Development and validation of a model to predict incident chronic liver disease in the general population: The CLivD score

Åberg F, Luukkonen PK, But A, Salomaa V, Britton A, Petersen KM, Bojesen SE, Balling M, Nordestgaard BG, Puukka P, Männistö S, Lundqvist A, Perola M, Jula A, Färkkilä M

Table of contents

Description of data and statistical analyses

FINRISK and Health 2000 cohort description

The methods, measurements and protocols used in the FINRISK and Health 2000 studies have been essentially the same over time and are similar to those in the Health 2000 Survey [1,2]. Data were collected from each participant at baseline via interviews (Health 2000), questionnaire and health examination by trained physicians and nurses (Health 2000) and trained nurses (FINRISK) using standardized procedures of the MONICA [3] and European Health Risk Monitoring projects [4]. Blood samples collected at baseline for a wide spectrum of laboratory measurements were handled using a standardized protocol. Detailed descriptions of study protocols have been published previously [1,2]. All participants provided signed informed consent, and the studies were approved by the Coordinating Ethical Committee of the Helsinki and Uusimaa Hospital District (previously studies also approved by the institutional review board of the National Public Health Institute, both in Helsinki, Finland). The FINRISK and Health 2000 sample collections were transferred to THL Biobank in 2015 after approval of the Coordinating Ethical Committee of the Helsinki and Uusimaa Hospital District.

Follow-up data were obtained from several national registers through linkage using the unique personal identity code assigned to all Finnish residents. Data for hospitalizations were obtained from the Care Register for Health Care (HILMO), which covers all hospitalizations in Finland since 1969. One or several ICD-diagnoses are assigned to each hospitalization at discharge; these diagnosis codes are systematically recorded in the HILMO register. Data for malignancies were

obtained from the Finnish Cancer Registry, with nationwide cancer records since 1953. Vital status and cause of death data were obtained from Statistics Finland. In Finland, each person who dies is by law assigned a cause of death (in accordance with the ICD) to the official death certificate, issued by the treating physician based on medical or autopsy evidence, or forensic evidence when necessary; the death codes are then verified by medical experts at the register and recorded according to systematic coding principle. Data collection to all these registries is mandatory by law and general quality is consistent and virtually 100% complete [5,6].

Statistical analyses

The primary variables of interest were objective, readily available, simple and reproducible factors identified *a priori* based on previously published data, clinical rationale and their ease of use in primary care settings; however, we were limited to factors available in the dataset.

In the derivation cohorts, respondents were asked to report how often they consumed alcoholic beverages during the previous year and the average amount they consumed per week during the previous month. Average alcohol intake (grams per day) was calculated as the sum of the daily number of drinks multiplied by the average alcohol content per type of alcoholic beverage. In the Whitehall II data, respondents reported the number of drinks they consumed in the previous week, and we used the average alcohol intake reported over the follow-up visits (phases) 1-5. One drink was defined as 10 grams of ethanol in line with recent guidelines [7]. Participants were also asked if they had been abstinent their entire life (lifetime abstainer) or had used alcohol earlier and then stopped (current abstainer). Binge

drinking was defined as drinking 5 or more alcohol drinks per occasion. Respondents reported the number of times during the last 12 months that they consumed 5 or more drinks per occasion. Smoking status (active smoker, former or never smoker) and number of daily cigarettes were asked. Waist and hip circumference were measured using standard techniques as previously described [1,2]. Exercise was assessed by asking how often the subject performs leisure-time physical exercise for at least 20-30 minutes so that he/she is at least slightly out of breath and sweaty. Diabetes was defined either by a fasting serum glucose ≥7.0 mmol/L (126 mg/dL), taking diabetic medication, or by having a prior known diabetes diagnosis.

We developed two parallel models, one based on non-laboratory values only ($Model_{non\text{-}lab}$), and one based on the same variables and additionally including laboratory values ($Model_{lab}$).
Gamma-glutamyl transferase (GGT) was considered as a marker of liver damage and/or oxidative stress.

GGT was chosen as the primary analyte of interest over other common liver enzymes because of several reasons:

1) GGT is a stronger predictor of incident clinical liver disease than ALT or AST ir previous studies [8].
2) GGT shows highest sensitivity for liver disease above other liver tests and is a good predictor of liver disease and liver mortality [9].

3) GGT is a more sensitive detector of hepatic steatosis than ALT or AST, and contributes to several algorithms for the diagnosis of NAFLD (Fatty Liver Index and SteatoTest) and liver fibrosis (Fibrotest and Hepascore) [10].

4) GGT is an acknowledged trigger for further liver fibrosis assessment in fatty liver disease according to expert opinion [11].

5) GGT reflects body oxidative stress [12], which is implicated in the pathophysiology of chronic liver disease [13].

Modification of some variables on clinical grounds

Alcohol use was assessed as number of drinks (á 10g of ethanol) per week. Alcohol use was capped at 50 drinks per week, because higher intake is generally associated with considerable reporting bias (underreporting), there were few subjects with higher consumption in our dataset, higher intake is clearly associated with severe health risks (both liver-related and other diseases) regardless of our risk prediction model. This means that, in such persons, alcohol-reducing interventions are merited anyway. In addition, UK guidelines recommend liver evaluations for those with alcohol consumption of more than 50 drinks per week [14].

GGT was capped at 200 U/L, because higher values deserve further evaluation regardless of our risk prediction model, and there were few subjects with GGT >200 U/L in our cohorts, thus resulting in substantial uncertainty around risk estimates when GGT is above 200 U/L.

Imputation of missing baseline values

Baseline data with ≤5% missingness were imputed by 5 multiple imputations using the predictive mean matching method in the mice package in R-software for the following continuous variables waist circumference, waist-hip ratio, body mass index, alcohol use (drinks/week), HDL-cholesterol, GGT, triglycerides, and non-HDL-cholesterol. Missing values were predicted based on these same variables as well as age, sex, diabetes, smoking status, and the liver outcome.

In the derivation cohorts, data on exercise, binge drinking, alanine aminotransferase (ALT), aspartate aminotransferase (AST), and homeostasis model assessment of insulin resistance (HOMA-IR) were missing in >15% because these variables had not been assessed in all sub-cohorts. These variables were excluded from initial variable selection, but later tested in the complete-case dataset whether they improved model fit and performance. These variables were not imputed due to the high missingness rate.

Model building

Candidate variables were tested for association with liver outcomes by univariate and multivariable Cox regression analyses with incident liver disease as the outcome. The proportional hazards assumption of the Cox models was checked using Schoenfeld residuals, and no violations were detected. Continuous predictors were assessed for possible nonlinear relationship with outcome using restricted cubic splines with degrees of freedom selected using the Akaike Information Criterion (AIC; a smaller AIC-value is better). These procedures were repeated for the multivariable model. To limit collinearity, among variables with a Spearman correlation coefficient

of >0.6 (Supplementary figure 1), we chose the variable judged to be clinically more important. In addition, in the multivariable model, we tested whether model fit was improved by replacing these variables with the omitted correlated variables. Multicollinearity was also assessed with the variance inflation factors. Predictors of liver outcomes with a univariate P<0.2 were examined in the multivariable analysis, and the final model was selected by backward stepwise elimination separately with P>0.05 or AIC as removal criteria, except for age, which was retained in the model as a measure of exposure time regardless of statistical significance, assuming that older individuals have longer exposure times. Improvement in model fit was subsequently tested by re-insertion one at a time of the removed predictors. Two-way interactions among the variables in the final model and with sex were investigated and included in the final model if they improved model fit. We assessed model fit and model performance with AIC and and Wolbers' C-statistic, respectively, and compared nested models using the likelihood ratio test.

Using cause-specific Cox regression considering death without liver disease as a competing-risk event, we assessed the performance of the final model in terms of discrimination (Wolbers' C and time-dependent AUC) and calibration. Unlike the commonly used Harrell's C-statistic, Wolbers' C and time-dependent AUC account for the competing risk of death without liver disease [15]. Calibration was assessed by comparing the predicted risk of incident liver disease with the observed risk using calibration plot. We did internal validation to correct measures of predictive performance for optimism (over-fitting) by bootstrapping 200 samples of the derivation data starting from the multivariable variable selection. Bootstrap estimates of calibration accuracy were performed using the calPlot function of R's pec-package.

$Model_{lab}$ was compared to $Model_{non\text{-}lab}$ by reclassification, which compares the models' abilities to correctly classify cases into correct risk categories (Table S7 and 8). Estimates of 15-year absolute risks of developing a liver outcome based on the individual risk-factor profile were visualized by color-coded scoring sheets (nomograms).

External validation was performed by calculating a risk score for each person in the validation cohorts using the individual predictors and the respective Cox regression coefficients as estimated in the derivation cohort. We then fitted a cause-specific Cox regression with this risk score as a single covariate, and examined the performance of this model by Wolbers' C-statistic, time-dependent AUC, and calibration plots and stratified subjects in risk groups using the same cutoff-values of the risk score as in the derivation cohort. To assess the extent of clinical transportability of the prediction model, we quantified the relatedness between the derivation and validation dataset, and the extent to which they share common predictor effects according to Debray et al [16].

Stepwise variable selection

The following variables were initially considered: age, sex, waist circumference (WC), waist hip ratio (WHR), body mass index (BMI), GGT, alcohol use (drinks/week), diabetes, alcohol status, smoking group and smoking status.

Of the anthropometric measures, we chose WHR because it has been shown to be superior to WC and BMI in discriminating risk for liver disease in the general

population in several studies [17-19]. However, we later analyze whether WC or BMI brings added value to model performance.

In addition, lipid levels (LDL, HDL, non-HDL and triglycerides) were used in computation of missing values, but were not included in the multivariate models due to uncertainty regarding the causal pathways for liver disease.

Variable selection for Model$_{non\text{-}lab}$

Multivariate modeling: Based on the analyses above, we started with the following variables: age, sex, diabetes, WHR, alcohol use, alcohol status and smoking status.

Stepwise backward elimination using the Akaike Information Criterion for variable selection: Factors in Final Model: age, WHR, alcohol use, diabetes, smoking status. Alcohol status and sex were removed from the model.

Alternative stepwise backward elimination using P-value <0.05 for variable selection: the same variables were chosen in the final model.

Testing linearity of predictors: only alcohol use was significantly non-linear.

Forward procedure: does insertion of alcohol status or sex one by one improve model fit?: Inclusion of sex led to a significant improvement (P=0.03, likelihood ratio test, LRT) of model fit. There was no improvement after inclusion of alcohol status.

Variable selection for Model$_{lab}$

Multivariate modeling: Based on the analyses above, we started with the following variables: age, sex, diabetes, WHR, alcohol use, alcohol status, smoking status, and GGT.

Stepwise backward elimination using the Akaike Information Criterion for variable selection: Factors in Final Model: age, WHR, alcohol use, diabetes, smoking status, and GGT. Alcohol status and sex were removed from the model.

Alternative stepwise backward elimination using P-value <0.05 for variable selection: the same variables were chosen in the final model.

Testing linearity of predictors: only alcohol use was significantly non-linear.

Forward procedure - does insertion of alcohol status or sex one by one improve model fit?: There was no improvement (P>0.05, LRT).

Improvement in model fit in various alternative scenarios

Anthropometric measures

Considering WC instead of WHR provided slightly poorer results for $model_{lab}$ performance/fit (C-statistic 0.779 vs 0.786, AIC 4070 vs 4053, p<0.001 by LRT) and $model_{non\text{-}lab}$ performance/fit (C-statistic 0.808 vs 0.809, AIC 3943 vs 3938, p<0.001 by LRT).

