HOSEA Aim I – Report

Simon Fontaine

July 12, 2022

Calibration & threshold

Some comments When working with the first iteration of data (6.6M controls), we found that the best threshold was in the upper 100s. Now, when trained using the 10M controls, the best threshold seems to be more around the lower 100s in comparison (see table below). That was to be expected since the "baseline" incidence changed between the two training sets. In particular, there is a factor of 2/3 between these two datasets in terms of incidence, so we should expect a similar change in the threshold. For example, in order to get 80% TPR, we previously needed a threshold between 140 and 160 (/100,000) while we now require a threshold between 100 and 105. There is no simple link between the two, but the 2/3 factor is a good rule of thumb.

We have also seen a similar phenomenon in the past: when applying the model to a testing set that was badly constructed (wrong incidence compared to the training set), we were severely miscalibrated because we had a mismatch between our training and testing incidence. The thing is that our predicted risks were correct, but the testing set contained 3x too many cases so it seemed we were miscalibrated.

This speaks to the broader issue of the generalizability of the method, which was brought up by Joel. For the model to be generalizable, we must assume that the relationship between predictors and the outcome is the same in the new population. More specifically, the model is learning a map $\text{logit}P[Y = 1 \mid X] = f(X)$ for some function f (the trees), so we need this to hold in the new population as well, which might or might not be true. With our various sensitivity analyses, we found that this map was good for most values of the predictors X (e.g. we split the testing set in multiple ways and still found decent performance in all cases). That means our learned f is good for most X *that we observed*, but we cannot provide similar guarantee for other X.

Now, the predicted risk should still hold for population with different sampling incidence, provided they share the same relationship $f(X)$. This is independent of the distribution of features! For example, the predicted risk should be valid for the whole population too (50-50 females, younger, etc.) as long as we can assume the same f.

One important note here is that these values are not annualized, these are the rates over the 14-year span. If we think about our sample, we looked aver 14 years and collected all cases so our estimated risk is really over 14 person-years. It might be better to report the rates /14 in that case.

Threshold	TPR	PPV	Det. prev.
10	98.63	0.18	62.08
20	96.07	0.22	49.30
30	93.61	0.25	41.32
40	91.75	0.29	35.63
50	89.72	0.32	31.34
55	88.49	0.33	29.54
60	87.22	0.35	27.93
65	86.52	0.36	26.48
70	85.75	0.38	25.16
75	84.91	0.39	23.97
80	83.82	0.41	22.87
85	82.73	0.42	21.86
90	82.41	0.44	20.92
95	81.54	0.45	20.06
100	80.84	0.47	19.25
105	79.92	0.48	18.50
110	79.26	0.49	17.80
115	78.55	0.51	17.15
120	77.78	0.52	16.54
125	76.87	0.53	15.96
130	76.45	0.55	15.42
135	75.92	0.56	14.92
140	75.18	0.58	14.44
145	74.69	0.59	13.98
150	74.17	0.61	13.54
155	73.50	0.62	13.13
160	72.59	0.63	12.74

Threshold	TPR	PPV	Det. prev.
165	71.88	0.65	12.36
170	71.71	0.66	12.02
175	71.29	0.68	11.67
180	70.80	0.69	11.36
185	70.20	0.71	11.05
190	69.85	0.72	10.76
195	69.29	0.73	10.48
200	68.90	0.75	10.21
210	68.06	0.78	9.72
220	66.83	0.80	9.26
230	66.02	0.83	8.83
240	65.29	0.86	8.44
250	64.65	0.89	8.07
260	63.95	0.92	7.73
270	63.07	0.95	7.41
280	62.34	0.97	7.10
290	61.60	1.00	6.82
300	60.93	1.03	6.56
325	59.49	1.11	5.97
350	58.13	1.18	5.46
375	56.97	1.26	5.02
400	55.77	1.33	4.64
425	54.72	1.41	4.30
450	53.46	1.48	4.00
475	52.51	1.56	3.73
500	51.63	1.64	3.49
1000	40.05	3.32	1.34

Akbar's reference Looking at the paper that Akbar sent me, here is the passage I think was relevant. From *Deep Learning Model to Predict Hepatocellular Carcinoma in Patients With Hepatitis C Cirrhosis* by Ioannou et al. (2020):

> **Prioritizing Patients for HCC Screening Outreach Interventions.** We envision that risk stratification models could be used to prioritize the patients with the highest risk for screening outreach interventions. Using the RNN model, we determined that 90% of all HCC diagnoses in the following 3 years occurred in samples with the mean (SD) highest 66% (1.2%) of risk scores, whereas 80% of HCCs occurred in samples with the mean (SD) highest 51% (1.5%) of risk scores. Thus, using the RNN model, we could potentially target the top 51% of samples with the highest HCC risk scores, in which 80% of all HCCs occurred,

or the top 66% of samples with the highest HCC risk scores, in which 90% of all HCCs occurred. In contrast, the proportions that would need to be screened to include 80% or 90% of patients who would be diagnosed with HCC were much greater using the longitudinal LR and cross-sectional LR models (Table 2).

We see that the reporting never uses risk scores directly only some cumulative performance metrics: e.g., if we want 90% TPR, we need to consider the top 66% risk scores. Actual calibration was only reported using tertiles and the brier score.

Some comparison to known incidence rates. We have a 14-year incidence rate of 11,395 / 10,268,282 = 110.97/100,000, which gives a 7.92/100,000 person-year incidence rate within our population. This is obviously not general because we have age and sex biases, at the minimum. In particular, we can compare this to the male incidence rates of EAC: 0.5/100,000 for <50, 9/100,000 for 50-64 yo, 26/100,000 for 65+ and 5/100,000 overall. Understanding that we have mostly older men, and adding in the EGJAC cases, we can reconcile these values with our predicted risks.

