Supplementary Material to

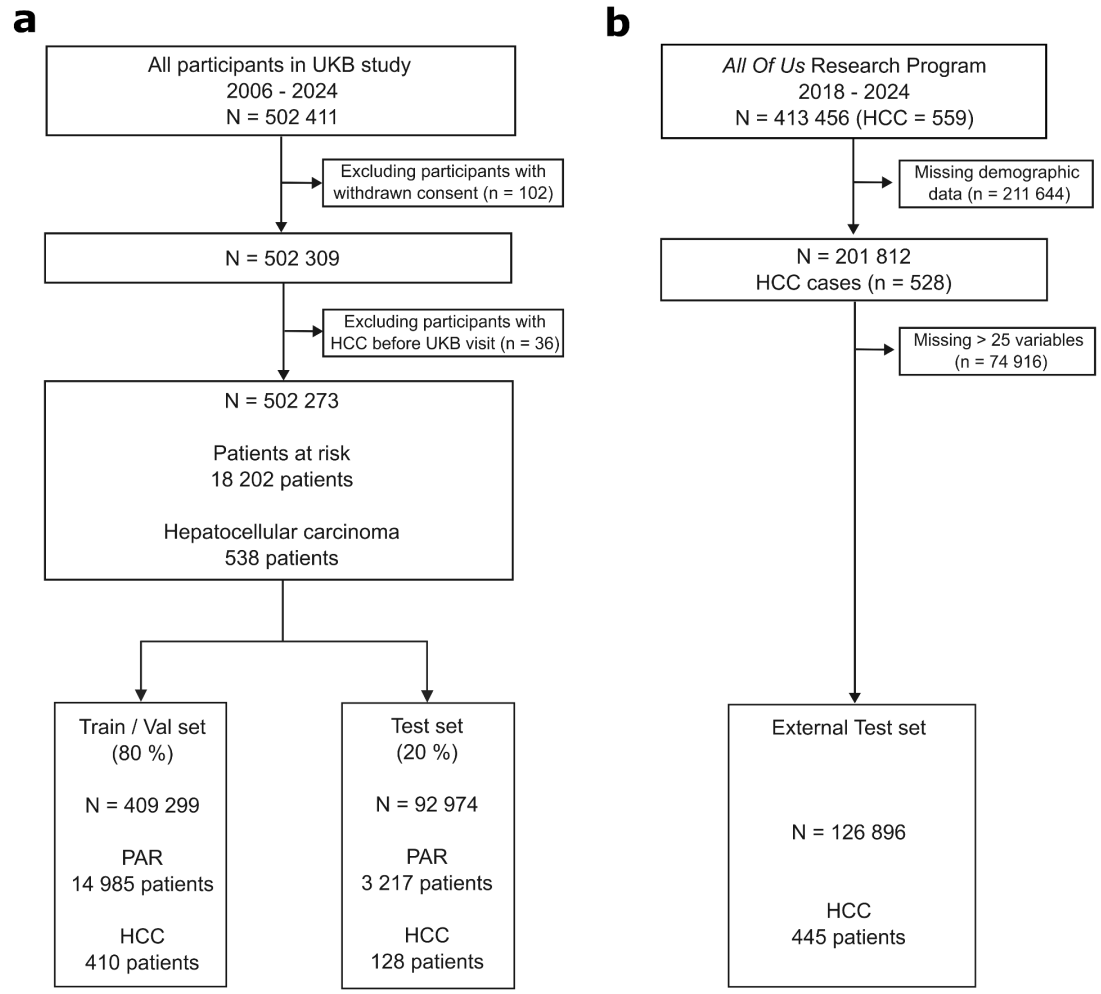
Machine learning predicts liver cancer risk from routine clinical data: a large population-based multicentric study

Jan Clusmann (1, 2), Paul-Henry Koop (1), David Y. Zhang (3,4), Felix van Haag (1), Omar S. M. El Nahhas (2, 5), Tobias Seibel (1), Laura Žigutytė (2), Apichat Kaewdech (6), Julien Calderaro (7, 8, 9, 10), Frank Tacke (11), Tom Luedde (12), Daniel Truhn (13), Tony Bruns (1), Kai Markus Schneider (1, 2, 14, 15), Jakob N. Kather (2, 14, 16, ‡), Carolin V. Schneider (1, ‡, *)

1. Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany
2. Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany
3. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA
4. Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA.
5. StratifAI GmbH, Dresden, Germany
6. Gastroenterology and Hepatology Unit, Division of Internal Medicine, Faculty of Medicine, Prince of Songkla University, Songkhla 90110, Thailand
7. Université Paris Est Créteil, INSERM, IMRB, F-94010, Créteil, France
8. Assistance Publique-Hôpitaux de Paris, Henri Mondor-Albert Chenevier University Hospital, Department of Pathology, Créteil, France
9. Inserm, U955, Team 18, Créteil, France
10. European Reference Network (ERN) RARE-LIVER, Créteil, France
11. Department of Hepatology and Gastroenterology, Charité - Universitätsmedizin Berlin, Campus Virchow-Klinikum and Campus Charité Mitte, Berlin, Germany
12. Department for Gastroenterology, Hepatology and Infectiology, University Hospital Düsseldorf, Germany
13. Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Germany
14. Department of Medicine I, University Hospital Dresden, Dresden, Germany
15. Center for Regenerative Therapies Dresden (CRTD), Technische Universität (TU), Dresden, Germany.
16. Department of Medical Oncology, National Center for Tumor Diseases (NCT), Heidelberg University Hospital, Heidelberg, Germany

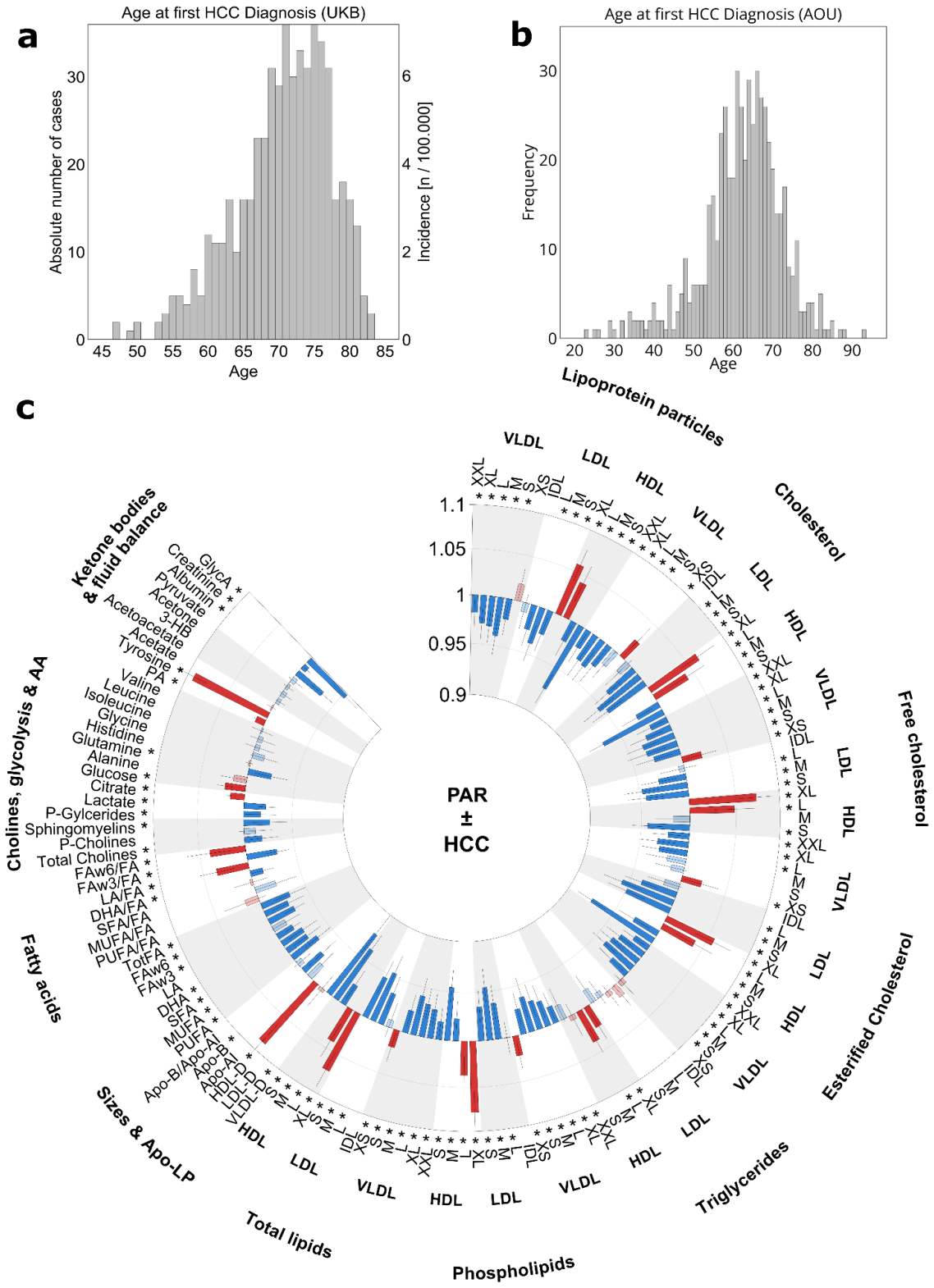
Table of Contents:

Supplementary Figures



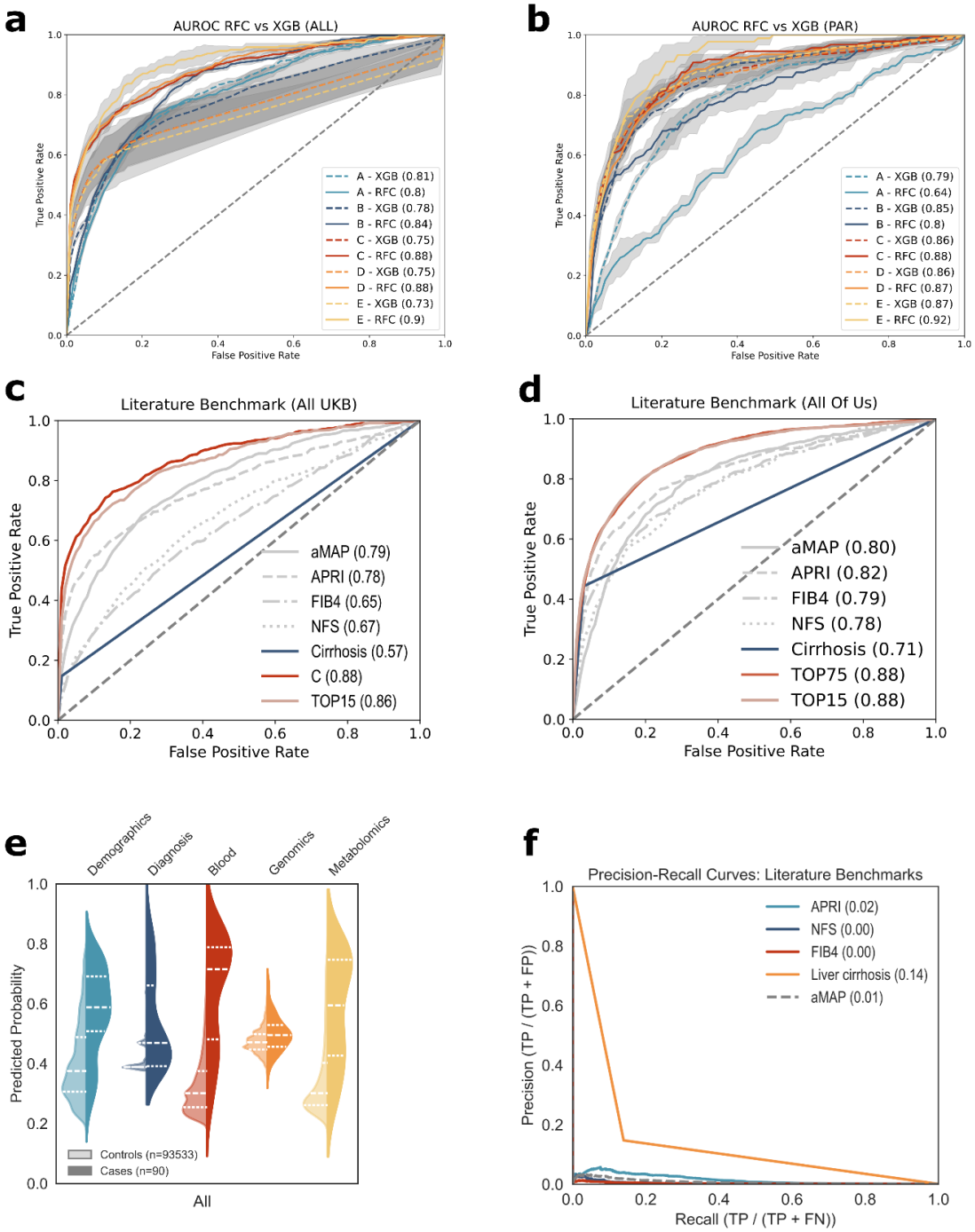
Supplementary Fig. 1: Data processing flowchart