An alternative $model_{lab}$ with BMI (non-linear) and WHR was similar (C-statistic 0.810, AIC 3914) to a model with WHR without BMI (C-statistic 0.809, AIC 3938, p<0.001). Similar results were found for $model_{non\text{-}lab}$ (C-statistic 0.790 vs 0.786, AIC 4034 vs 4053, p<0.001). A model with BMI vs a model with WHR were fairly similar, but with a small advantage for WHR ($model_{lab}$: C-statistic 0.808 vs 0.809, AIC 3927 vs 3938, p<0.001, $model_{non\text{-}lab}$: C-statistic 0.778 vs 0.786, AIC 4062 vs 4053, p=0.446).

Furthermore, based on the relationship between BMI and risk for incident liver disease in our multivariate models (Supplementary figure 2), it seems that the added value from BMI comes solely from underweight status, possibly reflecting pre-existing illness. In other words, the risk effect from obesity is already captured in the covariates WHR or WC, but these covariates may not sufficiently capture the risk effect from underweight. We also analyzed whether a dichotomous covariate reflecting underweight (yes vs no) stratified by WHO's definition of underweight (BMI <18.5 kg/m^2) improve the models. However, there were only 83 subjects with BMI <18.5 kg/m^2 and only 2 liver events in this subgroup. Similarly, model performance did not improve after exclusion of subjects with BMI <18.5 kg/m^2. Based on this, and

considering that measures of abdominal obesity have previously been shown to be stronger predictors of liver disease than BMI [17-19], we leave BMI out from the models altogether.

Smoking

Considering smoking group (which includes also the amount of smoking; Table S3) instead of smoking status (which only considers whether the subject is a current smoker) did not improve model performance/fit ($model_{lab}$: C-statistic 0.807 vs 0.809, AIC 3944 vs 3938, p<0.001; $model_{non\text{-}lab}$: C-statistic 0.784 vs 0.786, AIC 4057 vs 4053, p<0.001). The model with the simpler smoking status was better.

Additional variables

Inclusion of additional candidate variables was tested in the complete-case dataset. Model performance/fit was not significantly improved by inclusion of binge drinking ($model_{lab}$ p=0.57, $model_{non\text{-}lab}$ p=0.12), exercise ($model_{lab}$ p=0.40, $model_{non\text{-}lab}$ p=0.18), or HOMA-IR ($model_{lab}$ p=0.07). The model with GGT was significantly better than a model with ALT ($model_{lab}$ AIC 1131 vs 1203, p<0.001), or AST ($model_{lab}$ AIC 1054 vs 1101, p<0.001).

Influence of gender

We first tested whether the interaction term between sex and the other predictors was significant in age-adjusted Cox regression analysis for incident liver events (Table S5), and plotted the age-adjusted non-linear interaction effects between sex and key predictors using splines (Supplementary figure 3). We then tested whether

the interaction term between sex and the other predictors were significant in the multivariate models (Table S6).

Inclusion of the interaction terms between sex and GGT, and between sex and smoking status alone or together improved model fit significantly in $Model_{lab}$ ($p<0.05$, LRT). Inclusion of the interaction term between sex and smoking status improved model fit significantly in $Model_{non\text{-}lab}$ ($p=0.019$, LRT).

As the interaction effect between sex and GGT was particularly profound at low levels of GGT, we also tested an interaction term between GGT <25 U/L (yes or no) and sex in $Model_{lab}$. However, inclusion of GGT and the interaction term GGT<25 * sex did not improve model fit compared to inclusion of GGT and the interaction term GGT * SEX (P=0.338).

We found no significant interaction between alcohol use and other variables in the model. Also, considering WC instead of WHR did not improve model fit when tested separately in women or men. Hazards ratios with 95% confidence intervals for each covariate in $Model_{lab}$ separately for men and women are shown in Supplementary figure 4.

Rank-hazard plots to visualize the relative importance on a population level of covariates in $Model_{lab}$ are shown in Supplementary figure 5.

Final models

The final models include the following factors:

Model$_{lab}$: age, WHR, alcohol use (spline variable), GGT, diabetes, smoking status, sex*GGT, sex*smoking status.

Model$_{non\text{-}lab}$: age, WHR, alcohol use (spline variable), diabetes, smoking status, sex*smoking status.

Alcohol use remained significantly non-linear in both models.

We then checked possible multicollinearity using the variable inflating factor (VIF) method (excluding interaction terms):

Covariate	VIF
Age	1.11
WHR	1.87
Alcohol use	1.46
GGT	1.30
Diabetes	1.06
Smoking status	1.14
Sex	1.77

All VIF values are low (<2); therefore, no significant multicollinearity seems to exist.

Risk stratification

Risk stratification based on Model$_{lab}$

Based on the model's prognostic index risk score distribution, subjects were classified into risk groups defined by an estimated 15-year cumulative probability of liver events of <0.5%, 0.5-4%, 5-9% and ≥10%, respectively. These risk groups are called "minimal risk", "low risk", "intermediate risk", and "high risk" (Supplementary figure 7-8).

We used 15-year risk as it is known that it takes on average this time for clinical liver endpoints to develop from early-stage liver disease (fibrosis stage 0-1) [20], and risk stratification on a shorter time scale may thus lead to suboptimal discrimination. This is different from studies with risk stratification based on the severity of subclinical disease (fibrosis scores) rather than pathophysiologic risk factors of disease (present models).

Risk group	No events	Liver event	Non-liver death	Proportion (liver event / non-liver death)
Minimal risk	12074	23	634	0.036
Low risk	10525	123	1320	0.093
Intermediate risk	508	27	127	0.213
High risk	269	49	81	0.605

The high-risk group comprising 1.5% of the entire population was able to capture 22% of all liver events within 15 years. Considering the high- and intermediate-risk groups together, the corresponding values were 4% and 34%, respectively.

Risk stratification based on Model$_{non\text{-}lab}$

Subjects were classified into risk groups as above (Supplementary figure 9).

Risk group	No events	Liver event	Non-liver death	Proportion (liver event / non-liver death)
Minimal risk	10257	24	550	0.044
Low risk	12195	127	1396	0.091
Intermediate risk	731	48	161	0.298
High risk	193	23	55	0.418

The high risk group comprises only 1% of the entire population, but 10% of all liver events within 15 years occurred in this risk group. Considering the high and intermediate risk groups together, the corresponding values are 5% and 32%, respectively.

External validation

The distribution of the risk scores (linear predictor) in the derivation and validation cohorts are shown in Supplementary figure 10.

Relatedness between the derivation cohort and Whitehall II cohort

Besides statistical reproducibility, we also assessed the extent of clinical transportability of $Model_{nonlab}$ by quantifying the relatedness between the derivation and Whitehall II validation dataset and the extent to which they share common predictor effects [16]. Relatedness was examined by fitting a binary logistic regression model, a membership model, to predict the probability of an individual belonging to the derivation dataset. We assessed discriminative ability of this membership model, which included age, sex, WHR, diabetes, alcohol use, smoking status, follow-up time and liver event (yes/no) as independent variables, by means of concordance (C) statistic (Supplementary figure 14). Here, a higher C-statistic indicates a lesser extent of relatedness between the datasets.

The extent to which the datasets share common predictor effects was examined by assessing the relative difference in spread (standard deviation) and the difference in mean of the model's linear predictor in the validation dataset as compared to the derivation dataset. When the derivation and validation samples have a very similar case mix, external validation provides results similar to internal validation, and, thus, adds little additional value. A higher heterogeneity in predictor-outcome associations, i.e. higher variability of the linear predictor, indicates better discriminative ability of the model. A difference in the mean of the linear predictor between the derivation and validation samples reflects the difference in the predicted frequency of the outcome,

with a large difference being indicative of the model's calibration-in-the-large in the validation sample.

A similar relatedness analysis for the CCHS cohort could not be performed because none of the researchers had at the same time access to raw data from both the derivation and validation cohorts.

The risk score equation

The equation in R software language

data = name of the R data

AGE = age in years (40-70 years)

SEX = men = 1; women = 2

WHR = waist hip ratio

ALCOHOL = number of weekly drinks (1 drink = 10 g ethanol)

Set to = 50, if >50 drinks per week

GGT = gamma glutamyltransferase (U/L)

Set to = 200 U/L if >200 U/L

DIABETES = yes = 1; no = 0

SMOKING = current smoker = 1; never/previous smoker = 2

Model$_{lab}$

```
data$modellab <- (-6.7922721 + 0.044744302* data$AGE + 0.32961593*(
data$WHR*10) + 0.19860813* data$ALCOHOL -
0.0082096868*pmax(data$ALCOHOL-0.1,0)^3
+0.010575035*pmax(data$ALCOHOL-1,0)^3 -0.002004756*pmax(data$ALCOHOL-
3,0)^3 -0.00033998925*pmax(data$ALCOHOL-9,0)^3 -2.0602882e-
05*pmax(data$ALCOHOL-33,0)^3 +0.011813962* data$GGT
+0.18721469*(data$SEX=="2") +0.55249734*(data$DIABETES=="1") +
0.74679941*(data$SMOKING=="1") +0.0054325769* data$GGT*(data$SEX=="2") -
0.64903176*( data$SEX=="2")*( data$SMOKING=="1"))
```

Model$_{non\text{-}lab}$

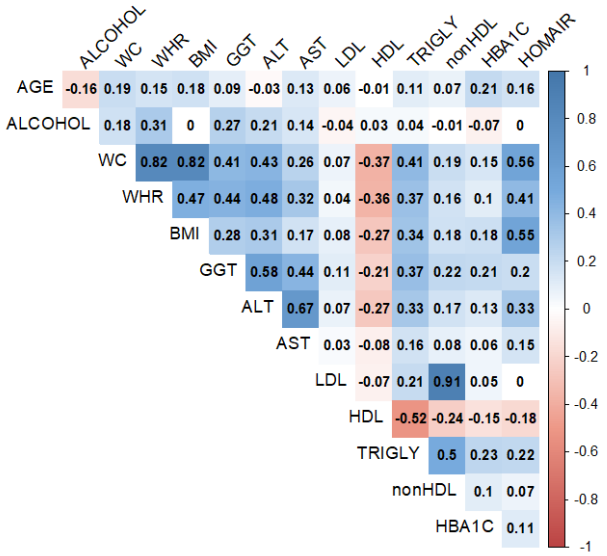
```
data$modelnonlab <- (-8.0940103 +0.044177151* data$AGE +0.48927753*(
data$WHR*10) +0.19222894* data$ALCOHOL -
0.00015029544*pmax(data$ALCOHOL-0.1,0)^3 -
0.0021265611*pmax(data$ALCOHOL-1,0)^3
+0.0029832769*pmax(data$ALCOHOL-3,0)^3 -
0.00068765143*pmax(data$ALCOHOL-9,0)^3 -1.8769011e-
05*pmax(data$ALCOHOL-33,0)^3 +0.69669285*(data$DIABETES=="1")
+0.75968055*(data$SMOKING=="1")+ 0.63248362*(data$SEX=="2") -
0.59146649*(data$SMOKING=="1")*(data$SEX=="2"))
```

Cutoff values for risk groups

	$Model_{lab}$	$Model_{non\text{-}lab}$
Minimal (15-yr risk <0.5%):	< -0.258	< -0.412
Low (15-yr risk 0.5-4%):	-0.259-2.066	-0.413-1.912
Intermediate (15-yr risk 5-9%):	2.067-2.784	1.913-2.632
High (15-yr risk ≥10%):	≥ 2.785	≥ 2.633

Supplementary figures

(A)



(B)

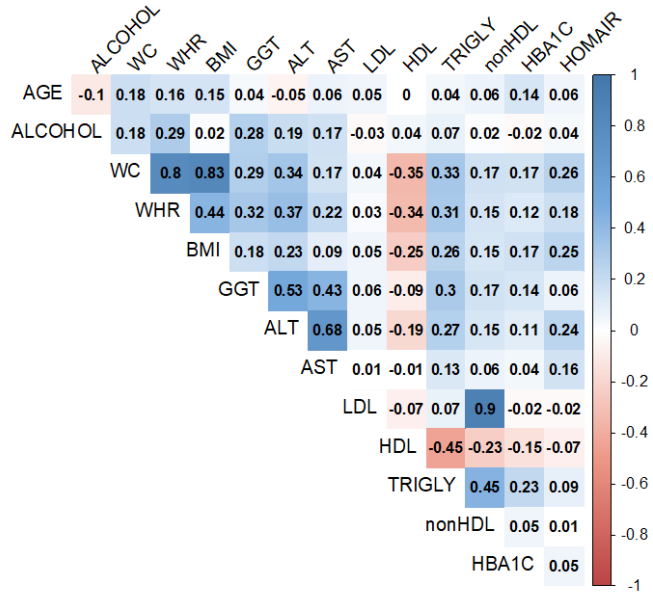


Fig. S1. Correlations between continuous variables measured at baseline. Blue indicates positive correlation coefficients, and red negative coefficients. (A) Spearman correlation and (B) Pearson correlation).

