ShepBERT: An OMOP-Portable EHR Transformer Pipeline for Cancer Risk Prediction in NHPI-Serving Health Systems

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Native Hawaiian and Pacific Islander (NHPI) populations face disproportionately high cancer incidence and late-stage diagnosis, yet remain systematically absent from the data used to train clinical risk models. We present ShepBERT, an OMOP-portable electronic health record (EHR) transformer *pipeline* and leakage-control methodology for cancer risk prediction, designed for deployment in health systems serving NHPI populations. The pipeline pre-trains a BERT-style encoder on MIMIC-IV structured events (364,627 patients, ∼680M OMOP-harmonized events) and fine-tunes it for 1-year cancer risk, and contributes: (i) an OMOP-standardized single-encoder pipeline that transfers from public pretraining to multi-cancer risk without architecture changes; (ii) a three-pass censoring audit (clinical seed list, frequency enrichment, feature-attribution timing) that removes cancer-pathway shortcut codes an EHR transformer would otherwise exploit; and (iii) an interpretability-based diagnostic (Integrated Gradients, EG-SHAP, occlusion) that surfaces what drives each prediction. We validate the pipeline end-to-end on 20,000 synthetic patients with Hawaii demographics across four cancers (breast, colorectal, lung, prostate). The synthetic cohort has no identifiable NHPI subgroup (its only race codes are Asian, White, and Black), so the reported AUROCs are pipeline diagnostics rather than clinical performance claims: they correctly show that an age+sex baseline meets or exceeds the sequence model—discrimination on this synthetic cohort is demographic and no genuine sequence signal survives censoring. Real, NHPI-specific evaluation is the explicit next step; because the pipeline transfers without code changes to a partner referral center, it is positioned for that evaluation.

Keywords: Electronic health records · Self-supervised pretraining · Transfer learning · Cancer risk prediction · Model interpretability

1 Introduction

Cancer is a leading cause of death in Pacific Island communities, yet these populations remain underrepresented in clinical machine learning research. Native Hawaiian women experience breast cancer incidence of 175 per 100,000—well

above the national rate of 124 [1]. The machine learning models commonly deployed for cancer risk stratification, however, are trained predominantly on data from large mainland medical centers that share little demographic overlap with Pacific Island health systems.

Clinical machine learning has increasingly turned to large-scale pretraining of transformer-based models on longitudinal EHR data. BEHRT [5] demonstrated that BERT-style masked language modeling over sequences of OMOP-coded clinical events learns rich representations that transfer to a wide range of clinical prediction tasks. CORE-BEHRT [7] refined this approach with improved temporal embeddings, a more stable training procedure, and the SARA fine-tuning framework—applied to multitask cancer survival prediction in a Danish national cohort. Med-BERT [9] and CLMBR [10] extend the same paradigm to structured EHR codes and timeline representations, respectively.

What is missing is a model designed for and validated on a Pacific Island population: models trained on mainland cohorts inherit demographic and practice patterns that, applied to NHPI patients with distinct risk profiles and utilization, risk miscalibrated predictions that could worsen rather than reduce disparities. We present ShepBERT, which addresses three problems: (1) no large-scale NHPI dataset exists for pretraining, so we transfer from a public cohort; (2) transfer requires a common vocabulary—we use OMOP to bridge MIMIC-IV, Synthea, and the target partner-site Epic system; and (3) cancer risk prediction requires careful handling of data leakage, class imbalance, and temporal truncation biases absent in general clinical prediction.

This paper describes: (1) an **OMOP-standardized single-encoder pipeline** (ShepBERT) that transfers from MIMIC-IV pretraining to 1-year risk across four cancer types on 20,000 synthetic patients with Hawaii demographics, without architecture changes; (2) a **three-pass censoring audit** (clinical seed list, frequency enrichment, feature-attribution timing) that eliminates shortcut learning from cancer-pathway codes; and (3) an **interpretability analysis** (Integrated Gradients, EG-SHAP, occlusion) showing that discrimination on the synthetic cohort is demographic, supported by `CancerDataset`, which generates annual rolling prediction samples via age-based truncation without storing redundant sequences.

