 random restarts in log-parameter space, with a softplus reparameterization keeping every parameter positive. Runs are weighted equally; the per-run weight $w = rm$ of Sedova et al. (2026) spans a ∼25,000-fold range on the 40M and 100M pools, where r reaches roughly 250, and lets a few heavily repeated points dominate. Uniform weighting attains $R^2 = 0.9216$ across the 200 grid points, with RMSE 0.190 and a maximum absolute error of 0.702 in validation loss.

Fitted parameters. $E = 3.7597$, $C = 7.4238 \times 10^{22}$, $\beta = 2.9349$, $B = 231.91$, $\delta = 0$, $\alpha = 0.17929$, $r_1 = 2.8071$, $r_2 = 4.3129$, $\tau = 54.802$, $\gamma = 0.89978$, with N in parameters and all token counts in tokens. Three of these carry the section's argument. The asymptotic ceilings $1 + r_1 = 3.81$ and $1 + r_2 = 5.31$ effective passes, together with the slower decay of returns on the generic corpus (e-folding every 4.31 passes against 2.81), say the generic corpus can absorb roughly 39% more effective repetition than the scarce target corpus, which is why the optimum shifts budget toward CLAIM as it grows. The exchange rate $\tau = 54.8$ says one MS-EHR token is worth about 55 CLAIM tokens inside D_{eff}, so the small corpus remains valuable despite its low optimal share. Because δ fits to zero, the data term reduces to $B/D_{\text{eff}}^{\alpha}$ and model size enters only through C/N^{β}.

The grid. Figure 7 shows the sweep used to fit the scaling law. With a small target dataset, validation loss begins to rise as larger training budgets lead to excessive repetition. Reducing the target mixture share delays this increase to higher training budgets. The 1B pool is not exempt: at $m = 1$ its final

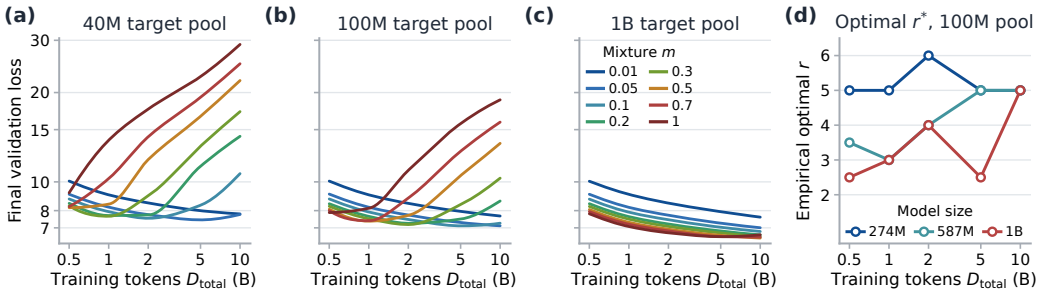


Figure 7: The repetition grid behind the fit in Appendix C: 200 runs, five configurations of model size and target-pool size crossed with eight mixture shares and five budgets. Panels (a)–(c): *final* validation loss against training budget on target pools of 40M, 100M and 1B unique tokens, one curve per mixture share m on a shared loss axis. The law is fitted on each run's *best* validation loss instead (Appendix C.1), which is lower wherever heavy repetition causes late-stage overfitting. Panel (d): empirical optimal repetition r^* against budget for three model sizes on the 100M pool; r^* is the same under either choice of loss.

loss rises from 6.533 at $D = 5$B to 6.632 at $D = 10$B, although every curve with $m \leq 0.7$ still falls across the whole sweep. In the fitted law, the saturating discount g limits each corpus to $1 + \rho$ effective passes, and the target corpus's saturation budget scales with D_{T}/m, so smaller pools or larger mixture shares saturate earlier, whereas larger pools or smaller shares sustain improvements over a wider budget range.

Scaling under data constraints. Scaling studies characterize how model size, data, and compute affect language-model performance (Kaplan et al., 2020; Hoffmann et al., 2022). Similar patterns have been reported for EHR foundation models (Zhang et al., 2025), including Curiosity models trained on Epic Cosmos (Waxler et al., 2025). EHR-FM studies scaling across aligned healthcare datasets and finds that incorporating cross-site data can outperform repeating limited single-site data (Sun et al., 2026). Studies of data-constrained language models characterize diminishing returns from repeated data (Muennighoff et al., 2025), while mixture scaling laws guide the allocation of training tokens across data sources under these constraints (Sedova et al., 2026). HealthSAW's law follows the mixture formulation: it discounts each corpus by its own repetition and chooses the share of MS-EHR against CLAIM, and it extends scaling analysis from health foundation models to a health world model.

D Disease Risk Prediction

D.1 Disease Risk Tasks

Endpoints and eligibility. Table 9 defines the 16 diagnosis endpoints, washout periods, and outcome windows shared by MS-EHR and SS-EHR. For anchor day a, washout period W, and outcome-window width D, model inputs end at a, the washout period is $(a, a + W]$, and the outcome window is $(a + W, a + W + D]$. Eligible anchors are active event-days with at least 365 days of recorded history, at least two active event-days through the anchor, records extending to $a + W$, and no prior task-family exclusion evidence. Breast-cancer cohorts are female-only; no other common age or sex restriction is imposed. Patients with ambiguous duplicate encoded event-days are excluded. Table 10 lists the distinct prior-history and washout-exclusion code sets.

Anchor sampling and labels. For each patient and task, eligible days are ordered by a deterministic hash and greedily sampled to retain at most ten anchors, separated pairwise by at least 365 days. This class-blind selection precedes label assignment; anchors removed during labeling are not replaced. Labels follow three rules in order:

1. A live task-family diagnosis during the washout period excludes the anchor.
2. A live target diagnosis in the outcome window assigns label 1.

Table 9: Disease endpoints and time windows, identical at both sites. ICD-10-CM codes omit dots; % denotes prefix matching. W is the post-anchor washout period and D the outcome-window width, in days. ACS: acute coronary syndrome; CAD: coronary artery disease; SAH: subarachnoid haemorrhage. Breast cancer uses a female-only cohort.

Task	Target prefixes	W	D	Positive days
Bladder cancer	C67%	180	550	181–730
Breast cancer	C50%	180	550	181–730
Colorectal cancer	C18%, C19%, C20%	180	550	181–730
Kidney cancer	C64%	180	550	181–730
Liver cancer	C22%	180	550	181–730
Lung cancer	C34%	180	550	181–730
Pancreatic cancer	C25%	180	550	181–730
ACS	I200%, I21%, I22%	90	640	91–730
CAD	I25%	180	550	181–730
SAH	I60%	90	640	91–730
Heart failure	I50%	180	550	181–730
Stroke (I61/I63)	I61%, I63%	90	640	91–730
Alzheimer disease	G30%	1095	365	1096–1460
Any dementia	F01%, F02%, F03%, G30%, G310%, G3183%	1095	365	1096–1460
Parkinson disease	G20%	1095	365	1096–1460
Vascular dementia	F01%	1095	365	1096–1460

3. Otherwise, the sample receives label 0 if the record reaches the window endpoint; samples with shorter follow-up are excluded.

One target diagnosis suffices for a positive label, without requiring follow-up to the window endpoint. A history marker applies to the following token in the same event; marked diagnoses count for prior exclusion but not for washout exclusions or positive labels. Model-specific context limits are applied to the retained prefix through the anchor.

Negative sampling. All labelable positives are retained. Within each task and partition, negatives are sampled with probability

$$q = \min\left(1, \frac{N_+(1-p)}{pN_-}\right),$$

where N_+ and N_- are the available positive and negative anchor counts. Here p is the fraction of nonempty patients in the entire site's corpus with a live target diagnosis anywhere in their record, including all partitions and both sexes for breast cancer. If either class is empty or $p \leq 0$, available negatives are retained. Tables 11 and 12 report sample, positive, and unique-patient counts by task and partition.

Table 10: Complete family-level exclusion patterns, shared by both sites. Prior exclusions apply through the anchor, including history-marked evidence; washout exclusions apply only to live diagnoses in $(a, a + W]$. Shared patterns are listed once. A trailing % denotes prefix matching; other codes are exact. Ranges include every consecutive numeric code or prefix between both endpoints, with leading zeros preserved. ICD codes omit dots; national procedure codes retain their source namespace.

Stage	Code system	Patterns
Cancer		
Prior and washout	ICD10CM	C00%–C43%, C45%–C76%, C4A%, C7A%, C7B%, C80%–C97%, D00%–D09%, D45%, D46%, D471%, D473%, D474%, D47Z1%, Z8600%
Prior only	ICD10CM	C77%–C79%, Z85%
Prior only	ATC	L01A%, L01BA03–L01BA05, L01BA13, L01BB%, L01BC%, L01C%, L01D%, L01E%, L01FA02–L01FA05, L01FA15, L01FB%, L01FC%, L01FD%, L01FE%, L01FF%, L01FG%, L01FX%, L01FY%, L01X%, L02BB04%–L02BB06%, L02BX%
Prior only	CPT	3824%, 774%, 7752%, 777%, 964%, 965%
Cardiovascular		
Prior and washout	ICD10CM	G450%–G453%, G458%, G459%, I200%, I21%, I22%, I25%, I50%, I60%, I61%, I63%, I671%, I69%, I702%, I703%
Prior only	ICD10CM	Z8673%, Z951%, Z9861%
Prior only	CPT	33510–33514, 33516–33519, 33521–33523, 33530, 33533–33536, 33572, 35301, 35501, 35506, 35508–35512, 35515, 35516, 35518, 35521–35523, 35525, 35526, 35531, 35533, 35535–35540, 35556, 35558, 35560, 35563, 35565
Prior only (cont.)	CPT	35566, 35570, 35571, 35583, 35585, 35587, 35601, 35602, 35606, 35612, 35616, 35621, 35623, 35626, 35631–35634, 35636–35638, 35642, 35645–35647, 35650, 35654, 35656, 35661, 35663, 35665
Prior only (cont.)	CPT	35666, 35671, 35681–35683, 35685, 35686, 35691, 35693–35695, 35697, 92920, 92921, 92924, 92925, 92928–92930, 92933, 92934, 92937, 92938, 92941, 92943–92945, 92972, 92973, 92975, 92977, 92980–92982
Prior only (cont.)	CPT	92984, 92995, 92996
Prior only	ICD10PCS	0210%–0213%, 0270%–0273%, 02C0%–02C3%
Prior only	HCPCS	C7533, C7569, C7571, C9600–C9608, S2205–S2209
Prior only	National procedure	33076A, 33076B, 33077A, 33077B, 33078A, 33078B, 33117B, 68023A, 68023B, 68024A, 68024B, 68025A, 68025B, 68053B, 68054B, 68055B, N26002, N26003
Neurological		
Prior and washout	ICD10CM	F01%–F03%, G10%, G122%, G20%, G21%, G231%–G233%, G30%, G310%, G3183%, G3185%
Prior only	ICD10CM	B20%, C71%, F20%–F25%, F28%, F29%, G912%
Prior only	ATC	N04BA%, N04BB%, N04BD%, N04BX%, N04CX%, N06DA%, N06DX01%, N06DX03%–N06DX05%, N07XX02%

Table 11: Disease-task cohorts on MS-EHR after negative sampling. Each split reports patient–anchor samples, positive samples, and unique patients. Patients are disjoint across splits but can contribute multiple anchors and appear in multiple tasks; patient counts must therefore not be summed across tasks. Task abbreviations are defined in Table 9.

	Train			Validation			Test		
Task	Samples	Positive	Patients	Samples	Positive	Patients	Samples	Positive	Patients
Bladder cancer	522,200	2,724	399,971	67,830	355	51,663	62,338	325	48,489
Breast cancer	539,304	11,071	345,897	69,541	1,435	44,279	68,735	1,414	43,956
Colorectal cancer	490,856	4,029	381,285	59,134	484	46,675	59,346	486	46,691
Kidney cancer	539,426	2,416	409,534	70,449	314	53,167	63,747	285	49,357
Liver cancer	522,660	1,769	400,093	60,844	205	47,468	59,235	201	46,691
Lung cancer	568,458	5,627	425,935	73,677	723	55,088	74,681	733	55,470
Pancreatic cancer	529,857	1,403	403,911	66,456	177	50,933	75,874	201	56,269
ACS	615,906	22,595	451,781	81,065	2,960	58,933	79,661	2,923	58,258
CAD	676,745	60,373	485,800	86,205	7,691	61,771	85,162	7,569	61,300
SAH	633,979	2,318	461,755	84,572	308	60,747	84,678	310	60,661
Heart failure	596,617	34,180	441,828	76,628	4,394	56,726	77,209	4,415	57,104
Stroke (I61/I63)	669,965	24,893	479,837	82,652	3,072	59,758	87,437	3,231	62,171
Alzheimer disease	342,387	2,534	276,448	43,571	322	35,314	45,707	338	36,478
Any dementia	501,130	10,926	370,121	64,924	1,418	47,687	63,830	1,395	46,971
Parkinson disease	386,900	2,457	304,295	46,314	295	37,035	46,311	295	36,899
Vascular dementia	394,739	1,561	308,947	45,621	181	36,301	59,990	238	45,027

Table 12: Disease-task cohorts on SS-EHR after negative sampling. Each split reports patient–anchor samples, positive samples, and unique patients. Patients are disjoint across splits but can contribute multiple anchors and appear in multiple tasks; patient counts must therefore not be summed across tasks. Task abbreviations are defined in Table 9.

	Train			Validation			Test		
Task	Samples	Positive	Patients	Samples	Positive	Patients	Samples	Positive	Patients
Bladder cancer	206,586	1,428	159,629	29,096	200	21,945	20,692	143	16,849
Breast cancer	155,654	4,182	108,615	20,061	536	14,020	18,271	487	13,078
Colorectal cancer	224,547	4,932	170,277	28,669	626	21,664	28,285	626	21,512
Kidney cancer	226,446	814	171,467	25,209	90	19,797	28,803	103	21,751
Liver cancer	254,792	5,557	186,817	32,301	705	23,701	33,988	740	24,545
Lung cancer	253,995	4,585	186,362	34,050	615	24,549	35,254	634	25,246
Pancreatic cancer	248,847	970	183,590	32,368	126	23,744	33,493	130	24,238
ACS	211,923	10,441	161,699	25,377	1,257	19,715	29,832	1,471	22,179
CAD	217,227	25,549	166,367	28,364	3,309	21,667	28,451	3,329	21,603
SAH	199,453	796	154,442	21,856	87	17,498	22,450	90	17,779
Heart failure	190,334	13,802	149,209	24,241	1,757	19,074	24,509	1,794	19,142
Stroke (I61/I63)	217,497	16,083	165,175	27,615	2,054	21,018	26,810	1,989	20,555
Alzheimer disease	159,266	1,659	128,091	21,296	224	16,933	21,100	221	16,760
Any dementia	188,735	7,607	146,701	26,264	1,067	19,942	23,212	930	18,170
Parkinson disease	133,556	1,532	110,932	18,428	209	15,144	16,217	187	13,635
Vascular dementia	139,702	988	115,020	19,660	140	15,915	13,951	98	11,989

D.2 Fine-Tuning and Baselines

Classifiers. HealthSAW concatenates the last event's three global outputs and applies a linear $4608 \to 2$ classifier. EHR-FM, SMB and Apollo use the last token's final-layer output with a $1536 \to 2$ classifier; SMB uses only its online backbone, and Apollo keeps bidirectional attention during SFT. Disease SFT updates the representation backbone and classifier, with unused generation heads frozen. Positive-class softmax gives the risk score.