a Data processing flowchart for UKB participants. Val = Validation. **b** Data processing flowchart for participants in "All Of Us" Research Program



Supplementary Fig. 2: Preprocessing of study participants.

a, b Age in years of participants at first-documented diagnosis of HCC. **a** for UKB and **b** for All Of Us Research Program. **c** Associations of metabolic biomarkers with development of HCC. Hazard ratios + 95% confidence intervals as 1-SD higher metabolic biomarker on the natural log scale, stratified by age, sex, BMI, compared to cohort with pre-diagnosed liver disease or elevated liver enzymes, jointly defined as patients-at-risk (PAR). *False discovery rate-controlled p < 0.01.



Supplementary Fig. 3: Additional performance metrics

a, b Model performance on the test set of UKB as receiver operating characteristic (ROC) curves and corresponding area under the curve (AUC) for either all participants (a) or patients with pre-diagnosed liver disease (b). Each line represents the performance of one majority vote model from five-fold cross-validation ± standard deviation of true positive rate, either for the random forest classifier (straight line) or extreme gradient boost (dotted line). **c** Performance of various established linear risk scores, cirrhosis as binary marker for all of UKB (England, Scotland, Wales), plotted against model performance of newly developed Model C and Model TOP15 for UKB test set performance (Scotland, Wales). **d** Risk scores as in c, but for AOU cohort. Models TOP75 and TOP15 applied to the whole AOU cohort. **e** Split violin distributions for prediction scores of separate models per modality for UKB cohort “All”. **f** Precision-recall curves for literature benchmarks in UKB “All” cohort TP = True Positives, FP = False Positives, FN= False Negatives.

Table of content for supplementary tables (see Supplementary_Material.xlsx)

Table	Description	Cohort
Supplementary Table 1	ICD10 Codes used for Patients at risk definition	UKB/AOU
Supplementary Table 2	ICD10 Codes defined as single-entity feature	UKB/AOU
Supplementary Table 3	ICD10 Codes defined as grouped feature	UKB/AOU
Supplementary Table 4	Assessment Features	UKB
Supplementary Table 5	Blood Count Features	UKB/AOU
Supplementary Table 6	Metabolomics Features	UKB
Supplementary Table 7	Single Nucleotide Polymorphisms Data	UKB
Supplementary Table 8	Mapping for self-reported diagnosis codes	UKB/AOU
Supplementary Table 9	At-risk subsets and respective HCC/control cohort sizes	UKB
Supplementary Table 10	Center splits	UKB
Supplementary Table 11	HCC cases per center	UKB
Supplementary Table 12	Mappings UKB_AOU MinMax Median Mean and Conversion	UKB/AOU
Supplementary Table 13	Table ICD	UKB
Supplementary Table 14	Table ICD	AOU
Supplementary Table 15	Table Blood Count/Biochemistry stratified for HCC and SEX	UKB
Supplementary Table 16	Table Blood Count/Biochemistry stratified for HCC and SEX	AOU
Supplementary Table 17	Metabolomics Stratified for HCC	UKB
Supplementary Table 18	Model Performances threshold-independent (AUROCs, AUPRCs)	UKB

Supplementary Table 19	Model performances threshold-dependent	UKB
Supplementary Table 20	Model Performances threshold-independent (AUROCs, AUPRCs)	AOU
Supplementary Table 21	Model performances threshold-dependent	AOU
Supplementary Table 22	Prediction Value Comparison	UKB/AOU
Supplementary Table 23	Delong Tests	UKB
Supplementary Table 24	Delong Tests	AOU
Supplementary Table 25	Feature Importances Model A-E + TOP75-TOP15 UKB	UKB