Demo: Combining Polygenic Risk and EHR Survival Models of Cardiac Patients in the Indian Context via Knowledge Distillation

Anonymous submission

Abstract

Polygenic risk scores (PRS) improve coronary artery disease (CAD) prediction in many populations but suffer from reduced portability and calibration when applied to underrepresented ancestries. We present a practical pipeline that transfers population-adapted genomic risk into a deployable electronic-health-record (EHR) survival model through censoring-aware knowledge distillation. Using IndiGenomes allele-frequency data to recalibrate GWAS-derived PRS for the Indian population, we build a genomic teacher that outputs risk scores and survival curves; these soft targets are distilled into a parsimonious Cox/AFT student that uses only routine clinical features collected from a longitudinal Maharashtra EHR cohort (N$\simeq$5,000). The student is trained with a combined loss that blends the Cox partial likelihood and an IPCW-weighted distillation term so censored observations contribute robustly. In cross-validated experiments, our distilled student improves discrimination and net reclassification vs. an equivalent EHR model without distillation, while retaining clinical interpretability and deployability. Ablations demonstrate that recalibration using IndiGenomes allele frequencies substantially enhances the benefit of distillation, underscoring the importance of population-specific genomic adaptation. Our approach provides a scalable route to embed genomic risk structure into routine clinical tools in resource-constrained settings

Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality in India, with coronary artery disease (CAD) showing earlier onset and faster progression than in Western populations (Prabhakaran, Jeemon, and Roy 2016; Gupta and Xavier 2018). Early risk prediction is critical, yet established clinical models such as Framingham or pooled cohort equations capture only conventional risk factors and generalize poorly to non-Western populations (Goff David et al. 2014).

Polygenic risk scores (PRS) aggregate the effects of numerous genetic variants to estimate inherited susceptibility to CAD (Khera et al. 2018; Inouye et al. 2018). While PRS improve prediction in European cohorts, their portability across ancestries remains limited due to differences in allele frequencies and linkage disequilibrium (Duncan et al. 2019). As a result, uncalibrated PRS often misestimate risk in South Asian populations, which are genetically diverse yet underrepresented in genome-wide association studies (GWAS) (Martin et al. 2019; Breedon et al. 2023).

Recent Indian genome initiatives such as *IndiGen* provide population-specific allele-frequency data that enable recalibration of PRS models (Jain et al. 2021). However, large-scale genotyping remains rare in Indian healthcare, where electronic health records (EHRs) offer the most comprehensive longitudinal clinical data (Sharma, Kumar, and Tyagi 2023). This gap limits the integration of genomic risk information into routine clinical prediction.

We address this challenge through a **censoring-aware knowledge distillation framework** that transfers genomic risk into EHR-based survival models. A *genomic teacher*, constructed by recalibrating GWAS-derived CAD effect sizes with IndiGen allele frequencies, outputs population-adapted risk and survival estimates. A *student survival model*, trained on longitudinal EHR data from approximately 5,000 cardiac patients in Maharashtra, learns to replicate these outputs while fitting observed time-to-event outcomes.

To accommodate censoring, the training objective combines the Cox partial-likelihood with an inverse-probability-of-censoring-weighted (IPCW) distillation term (Kvamme, Borgan, and Scheel 2019; Kuo et al. 2024), enabling effective learning from both observed and censored samples.

Preliminary results show that this population-adapted distillation improves discrimination and calibration compared to EHR-only baselines, while maintaining interpretability. The approach highlights how population-specific genomic adaptation using open-source Indian allele-frequency data can enable scalable, deployable, and ancestry-aware survival risk models.

Related Work

Polygenic Risk Scores for Coronary Risk.

Polygenic risk scores (PRS) aggregate the effects of thousands of genetic variants to quantify inherited susceptibility to common diseases such as coronary artery disease (CAD). Large biobank studies have shown that individuals in the highest decile of a CAD PRS can exhibit a multi-fold increase in event risk relative to the lowest decile, highlighting the utility of germline genetics for early risk stratification (Fahed and Natarajan 2023; Lindström et al. 2022).