Disease SFT. For the four EHR-pretrained models, disease SFT uses two epochs and learning rate 10^{-5} on MS-EHR, and four epochs and 3×10^{-5} on SS-EHR, with batch size 128. AdamW uses betas $(0.9, 0.999)$, weight decay 0.01, clipping 1.0 and the same schedule proportions. Validation AUROC selects MS-EHR checkpoints; validation AUPRC selects SS-EHR checkpoints. Histories retain the most recent events within a nominal 3,072-source-token budget; HealthSAW keeps up to 1,024 whole events and caps each historical block at 127 tokens.

For HealthSAW-3B in Appendix C, MS-EHR SFT uses one epoch and the final checkpoint, while SS-EHR uses four epochs and validation-AUPRC selection. Learning rates remain 10^{-5} and 3×10^{-5}, respectively, with global batch size 128 and the AdamW schedule above.

Additional baselines. LightGBM 4.6.0 and XGBoost 3.2.0 use the same sparse history-token counts, including age and sex, under a nominal 3,072-token budget that retains the whole boundary event. Both use learning rate 0.05, at most 31 leaves, up to 200 boosting rounds, and validation AUROC with early-stopping patience 50. LightGBM grows leaf-wise without a depth bound, with at least 20 samples and Hessian sum 0.001 per leaf. XGBoost uses histogram-based, loss-guided growth with maximum depth 5 and minimum child Hessian sum 1. Both use at most 255 histogram bins, full row and column sampling, no L1 or L2 penalty, unit positive-class weight, and 16 CPU threads per task. Absent sparse entries are treated as zero by LightGBM and as missing by XGBoost.

Each algorithm uses one configuration across MS-EHR and SS-EHR; within each run, validation AUROC selects the boosting iteration.

Qwen 3.5 uses the public Qwen 3.5-0.8B-Base weights without EHR pretraining. Its input contains demographics and event-day lines of clinical code descriptions, retaining recent history within 3,072 tokenizer tokens. The last text-token state feeds a linear $1024 \to 2$ head. It uses one epoch at 10^{-5} on MS-EHR (last checkpoint) and three at 4×10^{-5} on SS-EHR (best validation AUPRC), with batch 128 and the AdamW schedule above.

MedGemma 1.5 uses the public MedGemma 1.5-4B-IT weights, also without EHR pretraining: we keep its Gemma 3 language stack and drop the vision tower and multi-modal projector. Its input is the same rendering as Qwen 3.5's, retokenized with the MedGemma tokenizer and retaining recent history within 3,072 tokenizer tokens, and the last text-token state feeds a linear $2560 \to 2$ head. Because the backbone is 4B parameters, we adapt it with LoRA (rank 16, $\alpha = 32$, no dropout) on the attention and gated-MLP projections, freezing the released weights and training only the adapters and the head. Both datasets use the same configuration: one epoch at 10^{-4}, global batch 64, bf16 and best-validation-AUPRC selection, with the AdamW schedule above.

Zero-shot LLM baselines. Claude 4.5 Haiku and GPT-5.6 Luna receive clinical histories rendered as text under a 3,072-source-token budget and return a positive-class probability through a probability-only JSON schema, with provider-default sampling and no fine-tuning; requests run through the providers' batch APIs. Owing to API cost, each task is scored on a 4,000-row stratified subsample of its test split drawn at the split's own positive rate, whereas all other models use the full split.

D.3 Metrics

AUPRC is average precision. For PPV and sensitivity at 99% specificity, we group equal prediction scores and select the lowest attainable threshold with specificity at least 99%, applying the rule $\mathrm{score} \geq \mathrm{threshold}$. Sensitivity is evaluated over all ROC thresholds without interpolation. Operating points are determined on the test split, by the same rule for every model. If no nonempty predicted-positive set satisfies the constraint, PPV and sensitivity are zero.

D.4 PER-TASK RESULTS

Table 13 reports all four metrics for the 16 tasks, with MS-EHR and SS-EHR shown side by side.

Table 13: Per-task disease risk prediction on MS-EHR and SS-EHR; PPV@99 and Sens@99 at 99% specificity (Appendix D.3). Bold marks the largest value per dataset, task and metric, including displayed ties. The closing blocks give task-macro averages over all 16 tasks and within each disease group, including AUROC. * marks the zero-shot API baselines described in Appendix D.2. HealthSAW-1B† is pretrained at the MS-EHR share $m^* = 0.1077$, the optimal mixture ratio predicted by the scaling law (Appendix C); its data, update budget and fine-tuning are otherwise those of HealthSAW-1B.

	MS-EHR				SS-EHR			
Model	AUROC	AUPRC	PPV@99	Sens@99	AUROC	AUPRC	PPV@99	Sens@99
Bladder cancer								
Claude 4.5 Haiku*	0.7453	0.0237	0.0488	0.0952	0.6995	0.0164	0.0250	0.0000
GPT-5.6 Luna*	0.8494	0.1174	0.0714	0.1429	0.7797	0.0418	0.0930	0.1429
LightGBM	0.8632	0.0444	0.0719	0.1477	**0.9067**	0.4411	0.2705	0.5315
XGBoost	0.8563	0.0552	0.0719	0.1477	0.8779	0.2296	0.1800	0.3147
Qwen 3.5	0.8771	0.0594	0.0818	0.1692	0.7765	0.0282	0.0465	0.0699
MedGemma 1.5	0.8837	0.0558	0.0802	0.1661	0.7754	0.0197	0.0330	0.0490
Apollo	0.8760	0.0683	0.0962	0.2031	0.8672	0.1257	0.1535	0.2587
EHR-FM	0.9055	0.1169	0.1222	0.2646	0.8829	0.2548	0.1937	0.3427
SMB (JEPA)	0.9049	0.1069	0.1063	0.2246	0.8712	0.2905	0.1969	0.3497
HealthSAW-200M	0.8944	0.0901	0.1193	0.2585	0.8755	0.4911	**0.2730**	**0.5385**
HealthSAW-1B	**0.9120**	0.1147	**0.1255**	**0.2738**	0.8784	0.4769	0.2626	0.5105
HealthSAW-1B†	0.9015	0.1273	0.1225	0.2646	0.8527	0.4442	0.2491	0.4685
HealthSAW-3B	0.9066	**0.1380**	0.1246	0.2708	0.9012	**0.5218**	0.2705	0.5315
Breast cancer								
Claude 4.5 Haiku*	0.7319	0.0565	0.1333	0.0610	0.5419	0.0311	0.0000	0.0000
GPT-5.6 Luna*	0.7618	0.0986	0.1702	0.0976	0.5835	0.0349	0.0732	0.0280
LightGBM	0.7653	0.1176	0.2017	0.1202	0.7576	0.2278	0.3897	0.2320
XGBoost	0.7602	0.1156	0.2101	0.1266	0.6996	0.1261	0.2716	0.1355
Qwen 3.5	0.7629	0.0943	0.1726	0.0990	0.6223	0.0440	0.1026	0.0411
MedGemma 1.5	0.7697	0.0990	0.1874	0.1096	0.6204	0.0410	0.0435	0.0164
Apollo	0.7735	0.1172	0.2078	0.1245	0.7196	0.1388	0.2532	0.1232
EHR-FM	0.7976	0.1556	0.2540	0.1584	0.7611	0.2965	0.4272	0.2710
SMB (JEPA)	0.8011	0.1564	0.2570	0.1634	0.7501	0.2522	0.3924	0.2320
HealthSAW-200M	0.8023	0.1553	0.2525	0.1584	0.7957	0.4151	0.5230	0.3963
HealthSAW-1B	0.8041	0.1618	0.2588	0.1662	0.7897	0.4264	0.5178	0.3881
HealthSAW-1B†	**0.8076**	0.1608	0.2632	0.1697	0.7986	0.4280	0.5229	0.3984
HealthSAW-3B	0.8075	**0.1754**	**0.2667**	**0.1726**	**0.8073**	**0.4557**	**0.5426**	**0.4312**
Colorectal cancer								
Claude 4.5 Haiku*	0.7525	0.0237	0.0000	0.0000	0.6336	0.0339	0.0488	0.0112
GPT-5.6 Luna*	0.7626	0.0203	0.0000	0.0000	0.7156	0.0474	0.0714	0.0225
LightGBM	0.7732	0.0661	0.0798	0.1049	0.8315	0.2670	0.3741	0.2636
XGBoost	0.7729	0.0794	0.0898	0.1193	0.7924	0.1635	0.2813	0.1725
Qwen 3.5	0.7501	0.0217	0.0363	0.0453	0.7300	0.0558	0.0805	0.0383
MedGemma 1.5	0.7809	0.0522	0.0597	0.0720	0.7550	0.0565	0.0814	0.0383
Apollo	0.7952	0.0772	0.0919	0.1214	0.8076	0.1863	0.2941	0.1837
EHR-FM	0.8252	0.1002	0.1130	0.1523	0.8073	0.2938	0.3907	0.2827
SMB (JEPA)	0.8264	0.1105	0.1359	0.1893	0.8192	0.2897	0.3934	0.2859
HealthSAW-200M	0.8316	0.1137	0.1220	0.1667	0.8278	0.4248	0.4792	0.4058
HealthSAW-1B	**0.8350**	**0.1226**	**0.1431**	**0.1975**	**0.8429**	**0.4498**	0.4963	0.4313
HealthSAW-1B†	0.8311	0.1187	0.1327	0.1811	0.8415	0.4424	0.4926	0.4249
HealthSAW-3B	0.8259	0.1218	0.1331	0.1852	0.8121	0.4468	**0.4991**	**0.4377**

Continued on next page

Table 13 (continued)

Model	MS-EHR AUROC	MS-EHR AUPRC	MS-EHR PPV@99	MS-EHR Sens@99	SS-EHR AUROC	SS-EHR AUPRC	SS-EHR PPV@99	SS-EHR Sens@99
Kidney cancer								
Claude 4.5 Haiku*	0.7231	0.0400	0.0488	0.1111	0.5710	0.0043	0.0000	0.0000
GPT-5.6 Luna*	0.7196	0.0679	0.0488	0.1111	0.5725	0.0040	0.0000	0.0000
LightGBM	0.7851	0.0300	0.0509	0.1193	**0.9026**	0.2250	0.1705	0.5728
XGBoost	0.7879	0.0565	0.0593	0.1404	0.8634	0.2851	0.1355	0.4369
Qwen 3.5	0.7677	0.0472	0.0593	0.1404	0.7563	0.1037	0.0772	0.2330
MedGemma 1.5	0.7851	0.0487	0.0622	0.1474	0.6463	0.0064	0.0104	0.0291
Apollo	0.8020	0.0724	0.0705	0.1684	0.7947	0.0580	0.0860	0.2621
EHR-FM	0.8379	0.0954	0.0947	0.2316	0.8401	0.2961	0.1360	0.4369
SMB (JEPA)	0.8170	0.0733	0.0804	0.1930	0.8424	0.3047	0.1467	0.4757
HealthSAW-200M	0.8253	0.0963	0.0878	0.2140	0.8126	0.4292	0.1584	0.5243
HealthSAW-1B	**0.8403**	0.1030	0.0950	0.2316	0.8229	0.4870	0.1768	0.5922
HealthSAW-1B†	0.8380	0.1130	**0.1027**	**0.2526**	0.8623	0.4898	0.1725	0.5728
HealthSAW-3B	0.8313	**0.1135**	0.0986	0.2421	0.8707	**0.5388**	**0.1810**	**0.6117**
Liver cancer								
Claude 4.5 Haiku*	0.7657	0.0973	0.0930	0.2857	0.7648	0.1023	0.2353	0.1034
GPT-5.6 Luna*	0.8112	0.1512	0.1136	0.3571	0.7834	0.1485	0.1875	0.1034
LightGBM	0.8392	0.0681	0.0881	0.2836	0.9160	0.3798	0.4713	0.4000
XGBoost	0.8458	0.0916	0.0992	0.3234	0.8967	0.3499	0.4411	0.3541
Qwen 3.5	0.8479	0.1279	0.1156	0.3831	0.8677	0.2265	0.3439	0.2351
MedGemma 1.5	0.8587	0.1071	0.1141	0.3781	0.8625	0.2273	0.3529	0.2432
Apollo	0.8832	0.1477	0.1156	0.3831	0.9097	0.4898	0.5257	0.4973
EHR-FM	**0.9077**	0.1529	0.1291	0.4328	0.9090	0.4508	0.5074	0.4622
SMB (JEPA)	0.8965	0.1467	0.1311	0.4428	0.9089	0.4390	0.4977	0.4446
HealthSAW-200M	0.8958	**0.1569**	0.1311	0.4428	0.9135	0.5751	0.5626	0.5770
HealthSAW-1B	0.9030	0.1413	0.1315	0.4428	0.9245	0.5788	0.5654	0.5838
HealthSAW-1B†	0.9027	0.1547	0.1324	0.4478	0.8965	0.5742	0.5585	0.5676
HealthSAW-3B	0.9047	0.1501	**0.1355**	**0.4577**	**0.9291**	**0.6092**	**0.5755**	**0.6081**
Lung cancer								
Claude 4.5 Haiku*	0.7445	0.0585	0.1136	0.0769	0.7423	0.0573	0.0714	0.0417
GPT-5.6 Luna*	0.8215	0.0701	0.0930	0.1026	0.7991	0.0675	0.1136	0.0694
LightGBM	0.8642	0.1011	0.1545	0.1842	0.8397	0.2940	0.3810	0.3360
XGBoost	0.8571	0.0988	0.1467	0.1733	0.8001	0.1617	0.2654	0.1972
Qwen 3.5	0.8627	0.0949	0.1719	0.2087	0.7663	0.0636	0.1244	0.0773
MedGemma 1.5	0.8663	0.0945	0.1619	0.1937	0.7771	0.0710	0.1294	0.0804
Apollo	0.8748	0.1141	0.1752	0.2142	0.8162	0.1925	0.2716	0.2035
EHR-FM	0.8928	0.1411	0.1899	0.2360	0.8475	0.3286	0.3854	0.3423
SMB (JEPA)	0.8932	0.1617	0.1885	0.2333	0.8304	0.3334	0.3926	0.3517
HealthSAW-200M	0.8936	0.1451	0.1890	0.2347	0.8384	0.4595	0.4543	0.4543
HealthSAW-1B	**0.8974**	0.1505	0.1881	0.2333	0.8318	0.4982	0.4733	0.4890
HealthSAW-1B†	0.8972	0.1544	0.1925	0.2374	0.8588	0.5020	0.4750	0.4937
HealthSAW-3B	0.8972	**0.1665**	**0.2028**	**0.2565**	**0.8846**	**0.5638**	**0.5022**	**0.5505**
Pancreatic cancer								
Claude 4.5 Haiku*	0.6680	0.0082	0.0000	0.0000	0.6365	0.0131	0.0250	0.0625
GPT-5.6 Luna*	0.7518	0.0068	0.0000	0.0000	0.6784	0.0129	0.0250	0.0625
LightGBM	0.7682	0.0123	0.0169	0.0647	0.8012	0.3440	0.1570	0.4769
XGBoost	0.7883	0.0131	0.0270	0.1045	0.7830	0.2349	0.1144	0.3308
Qwen 3.5	0.7076	0.0127	0.0144	0.0547	0.5041	0.0042	0.0060	0.0154
MedGemma 1.5	0.7888	0.0087	0.0107	0.0398	0.6318	0.0061	0.0090	0.0231
Apollo	0.7910	0.0086	0.0133	0.0498	0.7136	0.0265	0.0431	0.1154
EHR-FM	0.8429	0.0297	0.0324	0.1244	0.8049	0.2021	0.1000	0.2846
SMB (JEPA)	0.8481	0.0199	0.0349	0.1343	0.8043	0.2152	0.1033	0.2923
HealthSAW-200M	0.8415	0.0231	0.0386	0.1493	0.8229	0.4033	0.1517	0.4538
HealthSAW-1B	0.8386	0.0253	0.0372	0.1443	**0.8264**	0.3744	0.1506	0.4462
HealthSAW-1B†	**0.8488**	0.0241	0.0406	0.1592	0.7877	0.4216	0.1612	**0.4923**
HealthSAW-3B	0.8463	**0.0410**	**0.0435**	**0.1691**	0.8191	**0.4446**	**0.1615**	0.4846