(A) (B)

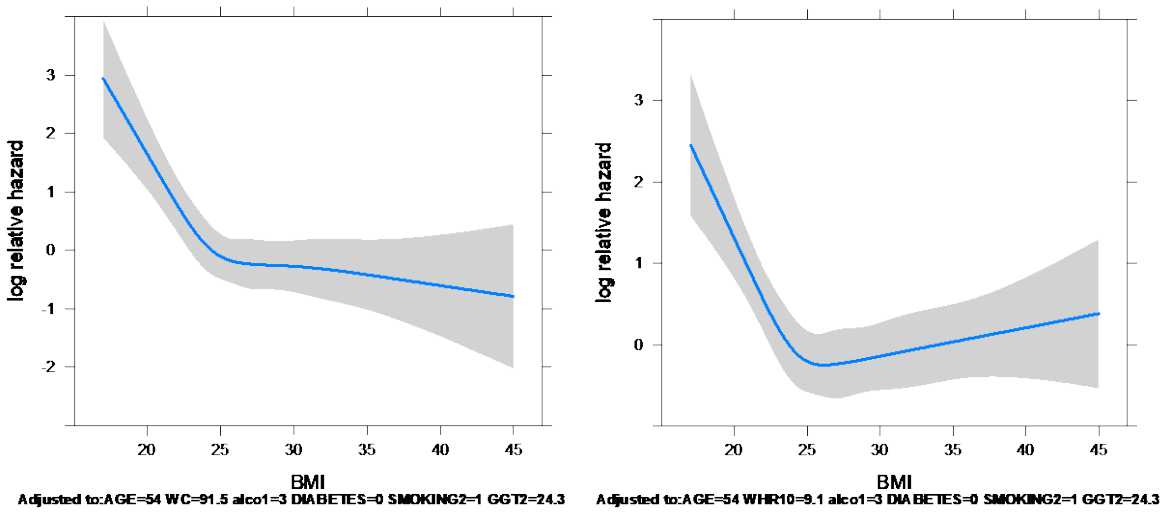


Fig. S2. The functional form of body mass index (BMI) in the multivariate model when adjusting for (A) waist circumference (WC) or (B) waist-hip ratio (WHR).

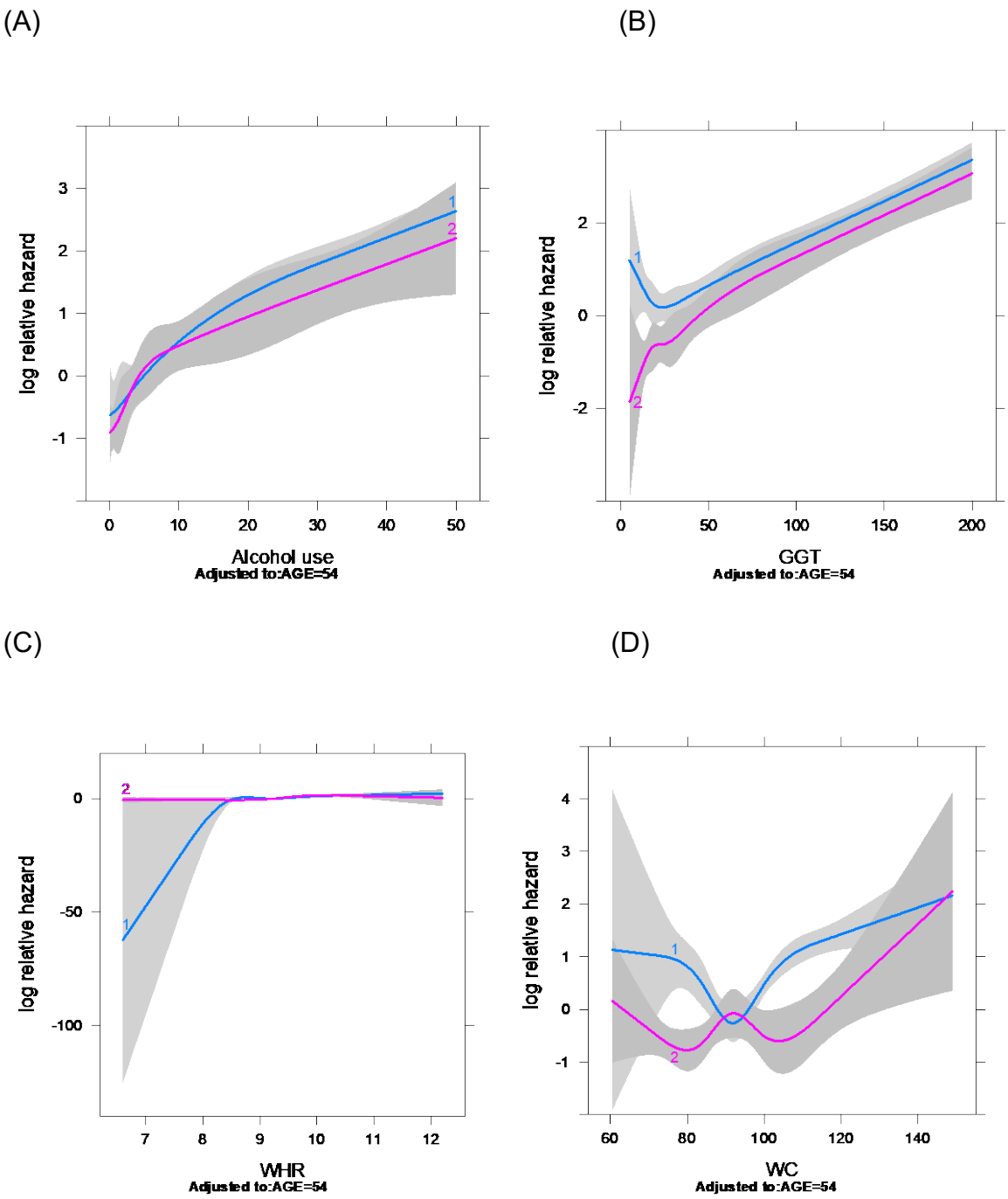


Fig. S3. Plots of the age-adjusted non-linear interaction effect between sex and key predictors using restricted cubic splines. (A) alcohol use (drinks/week), (B) gamma-glutamyltransferase (GGT), (C) waist-hip ratio (WHR), (D) waist circumference (WC). Blue (1) = men; red (2) = women.

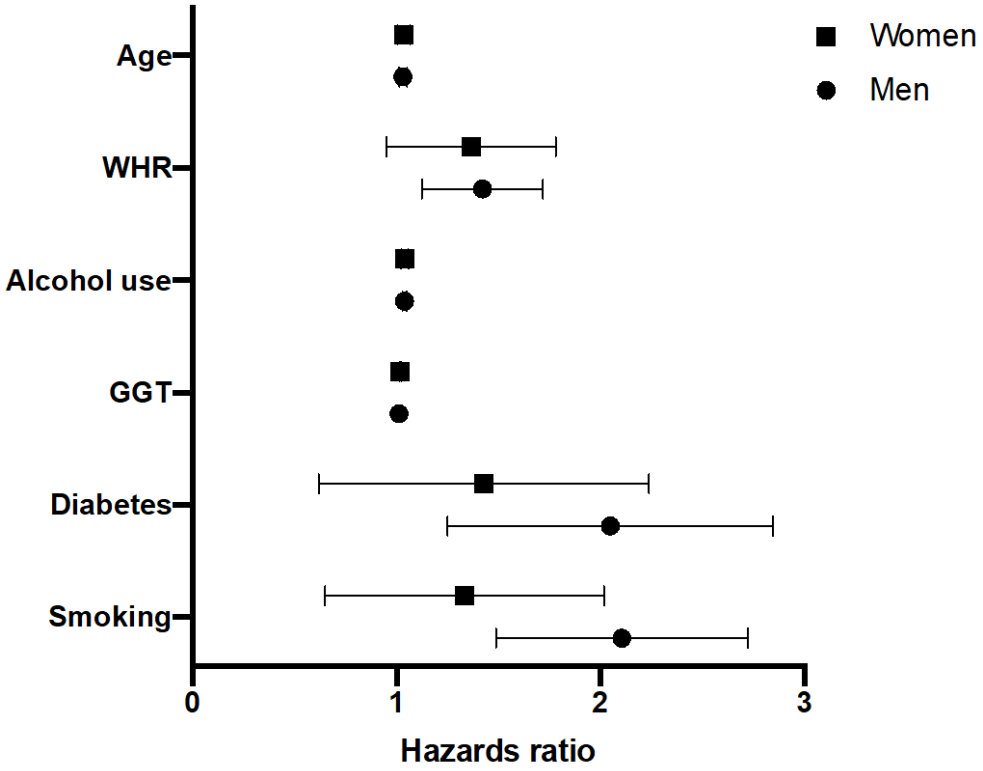


Fig. S4. Hazards ratios with 95% confidence intervals for each covariate in $Model_{lab}$ separately for men and women.

