A156 · Pan-cancer risk assessment with an EHR foundation model that predicts what happens next and when

Presenter: Asif Khan — Harvard Medical School

Authors: Asif Khan, Duncan Forster, Moshir Harsh, Chunlei Zheng, Daniel Ritter, Qi Wei, Debora Marks, Erica Warner, Allison Chang, Lecia Sequist, Tanya K. Sorensen, Jennifer Hadlock, Nathanael R. Fillmore, Chris Sander

Session: Day 2 · Fri Oct 2, 2026 · 1:15–2:15 PM

Longitudinal electronic health records contain signals of future disease, but irregular intervals between visits and incomplete follow-up complicate representation learning and risk estimation. We developed GenEHR, a decoder-only foundation model that jointly generates clinical events and inter-visit times. Each history combines demographic, diagnosis, medication, laboratory, and procedure tokens. Mixed-radix time encoding represents each gap with five compositional digits, retaining daily resolution for intervals up to 4,199 days with only 27 time-head logits. Cohort-specific models were trained at Providence, Mass General Brigham, US Veterans Affairs, UK Biobank, and All of Us.

Across five cohorts, next-visit retrieval AUROC was 0.90-0.99 and event-prevalence log-Pearson correlation was 0.93-0.99. In MGB, Providence, Veterans Affairs, and All of Us, ICD-chapter AUROC was 0.98-0.99, event co-occurrence correlation 0.92-0.97, and inter-visit-gap Spearman correlation 0.89-0.98.

We adapted the pretrained representation for pan-cancer risk using multiple prediction cutoffs per patient, a piecewise-exponential likelihood over observed follow-up, and gated low-rank adapters that share representation directions across cancers while retaining cancer-specific weights. In MGB, across 17 cancers and 6-60-month horizons, macro-AUROC increased from 0.677-0.736 with direct pretrained-logit scoring to 0.797-0.839 with a frozen-backbone risk head and 0.798-0.840 with gated adaptation. At the 60-month analytic operating point, gated adaptation achieved 7.83% trajectory-level PPV, corresponding to 12.8 alert rows per future cancer.

Contextual association graphs and attribution analyses recovered organ-specific and prediagnostic patterns, including pancreatic disease with dysglycemia, gynecologic conditions with ovarian cancer, cytopenias before myeloid leukemia, and seizures before brain cancer. These associations support model audit and hypothesis generation, but require prospective validation.

Pan-cancer risk assessment with an EHR foundation model that predicts what happens next and when

Asif Khan[1,3,4,*], Duncan Forster[1,3], Moshir Harsh[1,3], Chunlei Zheng[2], Daniel Ritter[1], Qi Wei[5], Debora Marks[1], Erica Warner[1,4], Allison Chang[1,4], Lecia Sequist[1,4], Tanya K. Sorensen[5], Jennifer Hadlock[5,+], Nathanael R. Fillmore[1,2,+], Chris Sander[1,3,4,+]

[1]*Harvard Medical School, Boston, MA, USA.* [2]*VA Boston Healthcare System, Boston, MA, USA.* [3]*Ludwig Center at Harvard, Boston, MA, USA.*
[4]*Massachusetts General Hospital, Boston, MA, USA.* [5]*Institute for Systems Biology, Seattle, WA, USA.*
[*]Presenting author. [+]Co-senior authors. Contact: asif_khan@g.harvard.edu

Background and model. Longitudinal electronic health records (EHRs) capture changes in patient health before diagnosis, but visits occur at irregular intervals and the time between events is clinically informative. We developed GenEHR, a decoder-only foundation model that generates both clinical events and inter-visit time. Each trajectory contains demographic and lifestyle tokens followed by diagnoses, medications, laboratory results, and procedures grouped into visits. Mixed-radix time encoding (MRTE) represents elapsed days as compositional digits at day, week, and longer timescales. This preserves day-level resolution without assigning a separate token to every possible interval. Cohort-specific models were trained within secure environments at Providence, Mass General Brigham (MGB), U.S. Veterans Affairs, UK Biobank, and All of Us.

