Pretraining Patient Foundation Models on Multimodal Patient Journeys

Daniel P. Jeong[1]* **Suhana Bedi**[2]* **Cliff Wong**[3] **Sheng Zhang**[3]

Jeya Maria Jose Valanarasu[3] **Qianchu Liu**[3] **Reuben Tan**[3] **Zelalem Gero**[3]

Jaspreet Bagga[3] **Juan Manuel Zambrano Chaves**[3] **Naoto Usuyama**[3] **Hanwen Xu**[4]

Roshanthi K. Weerasinghe[5] **Rom S. Leidner**[5] **Brian Piening**[5] **Carlo Bifulco**[5]

Tristan Naumann[3] **Hoifung Poon**[3]

[1] Machine Learning Department, Carnegie Mellon University
[2] Department of Biomedical Data Science, Stanford University
[3] Microsoft Research
[4] Department of Computer Science, University of Washington
[5] Providence Genomics

Abstract

Precision medicine requires integrating diverse data modalities such as structured electronic health records (EHRs), radiology images, digital pathology, and genomics to guide treatment decisions. Yet, many existing foundation models for longitudinal patient data remain unimodal, trained solely on structured codes or imaging modalities, which limits their clinical utility for highly complex conditions like cancer. In this work, we present the first scaling-law study for foundation models pretrained on *multimodal* patient journeys, using longitudinal structured records, CT scans, and whole-slide histopathology images from 2.3M cancer patients. We train our multimodal EHR Transformer (MEHRT) models to simultaneously process all modalities recorded across time via a multi-stage curriculum pretraining strategy, and introduce a new evaluation suite of six oncology prediction tasks (e.g., progression-free survival, metastasis) carefully defined with oncologists. MEHRT consistently outperforms state-of-the-art supervised baselines, e.g., achieving a +7% average improvement in AUROC over the best-performing baseline (CatBoost), and its performance scales with model size. When compared to its unimodal counterpart (EHRT), MEHRT shows modest yet consistent improvements in predictive accuracy and generative modeling capabilities, suggesting that multimodality can complement scaling. Finally, we discuss important limitations and practical lessons learned that inform future development of multimodal EHR foundation models.

*Equal contribution. Work conducted as an intern at Microsoft Research.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