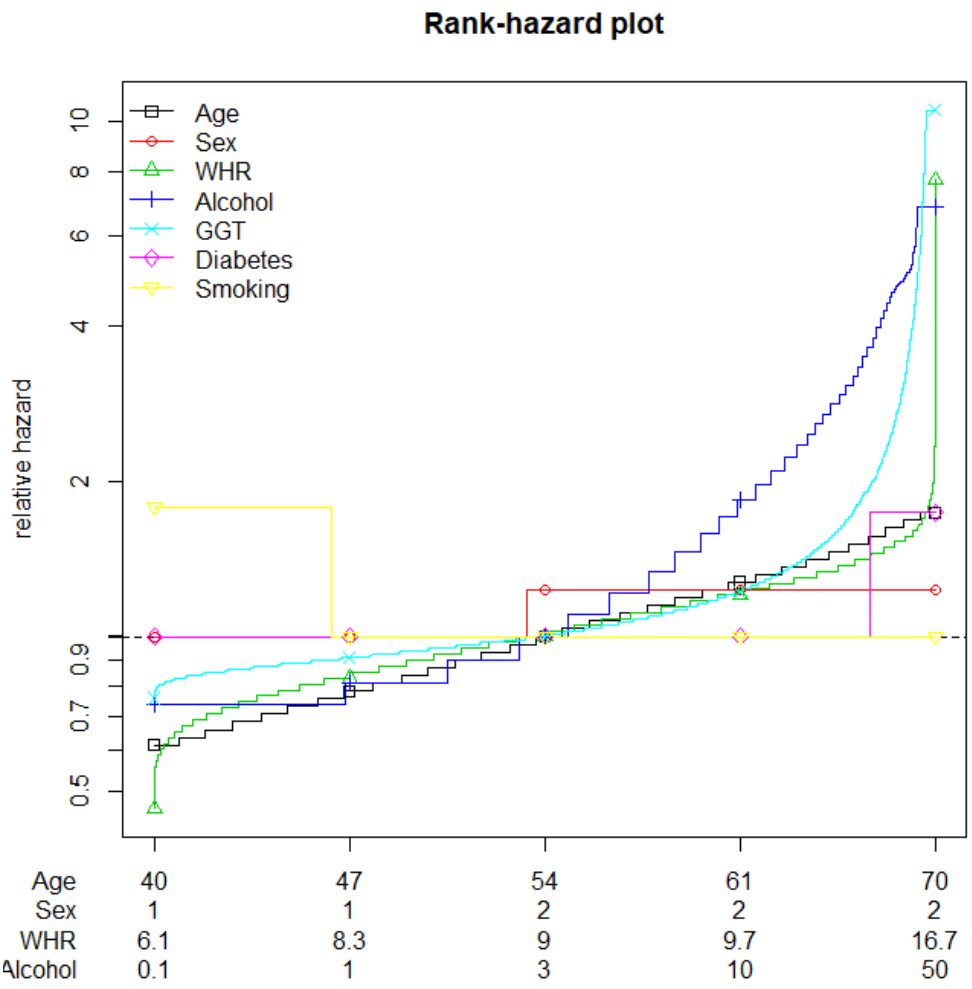


Fig. S5. Rank-hazard plots to visualize the relative importance on a population level of covariates in $Model_{lab}$. The key idea is to rank the covariate values and plot the relative hazard as a function of ranks scaled to interval (0-1). The relative hazard is the hazard plotted in respect to the reference hazard, which is set to the median of the covariate. Covariates, which are measured in different units, are scaled ranks to allow plotting them in the same graph. This takes into account not only the regression coefficients and statistical significance, but also the prevalence in the population of abnormalities in all the covariates.

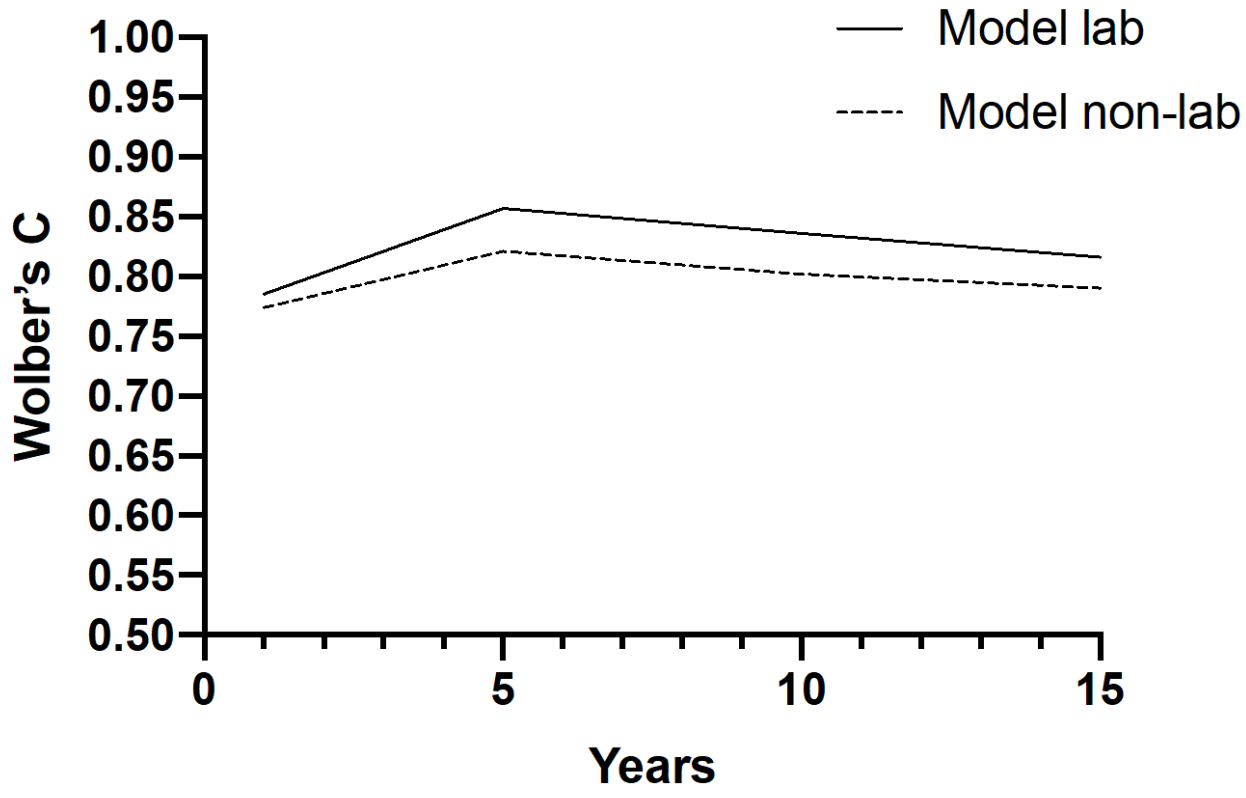


Fig. S6. Apparent Wolbers' C for $Model_{lab}$ and $Model_{non\text{-}lab}$ over 15 years of follow-up.

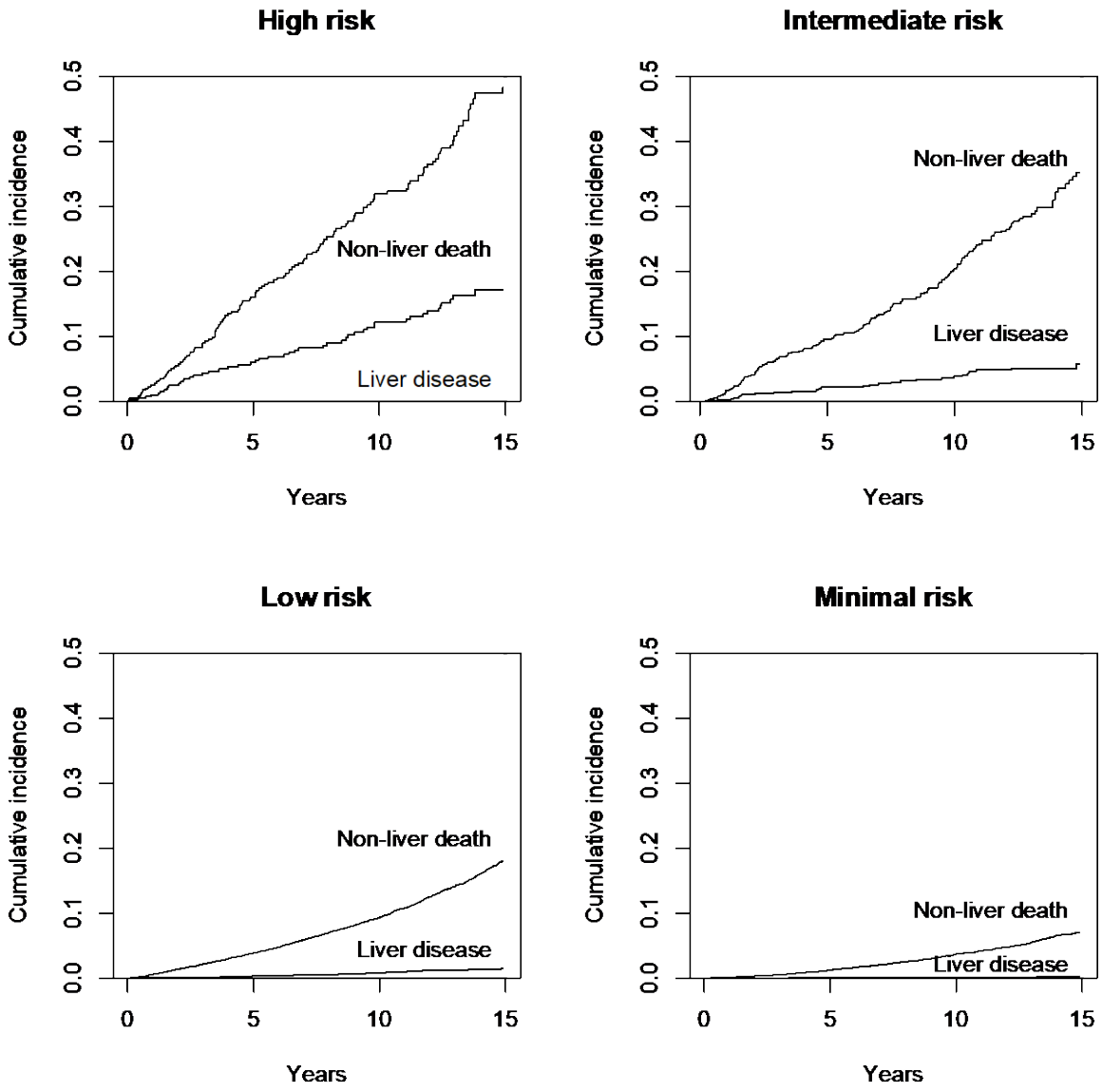


Fig. S7. Aalen-Johansen cumulative incidence of liver events stratified by risk group of $Model_{lab}$ in the derivation cohort, with non-liver death considered a competing event.

