Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Identifying cohorts at elevated risk of cancers using generative modeling of patient health states

Chris Sander

sander.research@gmail.com

Harvard Medical School https://orcid.org/0000-0001-6059-6270

Biological Sciences - Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11733688/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. D.S.M: Advisor - GenBio.ai, Dyno Therapeutics, Tectonic Therapeutic, Catalio Ventures; Cofounder - Seismic Therapeutic. Q.W., J.H.: Institutional research support for unrelated work from BMS, Gilead, Janssen, Delfina. L.V.S.: Institutional grant funding from AstraZeneca, Delfi Diagnostics. C.S.: SAB Cytoreason Ltd.

Identifying cohorts at elevated risk of cancers using generative modeling of patient health states

Asif Khan[1,3,4,5,8], Duncan T. Forster[1,3,8,#], Moshir Harsh[1,3,8,#], Chunlei Zheng[2,9#], Erica T. Warner[1,4], Daniel Ritter[1], Allison Chang[1,4], Qi Wei[5], Tanya K. Sorensen[6], Debora S. Marks[1,8], Lecia V. Sequist[1,4,*], Jennifer Hadlock[5,7,*], Nathanael R. Fillmore[1,2,9*], Chris Sander[1,3,4,8,9,*]

[1] Harvard Medical School, Boston, MA, USA
[2] VA Boston Healthcare System, Boston, MA, USA
[3] Ludwig Center at Harvard, Boston, MA, USA
[4] Mass General Brigham Cancer Institute, Boston, MA, USA
[5] Institute for Systems Biology, Seattle, WA, USA
[6] Providence Swedish, Seattle, WA, USA
[7]University of Washington School of Medicine, Seattle, WA, USA
[8] Broad Institute, Cambridge MA, USA
[9] OHSU Knight Cancer Institute, Portland, OR, USA
[9] Boston University School of Medicine, MA, USA

Abstract

While large language models are powerful generators of new text, forecasting disease progression from longitudinal health histories remains a challenging problem. We introduce GenEHR, an autoregressive generative model trained on electronic health records (EHRs) from millions of patients in diverse health systems. The model learns from real-world multimodal patient histories, including diagnoses, medications, lab values, and procedures while explicitly representing irregular intervals between visits to forecast future clinical events. We combine the general-purpose patient representation learned during foundational training with parameter-efficient supervised adaptation for the task of risk stratification for 17 cancer types. In five large EHR cohorts, supervised adaptation substantially improved prediction performance of a first cancer diagnosis within a five year horizon window. Our retrospective results support the prospective evaluation of GenEHR-CancerRisk as a clinical decision-support tool for prioritizing high-risk patients for screening and prevention of aggressive cancer types, such as pancreatic and ovarian cancer.

*Co-senior, #Co-second
Correspondence: asif.khan@hms.harvard.edu sander.research@gmail.com

Introduction

Global challenge of early cancer detection.

Cancer is a major global health challenge, and the stage at which it is detected is one of the strongest predictors of outcomes, healthspan, and survival. Once malignancy progresses to an advanced stage, therapeutic options are limited and outcomes are worse compared to treatment at an early, localized stage[1]. Screening can shift diagnosis earlier, but population-wide testing, even when recommended (e.g., by USPSTF), is constrained by cost, availability of tests, heterogeneity of true risk among the screen-eligible population, and the burden of false positives or poor uptake[2]. An immediate task is therefore to identify patients whose future risk is high enough to justify further investigation or closer surveillance to preempt disease progression. This requires models that use routinely collected, widely available longitudinal data, update patient's risk as their health progresses over time and provide the information as clinical decision support.

EHR histories contain distributed signals of future cancer risk.

Structured electronic health records (EHRs) offer a rich, detailed view of a patient's health. EHRs include a wide range of multimodal information such as diagnosis codes, medications, procedures, laboratory measurements, demographic factors, geocoding, and lifestyle factors collected during standard clinical care. Early signals relevant to cancer may be weak individually and distributed over years, and their information value depends on sequence and context. For example, iron-deficiency anemia can be a signal for colorectal cancer, while jaundice, evolving abdominal symptoms, and changes in glycaemic control may occur before pancreatic cancer[3–7]. These clinical events occur alongside epidemiological risk factors such as age, smoking, and obesity[8,9], whose predictive value may depend on their timing and clinical context. Models trained on longitudinal EHRs can test whether these features add information beyond demographic or established clinical risk factors.

EHR models for cancer risk prediction and patient health state representation.

Existing supervised approaches to pancreatic, lung, gastrointestinal, and pan-cancer risk prediction have shown that models trained on longitudinal EHR data can be used to risk stratify patients months to years before the diagnosis[10–14]. However, when labeled data are scarce, supervised model performance can be sensitive to differences between cohorts and healthcare systems. Several aggressive cancers are rare–for example, pancreatic cancer has an annual incidence of approximately 14 per 100,000 people in the U.S.(all ages, SEER[15]). Unsupervised representation learning on broader EHR data can instead learn general features of patient health that can then be adapted to specific diseases such as cancer, especially rare ones.

Earlier methods such as Med2Vec[16], Deep Patient[17], multilevel medical embedding (MiME[18]) learn representations of clinical events, visits, and patient histories. More recently, transformer based architectures such as BEHRT[19], MedBERT[20], CEHR-BERT[21], TransformEHR[22], have shown the benefit of pre-training for downstream tasks such as mortality,

readmission, and risk prediction[19]. Recent approaches such as CLMBR[23] predict sets of future

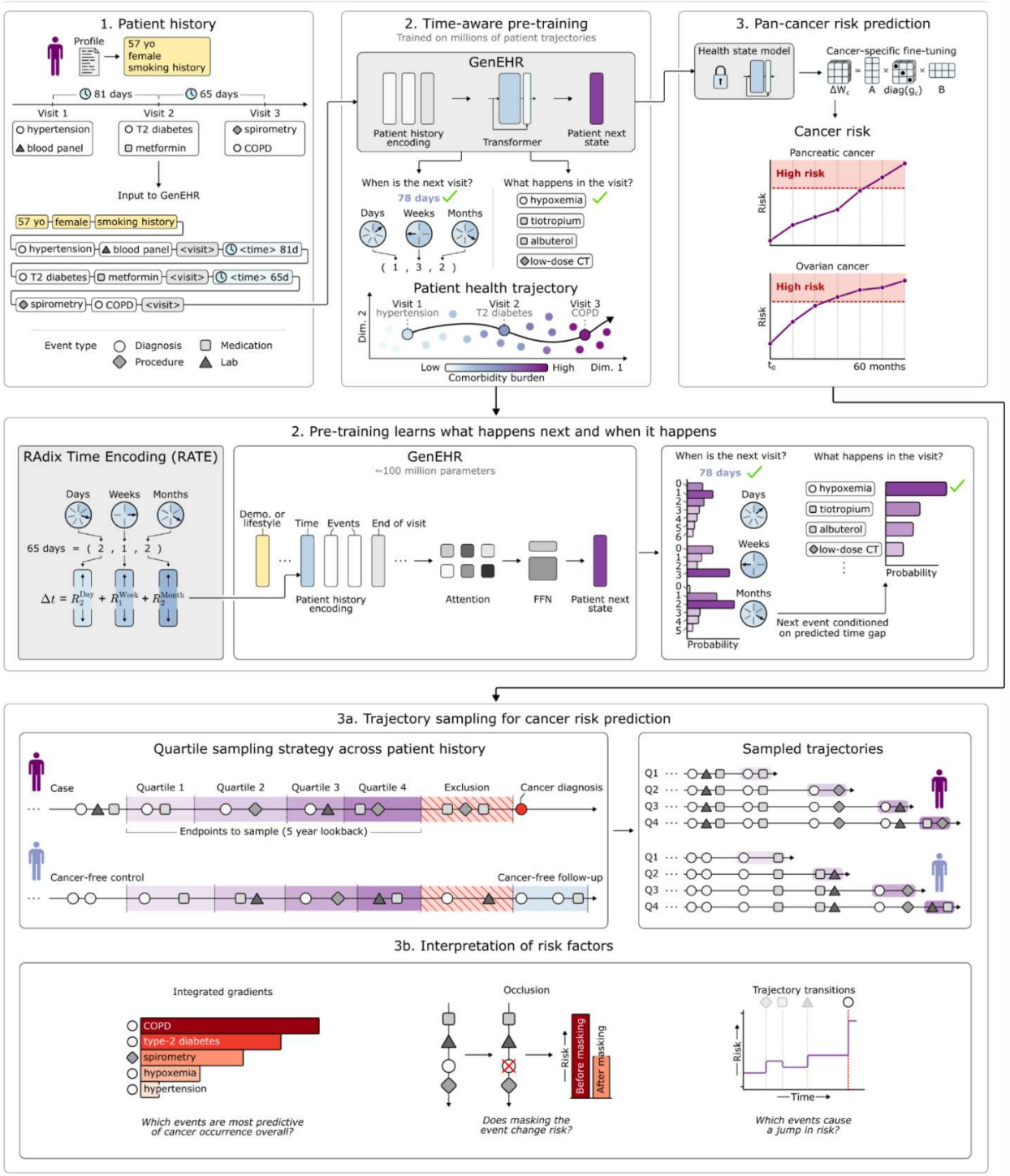


Figure 1: GenEHR: a foundation model of patient health state adapted for pan-cancer risk prediction. *1. Patient histories are tokenized as a combination of static (demographic and lifestyle) and dynamic (time-stamped events including diagnoses, medications, procedures, and laboratory results) variables. 2. GenEHR learns which clinical events occur next and after the end of visit, the five digits of the next inter-visit gap. 3a. Pre-trained GenEHR is combined with a piecewise exponential risk prediction head and cancer-gated low-rank adapters for pan-cancer*

risk stratification. For training and evaluation, the five year window preceding cancer diagnosis is divided into four equal size time bins and one trajectory endpoint is sampled from each bin, so as to assess risk at different points in a patient's history. Trajectories start at an earlier time and training and predictions are made at the trajectory endpoint. We exclude events recorded in a three month window before cancer diagnosis to reduce information leakage from diagnostic workup. For no-cancer controls we require five years of observed follow-up after the trajectory endpoint. 3b. Features contributing to predicted cancer risk. We use IG to measure the importance of each event within a patient trajectory (left). We select top medication and diagnosis events based on IG and and by masking them measure the change in cancer log-odds across the cohort (middle). For high risk patients, we sequentially score all sub trajectories to identify events that consistently raise or lower the cancer log odds (right).

codes, whereas Foresight[24], ETHOS[25], and CoMET[26] autoregressively generate future event sequences. CEHR-GPT[27] similarly uses autoregressive modeling to generate synthetic longitudinal EHRs. MOTOR[28] learns patient representation with a multi-task time-to-event objective but does not generate event sequences. SurvivEHR[29] models competing time-to-event distributions, while Delphi[30] models competing event rates; both can forecast trajectories. More recently, Apollo[31] combined longitudinal EHR with clinical text and images using masked auto-encoder training to learn patient representations, which were then applied to clinical forecasting tasks.