Continued on next page

Table 13 (continued)

Model	MS-EHR AUROC	MS-EHR AUPRC	MS-EHR PPV@99	MS-EHR Sens@99	SS-EHR AUROC	SS-EHR AUPRC	SS-EHR PPV@99	SS-EHR Sens@99
Acute coronary syndrome								
Claude 4.5 Haiku*	0.7427	0.0823	0.1915	0.0612	0.7167	0.1029	0.1915	0.0203
GPT-5.6 Luna*	0.7757	0.1097	0.2245	0.0476	0.7592	0.1590	0.3448	0.0964
LightGBM	0.8338	0.2016	0.3472	0.1396	0.8431	0.3062	0.4973	0.1903
XGBoost	0.8210	0.1928	0.3411	0.1358	0.8207	0.2536	0.4451	0.1543
Qwen 3.5	0.8340	0.1857	0.3145	0.1201	0.8058	0.1963	0.3403	0.0993
MedGemma 1.5	0.8379	0.1897	0.3140	0.1201	0.8008	0.1847	0.3230	0.0918
Apollo	0.8469	0.2259	0.3621	0.1478	0.8502	0.4280	0.6053	0.2950
EHR-FM	0.8668	0.2798	0.4285	0.1967	0.8725	0.5027	0.6702	0.3909
SMB (JEPA)	0.8690	0.2840	0.4224	0.1919	0.8766	0.4810	0.6519	0.3603
HealthSAW-200M	0.8625	0.2685	0.4100	0.1823	0.8848	0.6157	0.7319	0.5235
HealthSAW-1B	0.8717	0.3049	0.4465	0.2111	0.8887	0.6358	0.7391	0.5432
HealthSAW-1B†	0.8722	**0.3058**	**0.4577**	**0.2186**	0.8882	0.6476	0.7425	0.5547
HealthSAW-3B	**0.8732**	0.3025	0.4414	0.2073	**0.8968**	**0.6644**	**0.7496**	**0.5758**
Coronary artery disease								
Claude 4.5 Haiku*	0.8136	0.2766	0.4857	0.0899	0.7287	0.2343	0.3860	0.0470
GPT-5.6 Luna*	0.8357	0.3493	0.5556	0.1096	0.7382	0.2671	0.5205	0.0662
LightGBM	0.8576	0.3974	0.5940	0.1498	0.7800	0.3445	0.5702	0.1000
XGBoost	0.8412	0.3663	0.5728	0.1373	0.7621	0.3223	0.5477	0.0913
Qwen 3.5	0.8633	0.3945	0.5806	0.1418	0.7609	0.2969	0.4980	0.0739
MedGemma 1.5	0.8687	0.4104	0.5957	0.1509	0.7583	0.2963	0.5059	0.0772
Apollo	0.8739	0.4348	0.6339	0.1773	0.8038	0.4042	0.6462	0.1361
EHR-FM	0.8834	0.4716	0.6642	0.2025	0.8461	0.5096	0.7286	0.2025
SMB (JEPA)	0.8849	0.4791	0.6735	0.2113	0.8412	0.4886	0.7110	0.1847
HealthSAW-200M	0.8851	0.4787	0.6691	0.2070	0.8595	0.6184	0.8286	0.3602
HealthSAW-1B	0.8871	0.4895	0.6772	0.2146	0.8789	0.6534	0.8420	0.4019
HealthSAW-1B†	0.8874	0.4895	0.6763	0.2139	0.8613	0.6531	0.8475	0.4190
HealthSAW-3B	**0.8892**	**0.4987**	**0.6868**	**0.2239**	**0.8908**	**0.7025**	**0.8634**	**0.4767**
Subarachnoid haemorrhage								
Claude 4.5 Haiku*	0.6356	0.0104	0.0250	0.0667	0.5740	0.0079	0.0000	0.0000
GPT-5.6 Luna*	0.6035	0.0061	0.0000	0.0000	0.6857	0.0094	0.0250	0.0625
LightGBM	0.7682	0.0143	0.0277	0.0774	0.8491	0.4117	0.1679	0.5000
XGBoost	0.7582	0.0170	0.0420	0.1194	0.8176	0.2864	0.1186	0.3333
Qwen 3.5	0.6909	0.0092	0.0175	0.0484	0.5575	0.0047	0.0046	0.0111
MedGemma 1.5	0.7245	0.0101	0.0179	0.0484	0.6013	0.0060	0.0050	0.0111
Apollo	0.7625	0.0153	0.0288	0.0806	0.6785	0.0954	0.0708	0.1889
EHR-FM	0.7975	0.0394	0.0528	0.1516	0.8290	0.3385	0.1378	0.3889
SMB (JEPA)	**0.8216**	0.0351	0.0560	0.1613	0.7936	0.3141	0.1373	0.3889
HealthSAW-200M	0.8136	0.0446	0.0541	0.1548	**0.8628**	0.4844	0.1710	0.5111
HealthSAW-1B	0.8162	0.0620	0.0609	0.1742	0.8021	0.5281	0.1801	0.5444
HealthSAW-1B†	0.8151	**0.0756**	**0.0722**	**0.2097**	0.8324	**0.5515**	**0.1889**	**0.5667**
HealthSAW-3B	0.8161	0.0519	0.0605	0.1742	0.8388	0.5475	0.1832	0.5556
Heart failure								
Claude 4.5 Haiku*	0.8029	0.2122	0.3833	0.0830	0.7641	0.1749	0.2600	0.0273
GPT-5.6 Luna*	0.8338	0.3006	0.5195	0.1441	0.8051	0.2583	0.3934	0.0785
LightGBM	0.8883	0.3957	0.5776	0.2251	0.8660	0.4374	0.6303	0.2157
XGBoost	0.8748	0.3680	0.5572	0.2072	0.8516	0.4169	0.6229	0.2090
Qwen 3.5	0.8890	0.3822	0.5608	0.2100	0.8476	0.3753	0.5511	0.1533
MedGemma 1.5	0.8951	0.3964	0.5617	0.2104	0.8479	0.3655	0.5609	0.1616
Apollo	0.9036	0.4511	0.6192	0.2677	0.8783	0.5162	0.6895	0.2809
EHR-FM	0.9112	0.4742	0.6313	0.2815	0.8942	0.5294	0.6912	0.2832
SMB (JEPA)	0.9123	0.4772	0.6346	0.2856	0.8981	0.5376	0.7050	0.3010
HealthSAW-200M	0.9127	0.4773	0.6348	0.2854	0.8971	0.6151	0.7693	0.4220
HealthSAW-1B	0.9140	0.4868	0.6400	0.2920	0.9235	0.6753	0.7906	0.4777
HealthSAW-1B†	0.9143	0.4881	**0.6435**	**0.2972**	0.9237	0.6662	0.7830	0.4565
HealthSAW-3B	**0.9167**	**0.4959**	0.6392	0.2917	**0.9264**	**0.7046**	**0.8004**	**0.5072**

Continued on next page

Table 13 (continued)

Model	MS-EHR AUROC	MS-EHR AUPRC	MS-EHR PPV@99	MS-EHR Sens@99	SS-EHR AUROC	SS-EHR AUPRC	SS-EHR PPV@99	SS-EHR Sens@99
Stroke								
Claude 4.5 Haiku*	0.7633	0.1084	0.1556	0.0473	0.7192	0.1442	0.1778	0.0135
GPT-5.6 Luna*	0.7817	0.1398	0.2549	0.0878	0.7625	0.1817	0.2600	0.0404
LightGBM	0.8201	0.1708	0.2834	0.1031	0.8503	0.3600	0.5540	0.1549
XGBoost	0.8099	0.1586	0.2678	0.0953	0.8365	0.3383	0.5212	0.1357
Qwen 3.5	0.8276	0.1561	0.2480	0.0854	0.8292	0.2975	0.4619	0.1066
MedGemma 1.5	0.8306	0.1631	0.2584	0.0907	0.8220	0.2733	0.4161	0.0885
Apollo	0.8327	0.1938	0.3142	0.1188	0.8706	0.4415	0.6331	0.2152
EHR-FM	0.8529	0.2482	0.3835	0.1616	0.8939	0.5281	0.7113	0.3072
SMB (JEPA)	0.8539	0.2525	0.3863	0.1640	0.8931	0.5175	0.6965	0.2861
HealthSAW-200M	0.8534	0.2482	0.3806	0.1594	0.9028	0.6409	0.7819	0.4470
HealthSAW-1B	0.8554	0.2627	0.3997	0.1733	0.9098	0.6580	0.7898	0.4686
HealthSAW-1B†	0.8545	0.2647	0.3964	0.1705	0.9153	0.6752	0.7982	0.4932
HealthSAW-3B	**0.8573**	**0.2718**	**0.4054**	**0.1777**	**0.9206**	**0.6998**	**0.8070**	**0.5214**
Alzheimer's disease								
Claude 4.5 Haiku*	0.8592	0.0332	0.0000	0.0000	0.7917	0.0847	0.0930	0.0952
GPT-5.6 Luna*	0.8942	0.0707	0.0930	0.1333	0.8137	0.0461	0.0714	0.0714
LightGBM	0.8990	0.0646	0.0976	0.1450	**0.9282**	0.2973	**0.2852**	**0.3756**
XGBoost	0.9008	0.0723	0.1100	0.1657	0.9146	0.1815	0.2180	0.2624
Qwen 3.5	0.9065	0.0612	0.0907	0.1331	0.8918	0.0896	0.1303	0.1403
MedGemma 1.5	0.9036	0.0568	0.0683	0.0976	0.8980	0.0870	0.1333	0.1448
Apollo	0.9049	0.0688	0.0905	0.1331	0.8716	0.1087	0.1405	0.1538
EHR-FM	0.9212	0.1008	0.1431	0.2189	0.9049	0.2057	0.2491	0.3122
SMB (JEPA)	0.9210	0.0939	0.1293	0.1982	0.8747	0.1955	0.2210	0.2670
HealthSAW-200M	0.9142	0.0937	0.1238	0.1893	0.9079	0.3024	0.2643	0.3348
HealthSAW-1B	**0.9213**	0.1048	**0.1450**	**0.2249**	0.9187	0.3084	0.2581	0.3258
HealthSAW-1B†	0.9199	**0.1063**	0.1442	**0.2249**	0.9223	0.2997	0.2624	0.3348
HealthSAW-3B	0.9166	0.0919	0.1355	0.2101	0.9077	**0.3182**	0.2660	0.3394
Any dementia								
Claude 4.5 Haiku*	0.8838	0.1266	0.1875	0.1034	0.8240	0.2078	0.4242	0.1688
GPT-5.6 Luna*	0.9040	0.1801	0.2500	0.1494	0.8370	0.1938	0.3091	0.1062
LightGBM	0.9066	0.2147	0.3128	0.2036	0.9110	0.3179	0.4337	0.1828
XGBoost	0.9002	0.2036	0.3036	0.1950	0.8996	0.2906	0.4173	0.1710
Qwen 3.5	0.9153	0.2199	0.3142	0.2036	0.8886	0.2417	0.3333	0.1194
MedGemma 1.5	0.9174	0.2258	0.3173	0.2079	0.8887	0.2400	0.3403	0.1226
Apollo	0.9150	0.2260	0.3305	0.2208	0.8974	0.2888	0.4000	0.1591
EHR-FM	0.9228	0.2578	0.3577	0.2487	0.8883	0.3479	0.5022	0.2409
SMB (JEPA)	0.9242	0.2586	0.3574	0.2487	0.8927	0.3695	0.5011	0.2398
HealthSAW-200M	0.9240	0.2671	0.3581	0.2487	0.9018	0.4584	0.5866	0.3387
HealthSAW-1B	0.9242	0.2694	0.3607	0.2523	0.9234	0.4778	0.5854	0.3355
HealthSAW-1B†	0.9250	**0.2779**	0.3694	0.2616	0.9179	0.4542	0.5739	0.3215
HealthSAW-3B	**0.9256**	0.2724	**0.3726**	**0.2652**	**0.9247**	**0.5542**	**0.6361**	**0.4172**
Parkinson's disease								
Claude 4.5 Haiku*	0.7322	0.0171	0.0488	0.0400	0.6931	0.0467	0.0488	0.0435
GPT-5.6 Luna*	0.8479	0.0436	0.0488	0.0800	0.7628	0.0564	0.0714	0.0652
LightGBM	0.8338	0.0328	0.0476	0.0780	**0.8654**	0.2794	0.2593	0.2995
XGBoost	0.8263	0.0370	0.0631	0.1051	0.8397	0.1365	0.1579	0.1604
Qwen 3.5	0.8482	0.0324	0.0383	0.0610	0.7568	0.0317	0.0440	0.0374
MedGemma 1.5	0.8414	0.0283	0.0327	0.0508	0.8021	0.0453	0.0828	0.0749
Apollo	0.8256	0.0289	0.0497	0.0814	0.7981	0.0796	0.1117	0.1070
EHR-FM	0.8647	0.0545	0.0789	0.1322	0.8017	0.1603	0.2000	0.2139
SMB (JEPA)	0.8675	0.0511	0.0727	0.1220	0.7942	0.1659	0.1804	0.1872
HealthSAW-200M	0.8677	0.0528	**0.0800**	**0.1356**	0.8490	**0.3104**	**0.2773**	**0.3262**
HealthSAW-1B	0.8679	0.0540	0.0783	0.1322	0.8243	0.2784	0.2523	0.2888
HealthSAW-1B†	0.8672	**0.0561**	0.0788	0.1322	0.8505	0.2947	0.2582	0.2941
HealthSAW-3B	**0.8705**	0.0539	0.0746	0.1254	0.8502	0.3040	0.2694	0.3155