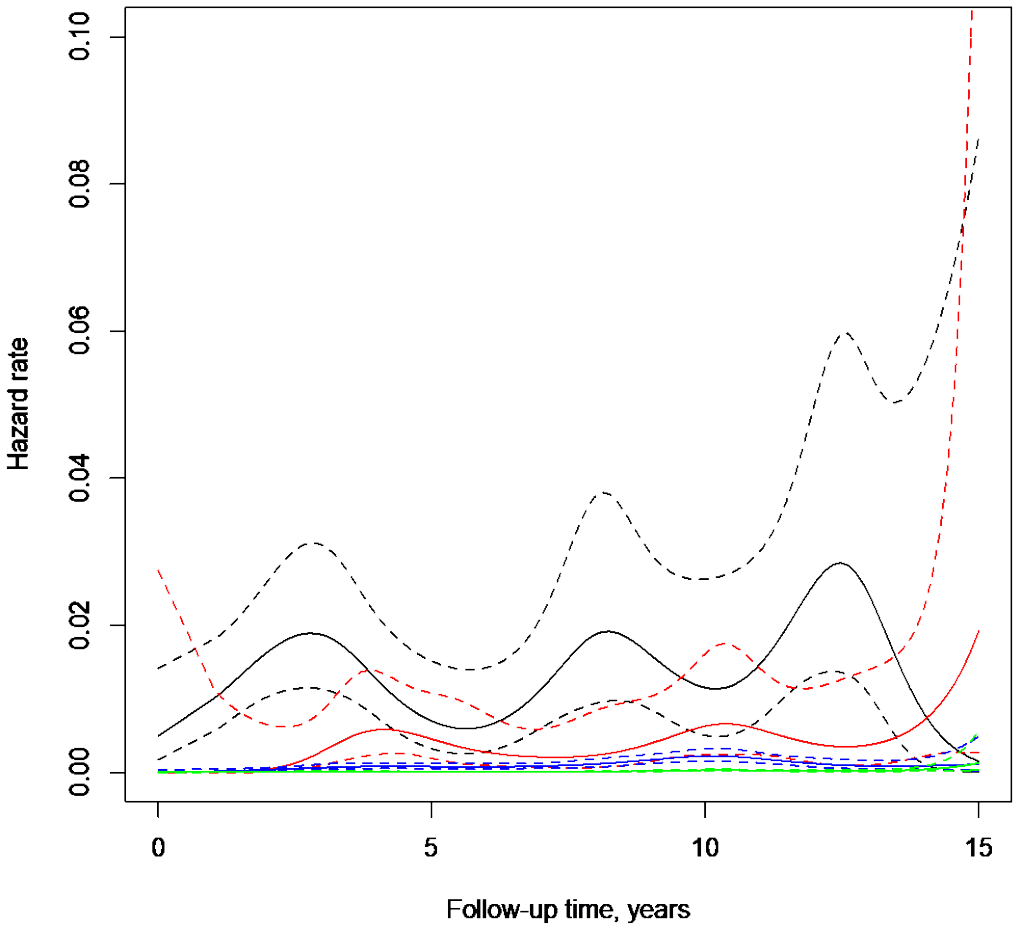


Fig. S8. Hazard rates are parallel between risk groups in $Model_{lab}$

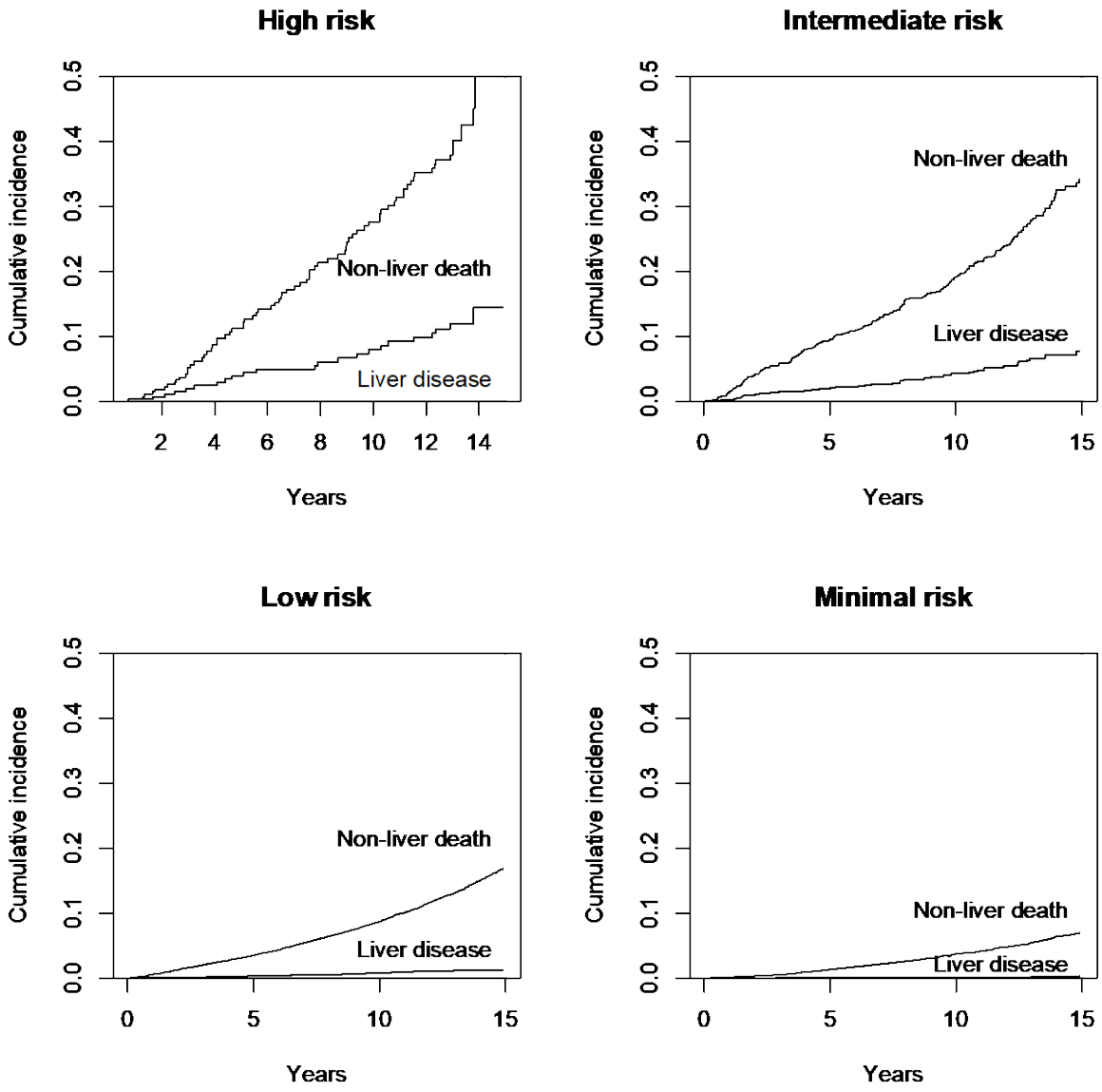
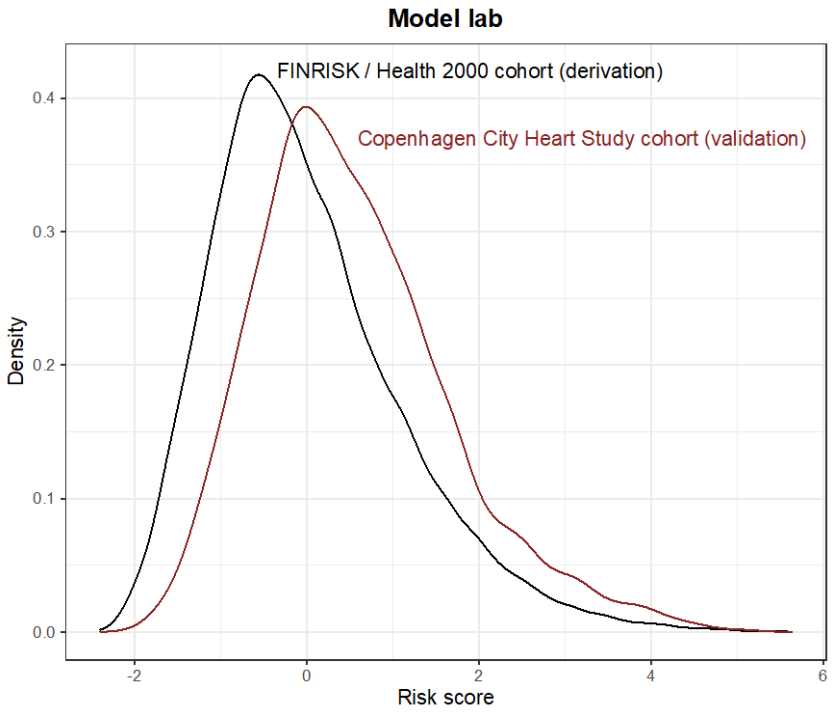


Fig. S9. Aalen-Johansen cumulative incidence of liver events stratified by risk group of $Model_{non\text{-}lab}$ in the derivation cohort, with non-liver death considered a competing event

(A)



(B)

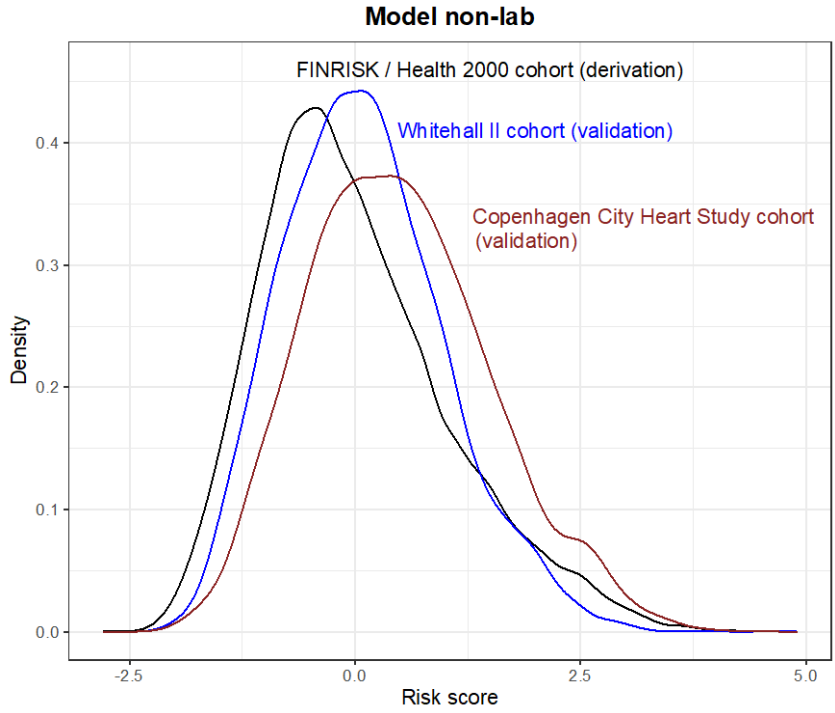
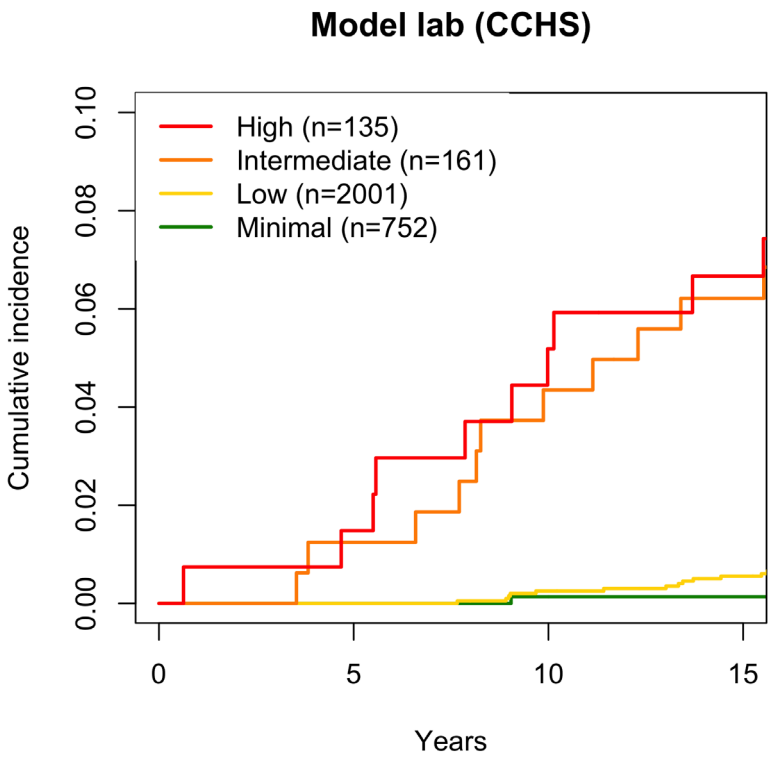
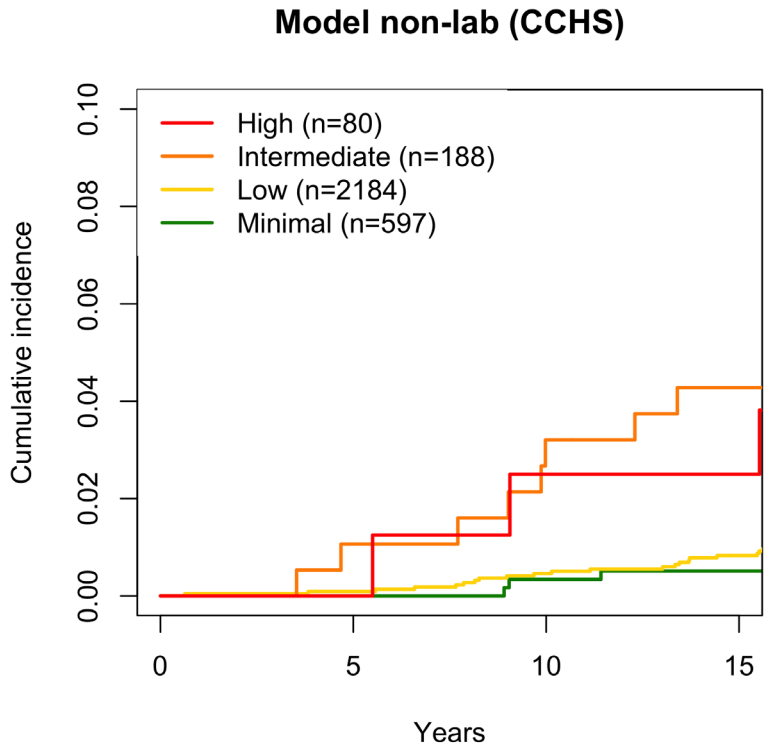