However, PRS models derived from European-ancestry genome-wide association studies (GWAS) often show degraded performance when transferred to non-European populations, primarily due to differences in allele frequencies, linkage disequilibrium structure, and environmental context (Mostafavi et al. 2019; Moreno-Grau et al. 2024; Lambert et al. 2023). Recent work demonstrates that even within a single ancestry, prediction accuracy varies across socio-economic and demographic strata, suggesting limits to current portability (Mostafavi et al. 2019; Martin et al. 2024). To mitigate these effects, multi-ancestry and cross-population adaptation methods have been proposed, such as polygenic transcriptome risk scores (PTRS) that rely on predicted gene expression rather than raw variants (Mancuso et al. 2021), or functional-variant prioritisation frameworks that improve trans-ancestral portability (Kelley et al. 2022).

Knowledge Distillation for Biomedical Prediction.

Knowledge distillation (KD) transfers knowledge from a complex "teacher" model to a simpler "student" model to improve generalisation or interpretability. Surveys such as (Mansourian et al. 2024; Moslemi et al. 2024) summarise the rapid evolution of KD from model compression to cross-domain knowledge transfer. In biomedicine, KD has been applied to multi-omics integration and sparse survival modelling, where it helps produce compact, interpretable learners without major loss in accuracy. For example, recent work by Wei et al. (Wei et al. 2024) introduces censoring-aware distillation losses for genomic survival tasks, demonstrating that KD can effectively handle right-censored time-to-event data.

Gap and Motivation.

While both PRS portability and knowledge distillation have been studied independently, their integration remains largely unexplored. Specifically, there is no existing framework that combines *population-adapted genomic risk* with *censoring-aware knowledge distillation* in an EHR-based survival model for underrepresented populations such as India. This gap motivates our proposed approach, which distils a population-recalibrated genomic teacher, based on IndiGen allele frequencies, into a deployable, interpretable EHR survival model.

Methodology

Genomic Teacher Construction

We begin by constructing a genomic teacher model that quantifies inherited CAD risk for the Indian population. Starting with GWAS summary statistics from large consortia (e.g., multi-ancestry CAD meta-analyses (Patel et al. 2023)), we build a PRS per individual (or per stratum) weighted by effect sizes. We then use allele-frequency data from the IndiGen consortium to recalibrate the PRS weights so that allele frequency differences between Indian and European populations are accounted for. The recalibrated PRS is either converted into a continuous risk score z_i or, when possible, a survival hazard estimate $\hat{S}_T(t \mid x_i)$ for each individual i.

Student Survival Model & Distillation

Our student model is a parsimonious survival model (e.g., Cox proportional hazards or Weibull/AFT) trained on routine EHR features $X_i^{(S)}$ from a longitudinal Indian cohort (N≃5,000). The student is optimised with a composite loss:

$$\mathcal{L} = \alpha\,\mathcal{L}_{\text{surv}}(X^{(S)}, t, \delta) + \beta\,\mathcal{L}_{\text{distill}}(f(X^{(S)}), z) + \gamma\,\mathcal{R}(\theta)\,,$$

where $\mathcal{L}_{\text{surv}}$ is the Cox partial-likelihood loss (or AFT negative log-likelihood) on survival time t, event indicator δ. The distillation term $\mathcal{L}_{\text{distill}}$ is defined as an IPCW (inverse-probability-of-censoring weighted) mean-squared error between student output and teacher target z. $\mathcal{R}(\theta)$ is a regularizer (e.g., L_1 for sparsity). At deployment, the student uses only clinical EHR inputs and no genomic data.