Continued on next page

Table 13 (continued)

Model	MS-EHR AUROC	MS-EHR AUPRC	MS-EHR PPV@99	MS-EHR Sens@99	SS-EHR AUROC	SS-EHR AUPRC	SS-EHR PPV@99	SS-EHR Sens@99
Vascular dementia								
Claude 4.5 Haiku*	0.8925	0.0338	0.0488	0.1250	0.8832	0.0774	0.1522	0.2500
GPT-5.6 Luna*	**0.9441**	**0.1115**	0.0714	0.1875	0.8878	0.1254	0.1702	0.2857
LightGBM	0.8876	0.0382	0.0613	0.1639	0.9212	0.4163	**0.2620**	**0.5000**
XGBoost	0.8966	0.0557	0.0672	0.1807	0.9201	0.2635	0.1930	0.3367
Qwen 3.5	0.8940	0.0403	0.0599	0.1597	0.9111	0.0670	0.0867	0.1327
MedGemma 1.5	0.8734	0.0312	0.0519	0.1344	0.9101	0.0966	0.1282	0.2041
Apollo	0.9021	0.0493	0.0701	0.1891	0.8822	0.0978	0.1482	0.2449
EHR-FM	0.9076	0.0578	0.0878	0.2395	0.9161	0.2405	0.2081	0.3673
SMB (JEPA)	0.9190	0.0669	0.0887	0.2437	0.9154	0.2613	0.1834	0.3163
HealthSAW-200M	0.9157	0.0679	**0.0919**	**0.2521**	0.9225	0.3854	0.2581	0.4898
HealthSAW-1B	0.9249	0.0703	0.0891	0.2437	0.9050	**0.4334**	0.2473	0.4592
HealthSAW-1B†	0.9124	0.0507	0.0816	0.2143	**0.9304**	0.4133	0.2459	0.4592
HealthSAW-3B	0.9145	0.0698	0.0917	**0.2521**	0.9030	0.3667	0.2459	0.4592
Macro over 16 tasks								
Claude 4.5 Haiku*	0.7661	0.0755	0.1227	0.0779	0.7053	0.0837	0.1337	0.0553
GPT-5.6 Luna*	0.8062	0.1152	0.1572	0.1094	0.7478	0.1034	0.1706	0.0813
LightGBM	0.8346	0.1231	0.1883	0.1444	0.8606	0.3343	0.3671	0.3332
XGBoost	0.8311	0.1238	0.1893	0.1548	0.8360	0.2525	0.3082	0.2372
Qwen 3.5	0.8278	0.1212	0.1798	0.1415	0.7670	0.1329	0.2019	0.0990
MedGemma 1.5	0.8391	0.1236	0.1809	0.1386	0.7749	0.1264	0.1972	0.0910
Apollo	0.8477	0.1437	0.2043	0.1676	0.8225	0.2299	0.3170	0.2141
EHR-FM	0.8711	0.1735	0.2352	0.2146	0.8562	0.3428	0.3899	0.3206
SMB (JEPA)	0.8725	0.1734	0.2347	0.2130	0.8504	0.3410	0.3819	0.3102
HealthSAW-200M	0.8709	0.1737	0.2339	0.2149	0.8672	0.4768	0.4544	0.4439
HealthSAW-1B	**0.8758**	0.1827	0.2423	0.2255	0.8682	0.4963	0.4580	0.4554
HealthSAW-1B†	0.8747	0.1855	0.2442	0.2285	0.8713	0.4974	0.4583	0.4574
HealthSAW-3B	0.8750	**0.1885**	**0.2445**	**0.2301**	**0.8802**	**0.5277**	**0.4721**	**0.4890**
Macro over cancer tasks (7)								
Claude 4.5 Haiku*	0.7330	0.0440	0.0625	0.0900	0.6557	0.0369	0.0579	0.0313
GPT-5.6 Luna*	0.7826	0.0760	0.0710	0.1159	0.7017	0.0510	0.0805	0.0612
LightGBM	0.8083	0.0628	0.0948	0.1464	0.8508	0.3112	0.3163	0.4018
XGBoost	0.8098	0.0729	0.1006	0.1622	0.8162	0.2215	0.2413	0.2774
Qwen 3.5	0.7966	0.0654	0.0931	0.1572	0.7176	0.0751	0.1116	0.1014
MedGemma 1.5	0.8190	0.0666	0.0966	0.1581	0.7241	0.0611	0.0942	0.0685
Apollo	0.8279	0.0865	0.1101	0.1806	0.8041	0.1739	0.2325	0.2348
EHR-FM	0.8585	0.1131	0.1336	0.2286	0.8361	0.3032	0.3058	0.3461
SMB (JEPA)	0.8553	0.1108	0.1334	0.2258	0.8323	0.3035	0.3033	0.3474
HealthSAW-200M	0.8549	0.1115	0.1343	0.2321	0.8409	0.4569	0.3717	0.4786
HealthSAW-1B	**0.8615**	0.1170	0.1399	0.2414	0.8452	0.4702	0.3776	0.4916
HealthSAW-1B†	0.8610	0.1219	0.1409	0.2446	0.8426	0.4717	0.3760	0.4883
HealthSAW-3B	0.8599	**0.1295**	**0.1436**	**0.2506**	**0.8606**	**0.5115**	**0.3903**	**0.5222**
Macro over cardiovascular tasks (5)								
Claude 4.5 Haiku*	0.7516	0.1380	0.2482	0.0696	0.7005	0.1328	0.2031	0.0216
GPT-5.6 Luna*	0.7661	0.1811	0.3109	0.0778	0.7501	0.1751	0.3087	0.0688
LightGBM	0.8336	0.2360	0.3660	0.1390	0.8377	0.3720	0.4839	0.2322
XGBoost	0.8210	0.2205	0.3562	0.1390	0.8177	0.3235	0.4511	0.1847
Qwen 3.5	0.8210	0.2255	0.3443	0.1211	0.7602	0.2341	0.3712	0.0888
MedGemma 1.5	0.8313	0.2339	0.3495	0.1241	0.7661	0.2251	0.3622	0.0860
Apollo	0.8439	0.2642	0.3916	0.1585	0.8163	0.3771	0.5290	0.2232
EHR-FM	0.8624	0.3026	0.4321	0.1988	0.8672	0.4816	0.5878	0.3145
SMB (JEPA)	0.8683	0.3056	0.4346	0.2028	0.8605	0.4678	0.5803	0.3042
HealthSAW-200M	0.8655	0.3035	0.4297	0.1978	0.8814	0.5949	0.6565	0.4528
HealthSAW-1B	0.8689	0.3212	0.4449	0.2130	0.8806	0.6301	0.6683	0.4872
HealthSAW-1B†	0.8687	**0.3247**	**0.4492**	**0.2220**	0.8842	0.6387	0.6720	0.4980
HealthSAW-3B	**0.8705**	0.3241	0.4466	0.2150	**0.8947**	**0.6638**	**0.6807**	**0.5273**

Continued on next page

Table 14: PanCan pancreatic cancer cohort after negative subsampling, split 80/10/10 at the patient level with stratification on whether a patient is positive at any anchor. Prevalence is the anchor-level positive rate within the split.

Split	Anchors	Patients	Positives	Prevalence
Train	13,975	9,033	2,339	16.74%
Validation	1,759	1,130	281	15.97%
Test	1,801	1,128	290	16.10%
Total	17,535	11,291	2,910	16.60%

Table 13 (continued)

	MS-EHR				SS-EHR			
Model	AUROC	AUPRC	PPV@99	Sens@99	AUROC	AUPRC	PPV@99	Sens@99
Macro over neurodegenerative tasks (4)								
Claude 4.5 Haiku*	0.8419	0.0527	0.0713	0.0671	0.7980	0.1042	0.1796	0.1394
GPT-5.6 Luna*	0.8976	0.1015	0.1158	0.1376	0.8253	0.1054	0.1555	0.1321
LightGBM	0.8818	0.0876	0.1298	0.1476	**0.9064**	0.3277	0.3100	0.3395
XGBoost	0.8810	0.0921	0.1360	0.1616	0.8935	0.2180	0.2465	0.2326
Qwen 3.5	0.8910	0.0885	0.1258	0.1394	0.8621	0.1075	0.1486	0.1075
MedGemma 1.5	0.8840	0.0855	0.1175	0.1227	0.8747	0.1172	0.1712	0.1366
Apollo	0.8869	0.0932	0.1352	0.1561	0.8623	0.1437	0.2001	0.1662
EHR-FM	0.9041	0.1177	0.1669	0.2098	0.8777	0.2386	0.2899	0.2836
SMB (JEPA)	0.9079	0.1176	0.1620	0.2032	0.8693	0.2481	0.2715	0.2526
HealthSAW-200M	0.9054	0.1204	0.1634	0.2064	0.8953	0.3641	0.3466	0.3724
HealthSAW-1B	**0.9096**	**0.1246**	0.1683	**0.2133**	0.8929	0.3745	0.3358	0.3523
HealthSAW-1B†	0.9061	0.1227	0.1685	0.2082	0.9053	0.3655	0.3351	0.3524
HealthSAW-3B	0.9068	0.1220	**0.1686**	0.2132	0.8964	**0.3858**	**0.3543**	**0.3828**

E Transfer to Unseen Health Systems

E.1 PanCan Pancreatic Cancer Cohort

Corpus. PanCan is absent from pretraining; its records arrive in the OMOP common data model. We convert 25,000 patients into the shared event format, yielding 2,281,075 daily events and 41,195,878 tokens from the condition, drug, procedure, measurement, observation, specimen, device, visit and death tables. Concepts are mapped to the shared vocabulary through a reverse mapping of the OMOP standard and source concepts; a concept with no representable token falls back to its coarse event-category token, and laboratory values use the pretraining bins.

Task. The pancreatic cancer task follows the construction of the MS-EHR cancer tasks (Appendix D.1): a target diagnosis (ICD-10-CM C25) in the outcome window 181–730 days after the anchor, the same eligibility and anchor rules, and negatives downsampled to the site's lifetime prevalence. Labels come from the OMOP condition records rather than from the converted tokens, and a negative must be observed through the end of the window according to the OMOP observation periods.

Train/validation/test split. The labeled pancreatic cancer cohort is split 80/10/10 at the patient level, stratified on whether a patient is positive at any anchor, so a patient's anchors never straddle a split. Table 14 gives the resulting counts.

Initialization. EHR-FM, SMB (JEPA), Apollo and HealthSAW start from the pretrained 1B CLAIM/MS-EHR checkpoints of §4.1; Qwen 3.5 starts from the public Qwen 3.5-0.8B-Base weights and MedGemma 1.5 from the released MedGemma-1.5-4B-it language backbone.

Tree baselines. Both tree rows read the same anchor histories as the neural models, featurized identically: bag-of-token counts over the anchor prefix with the `[sex, age]` tokens injected, a

2,048-token history budget and a 120,365-dimensional count vector. Each samples hyper-parameters at random, fits every configuration on the training split, and selects the configuration by validation AUPRC.

Results. Table 15 reports all four metrics for every model on the PanCan test split; Table 2 in the main text shows its AUPRC, PPV and sensitivity columns.

Table 15: All PanCan pancreatic cancer metrics. Test split: 1,801 anchors from 1,128 patients, 290 positive (16.1%). PPV and sensitivity at 99% specificity; largest value per column bold. MedGemma 1.5 is LoRA fine-tuned ($r = 16$); every other neural row trains all weights.

Model	AUROC	AUPRC	PPV	Sens.
LightGBM	0.7863	0.4608	0.6667	0.1034
XGBoost	0.7864	0.4545	0.6591	0.1000
Qwen 3.5	0.7207	0.3305	0.5833	0.0724
MedGemma 1.5	0.7672	0.4124	0.6739	0.1069
Apollo	0.7762	0.4333	0.6875	0.1138
EHR-FM	0.7961	0.4648	0.7059	0.1241
SMB (JEPA)	0.8059	0.4840	0.7115	0.1276
HealthSAW	**0.8060**	**0.5174**	**0.7887**	**0.1931**

E.2 EHRSHOT Few-Shot Generalization

EHRSHOT (Wornow et al., 2023) is a Stanford Medicine benchmark absent from pretraining. We use its six assignment-of-new-diagnoses tasks: acute myocardial infarction, celiac disease, hyperlipidemia, hypertension, lupus, and pancreatic cancer. Records are mapped to the shared vocabulary and event format; unmapped codes use their coarse event-category token, and laboratory values use the pretraining bins. Inputs end at the last event on or before the benchmark prediction time.