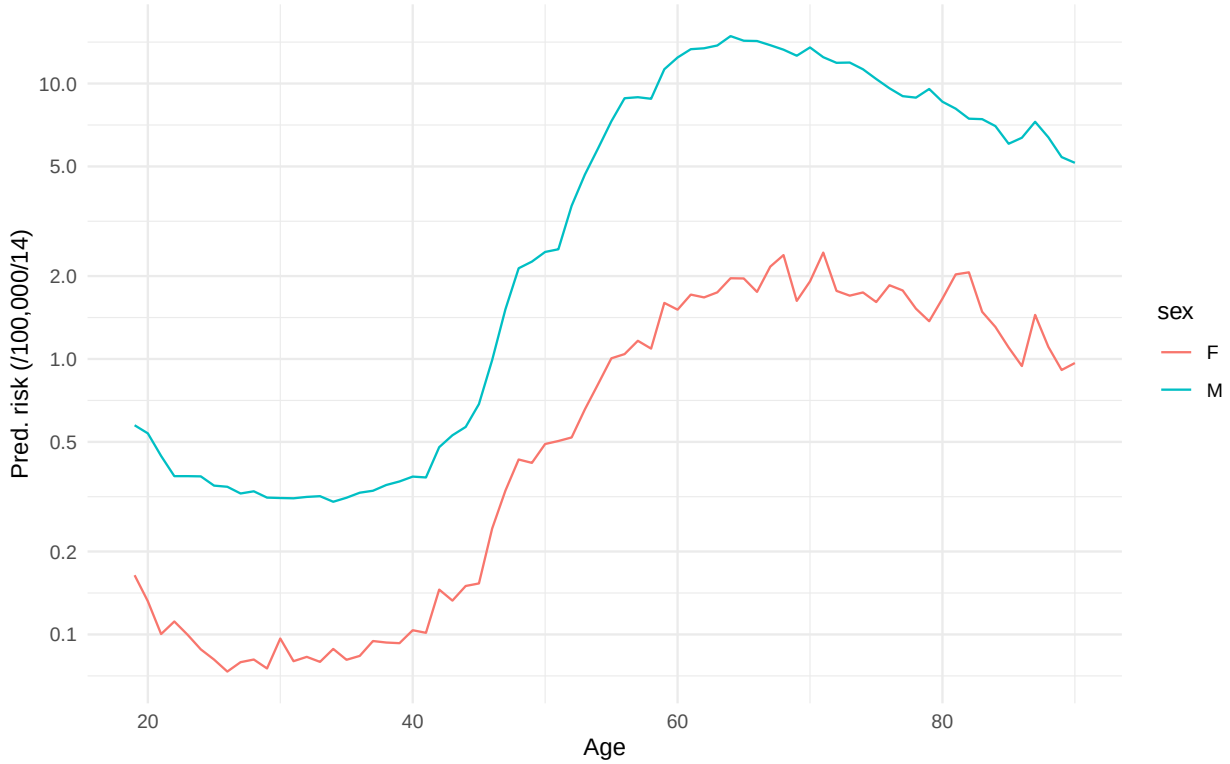
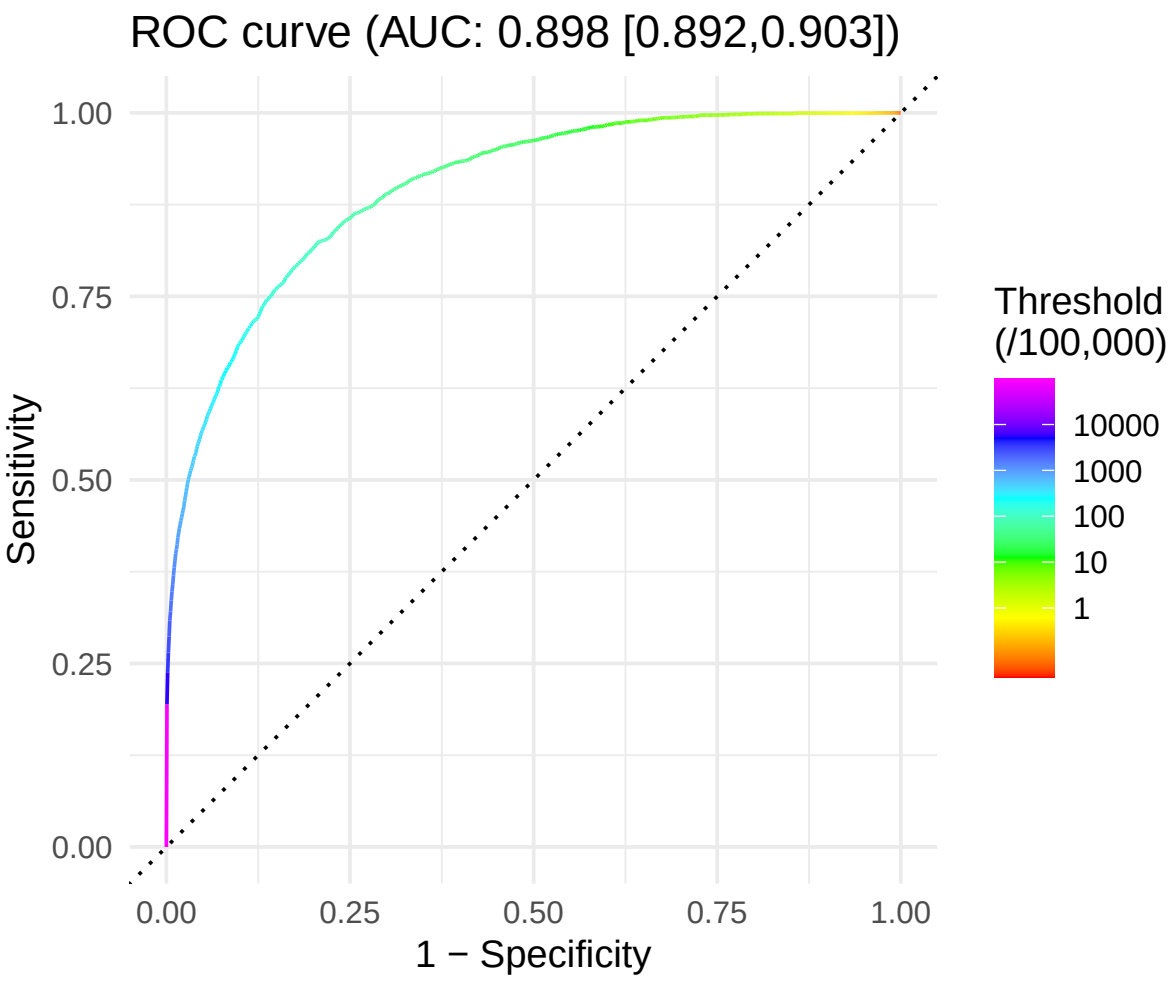
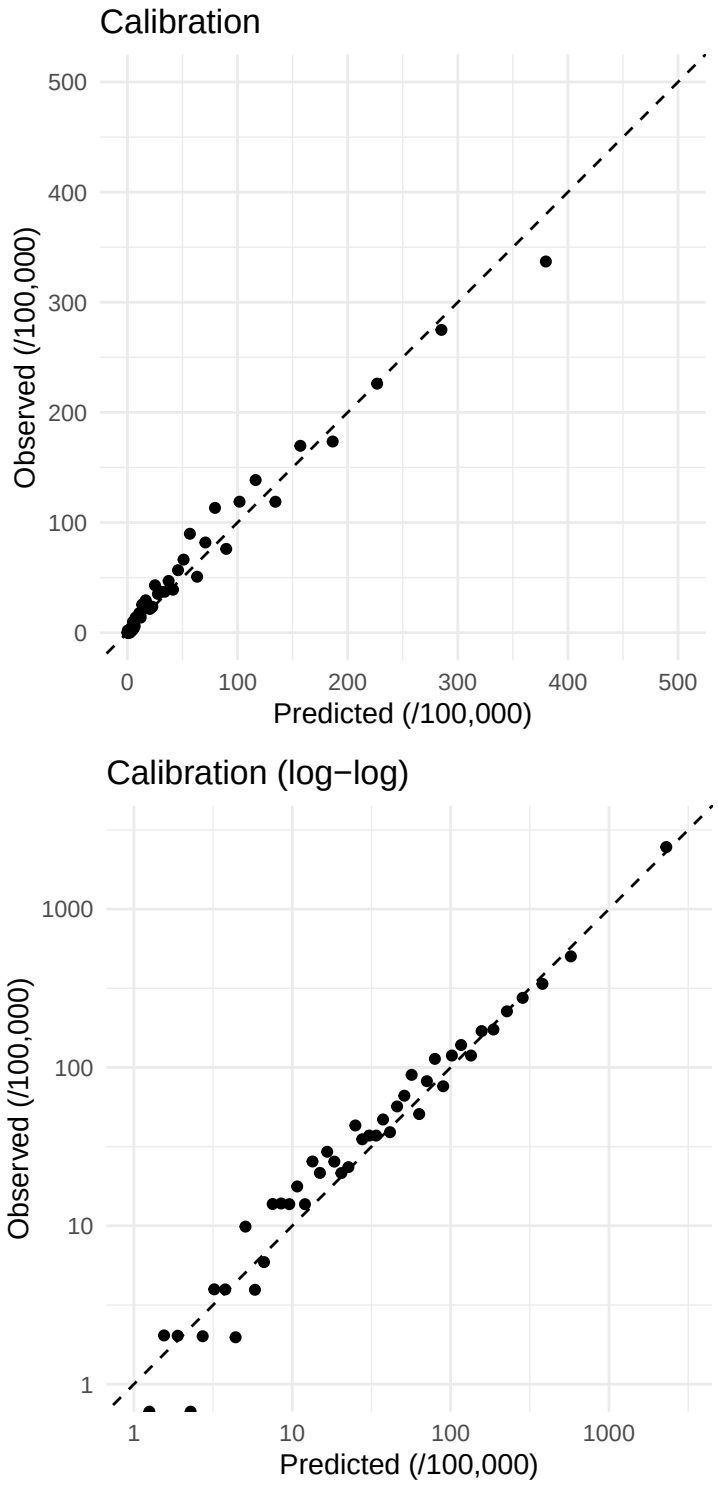


Figure 1: The average annalized risk is 7.6/100,000: for males, it is 8.2/100,000 and, for females, it is 0.7/100,000.


ROC curve (AUC: 0.898 [0.892,0.903])
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Threshold
(/100,000)
10000
1000
100
10
1

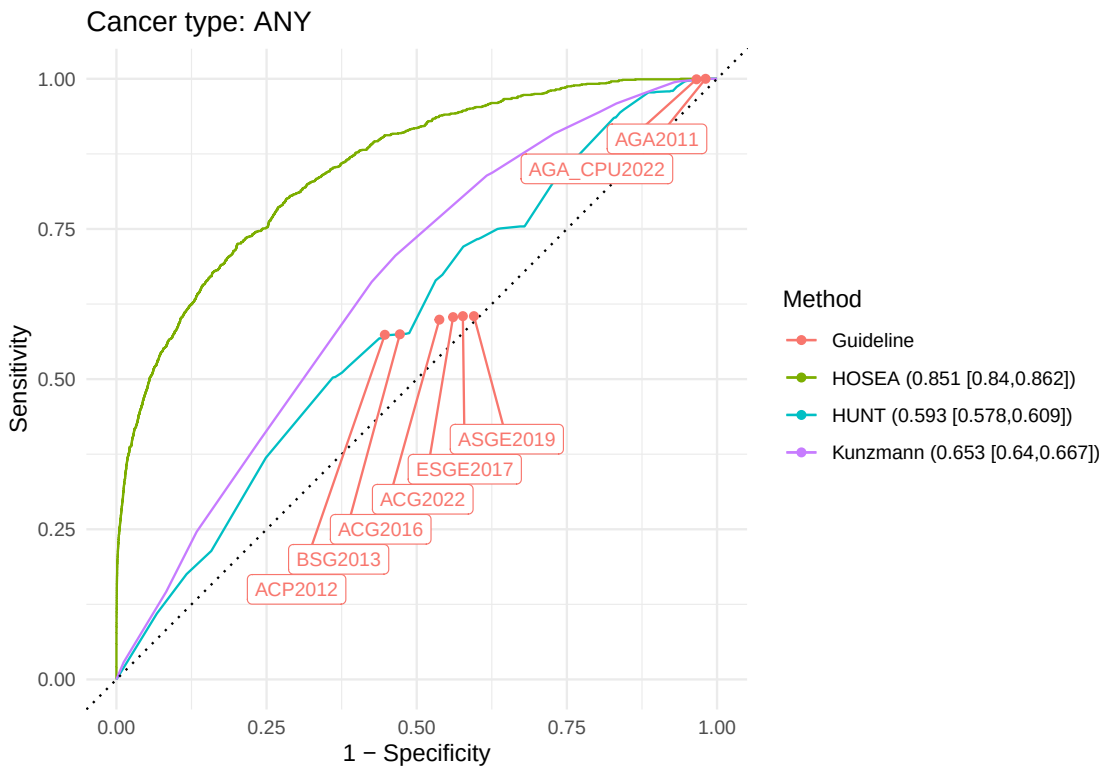


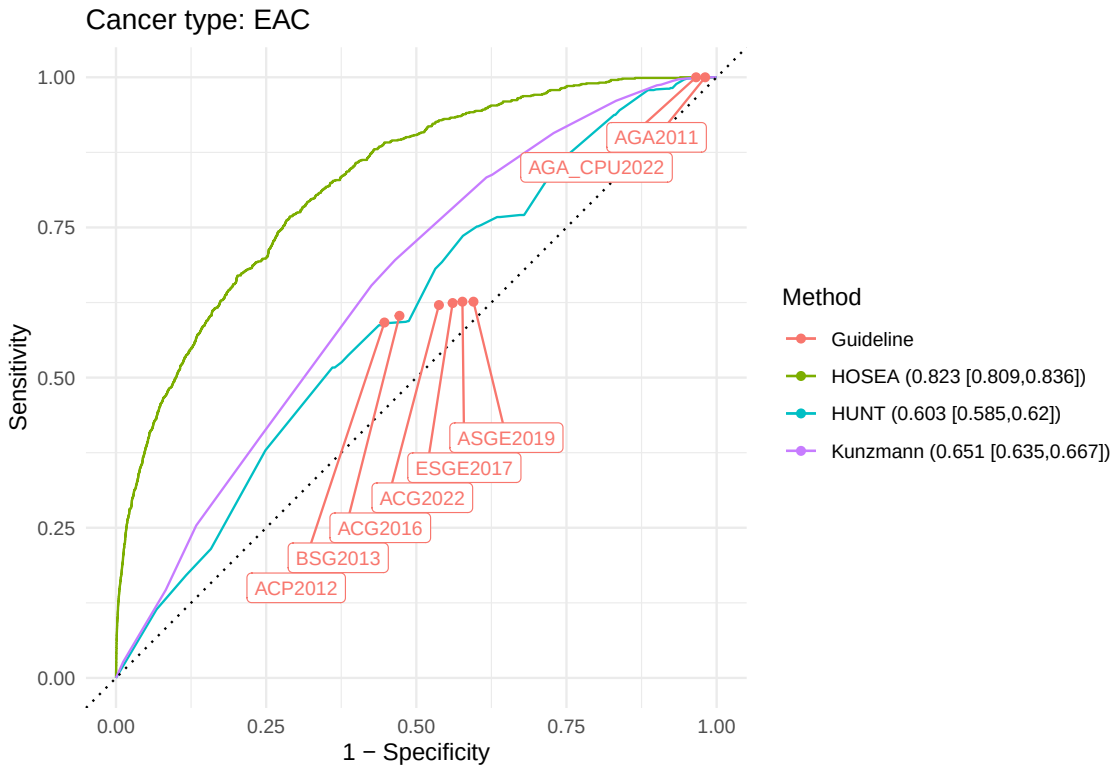
Each point represent 2% of the test data, split using predicted risk quantiles. The top plot is cropped on the right and top so we can focus on the more important region.