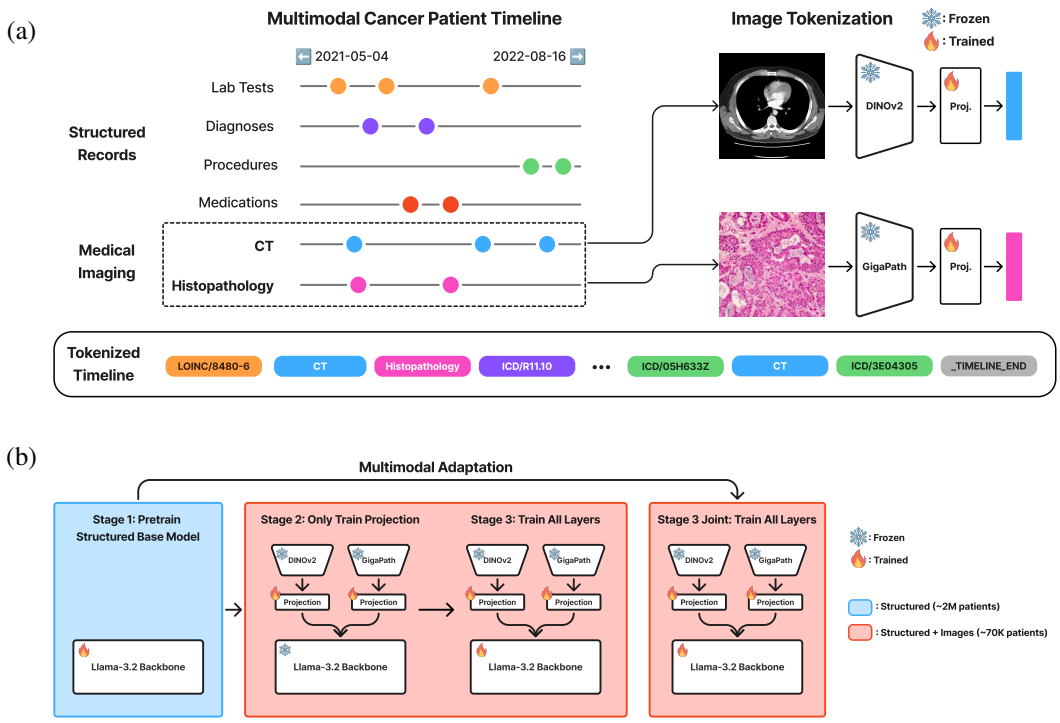


Figure 1: Overview of MEHRT pretraining (Section 2.1). **(a) Tokenization:** We treat each patient timeline as a temporally ordered sequence of tokens, converting each structured record and CT/histopathology image into a set of tokens. For the latter, we transform each image into visual tokens by feeding it into a modality-specific encoder. **(b) Pretraining:** We first pretrain a *structured-only* base model (EHRT; Stage 1) using 2.3 million patient timelines, then adapt it to handle *multimodal* timelines (MEHRT) using 70K timelines with images. For MEHRT training, we consider both a curriculum learning approach (middle), initially freezing the EHRT backbone (Stage 2), then training all parameters (Stage 3), and a joint pretraining approach (Stage 3 Joint; right).

1 Introduction

Precision medicine is inherently multimodal. For example, in an oncological setting, clinicians often need to reason over information gathered across longitudinal structured electronic health records (EHRs) (e.g., labs, procedures, diagnoses), computed tomography (CT) scans, and histopathology slides to diagnose disease, guide treatment decisions, and monitor disease progression. Each modality provides complementary information: EHRs capture patient comorbidities and treatment trajectories over time; CT imaging reveals tumor morphology and anatomical spread; and histopathology exposes cellular-level features critical for cancer grading and treatment response prediction. It is essential to understand the complex interactions between these modalities to accurately predict challenging oncology outcomes such as survival, disease recurrence, and other adverse events.

Despite this clinical reality, many existing EHR foundation models for patient journeys remain unimodal, trained exclusively on structured medical codes, free text, or individual imaging modalities (Steinberg et al., 2021; Kraljevic et al., 2024; Wornow et al., 2025; Rajamohan et al., 2025). Such lack of multimodality is especially limiting for highly complex and heterogeneous conditions like cancer, which also makes cancer patient cohorts an interesting testbed for studying the benefits of a multimodal foundation model. Moreover, existing models are often only trained on cohorts clinically distinct from cancer patients (e.g., patients in intensive care units or emergency departments (Johnson et al., 2023; Renc et al., 2024, 2025; Zhang et al., 2025)), further motivating their exploration in oncology. Meanwhile, recent works on vision-language models demonstrate that unimodal foundation models can be combined to train a single multimodal model (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023), but these approaches have not yet been extended to address the unique challenges of multimodal healthcare data, rich with complex temporal and cross-modal dependencies.

In this paper, we present the first scaling-law study for foundation models pretrained on multimodal patient journeys, exploring the opportunities and pitfalls in training such models for precision oncology. Using a large-scale longitudinal dataset with 2.3M cancer patients (1.6B structured tokens, 1.6M CT images, 170K histopathology slides) from the Providence Health System, we pretrain multimodal EHR Transformer models (MEHRT) at the 9M–982M scale to simultaneously process temporally ordered structured records, CT imaging, and whole-slide histopathology images (Figure 1a). In doing so, we consider a multimodal curriculum pretraining approach, where we first train a model only using structured records, then gradually adapt it to jointly leverage the rich context from timestamped medical images (Figure 1b). We then evaluate MEHRT on six prognostic tasks selected for clinical relevance under the guidance of attending oncologists (Section 2.2).

Overall, we find that MEHRT achieves substantial improvements over state-of-the-art baselines (e.g., gradient-boosted trees), and that its downstream performance improves with increasing model scale. MEHRT also shows modest yet consistent improvements over its structured-only base model (EHRT) in downstream prediction and generative modeling, suggesting that multimodality can complement scaling, although the different modalities may exhibit high overlap in information content.