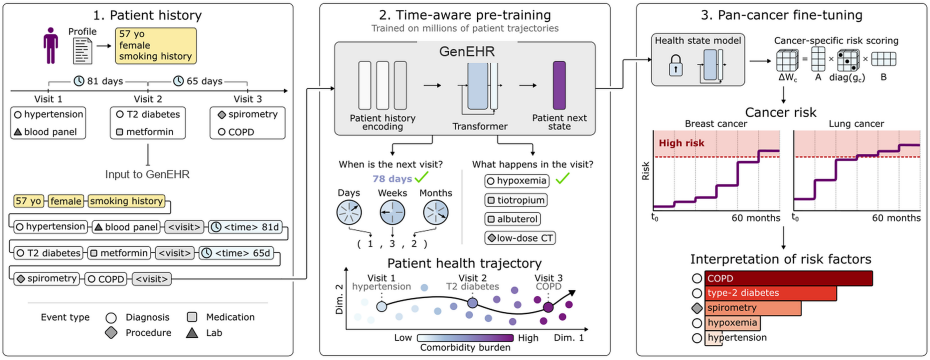

Figure 1. GenEHR represents patient history as clinical-event tokens and MRTE-encoded gaps, learns the evolving patient health state through next-event and time-gap prediction, and adapts this representation to horizon-specific pan-cancer risk.

Foundation-model evaluation. We evaluated generative performance independently in MGB, Providence, U.S. Veterans Affairs, All of Us, and UK Biobank. Across the five cohorts, next-visit retrieval AUROC ranged from 0.90 to 0.99 and generated event prevalence showed log-Pearson correlation of 0.93-0.99. In MGB, Providence, VA, and All of Us, ICD-chapter AUROC was 0.98-0.99, event co-occurrence log-Pearson 0.92-0.97, and inter-visit gap Spearman 0.89-0.98. Expected calibration error ranged from 0.07 to 0.14 across cohorts. These results show that the model captures clinical events, their relationships, and the intervals between visits.

Pan-cancer risk prediction. Next-event pretraining learns cancer-associated signals but is not directly optimized for the timing of a first cancer diagnosis or the duration of observed follow-up. In MGB, we removed events on and after diagnosis, applied a three-month exclusion window, and sampled up to four prediction cutoffs across the preceding five years. We fitted a piecewise-exponential likelihood from each cutoff until diagnosis or the end of recorded follow-up. The risk head combines the pretrained cancer-token signal with a patient state conditioned on each prediction horizon. Across 17 cancers and 6-60-month horizons, macro-AUROC ranged from 0.677 to 0.736 for direct pretrained-logit scoring, 0.797 to 0.839 for head-only training, and 0.798 to 0.840 for gated low-rank adaptation. Most of the improvement came from the follow-up-aware risk objective, while gated adaptation gave a small additional gain. At 60 months, gated adaptation increased trajectory-level PPV from 4.69% to 7.83% and reduced the alert-row-to-future-cancer ratio from 21.3 to 12.8.

Interpretation and conclusion. Contextual pointwise mutual information produced a cancer graph organized by organ system and linked diagnoses with related tests, medications, and procedures. Integrated gradients recovered pancreatic disease with dysglycemia, ovarian and pelvic disorders with endometriosis, cytopenia and marrow disorders before myeloid leukemia, and seizures and neurological events before brain cancer. Occlusion and visit-transition analyses further identified care-intensity signals that influence the fitted score. These findings show that GenEHR learns clinically meaningful patterns and can nominate temporal hypotheses for further study. They remain associations shaped by disease, documentation, and clinical practice. Prospective evaluation is required before clinical use.

References. [1] Kraljevic et al., *Lancet Digit Health* 2024. [2] Shmatko et al., *Nature* 2025. [3] Steinberg et al., ICLR 2024. [4] Hu et al., ICLR 2022.