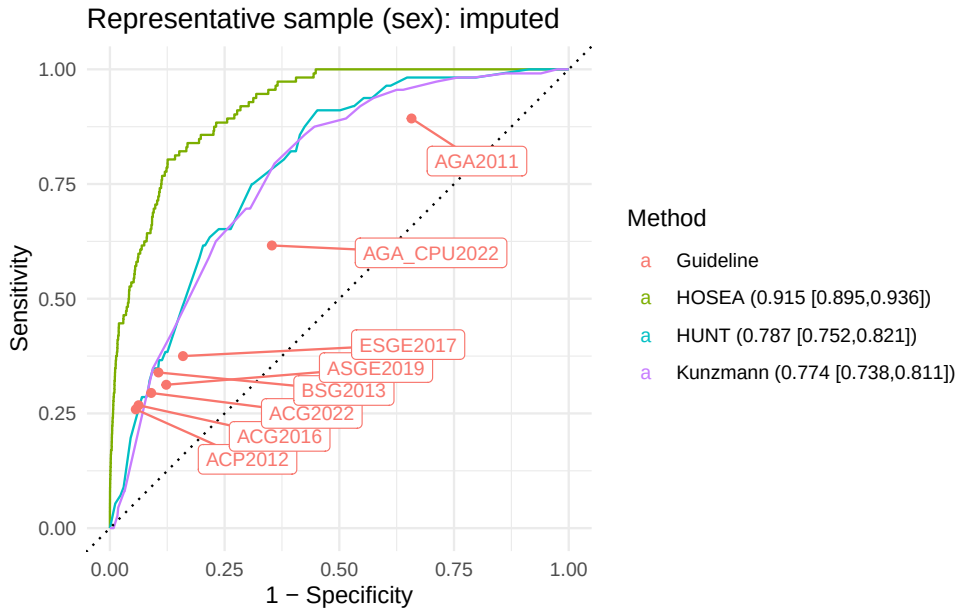
Comparison

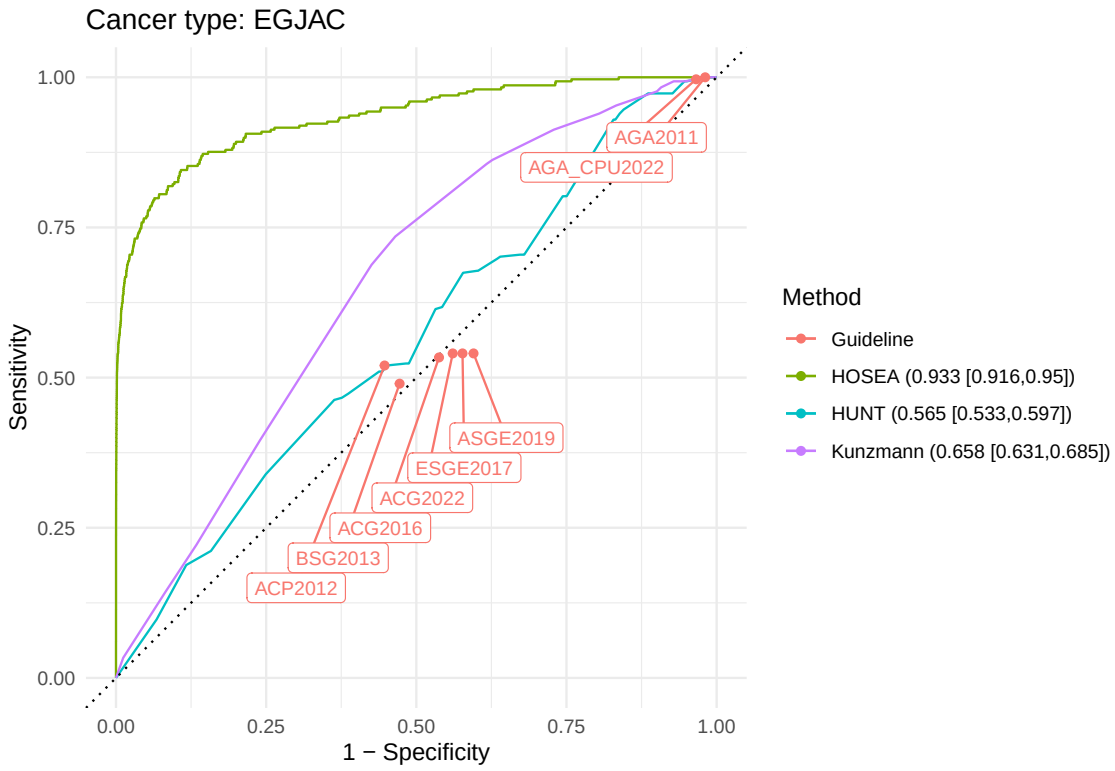
For these comparisons:

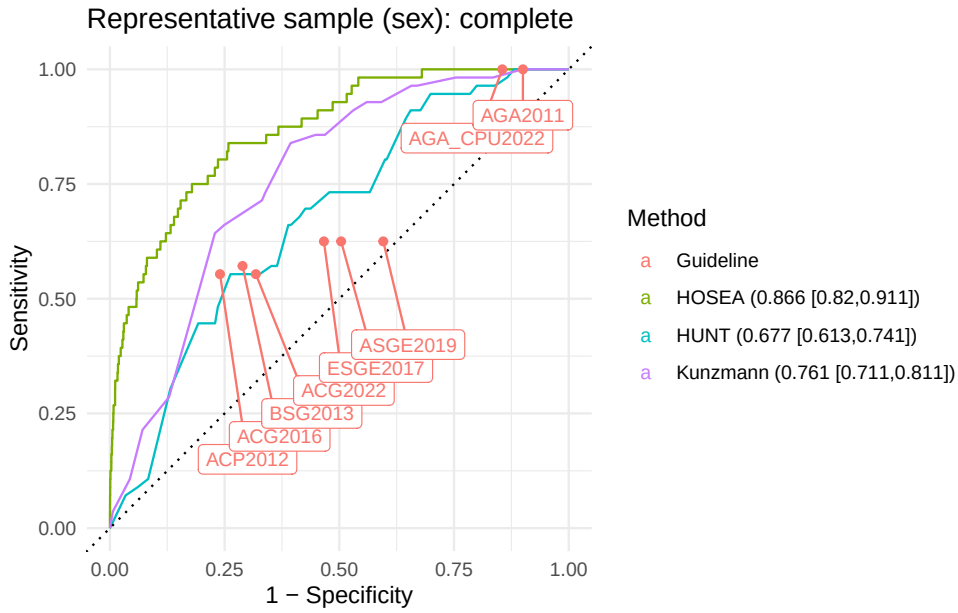
- Subset to test observations
- Filter out patients with missing information for HUNT/Kunzmann/Guidelines
- i.e., require age, BMI, race, smoking status, gerd, h2r/ppi
- 407K patients, 1192 cases (292/100,000)
- AUC + 95% CI using DeLong method




Cancer type: EAC
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
AGA2011
AGA_CPU2022
ASGE2019
ESGE2017
ACG2022
ACG2016
BSG2013
ACP2012
Method
Guideline
HOSEA (0.823 [0.809,0.836])
HUNT (0.603 [0.585,0.62])
Kunzmann (0.651 [0.635,0.667])


Representative sample (sex): imputed
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
AGA2011
AGA_CPU2022
ESGE2017
ASGE2019
BSG2013
ACG2022
ACG2016
ACP2012
Method
a Guideline
a HOSEA (0.915 [0.895,0.936])
a HUNT (0.787 [0.752,0.821])
a Kunzmann (0.774 [0.738,0.811])


Cancer type: EGJAC
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
AGA2011
AGA_CPU2022
ASGE2019
ESGE2017
ACG2022
ACG2016
BSG2013
ACP2012
Method
Guideline
HOSEA (0.933 [0.916,0.95])
HUNT (0.565 [0.533,0.597])
Kunzmann (0.658 [0.631,0.685])

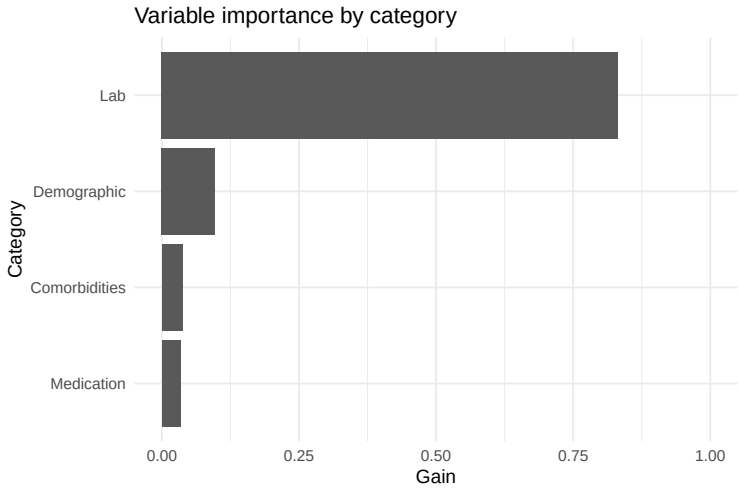

Representative sample (sex): complete
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
AGA2011
AGA_CPU2022
ASGE2019
ESGE2017
ACG2022
BSG2013
ACG2016
ACP2012
Method
a Guideline
a HOSEA (0.866 [0.82,0.911])
a HUNT (0.677 [0.613,0.741])
a Kunzmann (0.761 [0.711,0.811])