Our main contributions can be summarized as follows:

- **Multimodal EHR foundation model for cancer:** We train one of the first (to our knowledge) multimodal EHR foundation models for cancer, which natively and simultaneously handles longitudinal structured records, CT scans, and whole-slide histopathology images.
- **Scaling multimodal pretraining:** We show that our pretrained multimodal models consistently outperform state-of-the-art supervised learning baselines (e.g., CatBoost) on oncology outcome prediction tasks, and that their performances scale predictably with model size.
- **Multimodal vs. unimodal pretraining:** We evaluate the benefits of multimodal over structured pretraining and observe that multimodality can complement scaling, leading to modest yet consistent improvements in downstream prediction and generative modeling.

2 Methods

We pretrain 14 MEHRT models across seven parameter scales (9M–982M) on longitudinal data from $\sim$2M cancer patients (Section 2.1), and then evaluate their performance on six oncology outcome prediction tasks via linear probing (Section 2.2).

2.1 Model Training

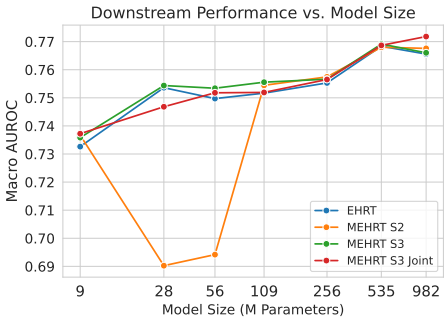
Data representation. We represent each patient p as a chronologically ordered sequence $\mathbf{s}_p = [s_1, s_2, \ldots, s_T]$, where s_t is a clinical event (e.g., "ICD/R11.10") at timestamp t. Structured records (ICD-10-CM, ICD-10-PCS, cancer-specific RxNorm medications, LOINC lab codes, demographic variables), CT scans, and histopathology slides are temporally aligned and concatenated to construct $\mathbf{s}_p$. We generate an 85–5–10 train–validation–test split by patient ID, ensuring that there is no train–test leakage of patient timelines and the distributions of token lengths and inter-event time intervals (Tables A1–A2) and included modalities (Figure A1) are similar across the three data subsets.

Tokenization and embedding. We construct a unified vocabulary $\mathcal{V}$ treating each structured event as a single token, except for ICD codes, for which we use 1–3 tokens based on their hierarchy (e.g., "ICD/R11.10" $\rightarrow$ ["ICD/R11", "ICD/10"]) following Renc et al. (2024). CT scans are encoded via a DINOv2-based (Oquab et al., 2024) foundation model pretrained on 1.6M oncology volumes ($\mathbf{e}_{\text{CT}} \in \mathbb{R}^d$), while histopathology slides are encoded using GigaPath (Xu et al., 2024) ($\mathbf{e}_{\text{WSI}} \in \mathbb{R}^{d'}$). After averaging across slices/tiles, the image embeddings are projected to $\mathbb{R}^{d_{\text{model}}}$ via learnable linear projections $\mathbf{W}_{\text{CT}} \in \mathbb{R}^{d_{\text{model}} \times d}$ and $\mathbf{W}_{\text{WSI}} \in \mathbb{R}^{d_{\text{model}} \times d'}$, and then concatenated to form the input embeddings $\mathbf{x}_p = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{T \times d_{\text{model}}}$ (Figure 1a). We also insert special time-interval token embeddings (representing e.g., 5 minutes, 6–12 hours, 1–2 years) to encode the temporal gaps between consecutive medical events in $\mathbf{x}_p$.

Model architecture and training objective. We pretrain models based on the Llama-3.2 architecture (Grattafiori et al., 2024) with the standard autoregressive language modeling objective (Radford et al., 2019), computing the cross-entropy loss only on structured tokens (not images). We implement