Inter-visit time intervals are essential for capturing the clinical signal.

Inter-visit time intervals carry information about a patient's health state. For example, a repeat blood test within weeks can signal an active health issue, whereas the same test after a year may reflect routine follow-up. Interleaving time tokens with clinical visits is a common approach, but existing methods generally encode time at a fixed resolution. Increasing temporal resolution expands the time vocabulary as well as context length. For example, representing gaps of up to ten years at day-level resolution[27] would require 3650 distinct tokens. Many exact gaps are sparsely observed, making it difficult to learn each token's representation and its relationship to other intervals. ETHOS[25] and CoMET[26] instead use a fixed vocabulary of interval tokens, but cannot flexibly capture fine-grained gaps. Such limitations can become problematic when precursor events cluster tightly before diagnosis or are separated by prolonged asymptomatic intervals. We introduce GenEHR (Fig 1) that decomposes each inter-visit time gap into a small set of digits at progressively longer time scales using RAdix Time Encoding (RATE). This allows the model to predict what happens next and when it happens at a day level resolution using a small set of time tokens.

Fine-tuning for cancer risk using registry-data and outcome-specific time-to-event modeling.

While generative models learn broad patterns of disease progression and healthcare practices, adapting them to a specific clinical task usually requires updating model parameters. Full fine-tuning is computationally expensive and can cause forgetting, where the foundation model drifts from its pre-trained state[32]. We therefore use parameter-efficient fine-tuning (PEFT)

with gated low-rank adaptation (LoRA), which freezes the pretrained model and learns a small set of adapter parameters in selected transformer layers[33]. We decompose each adapter into low-rank components shared across cancer types and a cancer-specific gate sets the strength of each component. Rare cancers can therefore benefit from the information in a shared direction instead of learning from their own few outcomes. We add multiple heads for pan-cancer risk prediction on top of GenEHR that learns to predict risk of all cancers simultaneously. Selected cancers are reported in Supplement (Tab S1). As diagnosis codes in the EHR are frequently recorded for the purpose of billing and reimbursement of tests and treatments rather than to actually confirm the presence of a diagnosis, and more than half of cancer diagnoses recorded are not confirmed by the registry (Fig S1c, S2c), we use the 'gold standard' cancer registry data for outcomes.

A unified framework for pan-cancer risk prediction.

We present GenEHR, a unified framework for generative modeling of patient trajectories and pan-cancer risk prediction for up to 17 cancer types (Table S1) using large-scale EHR data. We trained and evaluated cohort-specific GenEHR models using de-identified patient histories from Providence Health & Services (PROV), Mass General Brigham (MGB)[34], U.S. Veterans-Affairs (US-VA)[35], UK Biobank (UKBB)[36], and NIH All of Us Research Program[37]. Across these cohorts, we demonstrated broad applicability of GenEHR on a range of tasks such as conditional prediction, trajectory generation, time-gap fidelity, learned health-state representations, and contextual event relationships reflecting model belief about data. GenEHR was then adapted for pan-cancer risk prediction across pre-specified horizons using supervised time-to-event modeling. Beyond overall discrimination performance (AUROC), GenEHR clinical utility is demonstrated at a clinical operating point using positive predictive value (PPV, defined as the proportion of selected high-risk patients followed by a cancer diagnosis within the prediction horizon, standardized incidence ratio (SIR, defined as the ratio of observed cancer cases to those expected from age- and sex-specific population incidence over the same prediction horizon), and "suggested number to screen" (SNS, defined as the reciprocal of PPV). We interpreted key features contributing to model risk using contextual pointwise mutual information (cPMI), integrated gradients (IG)[38], and occlusion analysis. In summary, GenEHR is a time-gap aware foundation model of EHR specialized for targeted pan-cancer risk stratification, a step towards prospective clinical decision-making.



Figure 2: Datasets: large-scale multi-institutional EHR cohorts. *(A) Millions of patients: cohort size after site specific quality control filter (Supplement 1). (B-C) Age and sex distribution. Age distribution at the end of the patient record (left) and at first cancer diagnosis (right) indicates broad coverage. Age distributions were truncated above 89 to prevent patient re-identification. (D) Rich data: total number of unique data types and events across modalities (diagnoses, medications, procedures, and laboratory tests). (E) Diverse cancer prevalence by cancer type and cohort. (F) Broad geographic distribution of institutions in healthcare systems. Details in Supplementary Fig S1–S5.*

Results

Large-scale multi-institutional EHR cohorts.

We used longitudinal patient histories from five large-scale EHR data sources, including three U.S. health systems Providence Health & Services (PROV), Mass General Brigham (MGB), U.S. Veterans Affairs (US-VA)), and two national research cohorts, namely the U.K. Biobank (UKBB) and the NIH All of Us Research Program (AoU) (Fig 2). Access to patient data in secure protected computing environments was regulated by IRB protocols at MGB, US-VA, and PROV. Data harmonization, tokenization of clinical events, cohort-specific inclusion criteria, observation periods, and detailed data statistics are in the Supplement Methods.

Cancer outcomes. MGB, US-VA, PROV, and UKBB used linked cancer registries to avoid known inaccuracies resulting from billing practices[39]. Registry records serve as the gold standard as they are confirmed by registrar teams based on pathology reports and other clinical records. For AoU, which does not have linked cancer-registry data, we require at least two occurrences of cancer-specific ICD codes to define cancer cases.

Generative foundation model and multi-institutional training (GenEHR; Figure 1)

We represent each patient as a unified sequence of static (time-independent) and longitudinal (time sequence) tokens. The static prefix consists of demographic and lifestyle factors represented as discrete tokens. All events on the same date are grouped into a visit block, and we use a visit token to mark the end of a visit and a $< rate_gap >$ token to represent the time from that visit to the next visit. All token embeddings except padding and static prefix are learned as a part of the training process.

RATE learns intervisit intervals at day level resolution.

The time gap between visits in EHR data is irregular and influences the model representation and interpretation of patient health. For example, a repeat blood test within a week of an abnormal result may indicate active illness, whereas a test one year apart can be a part of a routine annual checkup. We therefore encode each inter-visit interval, using a mixed-radix digit encoding that is designed to capture the essence of different time scales. The first digit represents 1-7 days, the second counts 1-4 seven day periods, and subsequent digits represent progressively longer intervals. Using radices $(7,4,6,5,5)$ gives place values of $(1,7,28,168,840)$ days. For example, a gap of 75 days is encoded as $(5,2,2,0,0)$ because $75 = 5 \times 1 + 2 \times 7 + 2 \times 28 + 0 \times 168 + 0 \times 840$. Thus RATE preserves day-level resolution without requiring no-event[30] or a separate token for every possible interval[25]. Similar to events, all digits embeddings are learned. The composition of digit embedding is used at the respective $< rate_gap >$ position. Full details of tokenization in Supplement Methods.

Model training, inference, and evaluation process.

We train the model independently within each cohort using a combination of an autoregressive next event objective and a time gap objective. RATE heads predict the time gap

digits from the hidden state of the preceding visit token. During generation, we sample event tokens autoregressively from the event head. After a visit token is sampled, its hidden state is given to the RATE heads to sample the time digits. Their embeddings are then composed and inserted at the $< rate_gap >$ position, allowing events to be generated conditional on both the clinical history and the sampled elapsed time. Full details on tokenization and model formulation are in Supplement.

We evaluated predictive and generative performance together with the clinical structure of the learned representations. Conditional prediction asks whether held-out next events are ranked highly from an observed patient context at the end of a partial trajectory. Generation starts with a randomly sampled static-token (starting points), samples subsequent event tokens and RATE time gaps, and then comparison is done by comparing prevalence of events and time-gaps in generated and real data. We further examined whether the learned event and patient representation were organized by clinical category and disease burden. We also constructed a clinical belief graph using a cPMI matrix. Using context-dependent shifts in probability distributions for next event in subsequent visits, cPMI identifies recurrent, highly probable associations among clinical events within a cohort of held-out patients. Detailed evaluation scores are in the Supplement Methods with Table S1 providing the role and interpretation of each score. PROV and US-VA each used a fixed held-out retrospective cohort of one million patients, MGB used 500,000, All of Us used 50,000 and UKBB used 75,000 patients. The remaining cohort is further divided into training and validation sets. Final evaluation metrics were computed only on the retrospective held-out cohort.

We also carried out a scaling analysis by training multiple model architectures with increasing parameter counts and evaluating them on a common held-out dataset. Larger model size leads to a lower event perplexity, high next event accuracy and cancer-risk AUROC (Fig S14 MGB). Because of the computational cost, this analysis was limited to the MGB cohort.

GenEHR foundation model learns contextual patient representations useful for clinical forecasting and hypothesis generation.

We evaluated GenEHR on a range of tasks such as next visit forecasting, transitions in health state during disease progression, and downstream pan-cancer risk prediction.

Conditional prediction, calibration, and trajectory generation.

We evaluated next-visit retrieval at every eligible visit endpoint in the held-out patient trajectories, with results pooled across endpoints. Across cohorts, retrieval-AUC for next event in next visit, defined as the area under recall curve as a function of the fraction of the event vocabulary searched, ranged from 0.90 to 0.99 (Fig 3B). For the first occurrence of broader clinical categories such as the ICD chapter, one-vs-rest macro-AUROC across categories ranged from 0.98 to 0.99 (Fig 3B). Next, an autoregressive simulation tested whether the model captured the overall prevalence of events and time-gaps in observed data. Event prevalence in generated trajectories reproduced the observed prevalence in real-data with log-Pearson

A **How does GenEHR perform across evaluation metrics and cohorts?**

Metric	VA	Providence	MGB	UKBB	AllofUs	Range	Better
Top-1 retrieval accuracy	0.60	0.43	0.66	0.60	0.42	0–1	↑
Top-5 retrieval accuracy	0.78	0.65	0.85	0.72	0.61	0–1	↑
Retrieval MRR	0.68	0.53	0.75	0.66	0.51	0–1	↑
Event perplexity	33.34	36.90	6.75	109.58	73.16	≥1	↓
Mean token entropy	2.05	2.75	1.54	2.06	3.24	0–7.8 nats	↓
Median token entropy	1.97	2.74	1.55	2.02	3.23	0–7.8 nats	↓
Generation self-consistency	—	0.32	0.34	0.41	0.28	0–1	↕
Expected calibration error	0.14	0.08	0.07	0.11	0.13	0–1	↓
Age recovery, mean-pool probe (R^2)	—	0.96	0.90	0.53	0.84	≤1	↑
Effective rank (EOS embedding)	416	649	669	80	244	1–768	↑
Co-occurrence fidelity (Pearson-log)	0.92	0.95	0.97	0.53	0.90	0–1	↑
Utility TSTR AUROC (macro)	0.78	0.66	0.81	0.61	0.63	0.5–1	↑
Utility TRTR AUROC (macro)	0.77	0.77	0.81	0.70	0.79	0.5–1	↑