Fig. S10. The distribution of the risk scores (linear predictor) in the derivation and validation cohorts. (A) $Model_{lab}$, (B) $Model_{non\text{-}lab}$

(A)


Model lab (CCHS)
High (n=135)
Intermediate (n=161)
Low (n=2001)
Minimal (n=752)
Cumulative incidence
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10
15
Years

(B)


Model non-lab (CCHS)
High (n=80)
Intermediate (n=188)
Low (n=2184)
Minimal (n=597)
Cumulative incidence
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10
15
Years

(C)

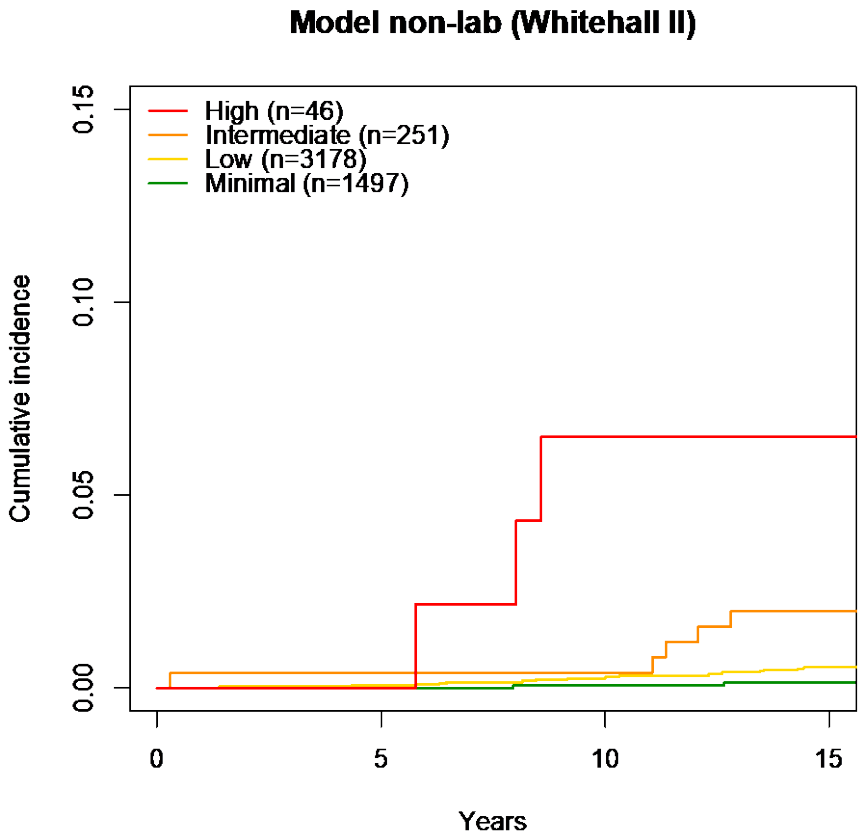


Fig. S11. Cumulative incidence of liver outcomes by risk group for (A) $Model_{lab}$ and (B) $Model_{nonlab}$ in the CCHS cohort, and (C) for $Model_{nonlab}$ in the Whitehall II cohort. Analyses performed using the Aalen-Johansen method considering death without liver disease as a competing risk event.

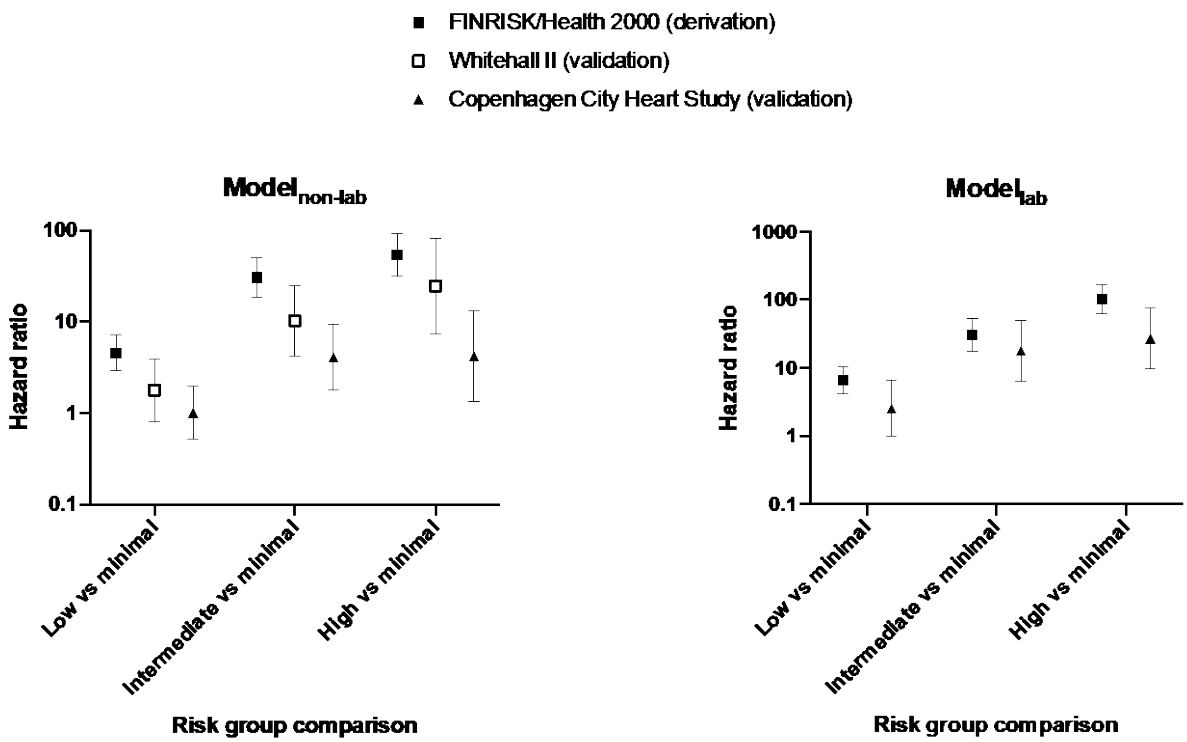
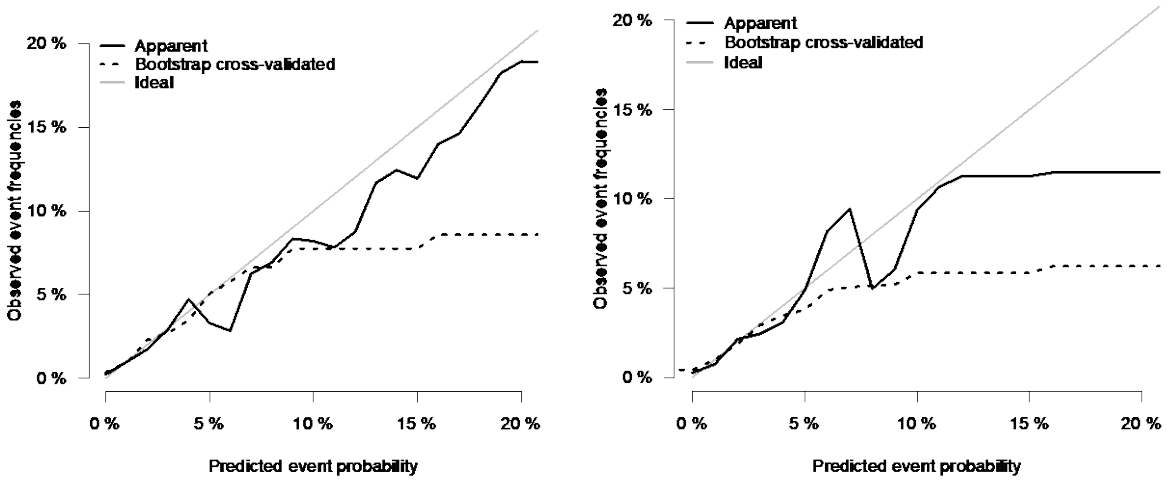


Fig. S12. Hazard ratios with 95% confidence intervals for comparisons of low, intermediate, and high risk groups against the minimal risk group separately in the derivation and validation cohorts.

(A) (B)



(C) (D)

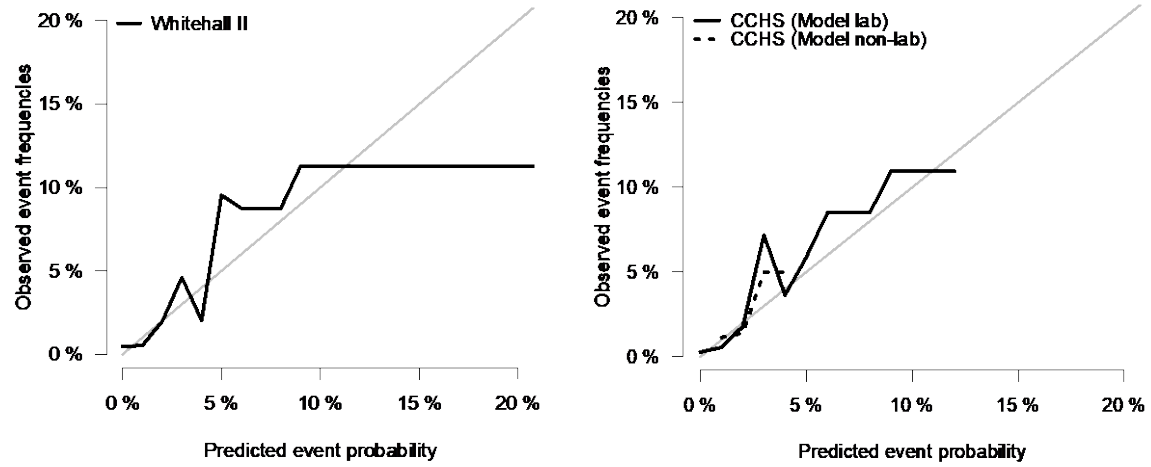


Fig. S13. Calibration plots for (A) $Model_{lab}$ and (B) $Model_{nonlab}$ in the derivation cohort showing the calibration between observed and predicted survival probability. The dotted black line corresponds to bootstrap cross-validated estimates. Calibration plots based on the external validation datasets; the (C) Whitehall II cohort and (D) the Copenhagen City Heart Study (CCHS) cohort.