Model	AKI	Meta.	Mort.	Neut.	Prog.	Trans.
CatBoost	0.702	**0.733**	**0.790**	**0.688**	**0.719**	**0.727**
XGBoost	0.705	0.720	0.786	0.618	0.715	0.701
Random Forest	0.678	0.715	0.769	0.674	0.712	0.682
Logistic Regression	**0.711**	0.686	0.752	0.669	0.694	0.710
MEHRT-9M	0.753	0.680	0.785	0.698	0.736	0.770
MEHRT-28M	0.769	0.687	0.789	0.709	0.747	0.780
MEHRT-56M	0.751	0.681	0.798	0.738	0.751	0.792
MEHRT-109M	0.763	0.718	0.794	0.722	0.748	0.766
MEHRT-256M	0.775	0.731	**0.809**	0.743	0.752	**0.810**
MEHRT-535M	**0.797**	**0.742**	0.804	**0.753**	0.757	0.803
MEHRT-982M	0.791	0.732	0.808	0.751	**0.762**	0.805

(a)



(b)

Figure 2: MEHRT and EHRT consistently outperform supervised learning baselines, and the improvements scale with model size. **(a)** AUROCs of MEHRT (Stage 3 Joint) against supervised baselines across six oncology outcome prediction tasks described in Section 2.2. The best baseline scores are in **blue**, and the best MEHRT scores are in **red** (see Table C1 for EHRT results). **(b)** Downstream performances (AUROC, macro-averaged across tasks) of MEHRT and EHRT both scale with model size, albeit with diminishing gains at larger model scales.

a multi-stage curriculum (Figure 1b): (i) **Stage 1** (EHRT): structured-only pretraining on 2M patients (10 epochs); (ii) **Stage 2** (MEHRT): from EHRT, train projection layers on 70K imaging patients (5 epochs); (iii) **Stage 3** (MEHRT): unfreeze all parameters, continue training (5 epochs); (iv) **Stage 3 Joint** (MEHRT): from EHRT, train all parameters on imaging patients (10 epochs).

2.2 Evaluation Framework

We evaluate our trained models on six binary classification tasks with 1-year prediction horizons from cancer treatment initiation: mortality, progression-free survival, metastatic progression, acute kidney injury (AKI), neutropenic fever, and blood transfusion requirements. For linear probing, we extract patient representations $\mathbf{h}_p \in \mathbb{R}^{d_{\text{model}}}$ via (i) token averaging using the "echo embedding" approach (Springer et al., 2025) or (ii) last token extraction, then train task-specific logistic classifiers $f(\mathbf{h}_p) = \sigma(\mathbf{w}^T\mathbf{h}_p + b)$. We compare against supervised learning baselines (CatBoost (Prokhorenkova et al., 2018), XGBoost (Chen and Guestrin, 2016), random forest (Breiman, 2001), logistic regression) trained on either structured data alone using "bag-of-events" features (Steinberg et al., 2021; Wornow et al., 2023) (Appendix B.1) or structured + imaging data with late-fusion ensembling (Appendix B.2). All hyperparameters are optimized via 5-fold cross-validation using area under the receiver operating characteristic curve (AUROC) as the primary evaluation metric (Appendix B.3).

3 Results

MEHRT and EHRT consistently outperform supervised learning baselines, and the improvements scale with model size. MEHRT consistently outperforms supervised learning baselines across all tasks, especially at the 256M+ scale (Figure 2a). MEHRT-535M, the best overall among MEHRT models in terms of test AUROC, achieves a **+7%** average AUROC improvement over the best supervised learning baseline with multimodal inputs (CatBoost), while EHRT-535M shows a **+5.3%** improvement over the best baseline with structured-only inputs (Table C1). Both MEHRT and EHRT achieve higher test AUROCs (macro-averaged across tasks) as we increase the number of parameters (Figure 2b), showing that the scaling behavior observed with structured EHR foundation models (Zhang et al., 2025; Waxler et al., 2025) extends to multimodal settings. Notably, joint pretraining (training all parameters throughout multimodal adaptation) achieves comparable scaling behavior to gradual adaptation (Stage 2 followed by Stage 3) for MEHRT.