For each task and label budget in Table 16, we follow the standard protocol description and fit logistic regression on frozen patient representations using the five released support-set repeats, evaluate on the complete test split, and average over repeats and then equally over tasks. HealthSAW-base and EHR-FM-base use identical probing procedures, with no EHRSHOT backbone training. CLMBR-T-base (Steinberg et al., 2021), logistic regression, GBM, and random forest values are the published benchmark results; count-based baselines use EHRSHOT's count features. The table supplies the values plotted in Figure 4.

Table 16: Few-shot transfer to EHRSHOT, averaged over the assignment-of-new-diagnoses tasks. Columns give the number of labeled training examples per class k. HealthSAW-base and EHR-FM-base are probed from frozen jointly pretrained checkpoints; the remaining rows are the benchmark's published results. Bold marks the best value per column and metric, including displayed ties.

Model	Training examples per class k										
	1	2	4	8	12	16	24	32	48	64	128
AUROC											
HealthSAW-base	**0.565**	**0.559**	**0.605**	**0.624**	**0.638**	**0.662**	**0.679**	**0.701**	**0.706**	0.710	0.728
EHR-FM-base	0.556	0.548	0.598	**0.624**	0.631	0.650	0.664	0.689	0.703	**0.712**	**0.735**
CLMBR-T-base	0.534	0.554	0.593	0.616	0.624	0.640	0.649	0.656	0.666	0.676	0.693
Logistic Regression	0.543	0.556	0.566	0.599	0.604	0.631	0.644	0.660	0.685	0.703	0.722
GBM	0.487	0.521	0.537	0.571	0.588	0.586	0.622	0.618	0.652	0.691	0.723
Random Forest	0.503	0.540	0.549	0.569	0.585	0.595	0.627	0.638	0.665	0.674	0.673
AUPRC											
HealthSAW-base	**0.098**	**0.097**	**0.106**	**0.108**	**0.117**	**0.124**	**0.137**	**0.146**	**0.153**	**0.152**	**0.179**
EHR-FM-base	0.091	0.088	0.096	0.104	0.110	0.116	0.127	0.132	0.136	0.143	0.157
CLMBR-T-base	0.076	0.086	0.097	0.100	0.104	0.114	0.118	0.125	0.121	0.128	0.138
Logistic Regression	0.075	0.079	0.083	0.082	0.086	0.092	0.099	0.103	0.113	0.121	0.131
GBM	0.061	0.063	0.071	0.086	0.087	0.086	0.098	0.110	0.128	0.141	0.156
Random Forest	0.067	0.074	0.076	0.084	0.084	0.092	0.102	0.106	0.119	0.131	0.137

F Treatment Outcome Simulation

F.1 Treatment Outcome Tasks

The 22 tasks follow the five outcome groups in §4.3, with 7 blood-count/coagulation, 6 renal-function/electrolyte, 3 hepatobiliary, 3 metabolic/endocrine and 3 inflammatory/immune tasks. Diagnostic eligibility, treatment codes, laboratory catalogs and baseline rules for every task are in the grouped tables below.

Treatment anchors. For each patient and task, we select the earliest treatment event satisfying the codes, history, diagnosis, and baseline criteria in Tables 17 and 18. Anchors require an earlier event, valid increasing chronology, an age within the range of the discretized age encoding, and no recorded death through the anchor. Inputs include the anchor event. Nine tasks require a qualifying baseline; thirteen use an optional baseline. Missing future targets do not trigger selection of a later anchor.

Baselines and targets. For tasks requiring a baseline, the latest compatible measurement in the specified lookback must meet its threshold; an earlier qualifying value cannot replace a newer nonqualifying one. Allopurinol/febuxostat, red-cell products, and magnesium sulfate extract adjacent code–value pairs and use an arithmetic-median baseline. Other tasks use complete laboratory tuples and the lower-median observed point. The target is the first strictly subsequent event with an eligible exact-point measurement in the outcome window. Multiple compatible values in that event are resolved by the lower-median actual record, with ties broken by test identifier, value identifier, and source occurrence. History-marked measurements and comparator or range boundaries are excluded. Samples without a target, or with death before or on the target event, are excluded; observed targets require no further follow-up.

Time and numerical labels. Elapsed time is reconstructed from event gaps and 64 sub-bins. A nominal L–H day outcome window corresponds to $[L-0.5, H+0.5)$ days after the anchor; a B-day lookback accepts elapsed time $< B + 0.5$, and minimum history H requires at least $H - 0.5$ days. Targets use the task's full numeric-bin catalog, including tail-bin representatives. Future event dates, test codes, and values are label metadata and are unavailable in the model input. Table 20 gives the outcome windows and test counts.

Cohort accounting. All eligible observed targets are retained without binary balancing, giving 110,413 training, 14,236 validation, and 13,921 test patient–task samples under the shared patient

partitions. Table 19 separates retained targets, missing targets, and death exclusions. Construction-time missing targets are excluded from the cohort; missing model predictions within that cohort use the fallback defined in Appendix F.2.

Table 17: Treatment codes and diagnostic eligibility for all 22 MS-EHR tasks. Codes match exactly; diagnoses use ICD-10-CM prefixes without dots. Parentheses give the live-diagnosis lookback, including the anchor; a dash denotes no additional requirement. The warfarin exclusion includes history-marked evidence. RA denotes rheumatoid arthritis and SLE systemic lupus erythematosus.

Treatment → outcome	Treatment codes	Diagnosis (lookback)
Blood/coagulation		
CABG → hemoglobin	CPT: 33510, 33511, 33512, 33513, 33514, 33516, 33517, 33518, 33519, 33521, 33522, 33523, 33533, 33534, 33535, 33536	–
Long-acting ESA → hemoglobin	ATC: B03XA02, B03XA03	N18 (365 d)
Radiation → hemoglobin	CPT: 77402, 77407, 77412	C (365 d)
Erythropoietin → hemoglobin	ATC: B03XA01	N18 (365 d)
Radiation → platelets	CPT: 77402, 77407, 77412	C (365 d)
Warfarin → INR	ATC: B01AA03	I48, I26, I80, I82 (365 d) Exclude Z952 throughout history.
RBC product → hemoglobin	HCPCS: P9016, P9021	–
Renal/electrolytes		
Nephrostomy → creatinine	CPT: 50432, 50433	–
Ureteroscopy → creatinine	CPT: 52352, 52353, 52356	–
Thiazide → sodium	ATC: C03AA03, C03BA04, C03BA11	–
Spironolactone → potassium	ATC: C03DA01	–
Magnesium sulfate → magnesium	ATC: B05XA05	–
Nephrectomy → creatinine	CPT: 50240, 50543, 50230, 50545	–
Hepatobiliary		
ERCP → bilirubin	CPT: 43264, 43274, 43276	K80, K831 (30 d)
Cholecystectomy → ALT	CPT: 47562, 47563, 47600, 47605	–
Terbinafine (dermatophytosis) → ALT	ATC: D01BA02	B35 (365 d)
Metabolic/endocrine		
Allopurinol/febuxostat → urate	ATC: M04AA01, M04AA03	–
SGLT2 inhibitor → HbA1c	ATC: A10BK01, A10BK03	E11 (365 d)
Parathyroid surgery → PTH	CPT: 60500, 60502, 60505	–
Inflammatory/immune		
Methotrexate (RA) → CRP	ATC: L04AX03, L01BA01	M05, M06 (365 d)
Leflunomide (RA) → ESR	ATC: L04AK01	M05, M06 (365 d)
Hydroxychloroquine (SLE) → C3	ATC: P01BA02	M32 (365 d)

Table 18: Laboratory codes and baseline eligibility for all 22 MS-EHR tasks. V counts native target classes; target units are listed. History is the minimum before the anchor; parentheses give baseline lookback days, including the anchor. Thresholds use the listed units; optional baselines do not gate inclusion. †: adjacent-pair arithmetic median; otherwise complete-tuple lower median.

Treatment → outcome	Accepted LOINCs; unit; V	History (d); baseline (lookback d)
Blood/coagulation		
CABG → hemoglobin	30313-1, 30350-3, 30351-1, 30352-9, 718-7 g/dL; $V = 40$	30; optional (365)
Long-acting ESA → hemoglobin	30313-1, 30350-3, 30351-1, 30352-9, 30354-5, 40719-7, 718-7 g/dL; $V = 40$	90; [3, 10) (14)
Radiation → hemoglobin	30313-1, 30350-3, 30351-1, 30352-9, 718-7 g/dL; $V = 40$	30; optional (365)
Erythropoietin → hemoglobin	30313-1, 30350-3, 30351-1, 30352-9, 30354-5, 40719-7, 718-7 g/dL; $V = 40$	90; [3, 10) (14)
Radiation → platelets	26515-7, 74464-9, 777-3, 778-1 $10^3/\mu$L; $V = 40$	30; optional (365)
Warfarin → INR	34714-6, 38875-1, 46418-0, 6301-6 INR; $V = 15$	90; [0.5, 10) (14)
RBC product → hemoglobin	30313-1, 30350-3, 30351-1, 30352-9, 30354-5, 40719-7, 718-7 g/dL; $V = 40$	30; [3, 8) (3)†
Renal/electrolytes		
Nephrostomy → creatinine	2160-0, 38483-4 mg/dL; $V = 31$	30; optional (365)
Ureteroscopy → creatinine	2160-0, 38483-4 mg/dL; $V = 31$	30; optional (365)
Thiazide → sodium	2947-0, 2951-2, 32717-1, 39791-9, 39792-7, 77139-4 mmol/L; $V = 32$	90; [135, 145) (14)
Spironolactone → potassium	2823-3, 6298-4, 77142-8 mmol/L; $V = 32$	90; [3.5, 5.5) (7)
Magnesium sulfate → magnesium	19123-9, 21377-7 mg/dL; $V = 24$	30; [0.5, 1.7) (3)†
Nephrectomy → creatinine	2160-0, 38483-4 mg/dL; $V = 31$	30; optional (365)
Hepatobiliary		
ERCP → bilirubin	1975-2, 42719-5 mg/dL; $V = 40$	30; optional (365)
Cholecystectomy → ALT	1742-6, 1743-4, 1744-2, 76625-3, 77144-4 U/L; $V = 35$	30; optional (365)
Terbinafine (dermatophytosis) → ALT	1742-6, 1743-4, 1744-2, 76625-3, 77144-4 U/L; $V = 35$	90; optional (365)
Metabolic/endocrine		
Allopurinol/febuxostat → urate	3084-1, 98981-4 mg/dL; $V = 32$	90; $\geq$6 (30)†
SGLT2 inhibitor → HbA1c	17855-8, 17856-6, 4548-4 %; $V = 32$	365; $\geq$7.5 (90)
Parathyroid surgery → PTH	2731-8 pg/mL; $V = 24$	30; optional (365)
Inflammatory/immune		
Methotrexate (RA) → CRP	1988-5, 105126-7, 48421-2 mg/L; $V = 32$	90; optional (365)
Leflunomide (RA) → ESR	30341-2, 4537-7 mm/h; $V = 24$	90; optional (365)
Hydroxychloroquine (SLE) → C3	4485-9 mg/dL; $V = 24$	90; optional (365)

Table 19: MS-EHR cohort accounting for all 22 treatment tasks. The 391,129 eligible patient–task anchors comprise 138,570 observed targets across the three splits, 250,827 missing targets and 1,732 death exclusions. Missing and death columns sum across splits; they describe cohort construction, not model-prediction fallback.

Treatment → outcome	Train	Valid.	Test	Missing	Death
Blood/coagulation					
CABG → hemoglobin	1,318	151	181	3,609	48
Long-acting ESA → hemoglobin	2,231	295	337	2,293	171
Radiation → hemoglobin	3,177	416	416	5,225	136
Erythropoietin → hemoglobin	3,016	394	418	2,984	193
Radiation → platelets	3,405	430	422	5,073	40
Warfarin → INR	10,695	1,393	1,338	7,744	90
RBC product → hemoglobin	17,243	2,189	2,181	9,756	0
Renal/electrolytes					
Nephrostomy → creatinine	957	120	112	1,308	24
Ureteroscopy → creatinine	1,323	150	179	7,874	9
Thiazide → sodium	14,128	1,783	1,695	69,804	126
Spironolactone → potassium	14,389	1,868	1,776	26,687	339
Magnesium sulfate → magnesium	24,967	3,302	3,137	38,309	0
Nephrectomy → creatinine	742	90	102	1,161	14
Hepatobiliary					
ERCP → bilirubin	1,127	150	173	2,066	14
Cholecystectomy → ALT	2,096	257	262	20,179	20
Terbinafine (dermatophytosis) → ALT	1,974	271	246	10,623	5
Metabolic/endocrine					
Allopurinol/febuxostat → urate	2,784	344	331	9,768	407
SGLT2 inhibitor → HbA1c	2,960	387	365	15,493	59
Parathyroid surgery → PTH	410	61	50	1,315	0
Inflammatory/immune					
Methotrexate (RA) → CRP	692	92	78	5,118	14
Leflunomide (RA) → ESR	382	51	71	1,462	11
Hydroxychloroquine (SLE) → C3	397	42	51	2,976	12
Total	110,413	14,236	13,921	250,827	1,732

F.2 Implementation

HealthSAW and EHR-FM roll out from their treatment-simulation checkpoints, aligned to MS-EHR as described in Appendix B.2.

Natural rollout. In patient simulation, the model generates actions and observations alike; we call this a *natural rollout*. HealthSAW retains a whole-event history within a 3,072-source-token budget, preserves the anchor's action block, and generates at most 1,024 future events with at most 127 tokens per block. EHR-FM retains at most 2,048 history tokens within a 3,072-token total context. Both use the rollout sampler of Appendix G.1, temperature 1 without top-k or nucleus filtering, and stop at the window end, death or resource limit, rather than the first numerical hit. A nominal day window $[a, b]$ is evaluated as $[a - 0.5, b + 0.5)$ to accommodate fractional generated times.

Numerical readout. Both models use the same generated-event matching rule. The scorer selects the first in-window event containing an accepted LOINC and a value from its compatible group; adjacency is not required. A candidate event containing history markers, comparators or conflicting tests invalidates the trajectory rather than triggering selection of a later event. Death is checked before accepting a measurement. Among multiple candidates, the lower-median sampled value selects the query position. Replay uses the observed prefix and generated continuation, with no actual future measurement supplied. The causal logits immediately before the selected value are normalized over the task's legal value group. The trajectory prediction is the lower posterior median: the first numeric-bin representative whose cumulative probability reaches 0.5.