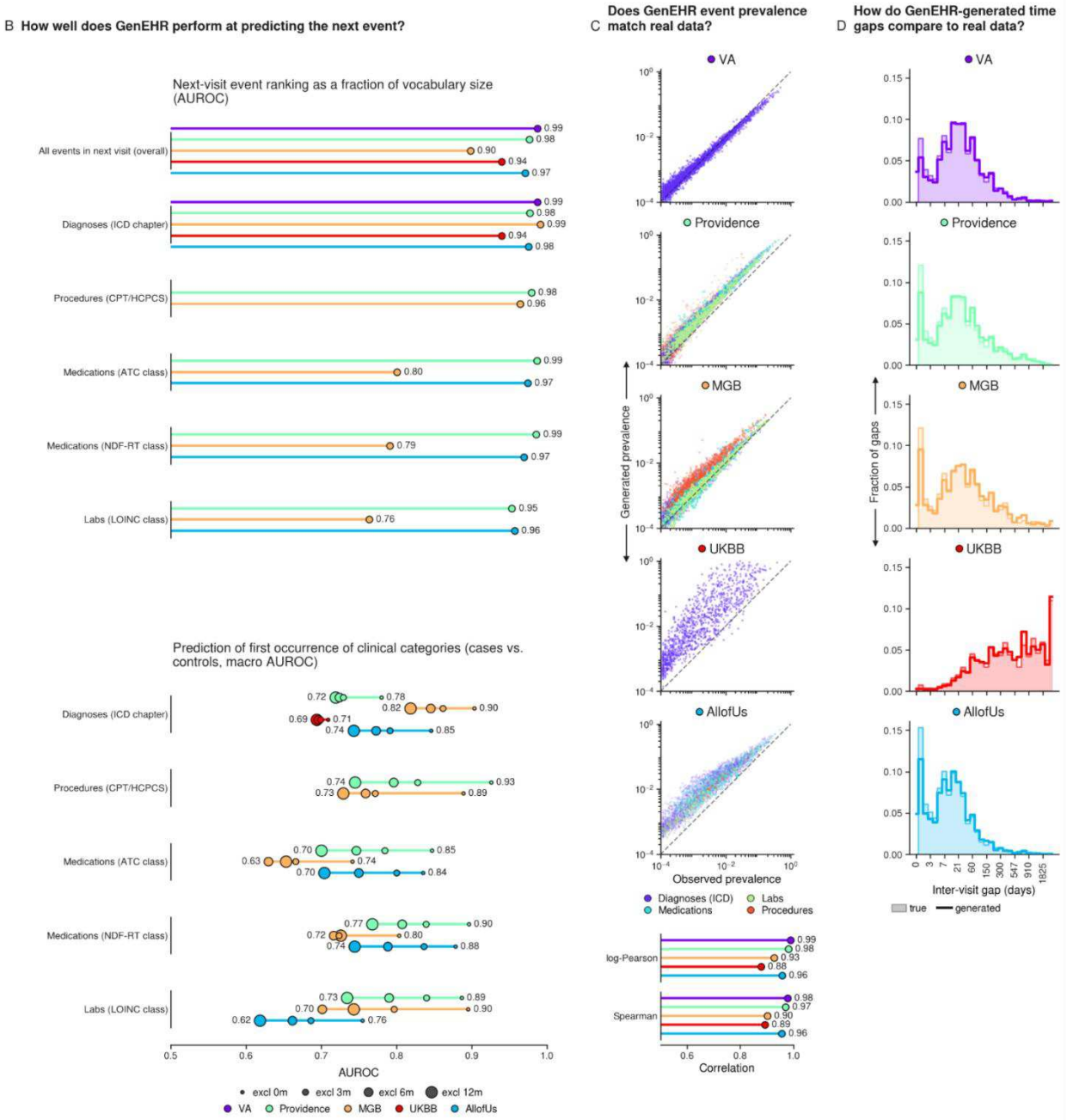


Figure 3: GenEHR foundation model learns contextual patient representations useful for clinical forecasting. *(A) Model performance of general forecasting on held-out patient histories.*

For any patient trajectory, GenEHR predicts the next clinical visit after a time gap and recovers events in the visit. Details of evaluation metrics (MRR, TSTR, TRTR) in Supplement Methods. (B) Top: next visit event recovery assesses the model ranking of a correct event as a fraction of the event vocabulary. Bottom: For prediction of next-event clinical categories, we aggregate event probabilities within each clinical category and evaluate performance using one-vs-rest AUROC scoring. Exclusion window ('excl 0-12m'): only evaluate trajectories without events in a certain time interval before the first occurrence of the clinical category. (C) Assessment of autoregressively generated trajectories: comparison of the prevalence of events in generated trajectories with the prevalence of events in the observed data; and log-Pearson and Spearman correlations of generated vs observed. (D) Distributions of observed and model-generated inter-visit time gaps. Together, (C-D) indicate how well events generated by the GenEHR model capture population prevalence of both clinical events and the temporal structure of patient histories.

correlation of 0.93 to 0.99 and Spearman correlation of 0.90 to 0.98 (Fig 3). Among the hundred most prevalent observed codes pairwise co-occurrence was the fraction of trajectories in which both codes appeared at least once, regardless of whether they occurred in the same visit. Its log-Pearson-correlation was 0.92-0.97, showing that generated trajectories preserved broad code-pair associations in addition to individual frequencies. Generated inter-visit time gaps also correlated strong correlation with observed time gaps, with Spearman correlation of 0.89 to 0.98.

Mean entropy of the next-event distribution quantified predictive uncertainty and ranged from 1.54 to 2.77 nats across cohorts. Lower entropy means that the probable next event is narrowed to fewer choices, whereas higher entropy means several events are plausible and the model is less certain which will occur. We used expected calibration error (ECE) to assess the reliability of next-event probabilities which ranged from 0.07 to 0.14 across cohorts. To assess how far in advance the model could predict ICD-chapter onset, we excluded events during 3-, 6-, and 12- month windows preceding chapter onset (bottom; Fig 3B). Performance suggests that the model learns from information present earlier in the patient trajectory rather than relying solely on events immediately preceding cancer diagnosis. We also carried out an ablation of time encoding by comparing RATE with ETHOS (a fixed set of time-gap tokens based on prior knowledge), relative-time encoding (a vocabulary with one token for each day up to ten years), no-time encoding (time tokens removed), and no-time-and-visit (both time and visit tokens removed). GenEHR with RATE performs best for predicting the next event in the next visit with the largest improvement in forecasting at longer time gaps (Fig S16b).

Across cohorts, GenEHR learned event and time-gap structure of patient histories with good accuracy and calibration, making it suitable for a broad range of downstream clinical tasks.

GenEHR learns patient health state progression.

We next asked whether the learned representations encode clinically meaningful differences between patients. We extracted patient-level representations from the penultimate layer of the model (using <eos> token) trained on the MGB cohort and visualized them in two dimensions using UMAP[40]. UMAP projection showed smooth gradients in Charlson comorbidity index (CCI)[41], with higher disease burden histories mapped to nearby regions (bottom left; Fig 4). Because the representation using <eos> token summarizes overall health but not its evolution,

we separately extracted the representation at each visit state. All visits across patients were projected into two dimensions and were connected using averaged displacement vectors within local regions of the UMAP (bottom right; Fig 4). The resulting visit path moved toward regions with higher average CCI and chronic-disease burden. This map shows a descriptive representation of longitudinal health trajectory over time. Results on other cohorts (MGB Fig S7; PROV S17, S19D; VA S25, S27D; UKBB S33, S35D; AoU S41, S43D).

Learned clinical event structure supports events-to-cancer hypothesis generation.

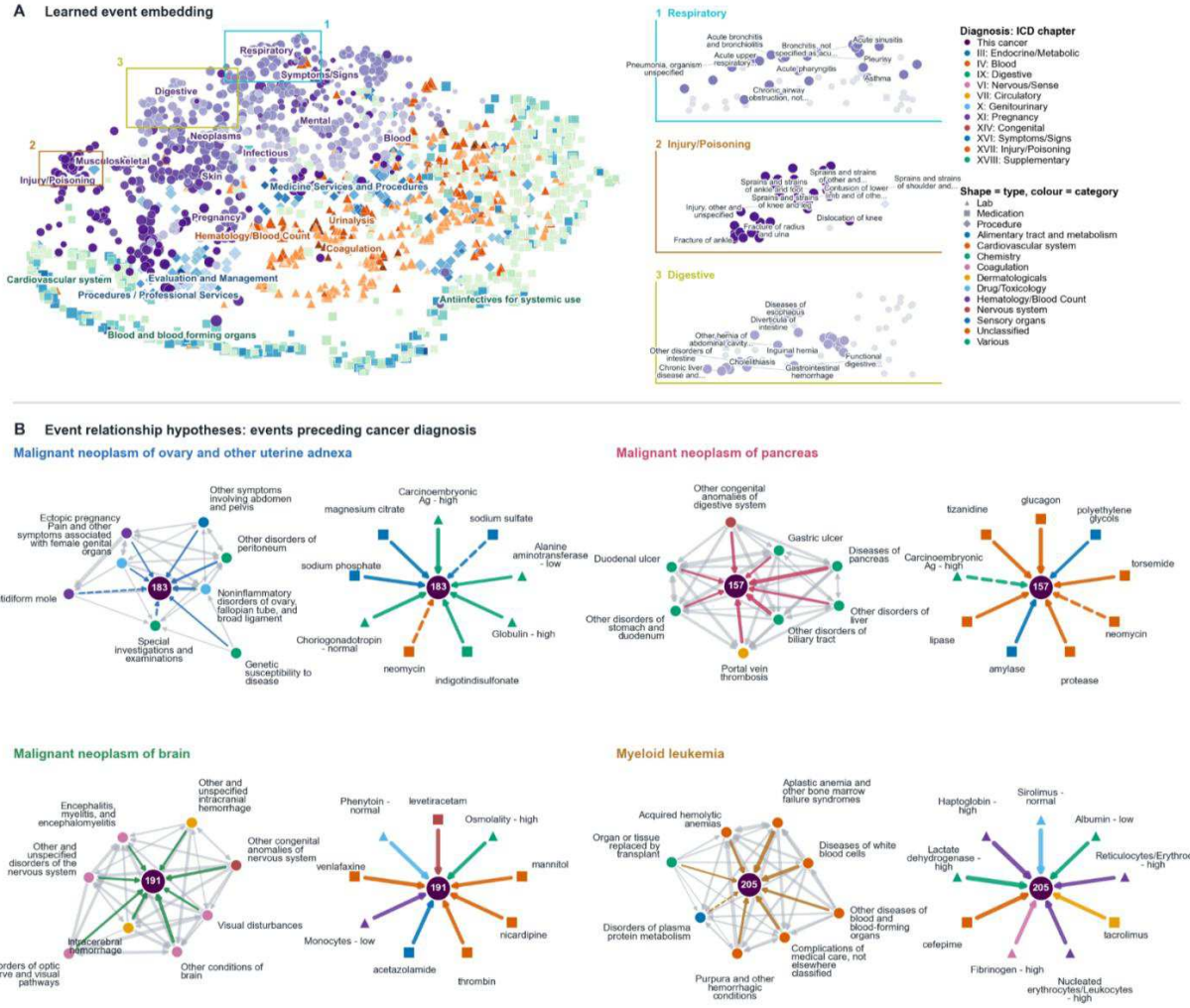
We first evaluated whether the event representations learned by GenEHR captured any information structure. We projected the learned event embeddings using UMAP (Fig 4A) and graphically coloured the events by ICD chapters, medication ATC class, lab assay group, and CPT procedure section. Although no ICD hierarchy information was used during training, events clustered by clinical categories with refined substructure, e.g., three magnified regions show separation of acute from chronic respiratory disease, sprains from fracture, and upper gastrointestinal symptoms from hernia and hepatobiliary disease.