MEHRT achieves modest yet consistent improvements over EHRT for predicting imaging-dependent outcomes and generative modeling. When compared against the unimodal EHRT on downstream tasks, MEHRT shows the most improvements (albeit modest) on oncology outcomes where imaging context is especially informative, i.e., mortality, metastasis, and progression-free

Task (AUROC)	EHRT	MEHRT	△ (MEHRT– EHRT)
AKI	**0.7759**	0.7750	-0.0009
Metastasis	0.6831	**0.6903**	+0.0073
Mortality	0.7989	**0.8093**	+0.0104
Neutropenia	**0.7489**	0.7377	-0.0112
Progression	0.7496	**0.7520**	+0.0024
Transfusion	**0.7754**	0.7747	-0.0007

(a)

Size (M)	EHRT	MEHRT (S2)	MEHRT (S3)	MEHRT (S3 Joint)
9	6.08	6.08	**5.92**	6.00
28	5.33	5.44	**5.27**	5.28
56	5.59	5.59	**5.46**	5.86
109	5.51	5.51	5.56	**5.43**
256	5.18	5.18	5.02	**5.01**
535	5.11	5.11	**4.91**	4.95
982	5.05	5.05	**4.82**	4.86

(b)

Figure 3: MEHRT shows modest yet consistent improvements over EHRT. **(a)** Test AUROCs at 256M scale, with largest gains in mortality (+0.0104) and metastasis (+0.0073). **(b)** Test perplexity across model scales (lower is better); S2 = Stage 2, S3 = Stage 3. Best results are marked in **bold**.

survival. For instance, at the 256M model scale, MEHRT (Stage 3) achieves higher AUROC precisely on these three tasks: +0.0104 on mortality, +0.0073 on metastasis, and +0.0024 on progression-free survival (Figure 3a, highlighted in gray). Beyond discriminative performance, MEHRT achieves consistent test perplexity reductions across all model scales (Figure 3b), e.g., decreasing from 5.18 to 5.01 at 256M parameters (3.3% improvement), with consistent improvement in perplexity with increasing model scale. These results suggest that multimodal pretraining can enable the model to better capture the underlying distribution of patient timelines.

4 Discussion and Conclusion

Our study demonstrates that pretraining on multimodal patient timelines, comprised of structured records, CT scans, and histopathology slides, can improve not only generative modeling capabilities but also downstream predictive performance. We find that such gains over unimodal approaches (i.e., only using structured records) are most pronounced for small-to-mid-sized models (9M–256M) but remain positive up to 982M parameters. These findings suggest that multimodality and scaling act as complementary levers for improving the effectiveness of clinical foundation models.

Meanwhile, we discuss these findings with the following caveats. First, while we explore multiple strategies for training multimodal EHR foundation models and observe relatively modest improvements compared to their structured-only counterparts, alternative architectures, pretraining strategies, and combinations of modalities (e.g., free-form clinical notes or multi-omics data) can potentially result in larger absolute gains. We leave an in-depth exploration of these design choices as future work. Second, the mixed downstream AUROC improvements from multimodal over structured-only pretraining (Figure 3a) illustrate the importance of evaluating on tasks that require information beyond that provided by structured records. While these results arguably suggest that structured information can be predictive of various outcomes, not all clinically relevant tasks are likely to be well-captured by such data alone. This underscores the need for carefully designing multimodal benchmarks that well-reflect the real-world clinical need for multimodal integration.

References

J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. Advances in Neural Information Processing Systems (NeurIPS), 2022.

L. Breiman. Random Forests. Machine Learning, 45(1):5–32, 2001.

T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi,

M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma. The Llama 3 Herd of Models. arXiv:2407.21783, 2024.

A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark. MIMIC-IV, A Freely Accessible Electronic Health Record Dataset. Scientific Data, 10(1), 2023.

Z. Kraljevic, D. Bean, A. Shek, R. Bendayan, H. Hemingway, J. A. Yeung, A. Deng, A. Baston, J. Ross, E. Idowu, J. T. Teo, and R. J. B. Dobson. Foresight: A Generative Pretrained Transformer for Modelling of Patient Timelines Using Electronic Health Records: A Retrospective Modelling Study. The Lancet, 6(4), 2024.

J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In International Conference on Machine Learning (ICML), 2023.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual Instruction Tuning. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (TMLR), 2024. ISSN 2835-8856.

L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems (NeurIPS), 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. In OpenAI Blog, 2019.

H. R. Rajamohan, X. Gao, W. Zhu, S.-L. Huang, L. Chen, K. Cho, C. M. Deniz, and N. Razavian. Foundation Models for Clinical Records at Health System Scale. International Conference on Machine Learning (ICML) 2025 Workshop on Foundation Models for Structured Data, 2025.