Aggregation and evaluation. For each patient, we take the lower empirical median across valid trajectory predictions from the same observed history. This gives a central estimate that is less sensitive to extreme trajectory-level predictions than the mean. Budgets $K \in \{1, 5, 10, 20\}$ use the first K saved trajectories, forming nested subsets.

When none of these trajectories yields a valid value, the prediction is the task's empirical training-target median. Missing individual trajectories are not filled before aggregation. Every method and budget is scored on all 13,921 test patient–task samples. For task j with test set T_j, target y_i, and prediction $\hat{y}_i^{(K)}$ after fallback,

$$\mathrm{MAE}_j(K) = \frac{1}{|T_j|} \sum_{i \in T_j} |\hat{y}_i^{(K)} - y_i|, \qquad \text{Macro WAPE}(K) = \frac{100}{22} \sum_{j=1}^{22} \frac{\mathrm{MAE}_j(K)}{|T_j|^{-1} \sum_{i \in T_j} |y_i|}.$$

Group macro WAPE applies the same formula to the tasks in that group; the overall mean weights all 22 tasks equally. MAE uses each task's native unit. Simulation coverage is the pooled fraction of test samples with at least one valid value before fallback.

Per-task EHR-FM SFT. A linear $1536 \to 1$ head reads EHR-FM's last token state and uses L1 loss and the MS-EHR SFT optimization settings in Appendix D.2, with microbatch and evaluation batch size 8. Targets are numeric-bin representatives, centered at the training-target median and scaled by their mean absolute deviation from it. Zero initialization makes the head predict that median. Predictions are continuous scalars in native units. Validation MAE is checked at initialization and three times per epoch; the lowest value selects the checkpoint, with ties retaining the earlier one.

F.3 Results

Table 3 gives group macro WAPE and simulation coverage for the main comparison. Table 20 gives all 22 tasks' per-task MAE for these methods and the constant baseline, organized into the same five outcome groups as the main results. Figure 8 shows coverage and macro WAPE across trajectory budgets; Table 21 gives the corresponding values under the full-test scoring protocol above. A constant predictor using each task's training-target median yields macro WAPE of 34.15%.

Table 20: Full-test post-treatment MAE on all 22 MS-EHR tasks, grouped as in Table 3. The task label gives the absolute-error unit and outcome window; pp denotes percentage points for HbA1c. Constant predicts the task training-target median. EHR-FM SFT uses a scalar L1 head; both simulation columns (EHR-FM and HealthSAW) use 20 trajectories with training-median fallback. Lowest MAE per task is bold. RA: rheumatoid arthritis; SLE: systemic lupus erythematosus; PTH: parathyroid hormone; CRP: C-reactive protein; ESR: erythrocyte sedimentation rate; C3: complement component 3.

Treatment → outcome	Test N	Constant	EHR-FM SFT (L1)	EHR-FM simulation	HealthSAW simulation
Blood/coagulation					
CABG → hemoglobin g/dL; 21–60 d	181	1.4971	1.4316	1.6551	**1.1514**
Long-acting ESA → hemoglobin g/dL; 30–60 d	337	1.1778	**1.1523**	1.2672	1.2346
Radiation → hemoglobin g/dL; 30–60 d	416	1.6601	1.1732	1.2626	**1.0211**
Erythropoietin → hemoglobin g/dL; 30–60 d	418	1.2654	**1.1614**	1.3319	1.1865
Radiation → platelets $10^3/\mu$L; 14–42 d	422	84.5694	61.7474	66.8892	**47.9281**
Warfarin → INR INR; 14–42 d	1,338	**0.7321**	**0.7321**	0.7635	0.7798
RBC product → hemoglobin g/dL; 3–7 d	2,181	0.7950	**0.7186**	0.8661	0.7841

Continued on next page

Table 21: Treatment simulation across trajectory budgets on all 22 tasks for the HealthSAW and EHR-FM rollouts. Simulation coverage is the share of samples with any valid value among the first K trajectories, before fallback. Macro WAPE uses all 13,921 test patient–task samples with training-median fallback. Best values per row are bold.

	Coverage (%) ↑		Macro WAPE (%) ↓	
Trajectories K	HealthSAW	EHR-FM	HealthSAW	EHR-FM
1	**57.52**	38.82	**31.7114**	33.2872
5	**89.66**	70.96	**30.2999**	32.7330
10	**95.80**	80.15	**27.9880**	32.2798
20	**98.48**	86.27	**27.2809**	30.7390

Continued from previous page

Treatment → outcome	Test N	Constant	EHR-FM SFT (L1)	EHR-FM simulation	HealthSAW simulation
Renal/electrolytes					
Nephrostomy → creatinine mg/dL; 14–42 d	112	0.7286	0.7119	**0.5873**	0.5991
Ureteroscopy → creatinine mg/dL; 14–42 d	179	0.4086	0.3902	0.2562	**0.1982**
Thiazide → sodium mmol/L; 14–42 d	1,695	2.9621	**2.5353**	2.8586	2.7774
Spironolactone → potassium mmol/L; 14–42 d	1,776	0.4569	**0.4130**	0.4352	0.4169
Magnesium sulfate → magnesium mg/dL; 3–7 d	3,137	0.2152	**0.1979**	0.2172	0.2049
Nephrectomy → creatinine mg/dL; 60–180 d	102	0.5323	0.5219	0.4365	**0.2978**
Hepatobiliary					
ERCP → bilirubin mg/dL; 14–42 d	173	1.8049	1.7356	1.8424	**1.6354**
Cholecystectomy → ALT U/L; 14–42 d	262	33.8661	33.5412	36.1353	**31.4417**
Terbinafine (dermatophytosis) → ALT U/L; 42–90 d	246	12.7975	12.1722	9.6321	**8.9085**
Metabolic/endocrine					
Allopurinol/febuxostat → urate mg/dL; 30–90 d	331	1.5918	1.4037	1.5399	**1.4017**
SGLT2 inhibitor → HbA1c pp; 56–90 d	365	1.0131	**0.9331**	0.9525	0.9996
Parathyroid surgery → PTH pg/mL; 14–90 d	50	87.6517	87.6900	72.0438	**65.4747**
Inflammatory/immune					
Methotrexate (RA) → CRP mg/L; 56–90 d	78	10.0141	9.9410	9.3779	**7.6441**
Leflunomide (RA) → ESR mm/h; 28–84 d	71	21.6207	21.5850	20.2028	**17.8445**
Hydroxychloroquine (SLE) → C3 mg/dL; 28–84 d	51	29.1293	28.9749	**18.3143**	25.2253
Total	13,921	–	–	–	–

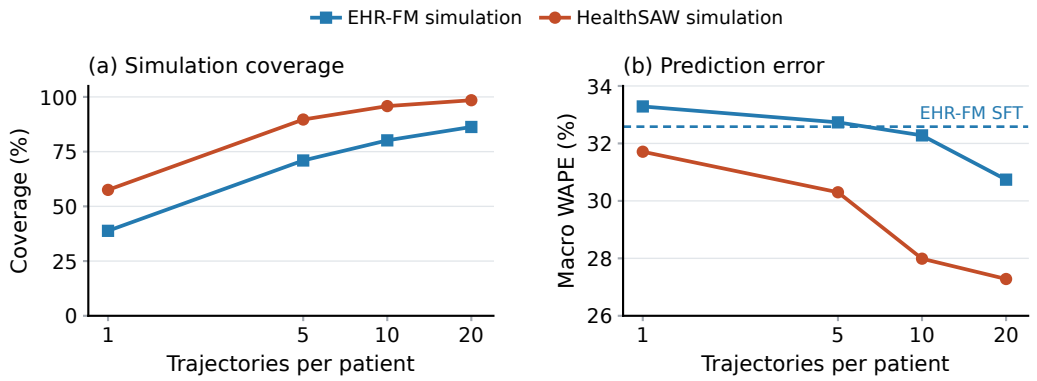


Figure 8: Treatment simulation across trajectory budgets on all 22 tasks. (a) Simulation coverage: the share of samples with at least one valid readout among the first K trajectories, before fallback. (b) Macro WAPE on all test samples, including training-median fallback; the dashed line is per-task EHR-FM SFT.

G Disease Simulation

The same simulation serves the 16 disease tasks of §4.2, with the generated future spanning two to four years rather than days to months. We evaluate on the same cohorts and outcome windows as per-task SFT, reading risk from the generated diagnoses. The main text reports the speed of this simulation (§4.4); this appendix gives its setup and accuracy at the original resolution, a temporally merged variant, and the inference cost of both.

G.1 Setup

Risk is the fraction of all requested trajectories containing a target diagnosis in the outcome window; incomplete and censored trajectories remain in the denominator. AUPRC is average precision. PPV and sensitivity use a tie-aware test ROC threshold with achieved specificity of at least 99%. Every reported rollout setting uses the same per-task trajectory count $K \in \{20, 50\}$ for prediction and timing, as listed in Table 26.

Generation budgets. EHR-FM uses a history prefix of at most 2,048 tokens within a total context of 3,072, without rolling the context. HealthSAW uses observation/action limits of 26/3 tokens for $K = 20$ tasks and 64/16 for $K = 50$. Its generation limits are 730 days and 256 future events for cancer and cardiovascular tasks, or 1,460 days and 512 events for neurological tasks.

Checkpoints. Both models roll out from their disease-simulation checkpoints, aligned to MS-EHR as described in Appendix B.2.

Sampling and stopping. All rollout samplers use temperature 1 without top-k or nucleus filtering. Natural HealthSAW samples coarse time and one of 64 conditional sub-bins, using the selected sub-bin midpoint. EHR-FM uses coarse-bin midpoints, with 1,461 days for the final open bin. HealthSAW stops at the horizon, recognized live death or event cap. EHR-FM also stops at an in-window target hit or patient-end token.

G.2 Results at the Original Resolution

At the original resolution, HealthSAW outperforms EHR-FM on every macro metric (Table 22), with 34.9% higher macro AUPRC, higher AUPRC on 13 of 16 tasks and the same ordering within every disease group. Appendix G.4 reports the corresponding inference costs.

Table 22: Disease risk prediction from original-resolution rollout on the 16 MS-EHR tasks, organized by disease group as in Table 1. Risk is the share of all requested trajectories, including incomplete ones, that contain the target diagnosis. Task-macro averages on the 0–1 scale; PPV and sensitivity at specificity $\geq 99\%$; the largest value per block and metric is bold. Group blocks average per-task values reported to four decimals. AUROC: Appendix G.2.

	MS-EHR		
Model	AUPRC	PPV	Sens.
All 16 tasks			
EHR-FM	0.0538	0.1377	0.0639
HealthSAW	**0.0726**	**0.1816**	**0.0965**
Cancer (7 tasks)			
EHR-FM	0.0383	0.0998	0.0830
HealthSAW	**0.0453**	**0.1298**	**0.1156**
Cardiovascular (5 tasks)			
EHR-FM	0.1032	0.2313	0.0569
HealthSAW	**0.1376**	**0.3042**	**0.0823**
Neurodegenerative (4 tasks)			
EHR-FM	0.0195	0.0872	0.0393
HealthSAW	**0.0387**	**0.1192**	**0.0808**

Table 24 reports all four metrics for the 16 MS-EHR tasks, including the merged settings of Appendix G.3. The macro AUROC omitted from Table 22 is 0.6066 for EHR-FM and 0.6371 for HealthSAW.

G.3 Temporal Merging

Procedure. Simulation at encounter resolution generates every intermediate visit over a multi-year horizon. Temporal merging instead divides the future into fixed windows of $M = 30$ or 90 days, right-closed and split at the washout boundary W and the horizon H so that no window straddles either; for $W = 180$, $H = 730$ and $M = 90$ the endpoints are 90, 180, 270, …, 720 and 730 days. All source events in a window form one merged event with the structure of an ordinary event: an elapsed-time slot that advances to the window endpoint, one observation block and one action block. The observed history stays at its original resolution. Merged blocks keep every diagnosis unit and complete code–value units; other units are sampled to a nominal 127-token budget per block and restored to clinical order. Both merged variants apply the alignment epoch to merged future targets with prefix losses masked, keeping the next-block objective: *shared* alignment draws one eligible anchor per training patient, with at least two prefix events and follow-up through a sampled task horizon, and yields one checkpoint per window width, the default setting; *per-task* alignment instead trains on the endpoint's own retained training anchors, whose merged futures end at that endpoint's horizon, giving one checkpoint per endpoint and window width (32 checkpoints for the 16 tasks) under the same one-epoch schedule. At inference the model generates one observation and one action block per window, nine windows for the two-year tasks and 18 for the four-year neurodegenerative tasks at 90 days (25 and 50 at 30 days), and risk is read from the generated windows exactly as at the original resolution.

Generation. Merged rollouts allow 127 tokens per observation and action block. For cancer and cardiovascular tasks, 30- and 90-day merging use 25 and 9 scheduled events, respectively; neurological tasks use 50 and 18. Schedules align separately to the washout boundary and outcome-window end, and elapsed time follows the fixed window endpoints.

Results. With shared alignment, the macro AUPRC of 30- and 90-day merging, 0.0723 and 0.0776, remains close to 0.0726 at the original resolution. Per-task alignment, one checkpoint per endpoint and window width, raises macro AUPRC further to 0.1001 (30 days) and 0.1110 (90 days) and sensitivity to 0.1187 and 0.1253. Table 23 lists all macro metrics for both variants.

Table 23: Predictive quality under temporal merging on 16 MS-EHR disease tasks. Metrics are task-macro averages; PPV and sensitivity use a specificity of at least 99%. Per-task rows use one checkpoint per endpoint. Best values within each merging group are bold.

HealthSAW	AUROC	AUPRC	PPV	Sens.
Original resolution	0.6371	0.0726	0.1816	0.0965
Shared alignment (random anchors)				
30 days	**0.6649**	0.0723	0.1528	0.0877
90 days	0.6582	**0.0776**	**0.1726**	**0.0897**
Per-task alignment				
30 days	0.7374	0.1001	0.1801	0.1187
90 days	**0.7402**	**0.1110**	**0.2030**	**0.1253**

Table 24: Per-task disease simulation on MS-EHR across the six rollout settings, on the 0–1 scale; PPV and sensitivity at specificity $\geq 99\%$; merge widths in days. Trajectory counts match within each task (Appendix G.1); values keep source precision.