2 Problem Formulation and Summary of Results

Annual rolling prediction. Let $H_p = \{(c_i, t_i)\}$ be patient p's history of OMOP concept–timestamp pairs, and let c index the target cancer type (we train one model per cancer). For each integer age a, let $H_p^{<a}$ be the events before age a after removing the cancer-specific *censored code set* $\mathcal{C}^{(c)}$ (Section 3); the task is

$$\hat{y}_{p,a}^{(c)} = \sigma\Big(f_\theta^{(c)}(H_p^{<a})\Big) \approx \Pr\big(\text{first diagnosis of cancer } c \text{ in } [a,\, a{+}1) \mid H_p^{<a}\big)\,, \quad (1)$$

where $f_\theta^{(c)}$ is the ShepBERT sequence model fine-tuned for cancer c and σ is the sigmoid. Patients with a prior diagnosis of cancer c are excluded; negatives require 18 months of follow-up after the prediction point.

Model. ShepBERT reads the patient's OMOP code history as a sequence of clinical event tokens. The encoder is fine-tuned for cancer risk by mean pooling over the 288-dimensional token embeddings followed by a linear head $\mathbb{R}^{288} \to \mathbb{R}^1$ that outputs the risk logit. As an interrogatable point of comparison on the same rows, an AIC-guided stepwise logistic regression on landmark features (age, sex, race and ethnicity, top-k code indicators, and utilization summaries) provides a tabular baseline that isolates how much of the discrimination is demographic (Section 4).

Summary of results. On 20,000 synthetic patients, pooled out-of-fold evaluation yields sequence-model AUROCs of 0.717 (breast), 0.459 (colorectal), 0.703 (lung), and 0.862 (prostate) (Table 1). Because the cohort is synthetic, these are not clinical performance claims but serve two diagnostic purposes: (i) *pipeline validation*—above-chance discrimination confirms the end-to-end system trains and evaluates correctly; and (ii) *censoring proof*—the drop from > 0.9 (uncensored) to the demographic-floor regime shows the censoring audit removes shortcut signal. An age+sex logistic-regression baseline meets or exceeds these AUROCs, so discrimination on the synthetic cohort is largely demographic; genuine EHR-sequence signal is the open question for the partner-site cohort.

Background. Masked language modeling [2] applied to clinical sequences teaches an encoder the statistical co-occurrence of clinical codes, yielding representations that transfer to downstream prediction. BEHRT [5] adapts this to longitudinal EHR data with age, visit-segment, and position embeddings, and CORE-BEHRT [7] refines it with continuous absolute-time embeddings, a more stable training recipe, and the SARA rolling-annual fine-tuning framework we adapt via `CancerDataset`; Med-BERT [9] and CLMBR [10] extend the paradigm to structured codes and timelines. ShepBERT builds on CORE-BEHRT's `BertEHRModel`, contributing an OMOP harmonization layer, cancer fine-tuning, a three-pass censoring audit, and the NHPI application. Harmonization to the OMOP Common Data Model [8] via the MEDS event standard [6] lets ShepBERT be pretrained on MIMIC-IV, validated on Synthea, and later fine-tuned on the partner-site Epic system without pipeline changes.

3 Implementation

3.1 Data Pipeline

The pipeline consists of five stages (Fig. 1).

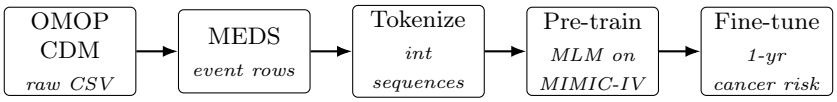

Fig. 1. The ShepBERT pipeline. OMOP records are harmonized to MEDS event streams and tokenized; a BERT-style encoder is pretrained on MIMIC-IV by masked language modeling and fine-tuned for 1-year cancer risk, with the three-pass censoring audit applied at the fine-tuning stage. An interrogatable stepwise logistic regression over landmark features runs on the same rows as a demographic baseline.

ETL, harmonization, tokenization, and pretraining. MIMIC-IV [4] (364,627 patients, ∼680M events) and Synthea [12] (20,000 patients) are converted to MEDS event streams via the open-source `meds_etl` tools, and a `CodeHarmonizer` resolves source codes (ICD-9/10, NDC, LOINC) to standard OMOP concepts (a 26,431-concept vocabulary) via the "Maps to" relationship. Each event stream is tokenized to a 512-token sequence with prepended gender/race background tokens, 24-hour-gap visit segmentation, and parallel age/position arrays. The CORE-BEHRT `BertEHRModel` is then pretrained by MLM (15% masking; AdamW, $\mathrm{lr} = 10^{-4}$; 25 epochs).

Fine-tuning for cancer risk. The pretrained encoder is fine-tuned for 1-year cancer risk using `CancerDataset`, which stores one compact sequence per patient and expands at access time: for each pre-stored row (p, a, y) it truncates patient p's sequence to events with age $\leq a$ and applies code censoring, avoiding ∼700,000 redundant materialized sequences.