Evaluation and Ablations

We recommend a 5-fold cross-validation scheme within the Indian EHR cohort. Key metrics include Harrell's C-index for discrimination, calibration plots at 1-, 3-, and 5-year horizons, and net reclassification improvement (NRI) compared to a baseline EHR model without distillation. Ablation experiments isolate the contributions of: (i) including distillation vs. not, (ii) recalibrating PRS with Indian allele frequencies vs. naïve transfer, and (iii) student regularisation for sparsity.

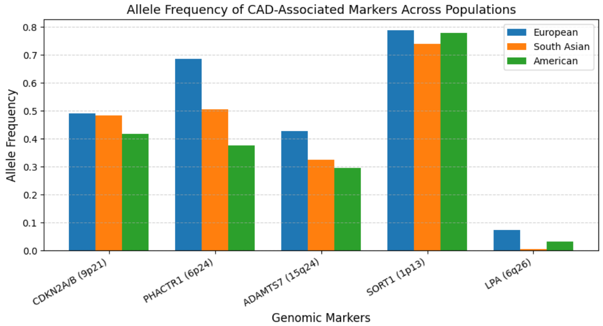

Figure 1: Allele frequency distribution of key CAD-associated loci across populations.

Demonstration

The proposed framework leverages well-established coronary artery disease (CAD)–associated loci. Our study is currently studying nearly 50+ well known alleles and their variants that are closely associated with cardiac risk. We are studying several public genomic datasets in the Indian contexts including the *IndiGenomes* project (Jain et al. 2021). Figure 1 presents the allele frequency distribution of five example variants near *CDKN2A/B* (9p21), *PHACTR1* (6p24), *ADAMTS7* (15q24), *SORT1* (1p13), and *LPA* (6q26) from the *IndiGenomes* dataset. The variation across European, South Asian, and American cohorts highlights the need for ancestry-specific calibration of polygenic risk scores to improve predictive accuracy in Indian populations. Variants near *CDKN2A/B* modulate vascular smooth muscle proliferation and atherosclerotic plaque stability; *PHACTR1* influences endothelial nitric oxide signaling; *ADAMTS7* promotes arterial wall remodeling; *SORT1* regulates hepatic lipid metabolism; and *LPA* strongly affects circulating lipoprotein(a) levels. These markers form the basis of the recalibrated polygenic risk score (PRS) that guides the genomic teacher model within our knowledge distillation pipeline.

During the conference demonstration, we will present the full end-to-end workflow, from recalibration of GWAS-derived PRS using *IndiGenomes* allele frequencies to training of a censoring-aware knowledge distillation model that integrates genomic risk with longitudinal EHR-based survival data. Interactive visualizations will showcase key steps, including recalibrated PRS distributions, student-predicted survival trajectories, and comparative performance metrics (C-index and calibration). The live session will highlight how population-specific genomic information can be distilled into interpretable, deployable EHR survival models, emphasizing scalability, adaptability, and relevance to precision-cardiology applications in resource-limited healthcare settings.

References

Breedon, J. R.; Marshall, C. R.; Giovannoni, G.; van Heel, D. A.; Dobson, R.; and Jacobs, B. M. 2023. Polygenic risk score prediction of multiple sclerosis in individuals of South Asian ancestry. *Brain Communications*, 5(2): fcad041.

Duncan, L.; Shen, H.; Gelaye, B.; Meijsen, J.; Ressler, K.; Feldman, M.; Peterson, R.; and Domingue, B. 2019. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature communications*, 10(1): 3328.

Fahed, A. C.; and Natarajan, P. 2023. Clinical applications of polygenic risk score for coronary artery disease through the life course. *Atherosclerosis*, 386.

Goff David, C.; Lloyd-Jones Donald, M.; Bennett, G.; Coady, S.; D'Agostino Ralph, B.; Gibbons, R.; Greenland, P.; Lackland Daniel, T.; Levy, D.; O'Donnell Christopher, J.; et al. 2014. ACC/AHA guideline on the assessment of cardiovascular risk. *Circulation*, 129(25_suppl_2): S49–73.