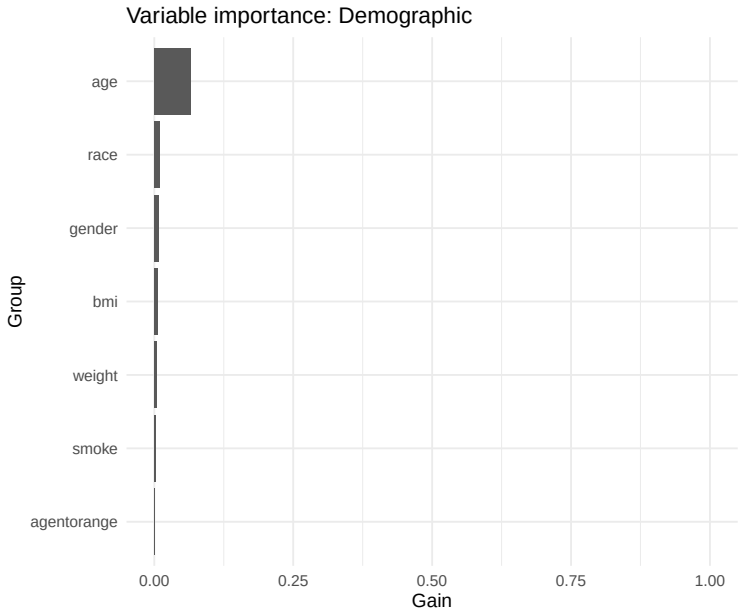
Gain Variable importance

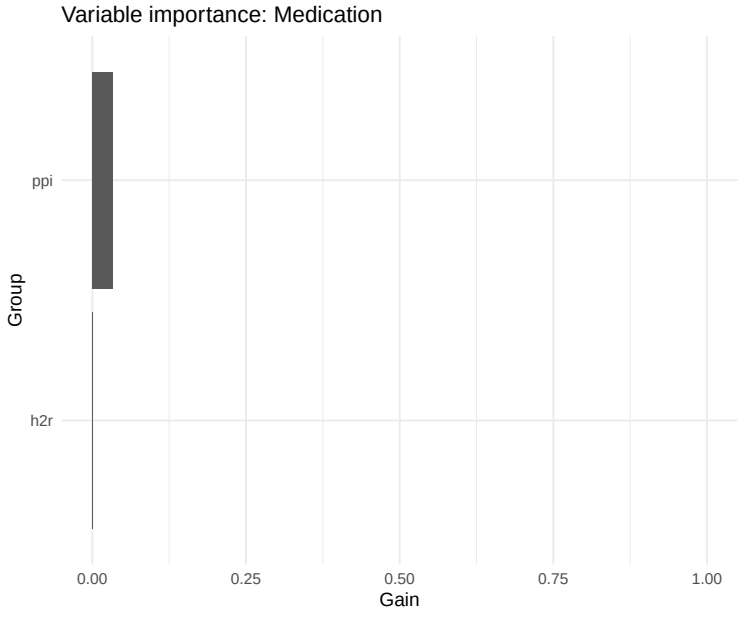
"Gain represents fractional contribution of each feature to the model based on the total gain of this feature's splits. Higher percentage means a more important predictive feature."

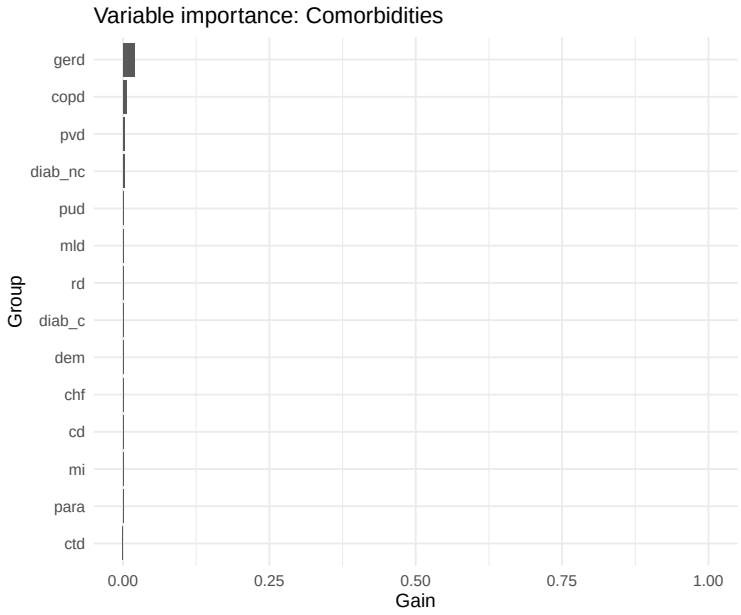
Some notes:

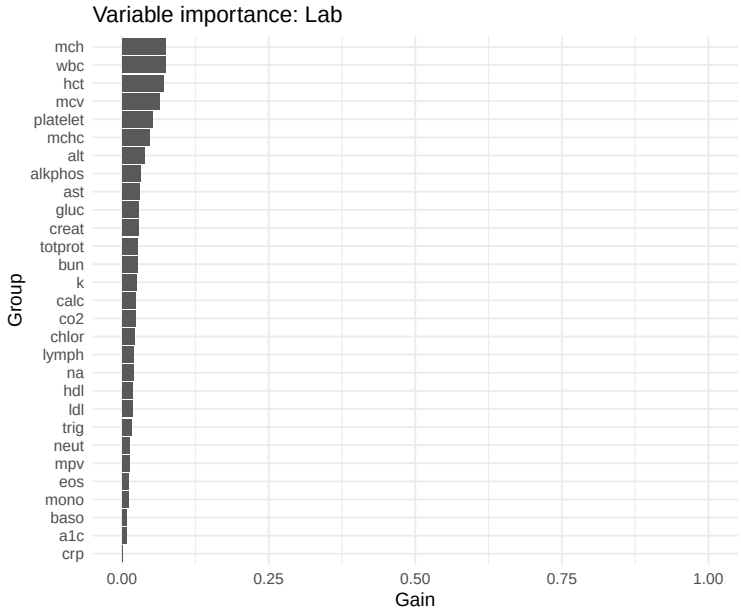
- These are additive, so we can compute the importance of a group of features.
- They sum up to 1
- We can only look at the relationship with the feature value (e.g., mostly positive, mostly negative) for single features




Variable importance: Demographic
age
race
gender
bmi
weight
smoke
agentorange
Group
0.00
0.25
0.50
0.75
1.00
Gain


Variable importance: Medication
ppi
h2r
Group
0.00
0.25
0.50
0.75
1.00
Gain


Variable importance: Comorbidities
gerd
copd
pvd
diab_nc
pud
mld
rd
diab_c
dem
chf
cd
mi
para
ctd
Group
0.00
0.25
0.50
0.75
1.00
Gain


Variable importance: Lab
mch
wbc
hct
mcv
platelet
mchc
alt
alkphos
ast
gluc
creat
totprot
bun
k
calc
co2
chlor
lymph
na
hdl
ldl
trig
neut
mpv
eos
mono
baso
a1c
crp
Group
0.00
0.25
0.50
0.75
1.00
Gain

SHAP Variable Importance

- As Gain, this is additive, but does not sum to 1
- Local measure, can be aggreagted using mean absolute value
- Understood as change in log-odds due to this variable ("Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value.")
- Scale to be inderstood as $\beta_j x_{ij}$ in logistic regression

$$\text{logit} P[Y_i = 1] = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip}$$

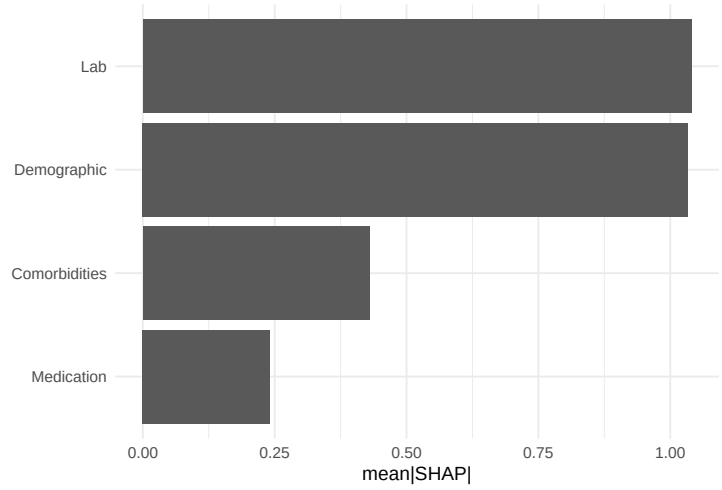
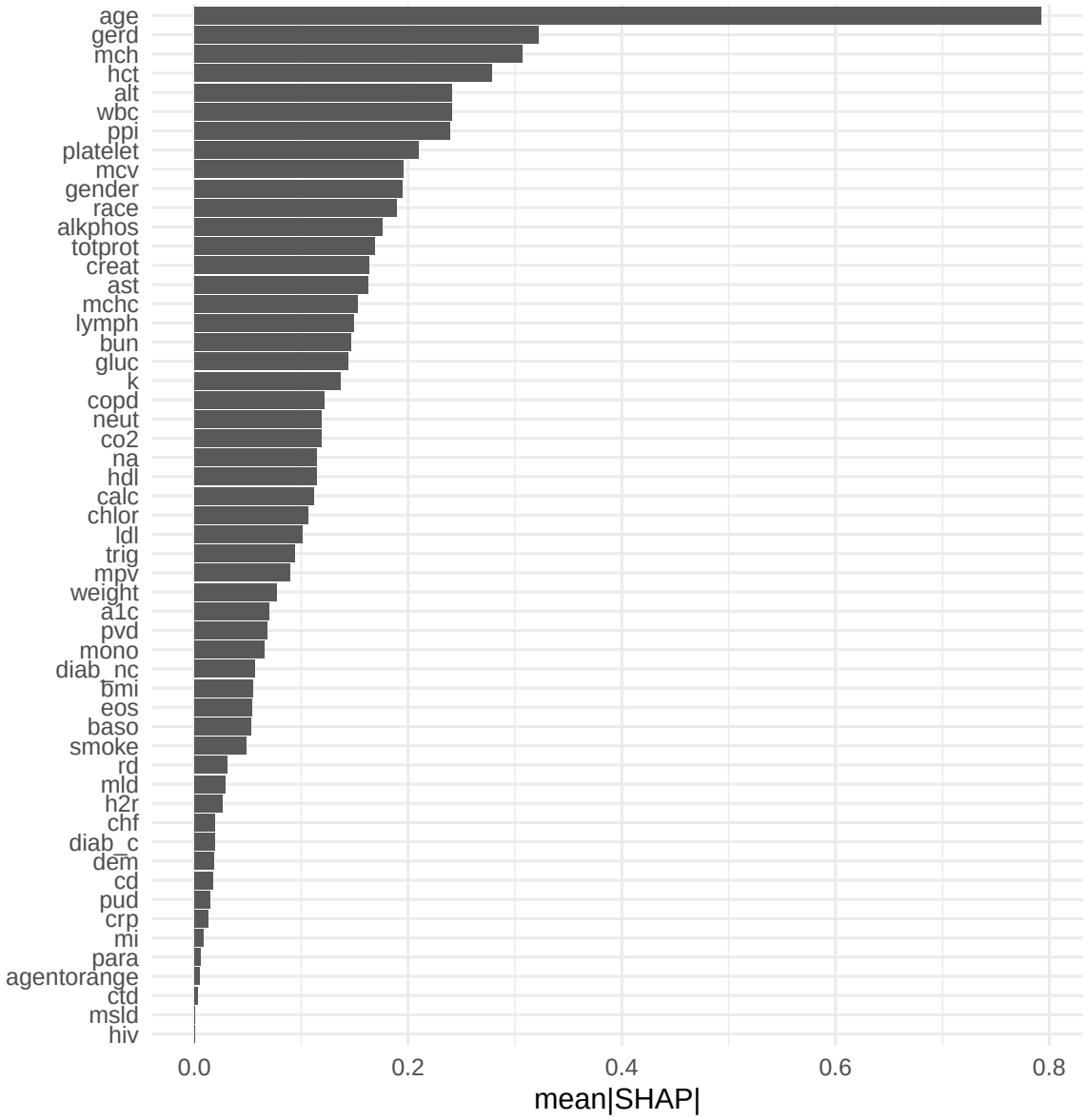