Code censoring. After truncation we remove a set of cancer-specific codes, preventing the model from inferring the diagnosis or its immediate consequences directly from the record. The censoring set is built in three passes. First, we seed the list with clinically known post-diagnosis codes: cancer diagnosis codes, chemotherapy drugs, staging procedures (TNM, lymph node biopsy), and radiation therapy. Second, we add codes with $> 100\times$ case/control enrichment in the Synthea cohort, verified against SNOMED, LOINC, and Synthea's module definitions. Third, we run EG-SHAP attribution on a trained model and audit the highest-ranked codes by their timing: a code concentrated before the diagnosis date but generated only for patients on the cancer-detection pathway is a *screening proxy* and is censored together with the post-diagnosis treatment codes. This pass identifies codes that frequency enrichment alone misses: for colorectal cancer, rectal polypectomy, fecal-occult-blood tests, and colonic polyps; for prostate cancer, digital rectal examination (the standard co-procedure with PSA, already censored).

Class imbalance and cross-validation. The 1-year cancer label prevalence is ≈0.05%–0.07% across all cancer types. We apply square-root inverse-frequency sampling (following SARA) to upsample the positive class during training; positive

class weight is not combined with the sampler to avoid double-correction. Five-fold patient-level cross-validation keeps all prediction rows for a patient in the same fold. Fine-tuning uses AdamW, lr $= 10^{-5}$, weight decay $= 0.01$, 15 epochs per fold with early stopping (patience 5) on validation AUROC.

Baseline, diagnostics, and evaluation. The tabular LR baseline uses AIC-guided forward-selection logistic regression with patient-clustered standard errors on the same landmark rows as ShepBERT (features: age, sex, race and ethnicity, top-k code indicators, and utilization summaries). We report AUROC and AUPRC (F1 and accuracy are uninformative at sub-1% prevalence); confidence intervals resample patients (1,000 replicates) to preserve within-patient correlation.

4 Results and Observations

Pretraining. On MIMIC-IV, the ≈5M-parameter model reaches top-1 / top-10 masked-token accuracy of 22.2% / 60.5% over a 26,431-entry vocabulary (chance $< 0.01\%$), reflecting strong clinical co-occurrence structure in the learned representations.

Censoring effect. A pre-censoring run inflates AUROC to > 0.9 for every cancer: the model predicts from post-diagnosis chemotherapy and staging codes that Synthea generates deterministically. The three-pass procedure returns each model to the regime where it must rely on pre-diagnostic signals (Table 1). Synthea validation confirms pipeline correctness, not clinical utility; the most significant limitation is that the synthetic cohort contains no identifiable NHPI subgroup (the only OMOP race codes present are Asian, White, and Black), so NHPI-stratified evaluation awaits the partner-site cohort.

Multi-cancer generalization. We fine-tuned a 1-year risk model for each of the four cancer types without modifying the architecture or training procedure (Table 1). Each model reuses the same MIMIC-IV-pretrained encoder; only the outcome label file and cancer-specific censoring list change. Swapping only the outcome file and censoring list transfers the encoder across four cancers, confirming ShepBERT as reusable infrastructure. The sequence model discriminates for prostate (0.862), breast (0.717), and lung (0.703); for colorectal it falls to chance (0.459) once three-pass censoring removes the screening-pathway shortcut—no genuine pre-diagnostic sequence signal survives.

The discrimination is demographic. Interrogatable step-forward logistic-regression baselines (same rows, folds, and censoring; patient-clustered SEs) isolate what drives discrimination (Table 2). The two-variable age+sex model matches or exceeds the full stepwise model for every cancer (0.86–0.95), so code

Table 1. ShepBERT sequence model across cancer types: 1-year risk on 20,000 Synthea patients. "Cases" is cancer-positive prediction rows.

Cancer	AUROC	AUPRC	Cases
Prostate	0.862	0.0033	131
Breast	0.717	0.0011	205
Lung	0.703	0.0010	119
Colorectal	0.459	0.0006	245

Table 2. Logistic-regression baselines on the same annual-landmark rows and folds as each ShepBERT model.