Next, we used GenEHR generation to construct a model-derived belief graph of event relations from held-out patient trajectories preceding cancer diagnosis. We excluded events on or after cancer diagnosis and within three months before it, to reduce prodromal signals. At each visit boundary, we used the clinical context through the current visit including an event A and the time gap to compute the probability distribution over the next event B. We then averaged the probability distribution over next event B across all positions and patients rather than looking at the most likely prediction. This conditional average was compared with the overall probability assigned to B across all contexts. A positive cPMI(A→B) therefore means that the model assigns a higher probability to B after A than expected from how often B generally occurs.

We mapped cPMI to a belief graph keeping sufficiently strong associations (Supplement Methods). Some edges in the graph reflect clinical prior knowledge (MGB Fig 4B, Fig S9a-c; PROV S19, S20a; VA S27, S28a; UKBB S35, S36a-b; AoU S43, S44a-b). Disease of the pancreas and disorders of biliary tract are top associations to pancreatic cancer followed by disorders of liver and gallbladder. These associations are linked to pancreatic inflammation, blockage of bile flow and may lead to the diagnosis of a previously undetected pancreatic tumor[42,43].

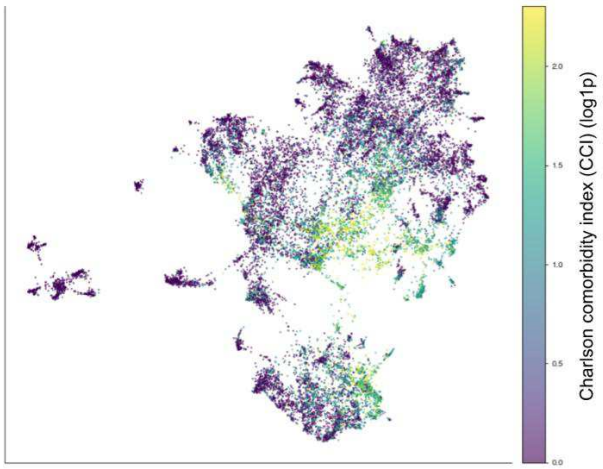
The brain cancer neighborhood captures both broad brain disease and treatment context. Visual and optic-nerve disorders and haemorrhage cluster with anti-seizure medicines (levetiracetam and phenytoin), and treatment as well as monitoring for brain swelling and increased intracranial pressure (mannitol and serum osmolality)[44,45]. By contrast, the myeloid-leukaemia neighborhood reflects abnormal blood-cell production and its diagnostic workup. Aplastic anaemia and other bone-marrow failure disorders are top preceding signals alongside white-cell abnormalities, purpura, elevated lactate dehydrogenase, and nucleated erythrocyte and reticulocyte measurements. This pattern is consistent with bone marrow

Event representations capture clinical structure and a context-dependent model-belief graph for hypothesis generation.



Patient representation over multiple visits captures comorbidity burden.

Patient map (end-of-trajectory) with co-morbidity gradient



Patient health state progression over visits

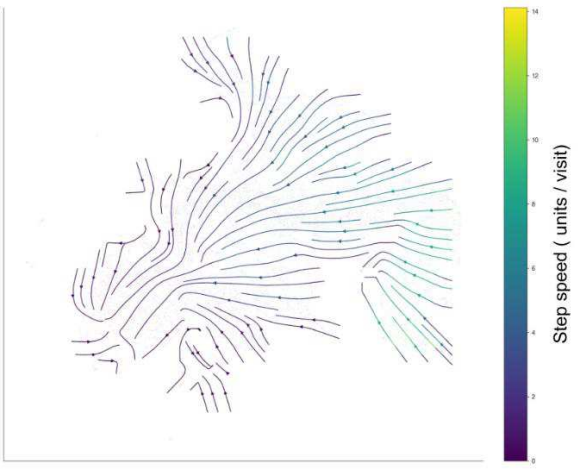


Figure 4: For the MGB dataset, other cohorts in Supplement. (Top) *Left:* ***(A)*** *UMAP projection of learned event embeddings using cosine distance between embedding vectors. Right: closeup of three regions numbered 1 to 3 shows GenEHR learns hierarchy of clinical events without explicit prior knowledge and without using text.* ***(B)*** *Event relationship preceding cancer diagnosis inferred using contextual pointwise mutual information (cPMI) between pairs of events for selected cancers (ICD codes 183, 157, 191, 205). For held-out patients whose preceding visit contains an event e_1, cPMI compares the average model probability assigned to e_2 normalized against the model averaged probability of e_2 across all positions and patients. In the cPMI belief graph nodes are clinical events and edges are informational links. Left: Diagnosis neighborhood of cancer with edges among preceding events, Right: Cross-modality hub - medications, procedures, and labs preceding cancer diagnosis. Dashed edges: events with weaker mutual information scores that are plausibly less specific to that cancer. (**Bottom)** UMAP neighborhoods of patients using their end-of-trajectory <eos> representation reflect gradient of comorbidity burden (left), and transitions along trajectories are informative about how the learned state of patients evolves over time using their <visit> representation over multiple visits. Flow (arrows) indicates how patient health evolves over visits (right) averaged over groups of held-out patients.*

dysfunction and increased cellular turnover. Aplastic anaemia may represent an alternative diagnosis or a condition that later progressed to myelodysplastic syndrome or acute myeloid leukaemia[46].

Since edges in cPMI represent learned model belief about the cohort they do not imply causality (Fig 4B). This model-belief graph can be constructed for any retrospective cohort, e.g., in different health systems, to analyze event relationships within patient context. The inferred edges are hypotheses learned from temporal patterns in clinical records, which may reflect disease progression, diagnostic work-up or healthcare use and not, or not necessarily, causal relationships. It is reassuring that the edges, derived in a data-driven mode, appear to agree with prior knowledge in many cases. However, the edges not explained by clinical practice may generate hypotheses for further epidemiological or experimental investigation.

Parameter efficient fine-tuning improves pan-cancer risk prediction(GenEHR-CancerRisk)

Generative pre-training learns how diagnosis, medications, lab tests, procedures, and time intervals change the probability of future cancer events, including cancer-associated events. However, the next-event objective does not directly estimate the time to a first cancer diagnosis. We therefore adapt the pretrained GenEHR model for pan-cancer risk prediction with a time-to-event objective that uses each patient's observed history and the period over which a new diagnosis could be observed. Selected cancer names and their definition are included in the Supplement (Table S1).

Trajectory sampling prevents data leakage and allows learning from earlier clinical histories.

We sampled multiple endpoints per patient to learn from earlier clinical histories. Within a five year window ending at the last eligible endpoint before cancer diagnosis, we divided time ordered patient trajectory into four approximately equal sized segments and sampled one cutoff

from each segment. To reduce the influence of immediate diagnostic work-up, we excluded events recorded during the three-month before the first cancer diagnosis. For cancer-free controls a reference date is assigned by randomly sampling from observed calendar dates of cancer diagnoses dates within each control's eligible observation window. Prediction endpoints for controls were then sampled from the preceding history with requirement of at least five years of subsequent follow-up. Each retained trajectory contained at least three clinical visits.(Fig 2b; Supplement Methods).

Fine-tuning using gated low rank adapters trained with time-to-event loss.

We introduce a pan-cancer risk head based on a piecewise-exponential model to estimate cancer risk in different horizons ending at 6, 12, 24, 36, and 60 months. For each cancer and horizon, the head combined the GenEHR cancer-token logit with a horizon-conditioned patient state. To obtain this state, we appended a visit token and the RATE encoding of the horizon endpoint to the observed trajectory and passed the augmented sequence through the GenEHR. The head is parameterized to learn a horizon baseline shared across cancers together with a cancer-specific deviation. The shared component stabilizes risk estimates for cancers with fewer outcomes, while the deviation captures cancer-specific rates. The interval specific rates are then integrated over observed at-risk time in the piecewise-exponential likelihood (Fig 2c; Supplement Methods).

We introduced LoRA[33] to combine shared representations with cancer-specific gates. LoRA specializes the model for a specific task by training a small set of additional parameters while not updating parameters of the pre-trained model. Specifically, the transformer backbone of GenEHR is frozen (not updated) except the additional inserted gated LoRA modules and the pan-cancer risk head trainable parameters. Each module learns a set of low-rank directions shared across cancers and a signed cancer-specific gate for each direction. This approach allows rare cancers with fewer outcomes to benefit from shared representation changes supported by other cancers while centered cancer deviations keep cancer specific information. (Fig 2c; Supplement Methods).

GenEHR-CancerRisk improves discrimination performance across cancer types.

We compared three approaches to pan-cancer risk scoring: (1) direct scoring using the frozen GenEHR cancer-token logits, (2) supervised training of only a pan-cancer risk head over the frozen backbone ('head only'), and (3) supervised joint fine-tuning of the risk head and cancer-type-specific gated LoRA parameters over the frozen backbone. All comparisons were carried out on the same patient trajectories for 17 cancer types in MGB, 17 in PROV, 13 in VA, 8 in All of US, and 8 in the UKBB, at prediction horizons of 6, 12, 24, 36, and 60 months after the end of the trajectory. In MGB, macro-AUROCs across horizons (see Methods) ranged from 0.71 to 0.77 for direct scoring, 0.80 to 0.84 for head only, and 0.80 to 0.84 for gated LoRA. The respective ranges were 0.60 to 0.67, 0.72 to 0.77, and 0.73 to 0.78 for PROV; 0.67 to 0.71, 0.80 to 0.81, and 0.82 to 0.83 for VA; 0.58 to 0.85, 0.67 to 0.82, and 0.68 to 0.84 for All of US; 0.54 to 0.80, 0.60 to 0.92, and 0.62 to 0.92 for UKBB. All individual cancer performances are in Fig 5.