Figure 2: This has mean updated using a representative sample; previously, I was using a sample that over represented cases so age was much less important.


age
gerd
mch
hct
alt
wbc
ppi
platelet
mcv
gender
race
alkphos
totprot
creat
ast
mchc
lymph
bun
gluc
k
copd
neut
co2
na
hdl
calc
chlor
ldl
trig
mpv
weight
a1c
pvd
mono
diab_nc
bmi
eos
baso
smoke
rd
mld
h2r
chf
diab_c
dem
cd
pud
crp
mi
para
agentorange
ctd
msld
hiv
0.0
0.2
0.4
0.6
0.8
mean|SHAP|

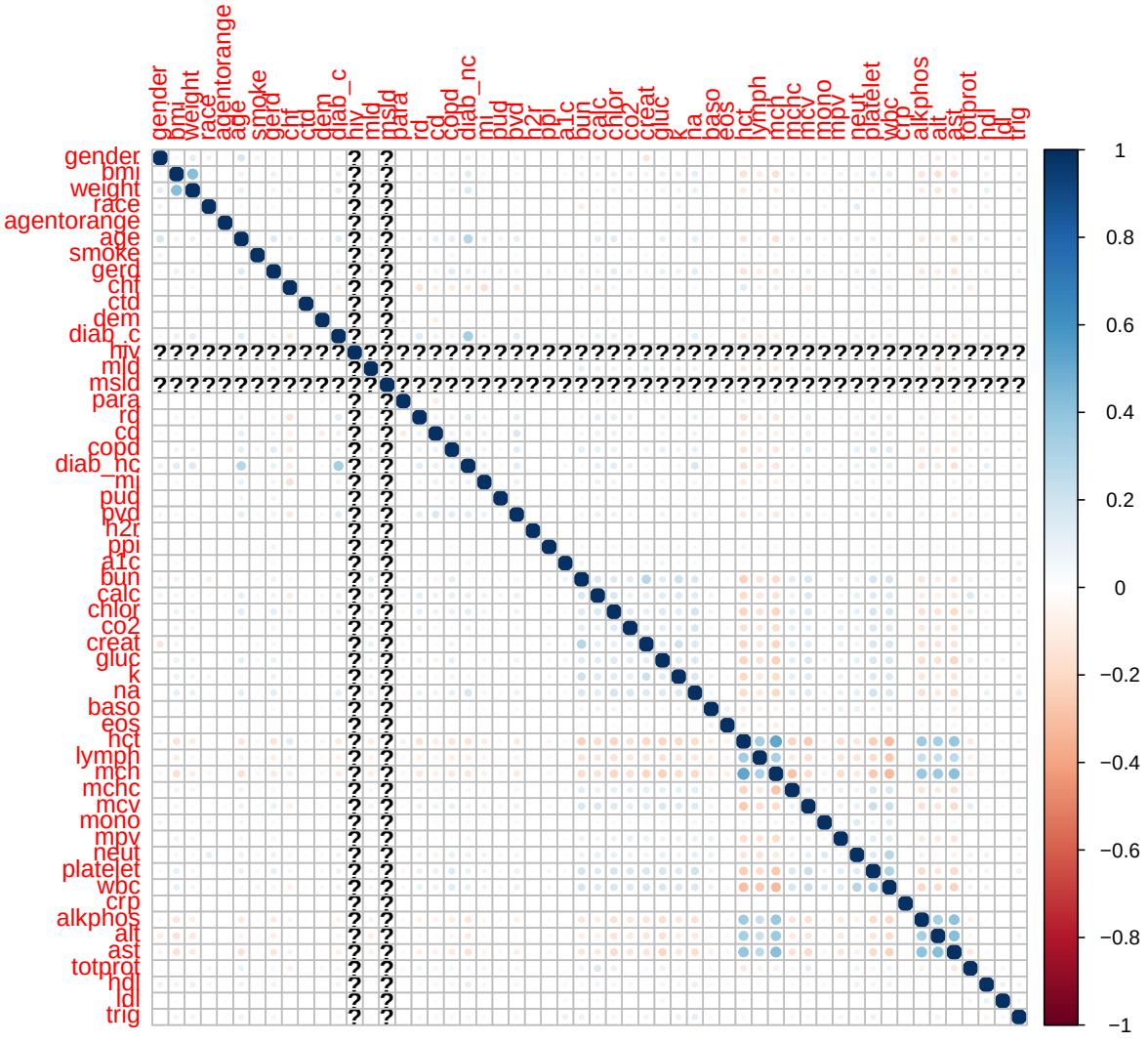
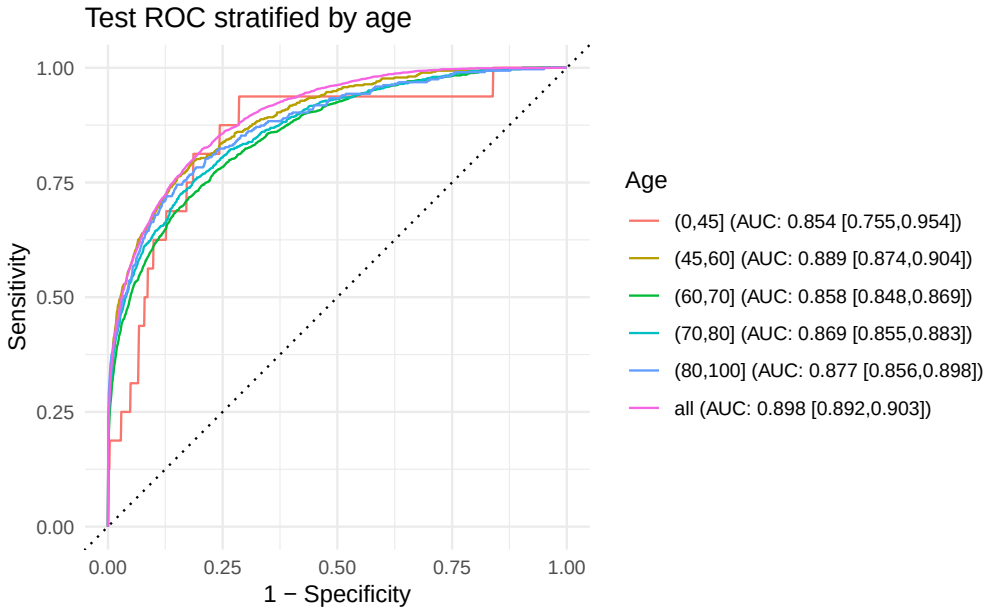


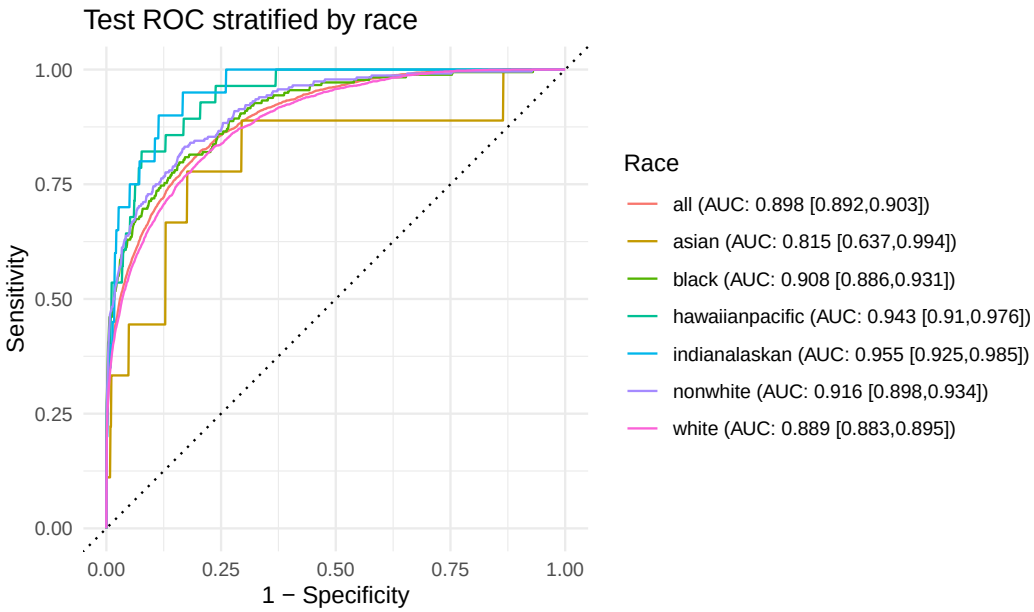
Figure 3: Correlation between group-level Shapley values. Largest absolute correlation in the following table. If two Shapley values are highly correlated, that means the underlying variables contribute essentially in the same way to the prediction and are therefore redundant.