Gupta, R.; and Xavier, D. 2018. Hypertension: The most important non communicable disease risk factor in India. *Indian heart journal*, 70(4): 565–572.

Inouye, M.; Abraham, G.; Nelson, C.; Wood, A.; Sweeting, M.; Dudbridge, F.; et al. 2018. Genomic risk prediction of coronary artery disease in nearly 500,000 adults: implications for early screening and primary prevention. bioRxiv 250712. 2018 Jan 19. *Publisher Full Text*.

Jain, A.; Bhoyar, R. C.; Pandhare, K.; Mishra, A.; Sharma, D.; Imran, M.; Senthivel, V.; Divakar, M. K.; Rophina, M.; Jolly, B.; et al. 2021. IndiGenomes: a comprehensive resource of genetic variants from over 1000 Indian genomes. *Nucleic acids research*, 49(D1): D1225–D1232.

Kelley, D. R.; et al. 2022. Enhancing portability of trans-ancestral polygenic risk scores via functional variant prioritization. *PLOS Genetics*, 18(12): e1011356.

Khera, A. V.; Chaffin, M.; Aragam, K. G.; Haas, M. E.; Roselli, C.; Choi, S. H.; Natarajan, P.; Lander, E. S.; Lubitz, S. A.; Ellinor, P. T.; et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, 50(9): 1219–1224.

Kuo, N. I.; Gallego, B.; Jorm, L.; et al. 2024. Ck4gen: A knowledge distillation framework for generating high-utility synthetic survival datasets in healthcare. *arXiv preprint arXiv:2410.16872*.

Kvamme, H.; Borgan, Ø.; and Scheel, I. 2019. Time-to-event prediction with neural networks and Cox regression. *Journal of machine learning research*, 20(129): 1–30.

Lambert, M.; et al. 2023. Principles and methods for transferring polygenic risk scores across populations. *Genome Medicine*, 15(1): 97.

Lindström, S.; et al. 2022. A polygenic risk score improves risk stratification of coronary artery disease: a large-scale prospective Chinese cohort study. *European Heart Journal*, 43(18): 1702–1711.

Mancuso, N.; et al. 2021. Polygenic transcriptome risk scores improve portability across human ancestry groups. *Genome Biology*, 22: 343.

Mansourian, A.; et al. 2024. A comprehensive survey on knowledge distillation. *OpenReview*.

Martin, A. R.; Kanai, M.; Kamatani, Y.; Okada, Y.; Neale, B. M.; and Daly, M. J. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4): 584–591.

Martin, A. R.; et al. 2024. Evaluation of polygenic risk scores across ancestries highlights limitations in generalizability. *Nature Genetics*, 56: 611–621.

Moreno-Grau, S.; et al. 2024. Polygenic risk score portability for common diseases across global populations. *Human Genomics*, 18(1): 41.

Moslemi, A.; et al. 2024. A survey on knowledge distillation: recent advancements. *Computer Science Review*, 52: 100579.

Mostafavi, H.; et al. 2019. Variable prediction accuracy of polygenic scores within an ancestry group. *eLife*, 8: e48376.

Patel, A.; et al. 2023. A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat. Med.* Https://www.nature.com/articles/s41591-023-02429-x.

Prabhakaran, D.; Jeemon, P.; and Roy, A. 2016. Cardiovascular diseases in India: current epidemiology and future directions. *Circulation*, 133(16): 1605–1620.

Sharma, P.; Kumar, T.; and Tyagi, S. 2023. A Study on Existing EHR Models Used for Validating the Clinical Records. In *International Conference On Innovative Computing And Communication*, 419–442. Springer.

Wei, X. S.; et al. 2024. Data tells the truth: A knowledge distillation method for genomic survival analysis by handling censoring. *Computers in Biology and Medicine*, 180: 108617.