Direct scoring showed that GenEHR alone can discriminate reasonably well between patients with and without cancer, without cancer-specific supervised time-to-event training. The substantial improvement in risk scores after fine-tuning showed that explicit supervision on diagnosis timing adds information beyond unsupervised direct scoring. Gated LoRA fine-tuning achieved the best overall discrimination across cohorts and prediction horizons (60-months horizon, Fig 5). Results for other horizon windows are included in the supplement (Fig S6a-d). Comparisons with alternative baseline methods on the MGB cohort are in Fig S16. The pan-cancer risk model is well-calibrated (Fig S12, S23, S31, S39, S47). We also report the breakdown of performance by age-bracket and sex (Fig S10b, S21b, S29b, S37b, S45b) together with sensitivity analysis of age-matched and length matched cases and controls (Fig S10a, S21a, S29a, S37a, S45a). A fine-tuned model performs best in prediction performance estimated on withheld data across all comparisons. For cancer with low prevalence, the model relied more on shared representations, while cancer-specific gates contribute more for cancers with more data(Fig S15).

Focus on highest-risk subgroups to facilitate targeted pre-cancer screening.

AUROC assesses prediction accuracy integrated over all operating points but does not tell us how well a model nominates the highest risk cohort for a realistic screening program. This requires choosing an operating threshold that minimizes false positives and maximizes true positives, and also accounts for capacity of the healthcare system and burden of follow-up testing.

We therefore evaluated SIR and SNS at a clinically relevant operating threshold N, typically a small number compared to the size of the population for high-risk focus (Fig 5). A lower SNS implies that fewer patients would need to be evaluated through an AI-informed screening program in order to find one cancer. This number differs from the standard ‘number needed to screen’ (NNS)[47] metric that assesses the strength of a given screening modality on its ability to reduce cancer mortality. The traditional NNS quantifies not only the sensitivity of a specific screening platform for early stage detection of a specific type of cancer in a specific screening cohort but is also meant to assess the ability to cure that cancer once detected. In contrast, our model estimates the risk enrichment for cancer diagnosis, but not the ultimate effect of therapeutic outcomes. Our suggested SNS estimates also do not take into account the performance of the clinical tests to be used in a newly configured screening program, as the datasets used for model training typically have only limited information about the clinical workup and no information about the number of patients seen with suspected cancer. One can use the suggested number to screen (or the model’s PPV at a given operational threshold) to compare different models, models applied to different cohorts and as a guide for the design of high-risk screening programs. SIR compares observed cancer cases in the high-risk cohort (as predicted by the model, evaluated on the withheld test set) and the expected number of cancer cases based on population wide incidence for a given age and sex. Higher SIR means better enrichment of high-risk cohorts compared to a baseline estimate based on population incidence.

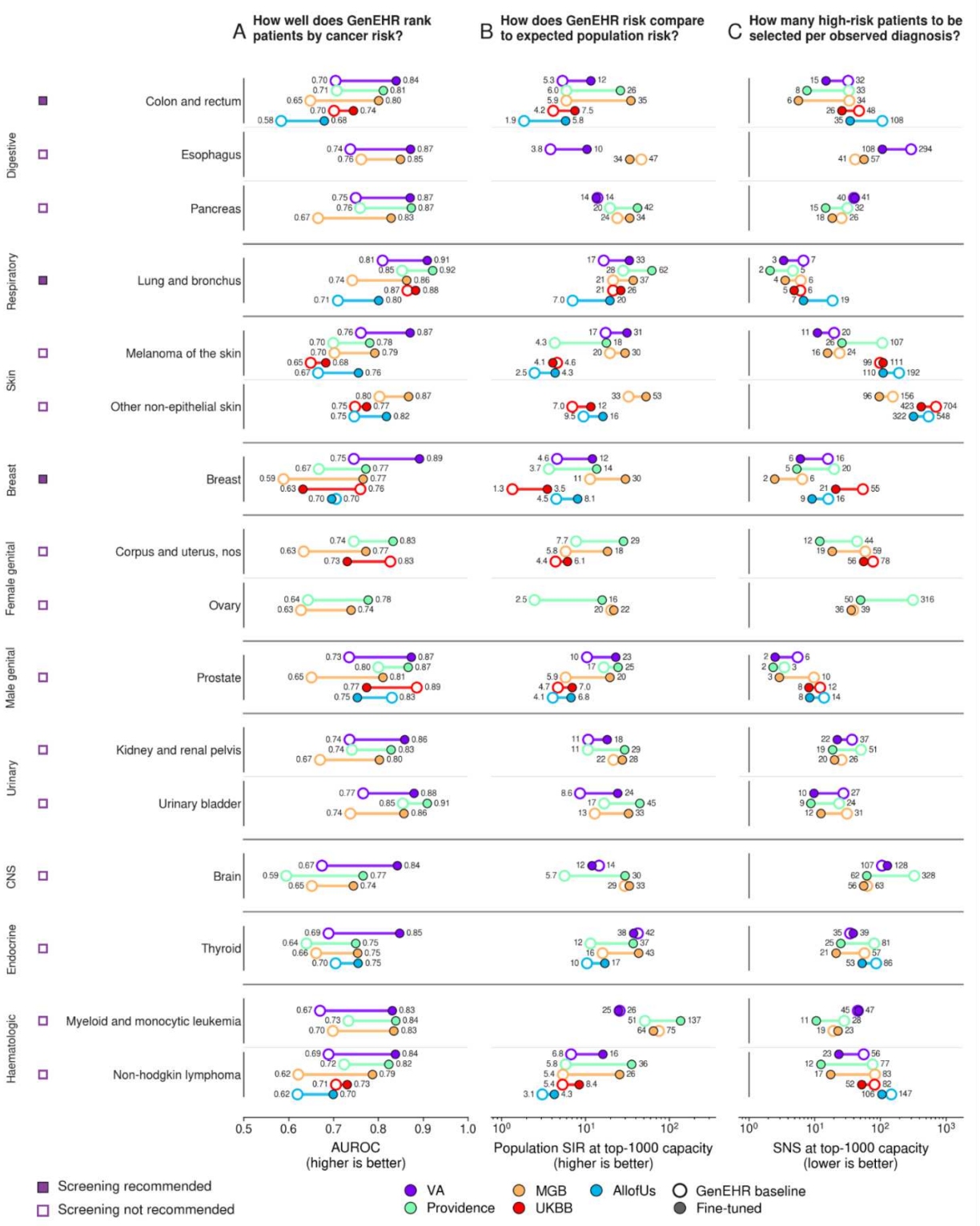


Figure 5: Performance of pan-cancer risk prediction (60 month prediction horizon). *(A) Comparison of prediction scores from just the pretrained GenEHR with that from fine-tuned GenEHR for 19 cancer types in 5 health system datasets. Pretrained GenEHR learns cancer-*

associated signals useful for ranking cases and controls without any supervision from the cancer registry. Fine-tuned GenEHR uses a time-to-event objective, which improves AUROC. (B) SIR (standardized incidence ratio; higher is better) quantifies the increase in prediction performance for the 1000 highest scoring trajectories relative to the population-wide demographic baseline incidence for a given age&sex subpopulation (SEER data[1]). Before computing SIR we adjust for differences in cancer prevalence of respective cohorts and real data (C) SNS (number suggested to screen for one detection; lower is better) here quantifies the estimate as to how many patients are to be enrolled for high-risk screening to find one cancer case. The estimate is based on training on actually detected cancer cases and evaluation on withheld test sets. In new screening programs the number of detected cases will of course depend on the performance of the screening modalities, which may be higher or lower compared to that of clinical work-up in the training sets. Cancers shown are those with 2000 or more patients in at least two cohorts. Results on other prediction horizons (6,12,24, and 36 months) are in Supplement Fig S6a-d.

At an operating threshold of N=1000 highest-risk patients, we report cancer-specific SIR, and SNS, across health system cohorts (Fig 5). At the 60-month prediction horizon, SIR, and SNS ranged from 18-224, 2-96 in MGB; 6-224, 2-136 in PROV; 3-91, 6-299 in VA; 3-27, 5-570 in UKBB; and 4-20, 6-322 in All of Us, respectively. For pancreatic, ovarian, and lung cancer, across cohorts SIR ranged from 34-48, 6-22, and 20-57, while SNS ranged from 13-18, 36-136, and 2-7 respectively. Because cancer incidence varies across cancer types and cohorts, we report SIR and PPV as a function of N for all cancer types across sites (Fig S11a-c; S22a-c; S30; S38; S45a-b). The SIR values are substantially higher than standard risk models, even those developed using curated data in specialized cohorts.

For nominated patients with correctly predicted cancer occurrence in the withheld test set, we further look up the time interval between the last clinical event in the observed patient history used for the prediction and the actual recorded cancer diagnosis date in the registry data. We report the distribution of these real-data intervals across cancer types and cohorts (Fig S6e-f). Many flagged cases might have been identified several months to years prior to actual diagnosis. At the 60-month horizon and an operating threshold of N=1000 across cancer types, median lead time ranged 5.5-17.6 months and max lead time 32.6-53.3 months in MGB; median 8.1-26.8 months and max 21.4-58.5 PROV; median 15.2-29.8 months and max 45.2-59.9 months in VA; median 18.9-24.2 months and max 53.2-59.9 months in UKBB; median 10.7-20.2 and 43.2-60 months in All of Us. For highly malignant cancer such as pancreatic cancer the median/max was 17/59.5 months in MGB, 9.3/29.8 in PROV, 15.9/45.2 months in VA. For ovarian cancer, 8.9/31.1 months in MGB and 26.8/47.1 in PROV. These results suggest the model may learn signals corresponding to future cancers which potentially were missed by standard care, suggesting the possibility of significantly widening the detection window, which may allow some patients to be diagnosed at an earlier stage of disease.

Interpretation of features contributing to cancer risk.

We carried out a three-step interpretability analysis (Supplementary Methods) to identify clinical events contributing most to cancer-risk prediction. The analysis used the fine-tuned model trained with a three-month exclusion and evaluated the 12 and 60-month prediction horizons. First, we used IG to measure the contribution score of each event to cancer log-odds. Rather than

using typical <pad> as the IG baseline, we used trajectories from multiple control patients. For each token in a given trajectory, we computed its attribution relative to each control baseline and averaged the scores. We then ranked events within each trajectory and recorded how often each event appeared in the top decile of attributions among true positive patients at the selected N=1000 operating threshold. A set of frequently contributing events shared across multiple cancers (Fig 6A;MGB) and cancer-specific events for pancreatic, ovarian, and lung cancer (Fig 6B;MGB). Second, we tested model sensitivity by masking top ranked medication and procedure tokens one at a time and recomputing log-odds. Finally, transition analysis assessed how often the predicted cancer log-odds increased or decreased when a medication or procedure code was recorded along the trajectory, quantified by $(n_{up} - n_{down})/(n_{up} + n_{down})$.