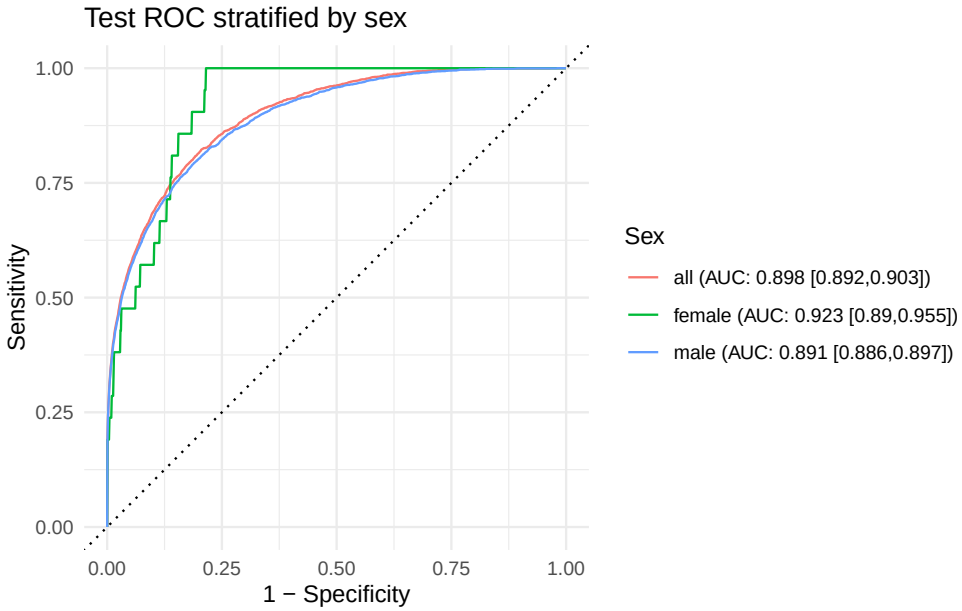
Feature pair		SHAP correlation
hct	mch	0.53
bmi	weight	0.43
alt	ast	0.42
mch	ast	0.42
alkphos	ast	0.41
hct	ast	0.39
mch	alkphos	0.38
mch	alt	0.36
hct	alkphos	0.35
hct	lymph	0.35
diab_c	diab_nc	0.34
hct	alt	0.34
alkphos	alt	0.33
mch	wbc	-0.32
lymph	mch	0.32
platelet	wbc	0.31
hct	wbc	-0.31
mch	mchc	-0.29
age	diab_nc	0.29
bun	creat	0.29

Table 1: Since the largest Shapley correaltion is fairly low, we can be reassured we do not have redundant variables.

Identity groups

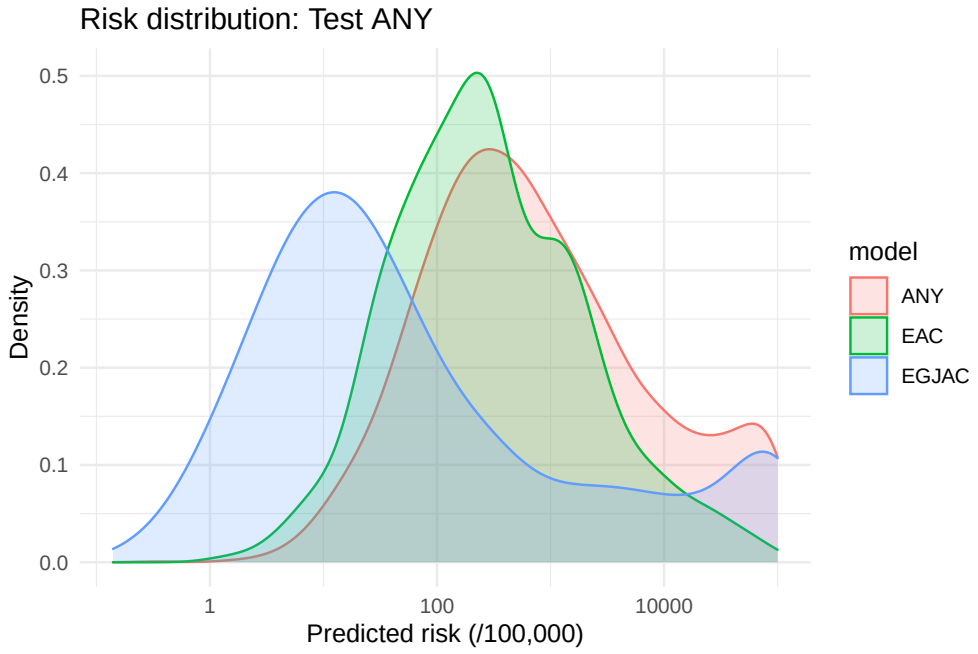


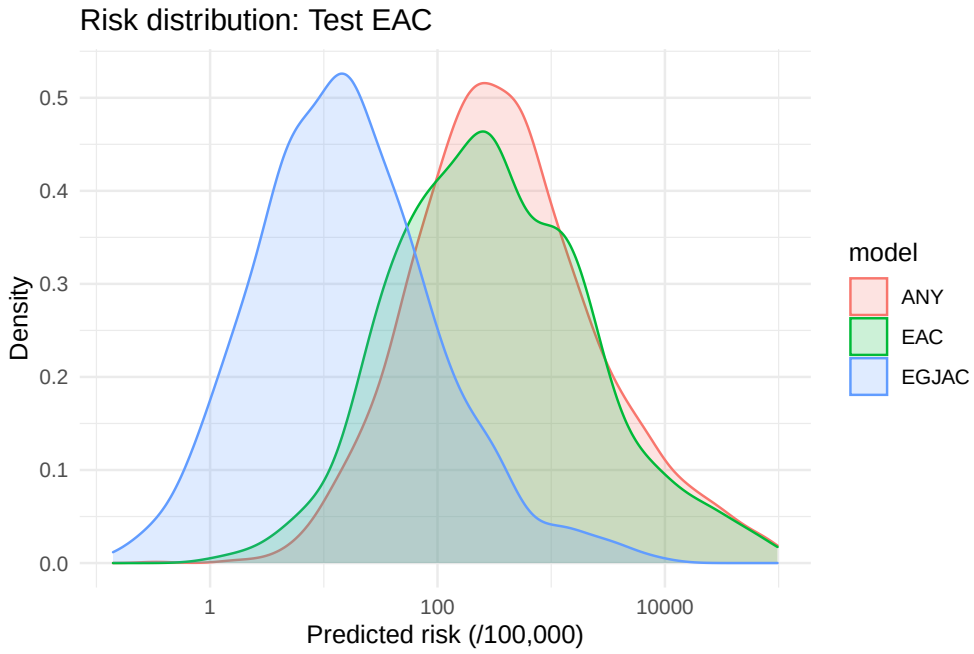


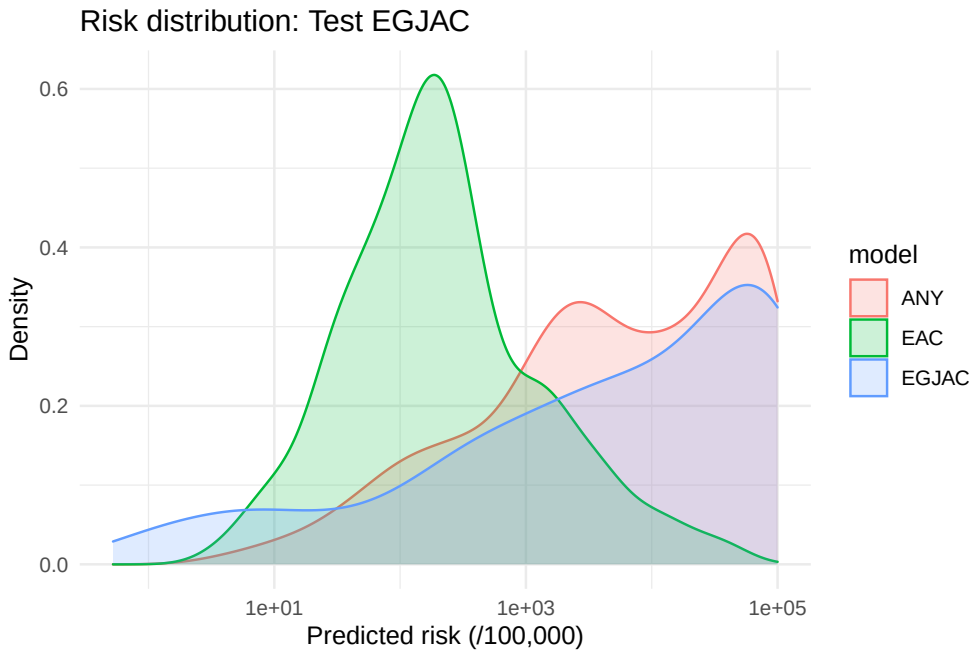

Test ROC stratified by sex
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Sex
all (AUC: 0.898 [0.892,0.903])
female (AUC: 0.923 [0.89,0.955])
male (AUC: 0.891 [0.886,0.897])

Cancertype

	Training		
Testing	ANY	EAC	EGJAC
ANY	0.898 [0.892,0.903]	0.879 [0.872,0.885]	0.955 [0.946,0.964]
EAC	0.852 [0.839,0.866]	0.858 [0.842,0.873]	0.836 [0.810,0.862]
EGJAC	0.800 [0.782,0.818]	0.747 [0.726,0.768]	0.949 [0.926,0.972]

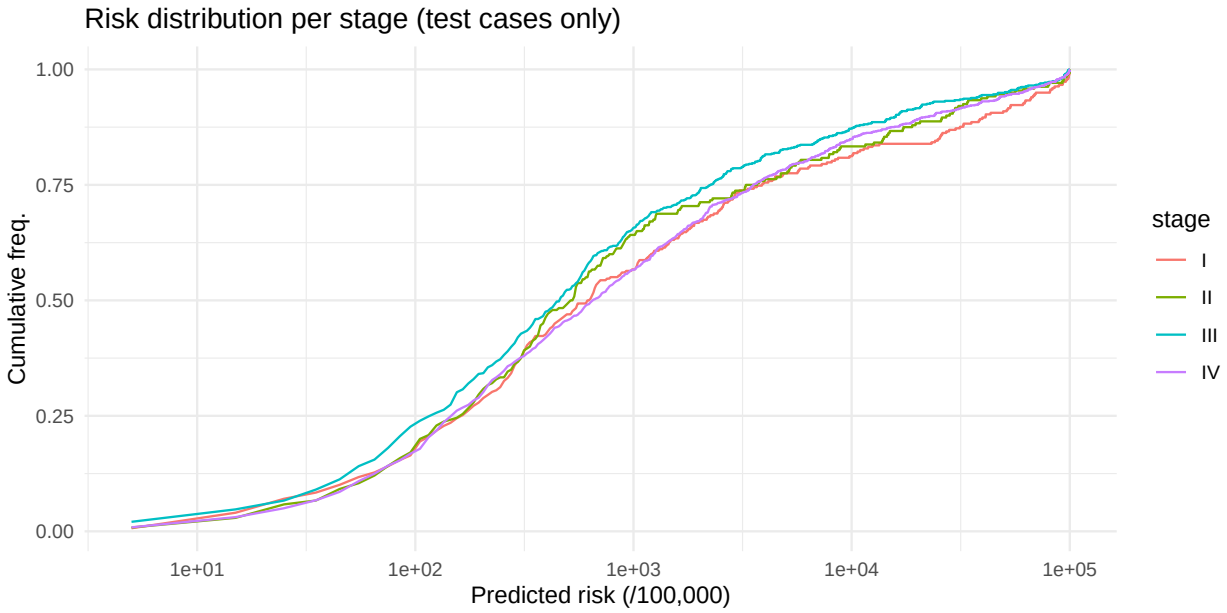



Risk distribution: Test EAC
Density
0.0
0.1
0.2
0.3
0.4
0.5
1
100
10000
Predicted risk (/100,000)
model
ANY
EAC
EGJAC


Risk distribution: Test EGJAC
Density
0.0
0.2
0.4
0.6
1e+01
1e+03
1e+05
Predicted risk (/100,000)
model
ANY
EAC
EGJAC

Cancer stage

Stage	Test. AUC	Nb. cases
Any	2810	0.897 [0.892,0.903]
I	298	0.904 [0.887,0.921]
II	240	0.904 [0.887,0.922]
III	631	0.887 [0.875,0.899]
IV	1041	0.907 [0.899,0.916]
I+	2254	0.900 [0.894,0.907]
II+	1956	0.900 [0.893,0.906]
III+	1716	0.899 [0.892,0.906]
IV+	1041	0.907 [0.899,0.916]



Lab variables interpretation

- I started to look at ways we could interpret the effect of the lab variables on the predictions
- I focused on `mch` for now to develop the approach
- During this, I noticed that the labels for maxdiff and mindiff were flipped ... This doesn't change anything about performance though.
- same thing for min and max ...