Setting	AUROC	AUPRC	PPV	Sens.
Bladder cancer				
EHR-FM	0.5710	0.0280	0.08197	0.06154
HealthSAW	0.5930	0.0380	0.16410	0.09846
HealthSAW, merge 30 (shared)	0.5760	0.0160	0.06731	0.04308
HealthSAW, merge 90 (shared)	0.6010	0.0220	0.05725	0.09231
HealthSAW, merge 30 (per task)	0.6800	0.0450	0.08889	0.11077
HealthSAW, merge 90 (per task)	0.6920	0.0630	0.08436	0.16923
Breast cancer				
EHR-FM	0.5950	0.0650	0.26129	0.05728
HealthSAW	0.6080	0.0750	0.22878	0.08769
HealthSAW, merge 30 (shared)	0.5960	0.0510	0.15242	0.05799
HealthSAW, merge 90 (shared)	0.6080	0.0650	0.17025	0.07284
HealthSAW, merge 30 (per task)	0.6560	0.0900	0.25000	0.10113
HealthSAW, merge 90 (per task)	0.6610	0.0970	0.23052	0.10042
Colorectal cancer				
EHR-FM	0.5850	0.0340	0.06355	0.07819
HealthSAW	0.5840	0.0460	0.12113	0.08848
HealthSAW, merge 30 (shared)	0.6120	0.0380	0.06145	0.06790
HealthSAW, merge 90 (shared)	0.5920	0.0480	0.07400	0.07613
HealthSAW, merge 30 (per task)	0.6840	0.0500	0.07182	0.08025
HealthSAW, merge 90 (per task)	0.7070	0.0620	0.08117	0.09671
Kidney cancer				
EHR-FM	0.6030	0.0290	0.08468	0.07368
HealthSAW	0.6250	0.0420	0.12281	0.09825
HealthSAW, merge 30 (shared)	0.6290	0.0180	0.04711	0.07719
HealthSAW, merge 90 (shared)	0.6000	0.0230	0.04742	0.08070
HealthSAW, merge 30 (per task)	0.6990	0.0340	0.05373	0.10877
HealthSAW, merge 90 (per task)	0.6940	0.0410	0.05788	0.12632
Liver cancer				
EHR-FM	0.6800	0.0770	0.11696	0.19900
HealthSAW	0.7010	0.0720	0.10953	0.26866
HealthSAW, merge 30 (shared)	0.7430	0.0910	0.10185	0.32836
HealthSAW, merge 90 (shared)	0.7260	0.0810	0.15823	0.24876
HealthSAW, merge 30 (per task)	0.7860	0.1180	0.12977	0.33831
HealthSAW, merge 90 (per task)	0.8090	0.1300	0.13663	0.34328

Continued on next page

Continued from previous page

Setting	AUROC	AUPRC	PPV	Sens.
Lung cancer				
EHR-FM	0.5990	0.0250	0.06397	0.05184
HealthSAW	0.6140	0.0330	0.13235	0.09823
HealthSAW, merge 30 (shared)	0.6480	0.0270	0.11515	0.05184
HealthSAW, merge 90 (shared)	0.6270	0.0270	0.10526	0.06003
HealthSAW, merge 30 (per task)	0.7320	0.0590	0.15266	0.12142
HealthSAW, merge 90 (per task)	0.7310	0.0640	0.21040	0.12142
Pancreatic cancer				
EHR-FM	0.5650	0.0100	0.02632	0.05970
HealthSAW	0.5760	0.0110	0.02979	0.06965
HealthSAW, merge 30 (shared)	0.5880	0.0090	0.02321	0.05473
HealthSAW, merge 90 (shared)	0.5570	0.0100	0.01080	0.03980
HealthSAW, merge 30 (per task)	0.6160	0.0090	0.01849	0.06468
HealthSAW, merge 90 (per task)	0.6370	0.0160	0.02544	0.07960
Acute coronary syndrome				
EHR-FM	0.6300	0.0660	0.15233	0.04482
HealthSAW	0.5790	0.0580	0.17620	0.05269
HealthSAW, merge 30 (shared)	0.6570	0.0790	0.20759	0.03182
HealthSAW, merge 90 (shared)	0.6460	0.0780	0.24368	0.03626
HealthSAW, merge 30 (per task)	0.7770	0.1360	0.26093	0.06534
HealthSAW, merge 90 (per task)	0.7760	0.1460	0.31941	0.08108
Coronary artery disease				
EHR-FM	0.7070	0.2290	0.48647	0.08310
HealthSAW	0.7510	0.3070	0.61001	0.13370
HealthSAW, merge 30 (shared)	0.7770	0.3010	0.55946	0.11560
HealthSAW, merge 90 (shared)	0.7770	0.3110	0.57187	0.12247
HealthSAW, merge 30 (per task)	0.8420	0.3700	0.58043	0.14110
HealthSAW, merge 90 (per task)	0.8500	0.3890	0.64921	0.12446
Subarachnoid haemorrhage				
EHR-FM	0.5810	0.0100	0.02356	0.02903
HealthSAW	0.5440	0.0050	0.01765	0.02903
HealthSAW, merge 30 (shared)	0.5660	0.0050	0.00943	0.02258
HealthSAW, merge 90 (shared)	0.5660	0.0060	0.03419	0.02581
HealthSAW, merge 30 (per task)	0.6650	0.0080	0.01544	0.02903
HealthSAW, merge 90 (per task)	0.6580	0.0090	0.02665	0.05484
Heart failure				
EHR-FM	0.6900	0.1520	0.36268	0.09332
HealthSAW	0.7430	0.2570	0.55451	0.16014
HealthSAW, merge 30 (shared)	0.8010	0.2750	0.52965	0.14972
HealthSAW, merge 90 (shared)	0.7990	0.2970	0.56784	0.16399
HealthSAW, merge 30 (per task)	0.8590	0.3340	0.53734	0.18414
HealthSAW, merge 90 (per task)	0.8590	0.3640	0.59430	0.19841
Stroke				
EHR-FM	0.6140	0.0590	0.13136	0.03435
HealthSAW	0.6000	0.0610	0.16246	0.03590
HealthSAW, merge 30 (shared)	0.6470	0.0690	0.14191	0.03312
HealthSAW, merge 90 (shared)	0.6460	0.0710	0.17478	0.05447
HealthSAW, merge 30 (per task)	0.7400	0.0990	0.18648	0.04612
HealthSAW, merge 90 (per task)	0.7460	0.1120	0.23701	0.05788
Alzheimer's disease				
EHR-FM	0.5700	0.0130	0.07965	0.02663
HealthSAW	0.6670	0.0270	0.06912	0.08876
HealthSAW, merge 30 (shared)	0.7200	0.0350	0.08732	0.09172
HealthSAW, merge 90 (shared)	0.6990	0.0350	0.09021	0.10355
HealthSAW, merge 30 (per task)	0.7850	0.0540	0.11066	0.16272
HealthSAW, merge 90 (per task)	0.7760	0.0580	0.11978	0.12722

Continued on next page

Continued from previous page

Setting	AUROC	AUPRC	PPV	Sens.
Any dementia				
EHR-FM	0.5910	0.0400	0.16185	0.04014
HealthSAW	0.7070	0.0890	0.24302	0.10609
HealthSAW, merge 30 (shared)	0.7570	0.1010	0.22222	0.12616
HealthSAW, merge 90 (shared)	0.7660	0.1250	0.31753	0.11039
HealthSAW, merge 30 (per task)	0.8160	0.1430	0.28723	0.11613
HealthSAW, merge 90 (per task)	0.8080	0.1620	0.33846	0.15771
Parkinson's disease				
EHR-FM	0.5750	0.0150	0.04119	0.06102
HealthSAW	0.6420	0.0190	0.08295	0.06102
HealthSAW, merge 30 (shared)	0.6540	0.0220	0.05128	0.06780
HealthSAW, merge 90 (shared)	0.6500	0.0200	0.07772	0.05085
HealthSAW, merge 30 (per task)	0.7270	0.0280	0.07756	0.09492
HealthSAW, merge 90 (per task)	0.7060	0.0250	0.06498	0.06102
Vascular dementia				
EHR-FM	0.5500	0.0100	0.06604	0.02941
HealthSAW	0.6610	0.0200	0.08163	0.06723
HealthSAW, merge 30 (shared)	0.6680	0.0210	0.06689	0.08403
HealthSAW, merge 90 (shared)	0.6720	0.0230	0.06085	0.09664
HealthSAW, merge 30 (per task)	0.7350	0.0270	0.05937	0.13445
HealthSAW, merge 90 (per task)	0.7320	0.0360	0.07123	0.10504

G.4 Inference Cost

Protocol. Each model–task case runs on one NVIDIA B200 with PyTorch 2.11.0 and CUDA 12.8. Natural-rollout models are paired on the same physical GPU; merging uses the same node and GPU type in a separate allocation. Each task reuses 128 test patients and anchors: 32 for warm-up and 96 for timing, with two timed repetitions.

Timing includes synchronized prefix preparation/prefill, allocation, generation, and CPU output processing and scoring. It excludes checkpoint loading and transfer, history scans, file writes, compilation, warm-up, and CUDA-graph setup. Seconds per patient include all requested trajectories. Reported mean runtimes weight tasks equally, and speedup is the ratio of those means. Per-task times are in Table 26.

Results. HealthSAW improves AUPRC from EHR-FM's 0.0538 to 0.0726 while reducing mean inference time from 5.700 to 1.781 s per patient, a $3.20\times$ speedup (§4.4; Table 25), and is faster on all 16 tasks. Temporal merging reduces runtime to 1.064 and 0.387 s per patient at 30 and 90 days, respectively; the 90-day setting is $14.71\times$ faster than EHR-FM and approximately $4.6\times$ faster than natural HealthSAW rollout, and per-task alignment runs at essentially the same cost.

Table 25: Disease-simulation quality and inference cost on the 16 MS-EHR tasks of Appendix G. Original rows roll out at encounter resolution; the 30- and 90-day rows roll out merged trajectories at that coarser resolution. AUPRC and runtime are task-macro averages; runtime covers all requested trajectories per patient on one B200 after warm-up, with trajectory counts matched per task. Speedups are ratios of task-mean times to EHR-FM.

Model	Resolution	AUPRC	s/patient	Speedup
Shared alignment: one checkpoint per resolution				
EHR-FM	Original	0.0538	5.700	1.00×
HealthSAW	Original	0.0726	1.781	3.20×
HealthSAW	30 days	0.0723	1.064	5.36×
HealthSAW	90 days	0.0776	0.387	14.71×
Per-task alignment: one checkpoint per endpoint and window				
HealthSAW	30 days	0.1001	1.062	5.37×
HealthSAW	90 days	0.1110	0.387	14.74×

Table 26: Per-task simulation runtime on MS-EHR in seconds per patient on one B200 after warm-up, including all requested trajectories. K is the trajectory count shared by all six settings; original-resolution rows use the shared alignment, merged rows shared or per-task alignment (Appendices B.2 and G.3). Means weight tasks equally. SAH: subarachnoid haemorrhage.

		Original resolution		Shared alignment		Per-task alignment	
Task	K	EHR-FM	HealthSAW	30 days	90 days	30 days	90 days
Bladder cancer	20	4.722	0.535	0.575	0.212	0.574	0.211
Breast cancer	20	4.761	0.531	0.578	0.212	0.574	0.213
Colorectal cancer	50	6.195	2.111	1.073	0.392	1.072	0.392
Kidney cancer	50	6.231	2.127	1.070	0.391	1.070	0.391
Liver cancer	50	6.573	2.185	1.080	0.393	1.069	0.393
Lung cancer	20	4.593	0.481	0.579	0.212	0.574	0.212
Pancreatic cancer	50	6.308	2.164	1.081	0.391	1.070	0.389
ACS	20	4.822	0.520	0.578	0.213	0.574	0.210
CAD	20	4.466	0.542	0.583	0.212	0.583	0.210
SAH	50	5.743	1.933	1.071	0.390	1.077	0.390
Heart failure	20	4.934	0.622	0.577	0.212	0.580	0.213
Stroke	20	4.228	0.519	0.584	0.212	0.585	0.212
Alzheimer's disease	50	7.460	4.557	2.141	0.780	2.142	0.782
Any dementia	20	5.522	1.056	1.147	0.420	1.154	0.415
Parkinson's disease	50	7.357	4.319	2.148	0.780	2.147	0.779
Vascular dementia	50	7.277	4.293	2.156	0.777	2.146	0.777
Mean	–	5.700	1.781	1.064	0.387	1.062	0.387

H Ablation Study

H.1 Configurations

All ablation classifiers use the per-dataset SFT settings in Appendix D.2; all comparisons use the main HealthSAW checkpoint as their reference.

Randomized block content. The random-blocks variant concatenates each event's observation and action clinical tokens, including value-bin tokens, applies a fixed permutation and splits the result at the original observation length, so both blocks draw on the same content without state–action meaning. The elapsed-time slot and demographics are not permuted, and each event's permutation is fixed across epochs. Both blocks use the union of the two vocabularies with unshared embeddings and output heads; widths, depths, corpus, schedule and 132,164 steps match HealthSAW.

Retained event content. HealthSAW and the main EHR-FM baseline start from their pretrained checkpoints at step 132,164 and are fine-tuned on SS-EHR for four epochs with learning rate 3×10^{-5},

global batch 128 and a 3,072-token context, selecting by validation AUPRC, at 100%, 75%, 50% and 25% of each event's code–value units. A unit is one clinical code together with its value bin when present, so an analyte and its bin are never separated; SS-EHR events carry 13.93 units on average against 7.79 in MS-EHR. Selection is deterministic, identical in every epoch and for both models, nested across levels, and applied alike to observation and action blocks. Realized retention is within 0.2% of target, and the number of events per window and of optimizer steps per epoch (1,613) is unchanged, so retention changes content but not training geometry.

Direct history access. The raw-history variant keeps the global Transformer output as decoder input and additionally lets the decoder attend causally to the raw tokens of all earlier blocks within the same patient window. Both models are pretrained for four epochs on CLAIM and the MS-EHR training partition; the raw-history model uses its final checkpoint at step 132,164, and disease SFT uses a 3,072-source-token context and global batch size 128.

Pretraining NLL normalization. Figure 6 plots every logged point; final losses average the last 20 common points, at steps 131,200–132,150, each itself an average of 50 optimizer steps. We convert the logged losses to estimated nats per original clinical token. Let E and P denote corpus event and patient counts, O and A the observation and action token counts, and $R = E + O + A$ the original token count. For HealthSAW,

$$\widehat{\ell}_{\text{local}} = \frac{(O+E)\ell_{\text{obs}} + (A+E)\ell_{\text{act}} + (E-P)\ell_{\Delta t}}{R}.$$

Each observation and action slot contributes an EOS target, including empty slots; elapsed time is not predicted across patient boundaries. The auxiliary sub-bin loss is excluded from this reported NLL, while remaining part of training. The corpus weights are 0.513, 0.616 and 0.129, respectively. The raw-history decoder uses demographic tokens but masks patient-start predictions, giving $\widehat{\ell}_{\text{global}} = (1 + P/R)\ell_{\text{global}}$, a factor of 1.0007. These corpus-level weights approximate per-batch target counts.