Across cancer types, cardiac dysrhythmias, ischaemic heart disease, hypertension, diabetes, lipid disorders, malignancy-screening encounters, abnormal images, and respiratory symptoms are some of the events that increase cancer log-odds (Fig 6A;MGB). For pancreatic cancer, top IG features include high glucose, intestinal disorders, diabetes, lipid disorders, and pancreatic disease contribute most to increasing log-odds and these patterns have been reported in patients with pancreatic cancer[6]. Masking abdominal or thoracic-computed tomography lowered the log-odds of pancreatic cancer at both 12 and 60 month horizons, which indicates imaging probably contributes to the risk score as a diagnostic workup for suspected cancer. Hydrochlorothiazide was associated with upward score transitions for pancreatic and ovarian cancer. Previous studies have associated thiazide use with ovarian cancer incidence, including a possible histotype-specific association[48,49]. In pancreatic cancer patients stopping antihypertensive treatment has been observed before cancer diagnosis[50] (Fig 6B;MGB). Since EHR data only records when hydrochlorothiazide appears in the trajectory, we cannot determine whether this reflects the drug, or a change in treatment. Ovarian cancer IG features included menopausal disorders, diabetes, family history, and general comorbidity (Fig 6C;MGB). For lung cancer, respiratory symptoms, pneumonia, pleurisy, and other lung disease appear as top events, while chest computed tomography and radiography show larger log-odds with increased risk (Fig 6D;MGB). Detailed results for other cancer types are in Supplementary Fig S13a-S13f and for other cohorts PROV in Fig S24, VA in Fig S32, UKBB in Fig S40, and AoU in Fig S48. Interpreted isolated features should be interpreted with caution, as they actually function in the full context of health trajectories and should not be considered causative. Some of the reported features do agree with prior knowledge, while others provide a basis for hypothesis generation which could be further investigated through causal inference and subsequent clinical trials of preventative interventions.

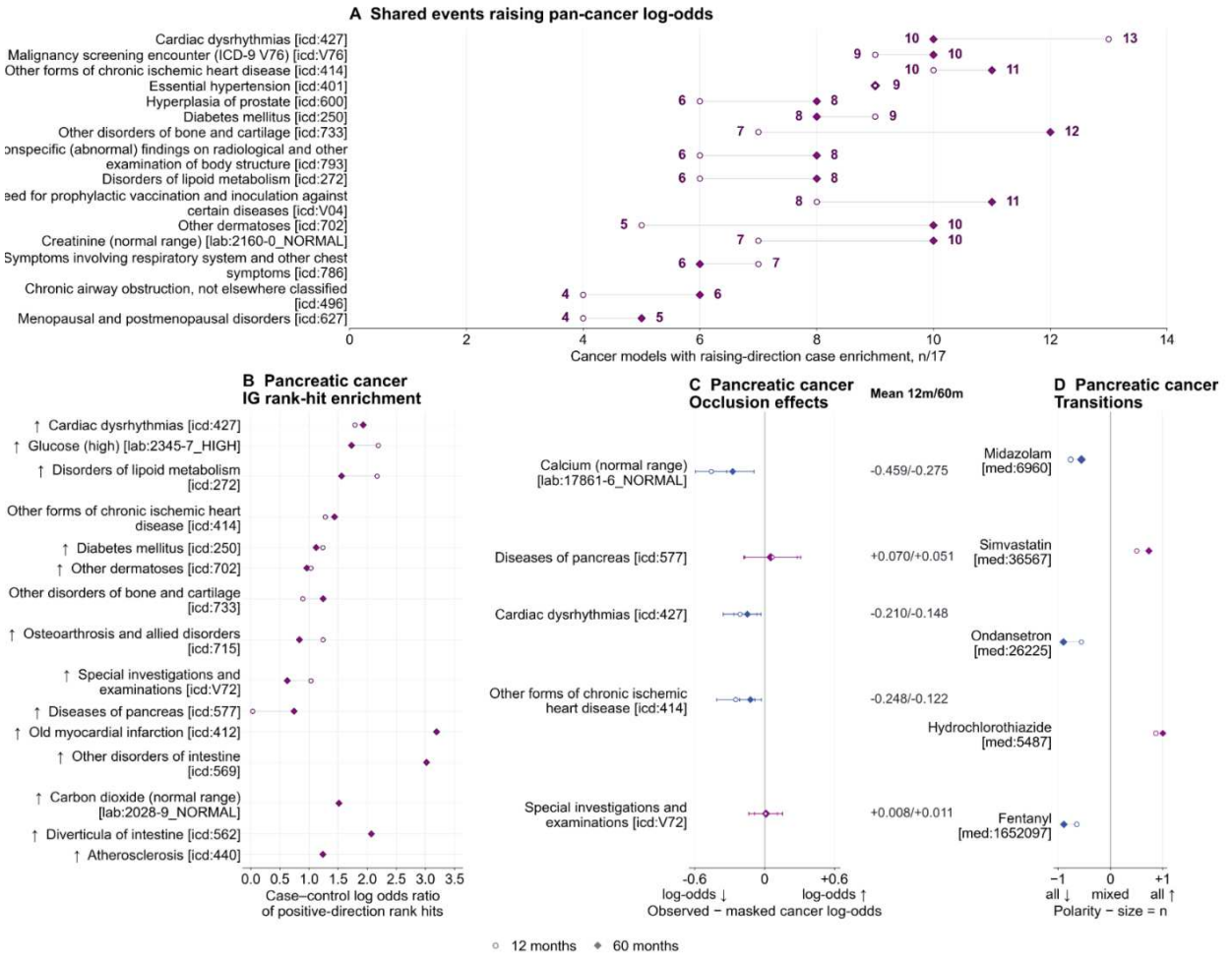


Figure 6: Interpretation of clinical events contributing to cancer-risk predictions in MGB. *IG scores computed for the 12- and 60-month cancer-risk models using a three-month exclusion window. (A) Cancer-general features: events which are enriched among top-decile across chosen cancer types (x-axis is the number of cancer types in which an event was consistently ranked high). (B-D) Cancer-specific features for pancreatic cancer. (B) Case-control log2 odds ratio of events ranked at top among cancer patients based on IG score. (C) We mask each eligible (in top decile of trajectory length) occurrence of an event and compute observed minus masked change in cancer log-odds averaged across trajectories. (D) We carry out transition analysis by incrementally adding events to a patient's trajectory and track transition in cancer log-odds (n_{up}) or decrease (n_{down}) and transition polarity is then given by* $(n_{up} - n_{down})/(n_{up} + n_{down})$. *Results on other cancer types in Supplementary* Fig S13a-S13f.

Discussion

GenEHR – an inter-visit gap aware foundation model for pan-cancer risk prediction.

We report a new pan-cancer EHR foundation model, with fine-tuning for disease risk prediction, with the clinical goals of improving cancer risk assessment and 'catching cancer early'.

The method combines the power of generative large models, analogous to large language models, with supervised adaptive learning for specific subtasks. Application of GenEHR to large-scale data sets in five different healthcare systems achieves reasonable predictive performance for the probability of a first cancer diagnosis within different horizons ending at 6, 12, 24, 36, and 60 months after the end of the trajectory used for the prediction. While the retrospective performance, based on evaluation in withheld test sets, is encouraging, real-world clinical implementation of screening programs assisted by the model will require prospective testing. We now discuss both strengths, limitations and challenges.

Clinical utility of risk stratification models.

An operating threshold trades false positive rate against true positive rate to minimize the potential harm of follow-up testing[51]. Choosing an optimal threshold is essential for nominating patients for AI-informed screening programs because screening tests and their downstream consequences have both financial and other resource implications. For effective resource utilization, we need to address key operational questions. How many individuals should be selected for AI-informed screening, and what proportion of those will eventually be diagnosed with cancer? We prioritize our evaluation of prediction performance using PPV, SIR and SNS. Here, SNS is an approximation to the number of nominated patients per subsequent cancer diagnosis and does not account for the performance of a screening test. SIR quantifies the risk enrichment within high risk stratum at a given operating threshold compared to a demographically adjusted baseline population incidence. These scores allow us to assess and compare the utility of various models for actionable intervention pathways by guiding patients toward targeted AI-informed screening programs, but admittedly do not capture the full spectrum of heterogeneity inherent in various available intervention pathways for individual cancer types.

GenEHR improvements over existing autoregressive EHR foundation models.

Recent generative EHR models differ in their handling of time intervals and the patient populations used for training. ETHOS[25] is trained on MIMIC[52] and evaluated as zero-shot, and CoMET[26] adopts ETHOS tokenization but is trained at a much larger scale on the Epic Cosmos database. Both represent time intervals using a fixed set of time-interval tokens, so fine-grained temporal resolution is not attained. Delphi[30] learns next event rates with a competing exponential model trained on first occurrence of ICD codes in the UKBB. Delphi handles time gaps using artificial no-event tokens and its own ablation evaluation shows that the performance depends on the rate at which these tokens are inserted. It also treats each level-3 ICD code as a separate competing outcome, without grouping codes that belong to the same disease. For example, colorectal cancer can be represented by any of the following codes C18, C19, C20, and C26; thus when the model predicts one of these codes, related codes may already be present in the patient's history and bias risk prediction. MOTOR[28] fits a piecewise exponential objective, but via a discriminative head rather than in a generative model of trajectories. GenEHR builds on these approaches by first learning time-interval aware patient health state representations through large-scale unsupervised pre-training. Given a patient's history up to the present, GenEHR can learn how past clinical events interact with the queried time gap to influence their relative cancer risk across future prediction horizons, whereas, a standard Cox model in this setting preserves

patients' risk ranking across horizons. The supervised stage uses cancer registry confirmed labels for pan-cancer risk prediction, excluding patients whose EHR-recorded cancer diagnoses are not confirmed in the registry to reduce noise from EHR coding practices.

Screening programs for early cancer intervention. Compared with our earlier work on risk stratification for pancreatic cancer[10,11], GenEHR affords an opportunity to design and test customized intervention programs for many tumor types. Cancer types differ in incidence, available screening tests, prevention protocols, and the balance between screening benefits and harms. U.S. Preventive Services Task Force (USPSTF) recommends population-wide screening only for breast[53], colorectal[54], cervical cancer[55], and lung cancer[56] typically conditioned on age (e.g., mammogram every two years for women above age 40; low-dose CT-scan for adults aged 50-80 years with smoking history). Key obstacles to USPSTF recommendations supporting screening for other cancers include the inability to accurately identify the population at sufficient risk and lack of efficient screening tools to adequately visualize or detect an early stage cancer before it metastasizes[57,58]. In addition, some existing screening criteria incompletely capture the at-risk population–for example, only about 50% of patients diagnosed with lung cancer would have qualified for screening prior to their diagnosis[59]. Use of GenEHR could help with the barrier of identifying populations at risk. AI-informed screening in a GenEHR-identified cohort could be analogous to enhanced screening methods currently recommended for those with inherited genetic variants (e.g. in BRCA1/2, Lynch syndrome)[60,61], family history of cancer or toxic exposure. GenEHR could be applied for real-world population risk stratification using widely available longitudinal data as it reaches or exceeds prediction performance over most of the specialized models without the need for specialized data.

Prospective evaluation of intervention programs for high-risk subpopulations.

Implementation of GenEHR in the clinic would require prospective trials in which risk stratification is followed by a carefully designed intervention program (Fig 7). These trials would test screening and preventive interventions in those at highest risk. The population benefit of AI-informed risk-based screening depends on the rate at which screen-nominated patients develop cancer, and whether screening platforms used in conjunction with the model can clinically detect cancer early enough to alter the disease outcomes. Cost-benefit analyses considering screening capacity, performance, costs, potential screening related harms, and expected gains in healthy life-span or survival will be necessary[62].