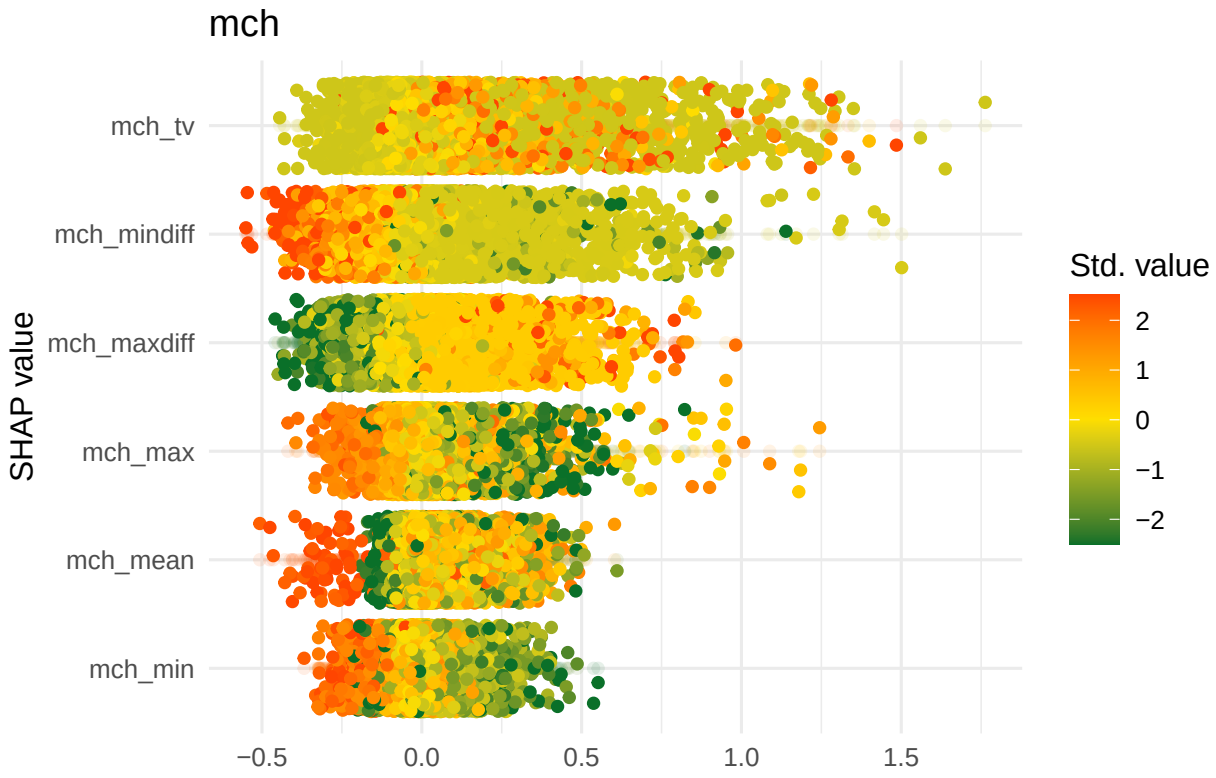


Figure 4: SHAP values for a sample with overrepresentation of cases ranked according to `mean|SHAP|` from top to bottom.

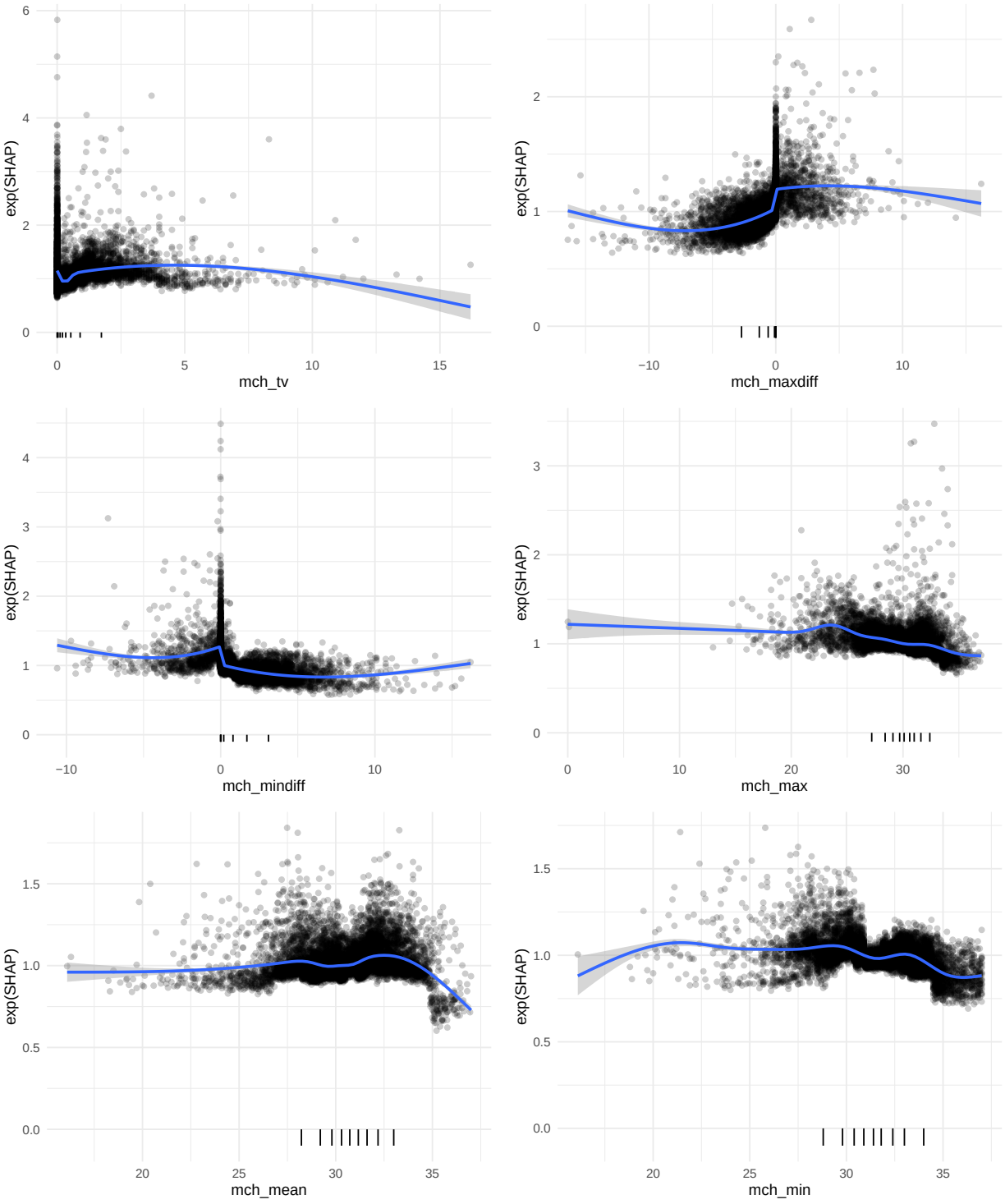


Figure 5: Very negative mindiff(maxdiff) means there is at least one large drop. Very positive maxdiff (mindiff) means at least one large positive jump. Small max (min) means always small. Large min (max) means always large.