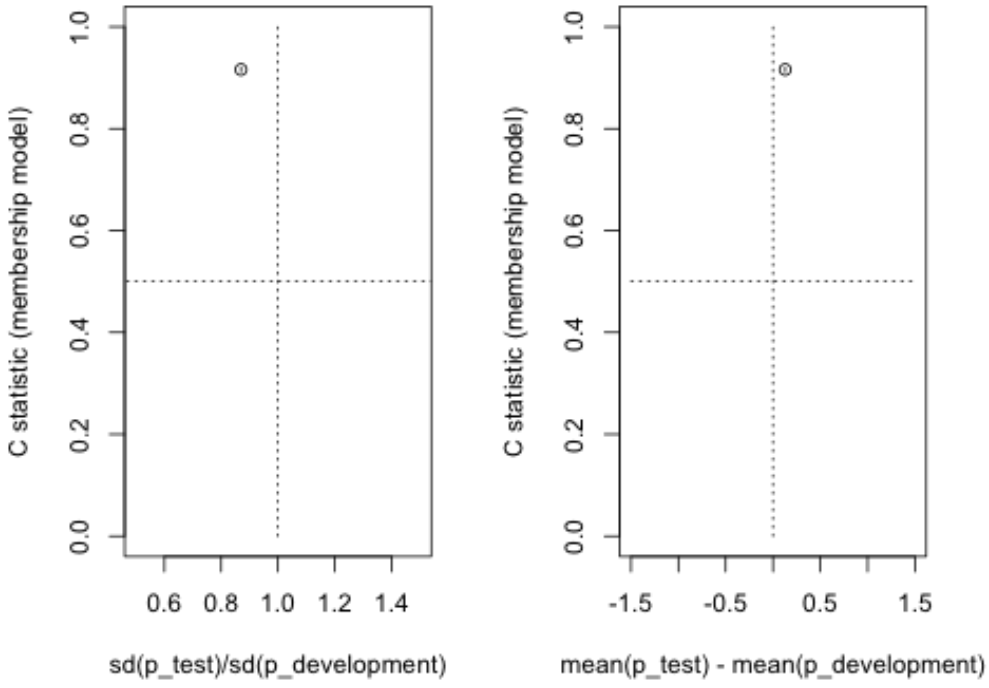


Fig. S14. Results of analysis on the relatedness between the derivation and Whitehall II validation dataset and the extent to which they share common predictor effects in the case of $Model_{non\text{-}lab}$. The membership model analysis showed that the derivation and validation samples were highly unrelated (C-statistic 0.92, 95% CI 0.91-0.92, y-axis) in terms of case mix. In addition, we observed a decreased variability in the model's linear predictor (left panel, x-axis) and a slightly increased mean value of the linear predictor in the validation sample (right panel, x-axis) as compared to that in the derivation sample.

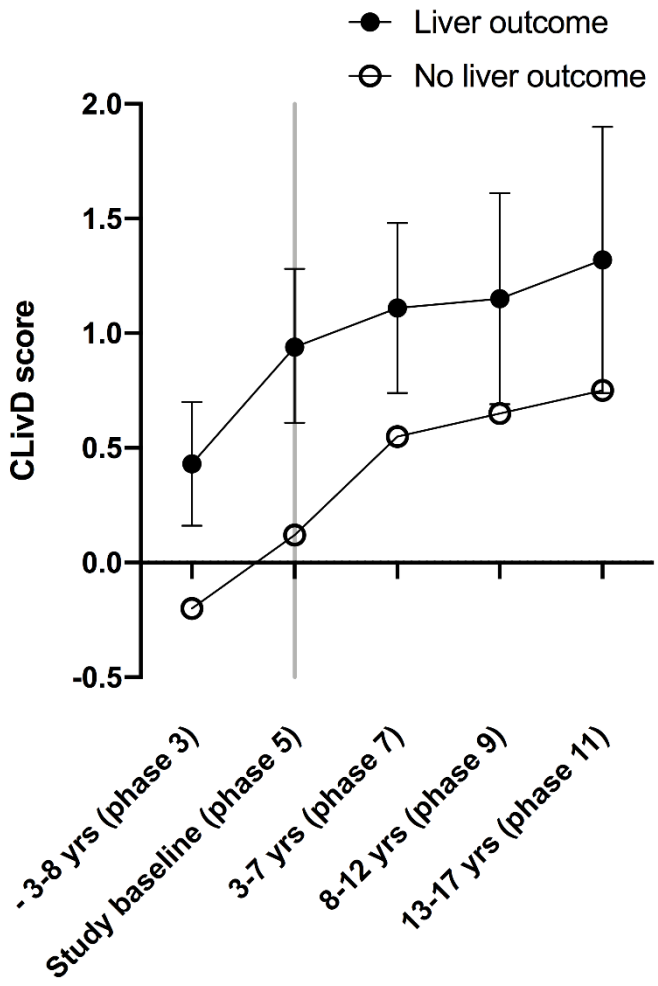
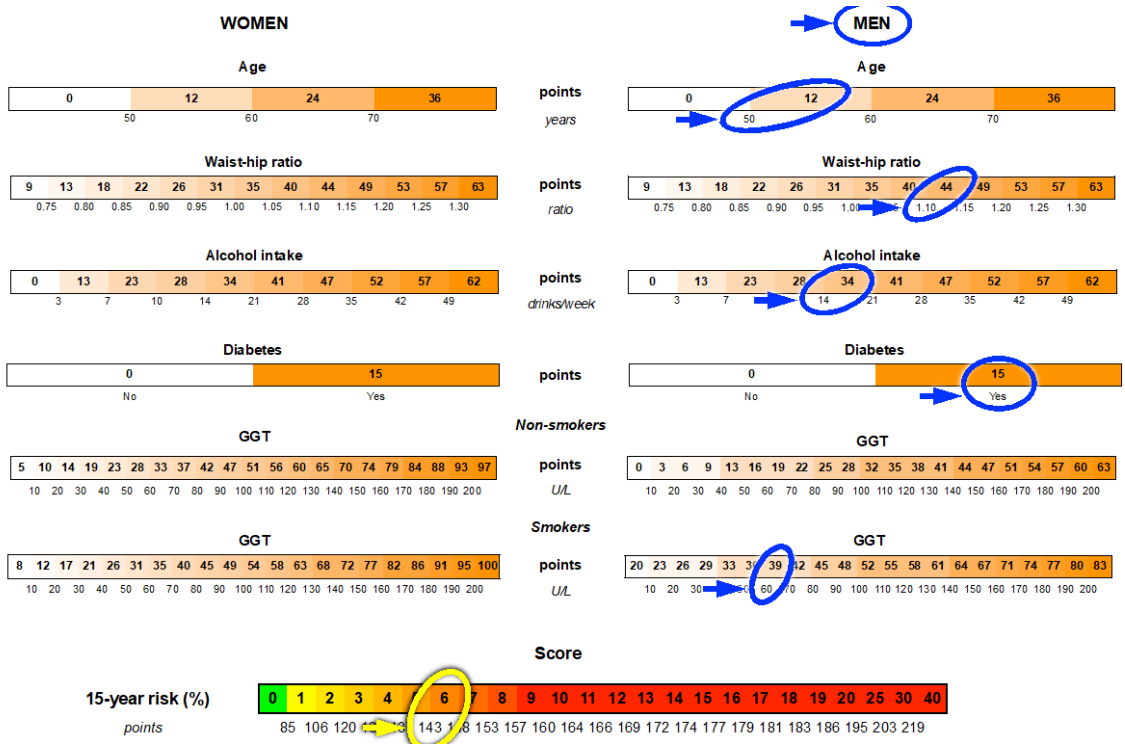


Fig. S15. Mean $Model_{non\text{-}lab}$ with 95% confidence intervals measured at the various follow-up examinations of the Whitehall II study separately for those who developed incident liver events and those who did not. Measurements are 3-8 years prior to baseline (phase 3 of the Whitehall II study), at baseline (phase 5), at 3-7 years post-baseline (phase 7), 8-12 years (phase 9), and 13-17 years (phase 11).

(A)


WOMEN
MEN
Age
points
years
Waist-hip ratio
ratio
Alcohol intake
drinks/week
Diabetes
No
Yes
Non-smokers
GGT
UI/L
Smokers
Score
15-year risk (%)
points

(B)

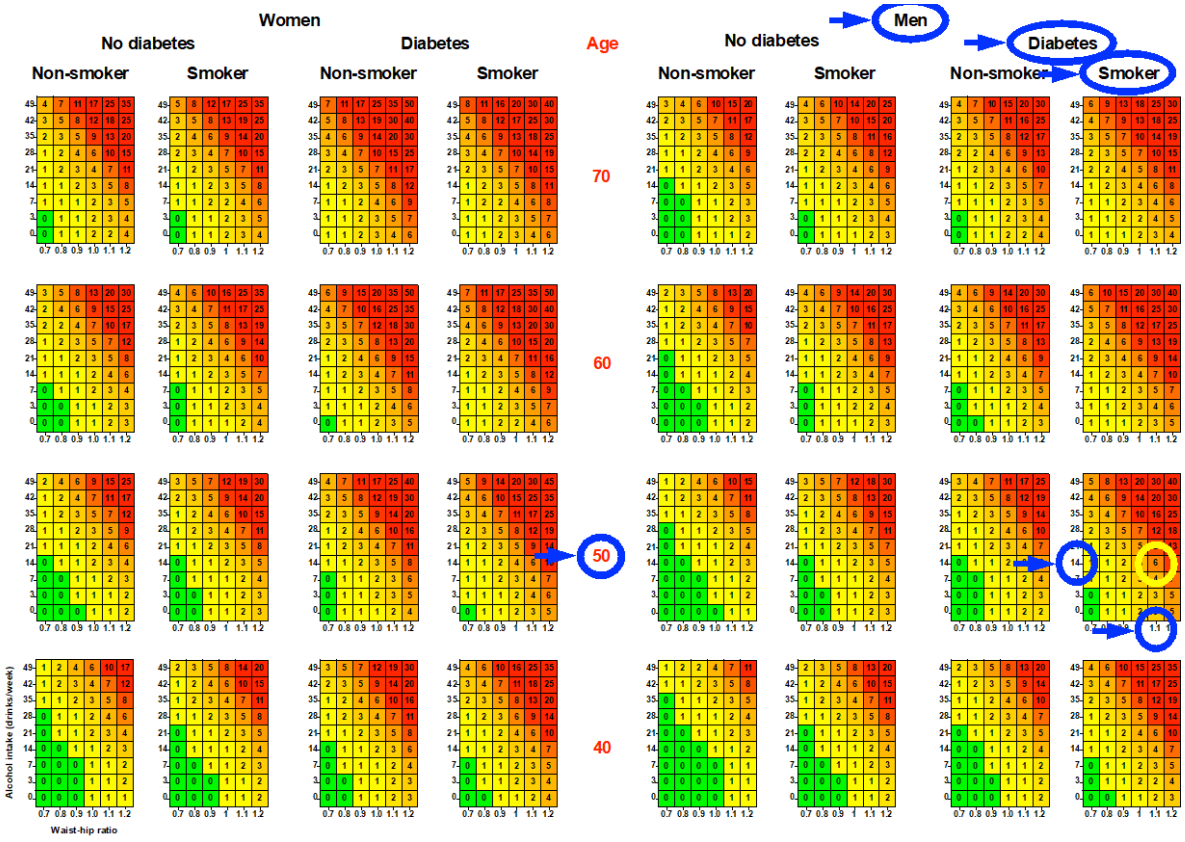

Women
Men
No diabetes
Diabetes
Non-smoker
Smoker
Age
70
60
50
40
Alcohol intake (drinks/week)
Waist-hip ratio

Fig. S16. Applying the nomograms in practice to evaluate an individual's 15-year risk for advanced clinical liver disease. Our example individual is a 52-year old male, who has diabetes, is a current smoker, drinks around 14 standard alcohol drinks per week, has a waist-hip ratio (WHR) of 1.1, and a gamma-glutamyltransferase (GGT) of 65 U/L. The blue circles indicate the characteristics of our example individual, while the yellow circles show the absolute 15-year risk estimate. In panel A, the user needs to sum the single points (12 + 44 + 34 + 15 + 39 = 144) and then convert the final points into the estimated 15-year risk (= 6 %).

