Machine Learning Predicts Hepatocellular Carcinoma Risk from Routine Clinical Data: A Large Population-Based Multicentric Study

Info

PUBLISHED DATE
2026-03-26

JOURNAL
Cancer Discovery

LINK
https://doi.org/10.1158/2159-8290.cd-25-1323

Authors

Jan Clusmann Paul-Henry Koop David Y. Zhang Felix van Haag Omar S.M. El Nahhas Tobias Seibel Laura Žigutytė Apichat Kaewdech Julien Calderaro Frank Tacke Tom Luedde Daniel Truhn Tony Bruns Kai Markus Schneider Jakob Nikolas Kather Carolin V. Schneider

Abstract

Hepatocellular carcinoma (HCC) is a highly fatal tumor, for which risk stratification is crucial yet remains challenging. In this study, we develop an interpretable machine learning (ML) framework for HCC risk stratification based on routinely collected clinical data. We utilize prospectively collected multimodal data from more than 900,000 individuals and 983 cases of HCC across two population-scale cohorts: the UK Biobank study (development) and the All of Us Research Program (external testing). We assess individual and cumulative contributions of data modalities, including demographics, lifestyle, health records, blood, genomics, and metabolomics. Our final random forest–based models significantly outperform all publicly available state-of-the-art risk scores on both internal and external test sets. We demonstrate robustness across ethnic subgroups, provide comprehensive interpretability, and release all code, model weights, and a web calculator for external validation and agentic integration. Our study presents PRE-Screen-HCC, a robust and interpretable ML framework for HCC risk stratification and early detection.

Significance:

Using data from population-scale cohorts, we develop and externally validate an ML framework for HCC risk stratification. Models trained on routine clinical data outperform published scores, perform on par with metabolomics and genomics, generalize across subgroups, and remain interpretable.