Cancer	ShepBERT	Age+sex LR	Full stepwise LR	No-demo. LR
Breast	0.717	0.862 ± 0.018	0.852 ± 0.009	0.560 ± 0.030
Colorectal	0.459	0.893 ± 0.014	0.891 ± 0.014	0.685 ± 0.032
Lung	0.703	0.870 ± 0.035	$0.828 \pm 0.090^{\dagger}$	0.607 ± 0.034
Prostate	0.862	0.949 ± 0.005	0.945 ± 0.016	0.620 ± 0.012

†Averaged over four folds; one fold hit numerical separation.

and utilization features add no out-of-fold signal above the demographic ceiling; removing demographics collapses every cancer to 0.56–0.69. The discrimination on the synthetic cohort is therefore demographic—genuine EHR-sequence signal is expected to separate only on real, heterogeneous data.

4.1 Interpretability and the Censoring Audit

We derived code importance using Integrated Gradients [11], cross-checked against Expected-Gradients SHAP (EG-SHAP) [3] and an occlusion test (rank correlation 0.84). The three methods concur that the strongest signals are demographic and social (female sex, race and ethnicity, social-determinant survey items) rather than clinical codes, consistent with the demographic ceiling in Table 2.

Censoring audit: catching what frequency misses. Colorectal shows why the attribution audit is needed beyond frequency censoring: before censoring, rectal polypectomy topped the attribution (EG-SHAP 3.33, six times lung's top demographic code) yet had only $\sim 2\times$ case/control enrichment—far below the $100\times$ threshold—because Synthea places it pre-diagnosis on the screening pathway. Frequency censoring missed it; the attribution audit caught it and the next-ranked screening-path codes. After censoring, the colorectal sequence path retains no meaningful signal (noise-level, non-cancer-specific attributions), consistent with the demographic floor.

5 Future Work and Conclusion

The primary next step is replacing Synthea with de-identified EHR data from the partner NHPI referral center; the pipeline requires no code changes (Epic → OMOP → MEDS → ShepBERT), and on real, heterogeneous data we anticipate genuine EHR-sequence signal above the demographic baseline, as the synthetic cohort's tabular dominance is an artifact rather than a clinical finding. Architecturally, we plan to replace the bidirectional backbone with a GPT-style autoregressive encoder that attends only to past tokens—causally clean at every prediction age without truncation, in the spirit of CLMBR [10] and the CORE-BEHRT GPT variant [7].

In summary, ShepBERT is a reusable, OMOP-standardized EHR risk-modeling infrastructure that, on synthetic data, validates end to end across four cancer types with a three-pass censoring audit that removes shortcut learning and an interpretability analysis showing the residual discrimination is demographic. Because it transfers without code changes to a partner NHPI referral center, it is positioned for the first NHPI-specific multi-cancer risk evaluation.

References

1. Deng, F., Hughes, C.K., Meng, L., et al.: The high and heterogeneous burden of breast cancer in Hawaii. Cancer Epidemiology, Biomarkers & Prevention **27**(6), 633–641 (2018). https://doi.org/10.1158/1055-9965.EPI-17-0946
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT pp. 4171–4186 (2019)
3. Erion, G., Janizek, J.D., Sturmfels, P., Lundberg, S.M., Lee, S.I.: Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nature Machine Intelligence **3**, 620–631 (2021). https://doi.org/10.1038/s42256-021-00343-w
4. Johnson, A.E.W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., Lehman, L.w.H., Celi, L.A., Mark, R.G.: MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data **10**(1), 1 (2023). https://doi.org/10.1038/s41597-022-01899-x
5. Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: BEHRT: Transformer for electronic health records. Scientific Reports **10**(1), 7155 (2020). https://doi.org/10.1038/s41598-020-62922-y
6. MEDS Working Group: Medical event data standard (MEDS). https://github.com/Medical-Event-Data-Standard/meds (2024)
7. Odgaard, M.F., Jensen, P.B., Bernth-Jensen, J., et al.: CORE-BEHRT: A carefully optimized and rigorously evaluated BEHRT (2024)
8. OHDSI Collaborative: The OMOP common data model. https://ohdsi.github.io/CommonDataModel/ (2019)
9. 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). https://doi.org/10.1038/s41746-021-00455-y

10. Steinberg, E., Jung, K., Fries, J.A., Corbin, C.K., Pfohl, S.R., Shah, N.H.: Language models are an effective representation learning technique for electronic health record data. Journal of Biomedical Informatics **113**, 103637 (2021). https://doi.org/10.1016/j.jbi.2020.103637
11. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML). pp. 3319–3328 (2017)
12. Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duda, C., Shura, K., Hartley, S., McLachlan, S.: Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association **25**(3), 230–238 (2018). https://doi.org/10.1093/jamia/ocx079