Supplementary tables

Table S1. Flow chart of inclusion and exclusion criteria in the derivation and validation cohorts.

Cohort	**Derivation**	**Validation**	
	Finrisk and Health 2000 studies	Whitehall II study, phase 5 participants	Copenhagen City Heart Study, fourth examination, age 40-70 years
Initial sample, n	41648	7860	3467
Exclusions, n			
Baseline diagnosis of chronic liver disease (ICD10: K70-K77, C22.0; ICD8/9: 570-573, 155.0)	299	4	39
Chronic viral hepatitis at baseline or during follow-up (ICD10: B18)	89	7	0
Current alcohol abstainers at baseline (had used alcohol earlier and then stopped)	1866	118	379

Age < 40 or > 70 years at baseline	13634	2673	0
Analytical sample, n	25760	5058	3049

Table S2. ICD-codes used to define liver outcomes in the derivation and validation cohorts. Only ICD-10 codes were used in the validation cohort.

ICD-10 code	**Diagnosis**
K70.1	Alcoholic hepatitis
K70.2	Liver fibrosis caused by alcohol
K70.3	Alcoholic cirrhosis
K70.4	Liver failure related to alcohol
K70.9	Liver disease caused by alcohol
K72.0	Acute liver failure
K72.1	Chronic liver failure
K72.9	Liver failure unspecified
K74.0	Liver fibrosis
K74.1	Liver sclerosis
K74.2	Liver fibrosis and sclerosis
K74.6	Liver cirrhosis unspecified
K76.7	Hepatorenal syndrome
I85.0	Esophageal varices with bleeding
I85.9	Esophageal varices without bleeding
C22.0	Hepatocellular carcinoma
ICD-9 code	
571.1	Acute alcoholic hepatitis
571.2	Alcoholic cirrhosis of liver

571.3	Alcoholic liver damage, unspecified
571.5	Cirrhosis of liver without mention of alcohol
571.8	Other chronic nonalcoholic liver disease
572.2	Hepatic encephalopathy
572.4	Hepatorenal syndrome
572.8	Other sequelae of chronic liver disease
456.0	Esophageal varices with bleeding
456.1	Esophageal varices without mention of bleeding
155.0	Liver cancer

ICD-8 code	
571.0	Alcoholic cirrhosis
571.8	Cirrhosis, other
571.9	Cirrhosis, unspecified
573.0	Hepatitis NUD
573.9	Other liver disease
155.01	Liver cancer

Table S3. Baseline variables initially included and number of missing data for each variable.

	N missing	%	Restrictions
HOMA-IR	16946	66	Available in FINRISK 1992 and 2002 and Health 2000 studies
Binge drinking	13806	54	Available in FINRISK 2002 and 2007 and Health 2000 studies
AST	13626	53	Available in FINRISK 2002-2012 studies
ALT	12956	50	Available in FINRISK 2002-2012 studies
Exercise	3979	15	Available in FINRISK 1992-2007 and Health 2000 studies
Alcohol use (drinks/week)	1332	5	
Smoking group	404	2	
Alcohol status *	346	1	
Smoking status	224	1	
LDL-cholesterol	114	<1	
GGT	94	<1	
Body-mass index	91	<1	
Triglycerides	83	<1	
HDL-cholesterol	82	<1	
Non-HDL-cholesterol	82	<1	

Waist-hip ratio	70	<1
Waist circumference	60	<1
Age	0	0
Sex	0	0
Diabetes	0	0

* Data on whether a subject was a lifetime abstainer, current abstainer or active alcohol user

Table S4. Univariate predictors of liver outcomes by Cox regression analysis.

Imputed dataset *	**HR (95% CI)**	**P**
Age	1.03 (1.01-1.05)	<0.001
Women	0.36 (0.27-0.48)	<0.001
Diabetes	2.78 (1.96-3.94)	<0.001
Waist circumference	1.04 (1.03-1.05)	<0.001
Waist-hip ratio per 1 SD change	2.04 (1.82-2.27)	<0.001
Body mass index	1.04 (1.01-1.07)	0.0052
Alcohol (drinks/week)	1.06 (1.06-1.07)	<0.001
Smoking status		
Never/former smoker	Reference	
Current smoker	2.80 (2.15-3.65)	<0.001
GGT	1.02 (1.02-1.02)	<0.001
Alcohol status		
Drinker	Reference	
Lifetime abstainer	0.30 (0.14-0.64)	0.0018
Smoking group		
Never smoker	Reference	
Former smoker	1.71 (1.19-2.45)	0.0036
Smoker, 0-9 cigarettes/day	2.45 (1.44-4.16)	<0.001
Smoker, 10-19 cigarettes/day	2.77 (1.80-4.25)	<0.001
Smoker, 20+ cigarettes/day	4.57 (3.20-6.52)	<0.001

Complete-case dataset

Exercise (20-30min slightly out of breath and sweaty)		
At least 2 times a week	Reference	
2-4 times a month	1.10 (0.79-1.54)	0.559
Less often	2.06 (1.48-2.85)	<0.001
ALT	1.02 (1.01-1.02)	<0.001
AST	1.01 (1.01-1.02)	<0.001
Binge drinking (5 or more drinks per occasion)		
Less often	Reference	
Monthly	2.99 (1.85-4.84)	<0.001
Weekly	7.09 (4.75-10.56)	<0.001
HOMA-IR	1.02 (1.01-1.02)	<0.001

* P-values and HRs are similar in both the complete-case dataset and in the imputed dataset.

Table S5. Interaction terms, adjusted for age.

	P (adjusted for age)
SEX * AGE	0.454
SEX * DIABETES	0.629
SEX * WHR	0.842
SEX * ALCOHOL	0.778
SEX * SMOKING STATUS	0.102
SEX * GGT	0.058

Table S6. Interaction terms in multivariable models.

	$Model_{lab}$	$Model_{non\text{-}lab}$
	P (multivariable)	P (multivariable)
SEX * AGE	0.785	0.591
SEX * DIABETES	0.474	0.542
SEX * WHR	0.934	0.912
SEX * ALCOHOL	0.800	0.995
SEX * SMOKING STATUS	0.080	0.071
SEX * GGT	0.020	

Table S7. Reclassification table for number of liver events occurring within 15 years from baseline. The 15-year risk cutoffs are 0.5%, 5%, and 10%.

	Model$_{non-lab}$			
Model$_{lab}$	< 0.5%	< 5%	< 10%	≥ 10%
< 0.5%	17	5	0	0
< 5%	7	91	19	1
< 10%	0	8	8	8
≥ 10%	0	13	25	22

Table S8. Net reclassification index (NRI) when comparing $Model_{lab}$ with $Model_{non\text{-}lab}$. NRI compares one model with another model for their ability to correctly classify cases into correct risk categories. The NRI of -0.117 means that reclassification worsens by approximately 12% when using $Model_{non\text{-}lab}$ instead of $Model_{lab}$; however, this difference was non-significant.

	Estimate	Lower 95% CI	Upper 95% CI
NRI	-0.117	-0.245	0.015
NRI+	-0.033	-0.158	0.084
NRI-	-0.084	-0.091	-0.055
Pr(Up|Case)	0.179	0.121	0.248
Pr(Down|Case)	0.212	0.148	0.288
Pr(Down|Control)	0.047	0.040	0.050
Pr(Up|Control)	0.131	0.095	0.138

Table S9. Number of subjects with incident liver outcomes, total number of subjects in each risk group, and cumulative incidence estimates by the Aalen-Johansen competing risk method.

	Derivation cohort		**CCHS**		**Whitehall II**	
$Model_{lab}$	**Events/subjects**	**15 yrs**	**Events/subjects**	**15 yrs**	**Events/subjects**	**15 yrs**
Minimal risk	23/12731	0.27%	5/752	0.14%		
Low risk	123/11968	1.48%	32/2001	0.56%		
Intermediate risk	27/662	5.69%	14/161	6.21%		
High risk	49/399	17.20%	14/135	7.67%		
$Model_{non\text{-}lab}$						
Minimal risk	24/10831	0.33%	11/597	0.51%	8/1497	0.09%
Low risk	127/13718	1.28%	39/2184	0.83%	29/3178	0.54%
Intermediate risk	48/940	7.75%	11/188	4.28%	12/251	1.99%
High risk	23/271	14.43%	4/80	3.50%	4/46	6.98%

Supplementary references

1. Aromaa A, Koskinen S. Health and functional capacity in Finland. Baseline results of the Health 2000 health examination survey. Publications of National Public Health Institute, Series B 12/2004. Helsinki, Finland. 2004.

2. Borodulin K, Tolonen H, Jousilahti P, et al. Cohort Profile: The National FINRISK Study. Int J Epidemiol 2018;47:696-696i.

3. MONICA Manual, Part III: Population Survey. Available at: https://www.thl.fi/publications/monica/manual/part3/iii-1.htm. Accessed December 8, 2021.

4. European Health Risk Monitoring (EHRM) Project. Available at: https://www.thl.fi/publications/ehrm/product3/title.htm. Accessed December 8, 2021.

5. Sund R. Quality of the Finnish Hospital Discharge Register: a systematic review. Scand J Public Health 2012;40:505-515.

6. Pukkala E. Biobanks and registers in epidemiologic research on cancer. Methods Mol Biol 2011;675:127-164.

7. European Association for the Study of the Liver. EASL Clinical Practice Guidelines: Management of alcohol-related liver disease. J Hepatol 2018;69:154-181.

8. McLernon DJ, Donnan PT, Sullivan FM, et al. Prediction of liver disease in patients whose liver function tests have been checked in primary care: model development and validation using population-based observational cohorts. BMJ Open 2014;4:e004837-2014.

9. McLernon DJ, Donnan PT, Ryder S, et al. Health outcomes following liver function testing in primary care: a retrospective cohort study. Fam Pract 2009;26:251-259.

10. Castera L, Friedrich-Rust M, Loomba R. Noninvasive Assessment of Liver Disease in Patients With Nonalcoholic Fatty Liver Disease. Gastroenterology 2019;156:1264-1281.e4.

11. Vilar-Gomez E, Chalasani N. Non-invasive assessment of non-alcoholic fatty liver disease: Clinical prediction rules and blood-based biomarkers. J Hepatol 2018;68:305-315.

12. Lee DH, Blomhoff R, Jacobs DR. Is serum gamma glutamyltransferase a marker of oxidative stress? Free Radic Res 2004;38:535-539.

13. Cichoz-Lach H, Michalak A. Oxidative stress as a crucial factor in liver diseases. World J Gastroenterol 2014;20:8082-8091.

14. Newsome PN, Cramb R, Davison SM, et al. Guidelines on the management of abnormal liver blood tests. Gut 2018;67:6-19.

15. Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics 2014;15:526-539.

16. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68:279-289.

17. Andreasson A, Carlsson AC, Onnerhag K, Hagström H. Waist/Hip Ratio Better Predicts Development of Severe Liver Disease Within 20 Years Than Body Mass Index: A Population-based Cohort Study. Clin Gastroenterol Hepatol 2017;15:1294-1301.e2.

18. Åberg F, Jula A. The sagittal abdominal diameter: Role in predicting severe liver disease in the general population. Obes Res Clin Pract 2018;12:394-396.

19. Schult A, Mehlig K, Bjorkelund C, Wallerstedt S, Kaczynski J. Waist-to-hip ratio but not body mass index predicts liver cirrhosis in women. Scand J Gastroenterol 2018;53:212-217.

20. Hagström H, Nasr P, Ekstedt M, et al. Fibrosis stage but not NASH predicts mortality and time to development of severe liver disease in biopsy-proven NAFLD. J Hepatol 2017;67:1265-1273.