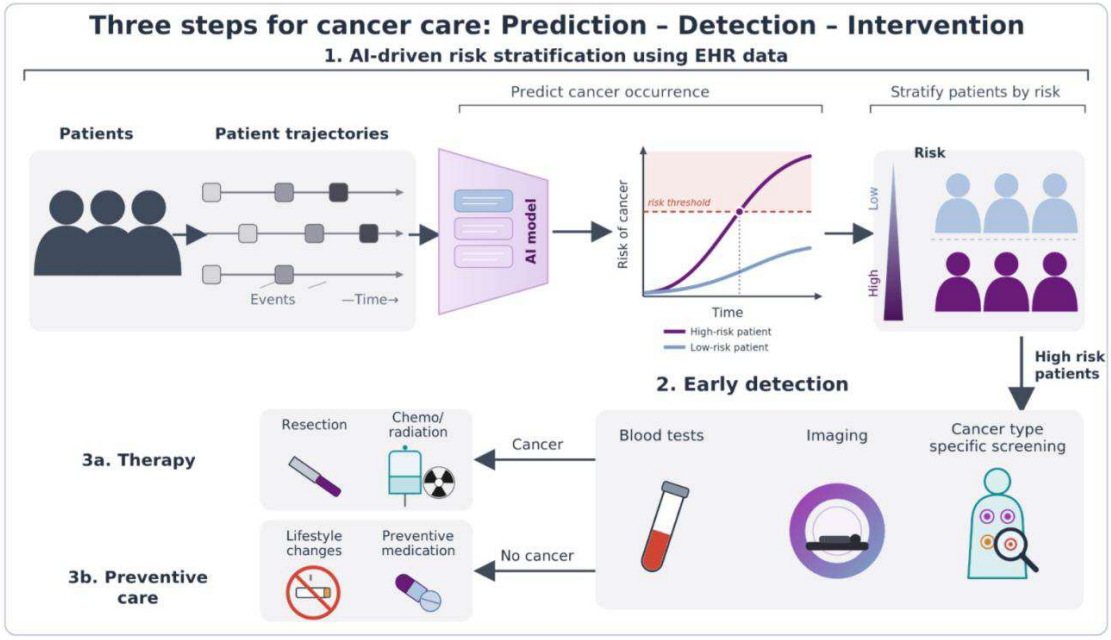


Figure 7: Three steps for getting AI risk stratification into clinical practice. *This work addresses the first step in a three step process potentially leading to patient benefit. Early detection in step 2 may be particularly effective in high-risk groups, due to affordable intense workup. If no cancer is detected in screening, the prediction of high risk may still be considered for enrolment in prevention programs.*

Global reach and prospective testing

In our previous work and now here, we have trained and evaluated cancer risk models in cohorts from different countries[10]. The ambition is to train a global model across healthcare systems via local collaborative agreements that provide governed access to protected data and secure computing. We can then successively apply adaptive training on genEHR using location-specific data. We have seen interest in collaborative efforts in clinical settings in, e.g., the US, Canada, the UK, Sweden, Denmark, Turkey and Germany. As such collaborations develop, in the next steps we intend to implement prospective silent evaluation, in which patients will be scored using genEHR without influencing their clinical care[63]. This phase should assess reliability, temporal drift, calibration, and whether retrospective performance holds in the tracked patient population. Because no intervention occurs, this provides a projected SNS that can be used to judge feasibility of operation threshold. These efforts would benefit patients and lighten the economic burden of cancer care.

Challenges and future outlook

Moving from retrospective risk prediction to clinical implementation requires models that are well calibrated and robust to data noise and dataset shift [64,65]. Current AI methods can handle some level of noise, but systematic errors or missing diagnoses can bias the predictions[66]. Recording accurate and complete diagnosis remains a major challenge in electronic health record

systems. While we have seen higher accuracy in the Danish clinical records, the accuracy of ICD codes in US healthcare systems remains challenging [67,68]. We have partly addressed this limitation by using cancer registry confirmed outcomes as the labels for cancer risk stratification. Future opportunities include inclusion of free text notes, and more accurate computational phenotyping combining domain-agnostic machine learning, ontologies, or large language models. This would likely incur increased resource use, but it could improve predictive performance and explainability of per-patient or population wide risks for malignancy [69,70]. An additional challenge within countries without a single payer system is that individuals often seek care within multiple different healthcare systems in sequence, and so integrating trajectories across these different settings can limit the model's predictive power for those individuals.

Clinical adoption also requires AI-supported screening programs to be acceptable to patients and clinicians, evaluated for clinical benefit and integrated into existing workflows[71]. Predictions from the model should be supported by interpretable reports that clinicians can review when guiding patients to appropriate screening programs[72,73]. AI methods typically perform differently across cohorts when certain demographics groups are unequally represented in training or data acquisition differs, e.g. in specialized cohorts. At the same time, cancer incidence can vary across populations because of environmental, socioeconomic, genetics, and other factors[74]. Performance of AI methods, even if well-balanced in training, should therefore be evaluated across demographic, socioeconomic, and environmental subpopulations with the aim to capture true population specific differences in risk from disparities introduced by data collection or healthcare access [74,75].

Cancer risk stratification from EHR data can also benefit from richer longitudinal data collected using standardized measurement units. For example, an annually repeated blood test with detailed profiles of DNA, proteins and metabolites, in which one can analyze changes rather than just snapshots, could be particularly informative [76,77]. Because molecular profiles of liquid biopsies vary substantially between healthy individuals, within-person changes over time may be more informative than a single measurement[78]. These year to year changes can capture an evolving health state and provide early evidence of changing disease state or risk.

In practice, clinical use of the genEHR-CancerRisk prediction tool is step one of three steps on the way to patient benefit: risk stratification, interception program, and early treatment (Fig 7)[79,80]. A full clinical program to "catch cancer early" is a challenge for a collaborative consortium involving AI experts, technology developers of medical tests, clinical professionals and patients.

Author contributions

Concepts: A.K., C.S.
Data acquisition and refinement: A.K., E.T.W, (MGB); C.Z., N.F. (VA); D.T.F., Q.W., J.H. (PROV); M.H., A.K. (UKBB); M.H., A.K. (AoU).
Methodology, evaluations, software system: A.K.
Software components: A.K., D.T.F., M.H., C.Z., D.R.
Model runs and evaluation: A.K. (MGB), D.T.F. (PROV), M.H. (UKBB, AoU), C.Z. (US-VA)
Results discussion: A.K., D.T.F., M.H., C.Z, D.R., D.S.M., E.T.W., J.H., N.F., L.V.S., C.S.
Epidemiology: E.T.W., J.H.
Clinical collaborators: E.T.W., L..V.S., A.C., T.S., J.H.

Supervision: J.H., N.F., L.V.S., C.S.
Displayed items: D.T.F., M.H., C.Z., A.K.
Manuscript: A.K., C.S.
Manuscript edits: all authors

Data and code availability

Open source, on github.
Academic license, restricted commercial use.
Model weights on request from each data-providing institution with the Sander lab at HMS.

Potential competing interests

D.S.M: Advisor - GenBio.ai, Dyno Therapeutics, Tectonic Therapeutic, Catalio Ventures; Cofounder - Seismic Therapeutic.
Q.W., J.H.: Institutional research support for unrelated work from BMS, Gilead, Janssen, Delfina.
L.V.S.: Institutional grant funding from AstraZeneca, Delfi Diagnostics.
C.S.: SAB Cytoreason Ltd.

Acknowledgements

We thank Gary Bader, Rob Grant, Brian Wolpin, Abdullah Kuziez, and members of Marks lab for discussions and guidance. We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data and cohort examined in this study. This research has been conducted using the UK Biobank Resource under application number 53995.

IRB protocols

Approved protocols at MGB, PROV and VA. For Providence Health and Services, procedures were reviewed and approved by the Institutional Review Board at PROV through expedited review (STUDY2019000389). Patient consent was waived because disclosure of protected health information for the study was determined to involve no more than a minimal risk to the privacy of individuals. For MGB, IRB approval was granted for access to the RPDR research database (Protocol 2024P000851)**.** For VA, IRB approval was granted for access to the VA Corporate Data Warehouse (CDW). For All of US and UK-Biobank standard approval for research use after application.

Funding for this work

ICI philanthropy (grant to Harvard Medical School (HMS)), US-CDMRP (grant #HT9425-23-1-0464 to HMS & the Boston VA Research Institute), Uniformed Services University/Henry Jackson Foundation #HU00012420063, subaward #67659, ISB (Institute for Systems Biology), and the VA Cooperative Studies Program. The views expressed are those of the authors and do

not necessarily reflect the position or policy of the Department of Veterans Affairs or the U.S. government. This project is sponsored by the Uniformed Services University of the Health Sciences (USU); however, the information or content and conclusions do not necessarily represent the official position or policy of, nor should any official endorsement be inferred on the part of, USU, the Department of Defense, or the U.S. Government.

References

1. U.S. Population Data - SEER Population Data. *SEER* https://seer.cancer.gov/popdata/index.html.

2. Pashayan, N., Morris, S., Gilbert, F. J. & Pharoah, P. D. P. Cost-effectiveness and Benefit-to-Harm Ratio of Risk-Stratified Screening for Breast Cancer: A Life-Table Model. *JAMA Oncol* **4**, 1504–1510 (2018).

3. Hamilton, W. *et al.* The importance of anaemia in diagnosing colorectal cancer: a case-control study using electronic primary care records. *Br J Cancer* **98**, 323–327 (2008).

4. Stapley, S. *et al.* The risk of pancreatic cancer in symptomatic patients in primary care: a large case-control study using electronic records. *Br J Cancer* **106**, 1940–1944 (2012).

5. Aggarwal, G., Rabe, K. G., Petersen, G. M. & Chari, S. T. New-onset diabetes in pancreatic cancer: a study in the primary care setting. *Pancreatology* **12**, 156–161 (2012).

6. Dayem Ullah, A. Z. M., Stasinos, K., Chelala, C. & Kocher, H. M. Temporality of clinical factors associated with pancreatic cancer: a case-control study using linked electronic health records. *BMC Cancer* **21**, 1279 (2021).

7. Fritz, C. D. L. *et al.* Red-flag signs and symptoms for earlier diagnosis of early-onset colorectal cancer. *JNCI Journal of the National Cancer Institute* **115**, 909 (2023).

8. Yuan, C. *et al.* The age-dependent association of risk factors with pancreatic cancer. *Ann Oncol* **33**, 693–701 (2022).

9. Klein, A. P. Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. *Nat Rev Gastroenterol Hepatol* **18**, 493–502 (2021).

10. Placido, D. *et al.* A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat Med* **29**, 1113–1122 (2023).

11. Zheng, C. *et al.* Pancreatic cancer risk prediction using deep sequential modeling of longitudinal diagnostic and medication records. *medRxiv* 2025.03.03.25323240 (2025) doi:10.1101/2025.03.03.25323240.