P. Renc, Y. Jia, A. E. Samir, J. Was, Q. Li, D. W. Bates, and A. Sitek. Zero Shot Health Trajectory Prediction Using Transformer. npj Digital Medicine, 7(256), 2024.

P. Renc, M. K. Grzeszczyk, N. Oufattole, D. Goode, Y. Jia, S. Bieganski, M. B. A. McDermott, J. Was, A. E. Samir, J. W. Cunningham, D. W. Bates, and A. Sitek. Foundation Model of Electronic Medical Records for Adaptive Risk Estimation. arXiv:2502.06124, 2025.

J. M. Springer, S. Kotha, D. Fried, G. Neubig, and A. Raghunathan. Repetition Improves Language Model Embeddings. In International Conference on Learning Representations (ICLR), 2025.

E. Steinberg, K. Jung, J. A. Fries, C. K. Corbin, S. R. Pfohl, and N. H. Shah. Language Models Are An Effective Representation Learning Technique for Electronic Health Record Data. Journal of Biomedical Informatics, 113:103637, 2021.

S. Waxler, P. Blazek, D. White, D. Sneider, K. Chung, M. Nagarathnam, P. Williams, H. Voeller, K. Wong, M. Swanhorst, S. Zhang, N. Usuyama, C. Wong, T. Naumann, H. Poon, A. Loza, D. Meeker, S. Hain, and R. Shah. Generative Medical Event Models Improve with Scale. arXiv:2508.12104, 2025.

M. Wornow, R. Thapa, E. Steinberg, J. A. Fries, and N. 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, 2023.

M. Wornow, S. Bedi, M. A. F. Hernandez, E. Steinberg, J. A. Fries, C. Re, S. Koyejo, and N. H. Shah. Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHRs. In International Conference on Learning Representations (ICLR), 2025.

H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B. J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, and H. Poon. A Whole-Slide Foundation Model for Digital Pathology from Real-World Data. Nature, 630: 181–188, 2024.

S. Zhang, Q. Liu, N. Usuyama, C. Wong, T. Naumann, and H. Poon. Exploring Scaling Laws for EHR Foundation Models. arXiv:2505.22964, 2025.

A Additional Details on Datasets

Table A1: Summary statistics for *all* 2.3M patient timelines with *only* structured tokens for the train, validation, and test sets.

	Train	**Validation**	**Test**
Number of Patients	1,972,558	116,105	232,323
Patient Timeline Length (Tokens)			
Mean / Std	677 / 1,515	676 / 1,451	674 / 1,430
Min / Max	2 / 319,477	2 / 77,727	2 / 107,562
Q1 / Q2 / Q3 quartiles	70 / 248 / 703	69 / 246 / 700	69 / 249 / 702
total timeline tokens	1,336,958,506	78,429,952	156,661,282
Mean Time Interval (Days)			
Mean / Std	13.223 / 45.982	13.293 / 46.101	13.269 / 46.039
Min / Max	0 / 396.485	0 / 395.999	0 / 396.737
Q1 / Q2 / Q3 quartiles	0.016 / 0.361 / 4.939	0.025 / 0.382 / 4.995	0.015 / 0.358 / 4.99

Table A2: Summary statistics for the 83K *imaging* patient timelines with *both* structured and imaging tokens for the train, validation, and test sets.

	Train	**Validation**	**Test**
Number of Patients	71,197	4,168	8,451
Patient Timeline Length (Tokens)			
Mean / Std	2,399 / 2,720	2,346 / 2,633	2,390 / 2,665
Min / Max	2 / 98,861	4 / 38,836	2 / 60,540
Q1 / Q2 / Q3 quartiles	701 / 1,642 / 3,185	697 / 1,599 / 3,122	691 / 1,652 / 3,220
total timeline tokens	170,815,403	9,778,512	20,199,961
Mean Time Interval (Days)			
Mean / Std	2.899 / 17.334	2.816 / 17.213	3.080 / 17.889
Min / Max	0 / 332.901	0 / 330.229	0 / 338.934
Q1 / Q2 / Q3 quartiles	0.002 / 0.012 / 0.305	0 / 0.008 / 0.311	0.003 / 0.012 / 0.343