Two-stage training. Two-stage alternatives first train separate observation/action encoder–decoder tokenizers on block reconstruction, then freeze the encoder mappings for autoregressive latent modeling. The continuous variant uses a 32-dimensional bottleneck and two tokenizer epochs, followed by four latent-model epochs with next-block MSE and time prediction; its local decoders train on detached tokenizer latents. VQ uses separate 4,096-entry codebooks with 32-dimensional vectors and reconstruction, codebook, and commitment losses for four epochs. Its four-epoch autoregressive model predicts code indices with cross-entropy and predicts time, reusing the tokenizer decoders. HealthSAW trains for four epochs with its end-to-end next-block objective.

H.2 Complete Results

Randomized block content. Table 27 gives the per-task values behind the random-blocks comparison of §4.4.

Retained event content. Tables 28 and 29 give the per-task values and absolute macro results behind Figure 5.

Direct history access. Table 30 gives per-task AUPRC, PPV, and sensitivity for the raw-history decoder, complementing the macro comparison in Figure 6.

Two-stage training. Table 31 adds AUROC and PPV to the two-stage comparison in Table 4.

Table 27: Disease prediction with randomized block content. The per-task rows report the random-blocks variant; the final rows compare its 16-task macro averages with the main HealthSAW model. PPV and sensitivity use 99% specificity. Main HealthSAW per-task results appear in Table 13.

	MS-EHR			SS-EHR		
Task	AUPRC	PPV	Sens.	AUPRC	PPV	Sens.
Bladder cancer	0.1296	0.1247	0.2708	0.3315	0.2062	0.3706
Breast cancer	0.1584	0.2558	0.1627	0.3518	0.4732	0.3265
Colorectal cancer	0.0964	0.1237	0.1708	0.3169	0.4052	0.3003
Kidney cancer	0.0833	0.0956	0.2351	0.4238	0.1667	0.5534
Liver cancer	0.1853	0.1322	0.4428	0.5351	0.5357	0.5176
Lung cancer	0.1615	0.1994	0.2510	0.3924	0.4253	0.4038
Pancreatic cancer	0.0228	0.0290	0.1095	0.2415	0.1260	0.3692
Acute coronary syndrome	0.2854	0.4255	0.1943	0.5688	0.6986	0.4460
Coronary artery disease	0.4868	0.6776	0.2130	0.5482	0.7725	0.2550
Subarachnoid haemorrhage	0.0283	0.0560	0.1613	0.4173	0.1553	0.4556
Heart failure	0.4837	0.6434	0.2958	0.5949	0.7429	0.3657
Stroke	0.2542	0.3859	0.1637	0.5862	0.7495	0.3731
Alzheimer's disease	0.0888	0.1335	0.2041	0.2529	0.2598	0.3303
Any dementia	0.2660	0.3713	0.2638	0.3892	0.5195	0.2581
Parkinson's disease	0.0536	0.0860	0.1458	0.1908	0.2233	0.2460
Vascular dementia	0.0596	0.0916	0.2479	0.3132	0.2114	0.3776
Macro over 16 tasks						
Random blocks	0.1777	0.2394	0.2208	0.4034	0.4169	0.3718
HealthSAW	**0.1827**	**0.2423**	**0.2255**	**0.4963**	**0.4580**	**0.4554**

Table 28: Per-task AUPRC on SS-EHR at each retention level of clinical code–value units (Figure 5). The larger value per task and level is bold.

	100%		75%		50%		25%	
Task	HealthSAW	EHR-FM	HealthSAW	EHR-FM	HealthSAW	EHR-FM	HealthSAW	EHR-FM
Bladder cancer	**0.4769**	0.2548	**0.2663**	0.1475	**0.1144**	0.1117	0.0379	**0.0640**
Breast cancer	**0.4264**	0.2965	**0.2305**	0.1546	**0.1017**	0.0957	**0.0653**	0.0565
Colorectal cancer	**0.4498**	0.2938	**0.2361**	0.1874	**0.1295**	0.1039	**0.0811**	0.0732
Kidney cancer	**0.4870**	0.2961	**0.3340**	0.1409	**0.2390**	0.0539	**0.0606**	0.0371
Liver cancer	**0.5788**	0.4508	**0.4169**	0.3544	**0.3035**	0.2830	**0.2618**	0.2555
Lung cancer	**0.4982**	0.3286	**0.2721**	0.2100	**0.1525**	0.1204	**0.0799**	0.0736
Pancreatic cancer	**0.3744**	0.2021	**0.2276**	0.0218	**0.1635**	0.0636	0.0359	**0.0509**
Acute coronary syndrome	**0.6358**	0.5027	**0.4316**	0.3830	**0.2946**	0.2933	**0.2113**	0.2093
Coronary artery disease	**0.6534**	0.5096	**0.4204**	0.3971	0.3246	**0.3444**	0.2968	**0.3152**
Subarachnoid haemorrhage	**0.5281**	0.3385	**0.2770**	0.1713	**0.1776**	0.1392	0.0821	**0.1312**
Heart failure	**0.6753**	0.5294	**0.5099**	0.4418	0.4233	**0.4265**	0.3871	**0.3982**
Stroke	**0.6580**	0.5281	**0.4670**	0.4133	**0.3596**	0.3521	0.2895	**0.3239**
Alzheimer's disease	**0.3084**	0.2057	**0.1961**	0.1351	0.1144	**0.1221**	**0.1155**	0.1032
Any dementia	**0.4778**	0.3479	**0.3341**	0.2906	0.2627	**0.2695**	0.2480	**0.2557**
Parkinson's disease	**0.2784**	0.1603	**0.1417**	0.1178	**0.0915**	0.0559	0.0557	**0.0591**
Vascular dementia	**0.4334**	0.2405	**0.2344**	0.1116	**0.1938**	0.0817	**0.0842**	0.0607
Macro	**0.4963**	0.3428	**0.3122**	0.2299	**0.2154**	0.1823	0.1495	**0.1542**

Table 29: Per-task sensitivity at 99% specificity on SS-EHR at each retention level of clinical code–value units. The larger value per task and level is bold.

	100%		75%		50%		25%	
Task	HealthSAW	EHR-FM	HealthSAW	EHR-FM	HealthSAW	EHR-FM	HealthSAW	EHR-FM
Bladder cancer	**0.5105**	0.3427	**0.3287**	0.2587	0.2028	**0.2448**	0.0909	**0.1678**
Breast cancer	**0.3881**	0.2710	**0.2177**	0.1478	**0.0883**	0.0719	0.0513	0.0513
Colorectal cancer	**0.4313**	0.2827	**0.2188**	0.2077	**0.1342**	0.1118	**0.0831**	0.0751
Kidney cancer	**0.5922**	0.4369	**0.4466**	0.2913	**0.3592**	0.1845	0.1456	**0.2039**
Liver cancer	**0.5838**	0.4622	**0.4135**	0.3622	0.3149	**0.3162**	0.2635	**0.2851**
Lung cancer	**0.4890**	0.3423	**0.2823**	0.2445	**0.1656**	0.1451	**0.0946**	0.0868
Pancreatic cancer	**0.4462**	0.2846	**0.2769**	0.0692	**0.2385**	0.1846	**0.1692**	0.1538
Acute coronary syndrome	**0.5432**	0.3909	**0.3175**	0.2563	**0.1876**	0.1788	**0.1210**	0.1156
Coronary artery disease	**0.4019**	0.2025	**0.1463**	0.1232	0.0841	**0.1012**	0.0733	**0.0811**
Subarachnoid haemorrhage	**0.5444**	0.3889	**0.3778**	0.3000	**0.3222**	0.2000	0.1889	**0.2333**
Heart failure	**0.4777**	0.2832	**0.2776**	0.2230	0.1979	0.1979	0.1728	**0.1895**
Stroke	**0.4686**	0.3072	**0.2443**	0.1895	**0.1604**	0.1438	0.1066	**0.1272**
Alzheimer's disease	**0.3258**	0.3122	**0.2398**	0.2217	0.1538	**0.2127**	**0.1674**	0.1403
Any dementia	**0.3355**	0.2409	**0.2054**	0.1677	0.1376	**0.1419**	0.1172	**0.1376**
Parkinson's disease	**0.2888**	0.2139	**0.2032**	0.1818	**0.1604**	0.1123	0.0695	**0.0802**
Vascular dementia	**0.4592**	0.3673	**0.3776**	0.2551	**0.3163**	0.2041	**0.2041**	0.1531
Macro	**0.4554**	0.3206	**0.2859**	0.2187	**0.2015**	0.1720	0.1324	**0.1426**

Table 30: Disease prediction with HealthSAW's raw-history decoder variant. The per-task rows report the raw-history variant; the final rows compare its 16-task macro averages with the main HealthSAW model. PPV and sensitivity use 99% specificity. Main HealthSAW per-task results appear in Table 13.

	MS-EHR			SS-EHR		
Task	AUPRC	PPV	Sens.	AUPRC	PPV	Sens.
Bladder cancer	0.1389	0.1298	0.2831	0.4086	0.2454	0.4615
Breast cancer	0.1627	0.2525	0.1605	0.4084	0.5070	0.3737
Colorectal cancer	0.1046	0.1304	0.1811	0.4375	0.4945	0.4313
Kidney cancer	0.0943	0.0960	0.2351	0.4869	0.1710	0.5728
Liver cancer	0.1485	0.1353	0.4577	0.5675	0.5514	0.5514
Lung cancer	0.1480	0.1849	0.2265	0.4690	0.4601	0.4637
Pancreatic cancer	0.0303	0.0423	0.1642	0.3949	0.1513	0.4539
Acute coronary syndrome	0.2833	0.4293	0.1974	0.6240	0.7334	0.5255
Coronary artery disease	0.4834	0.6720	0.2098	0.6324	0.8380	0.3899
Subarachnoid haemorrhage	0.0498	0.0666	0.1936	0.3920	0.1585	0.4667
Heart failure	0.4709	0.6239	0.2732	0.6500	0.7898	0.4755
Stroke	0.2560	0.3957	0.1702	0.6604	0.8011	0.5023
Alzheimer's disease	0.0838	0.1267	0.1923	0.2699	0.2454	0.3032
Any dementia	0.2694	0.3547	0.2459	0.4609	0.5780	0.3269
Parkinson's disease	0.0612	0.0929	0.1593	0.2147	0.2118	0.2300
Vascular dementia	0.0665	0.0916	0.2521	0.3471	0.2291	0.4184
Macro over 16 tasks						
Raw-history decoder	0.1782	0.2390	0.2251	0.4640	0.4479	0.4342
HealthSAW (local decoder)	**0.1827**	**0.2423**	**0.2255**	**0.4963**	**0.4580**	**0.4554**

Table 31: The two-stage comparison of Table 4 with AUROC and PPV included: 16-task macro averages per dataset under the same SFT protocol, with the main HealthSAW checkpoint as reference. Largest values per dataset are bold; PPV and sensitivity at 99% specificity (Appendix D.3).

	MS-EHR				SS-EHR			
Model	AUROC	AUPRC	PPV	Sens.	AUROC	AUPRC	PPV	Sens.
Continuous two-stage	0.8635	0.1574	0.2209	0.1915	0.8678	0.4787	0.4546	0.4453
VQ two-stage	0.8416	0.1241	0.1852	0.1428	0.8591	0.4449	0.4395	0.4141
HealthSAW	**0.8758**	**0.1827**	**0.2423**	**0.2255**	**0.8682**	**0.4963**	**0.4580**	**0.4554**

Table 32: Optimized pretraining step times on eight B200 GPUs, with global batch 256 and a 3,072-source-token context. Timing procedures are described in Appendix I.1.

Model	s/step ↓
EHR-FM	0.6500
HealthSAW (raw-history decoder)	0.3900
HealthSAW	0.3557

Table 33: Patient-representation throughput on SS-EHR in test rows per second on one B200, for four cancer tasks (Appendix I.2). Ratios are HealthSAW over EHR-FM; the mean row averages the four task ratios.

Retained content	Bladder	Breast	Lung	Colorectal	Mean ratio
25% HealthSAW	2,064	2,045	2,116	2,043	0.90×
EHR-FM	2,338	2,315	2,286	2,302	
50% HealthSAW	2,037	1,958	2,025	1,988	1.57×
EHR-FM	1,270	1,288	1,281	1,269	
75% HealthSAW	1,945	1,947	1,983	1,966	2.21×
EHR-FM	887	898	884	878	
100% HealthSAW	1,877	1,836	1,904	1,860	2.85×
EHR-FM	656	664	652	649	

I Training and Inference Efficiency

This appendix details the pretraining and patient-representation speedups of §4.4; the disease-simulation speedup is measured together with its rollouts in Appendix G.4.

I.1 Pretraining Throughput

All three models use eight B200 GPUs, a global batch of 256, microbatch size 32, a 3,072-source-token context, and mixed BF16 precision. The pretraining corpus and budget are shared: CLAIM plus the MS-EHR training partition, four epochs and 132,164 optimizer steps. Table 32 reports optimized step times after warm-up, excluding checkpoint writing. Nominal token throughput is the global batch times the source-token budget divided by step time. HealthSAW's 0.3557 s per step against EHR-FM's 0.6500 s gives the $1.8\times$ pretraining throughput of §4.4.

I.2 Patient-Representation Throughput

Both models reuse the SFT checkpoints of the retained-content experiment, so no training or checkpoint selection was done for timing. Each run scores a complete SS-EHR test set on one B200 under the SFT context limit, with FP32 weights, BF16 autocast and `torch.compile`; batch size is tuned per model and level on held-out rows and then fixed across tasks. One untimed pass warms the model and the compiled graphs, and the timed pass creates no new graphs. Throughput counts test rows per wall-clock second, including loading, transfers, the forward pass and the softmax, and excludes checkpoint loading, compilation and the warm-up pass. We time four cancer tasks; each level averages their throughputs equally, and the reported ratio is the mean of the four per-task ratios.

Table 33 gives the per-task rates behind Figure 5. EHR-FM processes about 816 tokens per patient at full content and 247 at a quarter, while HealthSAW encodes about 54 events, or 163 slots, at every level. At full content HealthSAW is $2.85\times$ faster, the $2.9\times$ of §4.4.