12. Wang, L. *et al.* Transformer-based deep learning model for the diagnosis of suspected lung cancer in primary care based on electronic health record data. *EBioMedicine* **110**, 105442 (2024).

13. Read, A. J., Zhou, W., Saini, S. D., Zhu, J. & Waljee, A. K. Prediction of Gastrointestinal Tract Cancers Using Longitudinal Electronic Health Record Data. *Cancers (Basel)* **15**, (2023).

14. Jung, A. W. *et al.* Multi-cancer risk stratification based on national health data: a retrospective modelling and validation study. *Lancet Digit Health* **6**, e396–e406 (2024).

15. Cancer of the Pancreas - Cancer Stat Facts. *SEER* https://seer.cancer.gov/statfacts/html/pancreas.html.

16. Multi-layer Representation Learning for Medical Concepts. https://dl.acm.org/doi/10.1145/2939672.2939823 doi:10.1145/2939672.2939823.

17. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports* **6**, 26094 (2016).

18. Choi, E., Xiao, C., Stewart, W. & Sun, J. MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. *Advances in Neural Information Processing Systems* **31**, (2018).

19. Li, Y. *et al.* BEHRT: Transformer for Electronic Health Records. *Scientific Reports* **10**, 7155 (2020).

20. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine* **4**, 86 (2021).

21. Pang, C. *et al.* CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. in *Machine Learning for Health* 239–260 (PMLR, 2021).

22. Yang, Z., Mitra, A., Liu, W., Berlowitz, D. & Yu, H. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat Commun* **14**, 7857 (2023).

23. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics* **113**, 103637 (2021).

24. Kraljevic, Z. *et al.* Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit Health* **6**, e281–e290 (2024).

25. Renc, P. *et al.* Zero shot health trajectory prediction using transformer. *NPJ Digit Med* **7**, 256 (2024).

26. Waxler, S. *et al.* Generative Medical Event Models Improve with Scale. (2025).

27. Pang, C. *et al.* CEHR-GPT: Generating Electronic Health Records with Chronological Patient Timelines. (2024).

28. MOTOR: A Time-to-Event Foundation Model For Structured Medical Records. https://iclr.cc/virtual/2024/poster/18777.

29. Gadd, C. *et al.* SurvivEHR: a competing risks, time-to-event foundation model for multiple long-term conditions from primary care electronic health records. *NPJ Digit Med* **9**, (2026).

30. Shmatko, A. *et al.* Learning the natural history of human disease with generative transformers. *Nature* **647**, 248–256 (2025).

31. Zhang, A. *et al.* A multimodal and temporal foundation model for virtual patient representations at healthcare system scale. (2026).

32. Chen, S. *et al.* Recall and learn: Fine-tuning deep pretrained language models with less forgetting. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020). doi:10.18653/v1/2020.emnlp-main.634.

33. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. (2021).

34. Nalichowski, R., Keogh, D., Chueh, H. C. & Murphy, S. N. Calculating the benefits of a Research Patient Data Repository. *AMIA Annu Symp Proc* **2006**, 1044 (2006).

35. Velarde, K. E. *et al.* An initiative using informatics to facilitate clinical research planning and recruitment in the VA health care system. *Contemp Clin Trials Commun* **11**, 107–112 (2018).

36. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

37. Ramirez, A. H. *et al.* The Research Program: Data quality, utility, and diversity. *Patterns (N Y)* **3**, 100570 (2022).

38. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. in *International Conference on Machine Learning* 3319–3328 (PMLR, 2017).

39. Hoopes, M. *et al.* Assessing Cancer History Accuracy in Primary Care Electronic Health Records Through Cancer Registry Linkage. *J Natl Cancer Inst* **113**, 924–932 (2021).

40. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018).

41. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* **40**, 373–383 (1987).

42. Kirkegård, J. *et al.* Acute pancreatitis as an early marker of pancreatic cancer and cancer stage, treatment, and prognosis. *Cancer Epidemiol* **64**, 101647 (2020).

43. Boulay, B. R. & Parepally, M. Managing malignant biliary obstruction in pancreas cancer: choosing the appropriate strategy. *World J Gastroenterol* **20**, 9345–9353 (2014).

44. Avila, E. K. *et al.* Brain tumor-related epilepsy management: A Society for Neuro-oncology (SNO) consensus review on current management. *Neuro Oncol* **26**, 7–24 (2024).

45. Sarin, R. & Murthy, V. Medical decompressive therapy for primary and metastatic intracranial tumours. *Lancet Neurol* **2**, 357–365 (2003).

46. Gurnari, C. *et al.* Clinical and Molecular Determinants of Clonal Evolution in Aplastic Anemia and Paroxysmal Nocturnal Hemoglobinuria. *J Clin Oncol* **41**, 132–142 (2023).

47. Rembold, C. M. Number needed to screen: development of a statistic for disease screening. *BMJ* **317**, 307–312 (1998).

48. Huang, T. *et al.* Hypertension, use of antihypertensive medications, and risk of epithelial ovarian cancer. *Int J Cancer* **139**, 291–299 (2016).

49. Tuesley, K. M. *et al.* Association between antihypertensive medicine use and risk of ovarian cancer in women aged 50 years and older. *Cancer Epidemiol* **86**, 102444 (2023).

50. Zhang, Y. *et al.* Pancreatic cancer is associated with medication changes prior to clinical diagnosis. *Nat Commun* **14**, 2437 (2023).

51. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* **26**, 565–574 (2006).

52. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Scientific Data* **3**, 160035 (2016).

53. US Preventive Services Task Force *et al.* Screening for Breast Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **331**, 1918–1930 (2024).

54. US Preventive Services Task Force *et al.* Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **325**, 1965–1977 (2021).

55. US Preventive Services Task Force *et al.* Screening for Cervical Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **320**, 674–686 (2018).

56. US Preventive Services Task Force *et al.* Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **325**, 962–970 (2021).

57. US Preventive Services Task Force *et al.* Screening for Pancreatic Cancer: US Preventive Services Task Force Reaffirmation Recommendation Statement. *JAMA* **322**, 438–444 (2019).

58. Final Recommendation Statement: Ovarian Cancer: Screening. https://www.uspreventiveservicestaskforce.org/uspstf/document/RecommendationStatementFinal/ovarian-cancer-screening (2018).

59. Yang, H. C. *et al.* Age-Based Screening for Lung Cancer Surveillance in the US. *JAMA Netw Open* **8**, e2546222 (2025).

60. Syngal, S. *et al.* ACG clinical guideline: Genetic testing and management of hereditary gastrointestinal cancer syndromes. *Am J Gastroenterol* **110**, 223–62; quiz 263 (2015).

61. Daly, M. B. *et al.* NCCN Guidelines® Insights: Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic, Version 2.2024. *J Natl Compr Canc Netw* **21**, 1000–1010 (2023).

62. Katki, H. A. & Bebu, I. A simple framework to identify optimal cost-effective risk thresholds for a single screen: Comparison to Decision Curve Analysis. *J R Stat Soc Ser A Stat Soc* **184**, 887–903 (2021).

63. Tikhomirov, L. *et al.* A scoping review of silent trials for medical artificial intelligence. *Nat. Health* **1**, 532–554 (2026).

64. Van Calster, B. *et al.* Calibration: the Achilles heel of predictive analytics. *BMC Med* **17**, 230 (2019).

65. Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21**, 345–352 (2020).

66. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* **361**, k1479 (2018).

67. Joiner, K. A., Lin, J. & Pantano, J. Upcoding in medicare: where does it matter most? *Health Economics Review* **14**, 1 (2024).

68. O'Malley, K. J. *et al.* Measuring diagnoses: ICD code accuracy. *Health Serv Res* **40**, 1620–1639 (2005).

69. Newton, K. M. *et al.* Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc* **20**, e147–54 (2013).

70. Kirby, J. C. *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* **23**, 1046–1052 (2016).

71. Vasey, B. *et al.* Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **377**, e070904 (2022).

72. Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. in *Machine Learning for Healthcare Conference* 359–380 (PMLR, 2019).

73. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* **3**, e745–e750 (2021).

74. Islami, F. *et al.* American Cancer Society's report on the status of cancer disparities in the United States, 2021. *CA Cancer J Clin* **72**, 112–143 (2022).

75. Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).

76. Price, N. D. *et al.* A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat Biotechnol* **35**, 747–756 (2017).

77. Tebani, A. *et al.* Integration of molecular profiles in a longitudinal wellness profiling cohort. *Nat Commun* **11**, 4487 (2020).

78. Dodig-Crnković, T. *et al.* Facets of individual-specific health signatures determined from longitudinal plasma proteome profiling. *EBioMedicine* **57**, 102854 (2020).

79. Minasian, L. M. *et al.* Study design considerations for trials to evaluate multicancer early detection assays for clinical utility. *J Natl Cancer Inst* **115**, 250–257 (2023).

80. Castle, P. E., Faupel-Badger, J. M., Umar, A. & Rebbeck, T. R. A Proposed Framework and Lexicon for Cancer Prevention. *Cancer Discov* **14**, 594–599 (2024).

81. Zeng, K., Bodenreider, O., Kilbourne, J. & Nelson, S. RxNav: a web service for standard drug information. *AMIA Annu Symp Proc* **2006**, 1156 (2006).

82. Kind, A. J. H. & Buckingham, W. R. Making Neighborhood-Disadvantage Metrics Accessible — The Neighborhood Atlas. *New England Journal of Medicine* https://doi.org/10.1056/NEJMp1802313 (2018) doi:10.1056/NEJMp1802313.

83. NHLBI Obesity Education Initiative Expert Panel on the Identification, Evaluation & Treatment of Obesity. Introduction. in *Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: The Evidence Report* (National Heart, Lung, and Blood Institute, 1998).

84. Site Recode ICD-O-3/WHO 2008 - SEER Data Reporting Tools. *SEER* https://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html.

85. Website. https://digital.nhs.uk/data-and-information/publications/statistical/cancer-registration-statistics/england-2023.

86. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer Normalization. (2016).

87. Friedman, M. Piecewise exponential models for survival data with covariates. *Ann. Stat.* **10**, 101–113 (1982).

88. Audibert, A., Amini, M. R., Usevich, K. & Clausel, M. Low-Rank Updates of pre-trained Weights for Multi-Task Learning. in *Findings of the Association for Computational Linguistics: ACL 2023* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2023). doi:10.18653/v1/2023.findings-acl.476.

89. Horvitz, D. G. & Thompson, D. J. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **47**, 663–685 (1952).

90. XGBoost. https://dl.acm.org/doi/10.1145/2939672.2939785 doi:10.1145/2939672.2939785.

91. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

92. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. (2019).

93. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020).

94. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. (2017).

95. Hoffmann, J. *et al.* Training Compute-Optimal Large Language Models. in *Advances in Neural Information Processing Systems 35* 30016–30030 (Neural Information Processing Systems Foundation, Inc. (NeurIPS), San Diego, California, USA, 2022).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GenEHRsupp.docx