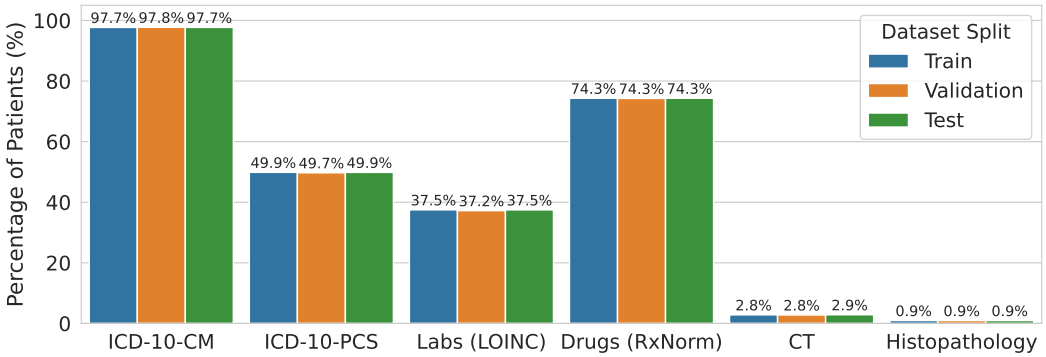


Figure A1: We ensure that the percentage of patients with each modality remain consistent across the train, validation, and test sets. Here, we show the distribution across all 2.3M patient timelines.

B Additional Details on Supervised Learning Baselines

Here, we provide additional details on (i) how we construct the "bag-of-events" features from longitudinal structured records; (ii) how we train the late-fusion baselines which take structured

Table A3: Number of train, validation, and test examples for each downstream prediction task.

Task	Train	Val	Test
AKI	7,314	376	875
Mortality	14,004	708	1,715
Metastasis	4,191	225	487
Neutropenia	8,824	446	1,036
Progression	8,730	455	1,073
Transfusion	8,561	438	1,017

records and all medical images as input; and (iii) what hyperparameter search spaces we consider for each supervised learning baseline.

B.1 Bag-of-Events Features from Structured Records

Given that all of the supervised learning baselines expect fixed-length vector-valued inputs, we convert the sequence of timestamped structured records for each patient into a vector of counts (hence, a "bag-of-events"). While this process removes any temporal information associated with the structured records, prior works demonstrate that it is often a strong baseline feature engineering approach for structured EHR data (Steinberg et al., 2021; Wornow et al., 2023, 2025).

We first extract the list of all unique medical codes (ICD-10-CM, ICD-10-PCS, RxNorm, LOINC) observed in the training set. Then, we remove all medical codes that occur in less than 1% of all patients in the training set to only retain the ones that are frequent enough to capture meaningful differences. For the ICD-10-CM and ICD-10-PCS codes, we only consider the highest levels of their hierarchical structure (i.e., the first three characters in each code) to reduce the sparsity of the resulting feature vector and avoid a blowup in the number of features. For example, for ICD-10-CM code "R11.10", we reduce it to "R11" before aggregating the counts of unique ICD-10-CM codes. After finalizing the list of C medical codes to use, we count the number of times each code occurs within each patient timeline $\mathbf{s}_p$ to obtain the feature vector $\mathbf{v}_p^{\text{structured}} \in \mathbb{Z}_+^C$ for each patient in the train, validation, and test sets. We also append the demographic features (gender, age at first event recorded in the patient timeline, and ethnicity) for each patient to $\mathbf{v}_p^{\text{structured}}$.

B.2 Late-Fusion Ensembling for Training Baselines with Multimodal Inputs

In order to train baselines that can also take in the CT and histopathology embeddings as input, we consider a late-fusion ensembling approach where we train a given baseline model (e.g., XGBoost) for each modality separately (i.e., one for structured records, one for CT embeddings, and one for histopathology embeddings), then average their predictions. We consider this approach over naïve concatenation of the "bag-of-events" features and image embeddings for the following two reasons. First, naïve concatenation of different modalities can lead to significant blowup in the number of features, which renders training an effective model more challenging given the relatively small number of labeled examples (Table A3). Second, different patient timelines contain a variable number of CT and histopathology embeddings (from one image to hundreds of images), with some patients entirely missing one of the two imaging modalities, which can lead to an inconsistent number of features across patients.

For each patient, we take the simple approach of averaging the embeddings of all images together to obtain one CT embedding $\mathbf{v}_p^{\text{CT}} \in \mathbb{R}^{1024}$ and one histopathology embedding $\mathbf{v}_p^{\text{WSI}} \in \mathbb{R}^{1536}$. For the train, validation, and test sets, we then concatenate the structured "bag-of-events" features as $\mathbf{V}^{\text{structured}} = [\mathbf{v}_1^{\text{structured}}, \ldots, \mathbf{v}_N^{\text{structured}}]$, the averaged CT embeddings as $\mathbf{V}^{\text{CT}} = [\mathbf{v}_1^{\text{CT}}, \ldots, \mathbf{v}_N^{\text{CT}}]$, and the averaged histopathology embeddings as $\mathbf{V}^{\text{WSI}} = [\mathbf{v}_1^{\text{WSI}}, \ldots, \mathbf{v}_N^{\text{WSI}}]$, where N denotes the number of patients in each set. We note that while the number of examples available for each modality can be different, we use the above notation for simplicity.

We train separate models for each input—$\mathbf{V}^{\text{structured}}$, $\mathbf{V}^{\text{CT}}$, and $\mathbf{V}^{\text{WSI}}$—via 5-fold cross-validation using the training and validation sets (Section 2.2). For each test example, we average the confidence scores from these three models to make the final prediction. If one of the imaging modalities is missing, we simply drop the corresponding model and use the remaining models for ensembling.

B.3 Hyperparameter Optimization

We optimize the hyperparameters of each supervised learning baseline via 5-fold cross-validation, using a combination of the training and validation sets. In the tables below, we detail the hyperparameter search spaces considered for each baseline. For linear probing (Section 2.2), we use the same search space considered for the logistic regression baseline.

Table B1: Hyperparameter search space for CatBoost.

Hyperparameter	Distribution
Number of Estimators	UniformInt(5,125)
Max Depth	UniformInt(2,12)
Learning Rate	LogUniform(1e-5,1)
L_2 Regularization	LogUniform(1,10)

Table B2: Hyperparameter search space for XGBoost.

Hyperparameter	Distribution
Number of Estimators	UniformInt(5,125)
Max Depth	UniformInt(1,10)
α	LogUniform(1e-16,1e2)
λ	LogUniform(1e-16,1e2)
Learning Rate	LogUniform(1e-7,1)

Table B3: Hyperparameter search space for random forest.

Hyperparameter	Distribution
Number of Estimators	UniformInt(5,100)
Max Depth	UniformInt(2,12)

Table B4: Hyperparameter search space for logistic regression.

Hyperparameter	Distribution
Inverse Regularization C	LogUniform(1e-10,1e10)

C Additional Results

Table C1: EHRT (ours) consistently beats supervised baselines in AUROC across all downstream prediction tasks. We mark the best score for the baselines in **blue** and EHRT in **red**.

Model	AKI	Metastasis	Mortality	Neutropenia	Progression	Transfusion
CatBoost	0.699	**0.726**	**0.781**	**0.692**	0.729	**0.754**
XGBoost	0.701	0.707	0.776	0.676	**0.730**	0.688
Random Forest	0.692	0.707	0.780	0.665	**0.730**	0.736
Logistic Regression	**0.726**	0.684	0.760	0.673	0.699	0.672
EHRT-9M	0.760	0.663	0.780	0.700	0.741	0.751
EHRT-28M	0.768	0.686	0.792	0.734	0.749	0.791
EHRT-56M	0.753	0.689	0.803	0.729	0.744	0.781
EHRT-109M	0.774	0.698	0.785	0.723	0.751	0.779
EHRT-256M	0.776	0.683	0.800	0.749	0.750	0.775
EHRT-535M	**0.802**	0.702	0.800	**0.752**	0.753	**0.801**
EHRT-982M	0.788	**0.704**	**0.802**	0.744	**0.759**	0.796