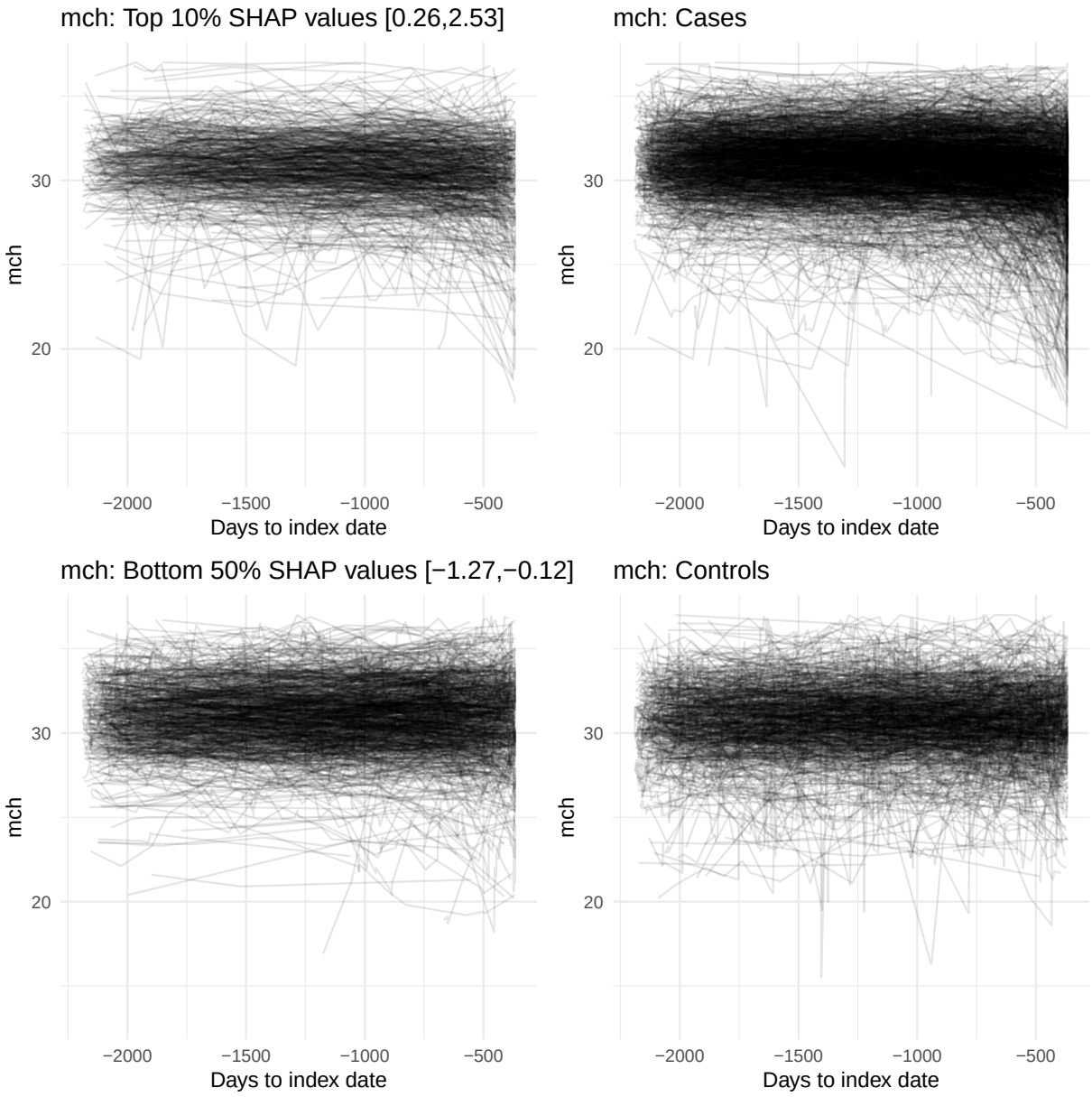


Figure 6: Sample of trajectories of mch lab results. Left: stratified by their aggregated SHAP values. Right: straified by case/control status. It is not obvious why some time series have low/high SHAP values just by looking at this, but we at least see that drop towards the end seem to be well separated.

Years prior

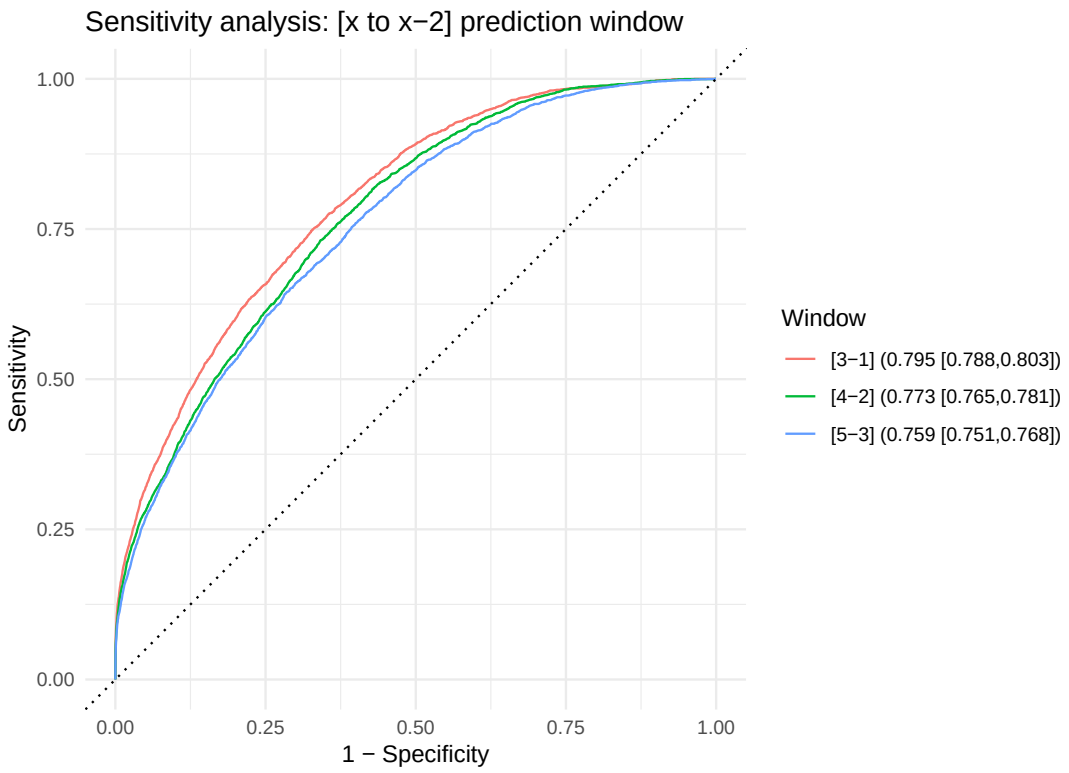


Figure 7: In this experiment, all three test set utilizes the same amount of data (2 years), but we predict farther in the future (1yr, 2yrs or 3yrs). As expected, we do worse and worse as we predict further in time, but only by a small margin. This seem to indicate are predictions are somewhat valid beyond the 1yr window we used.

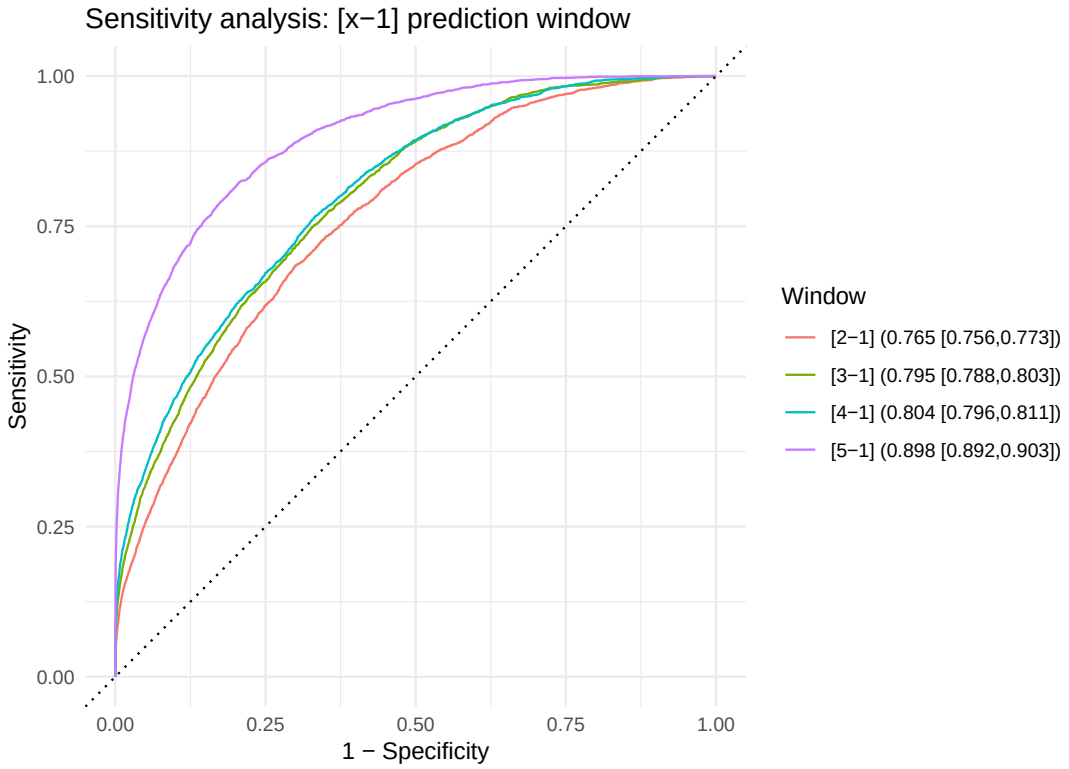


Figure 8: In this experiment, we decrease the amount of data (from 4yrs to 1yr) in the test set, but keep the prediction horizon to 1yr. As expected, performance decreases as the number of years decrease (we have less and less data). However, we notice a large jump in performance from 3yrs to 4yrs. A possible explanation is that our model is specifically trained for 4 years of data and the summary features change significantly. In particular, minimum and maximum statistics are highly sensitive to having more data. This also increases the amount of missing data.

EAC v EGJAC

- They have different predicted risk
  - EGJAC have higher predicted risk, much better separation from controls
  - Esp. for the region of interest, EAC has a lot more subject around the threshold
- Previous comparison of feature distribution was incorrect (was done after imputation)
  - see new results below
- There is a stark difference in proportion of missing values between EAC and EGJAC cases for HCT, LYMPH and MCH.
  - Almost always missing for EGJAC
- Some feature have different impact between the two models; EAC is generally very similar to ANY

	Mean (control)	Mean (EAC)	Mean (EGJAC)	pvalue.adj
black	0.169	0.041	0.089	0.000
gerd	0.229	0.481	0.397	0.010
chlor_mean	103.551	103.039	103.527	0.019
chlor_max	105.950	106.630	107.270	0.011
gluc_min	93.387	89.910	86.167	0.011
na_mean	139.252	138.515	138.856	0.048
na_max	141.373	141.611	142.023	0.048
mchc_mean	33.730	33.690	33.530	0.010
mchc_min	33.096	32.763	32.415	0.000
ldl_min	90.441	82.012	76.903	0.013

Table 2: Features with different means between EAC and EGJAC (at 5% level, BH).

	Prop. Control	Prop. EAC	Prop. EGJAC	pvalue.adj
baso_mindiff	0.543	0.620	0.650	0.042
baso_maxdiff	0.543	0.620	0.650	0.042
baso_tv	0.543	0.620	0.650	0.042
hct_mean	0.798	0.882	0.025	0.000
hct_min	0.798	0.882	0.025	0.000
hct_max	0.798	0.882	0.025	0.000
hct_mindiff	0.786	0.854	0.024	0.000
hct_maxdiff	0.786	0.854	0.024	0.000
hct_tv	0.786	0.854	0.024	0.000
lymph_mean	0.640	0.727	0.020	0.000
lymph_min	0.640	0.727	0.020	0.000
lymph_max	0.640	0.727	0.020	0.000
lymph_mindiff	0.560	0.636	0.018	0.000
lymph_maxdiff	0.560	0.636	0.018	0.000
lymph_tv	0.560	0.636	0.018	0.000
mch_mean	0.801	0.883	0.025	0.000
mch_min	0.801	0.883	0.025	0.000
mch_max	0.801	0.883	0.025	0.000
mch_mindiff	0.795	0.861	0.023	0.000
mch_maxdiff	0.795	0.861	0.023	0.000
mch_tv	0.795	0.861	0.023	0.000

Table 3: Features with different proportions of non-NAs between EAC and EGJAC (at 5% level, BH). It seems that EAC has similar level than Controls, but not EGJAC.

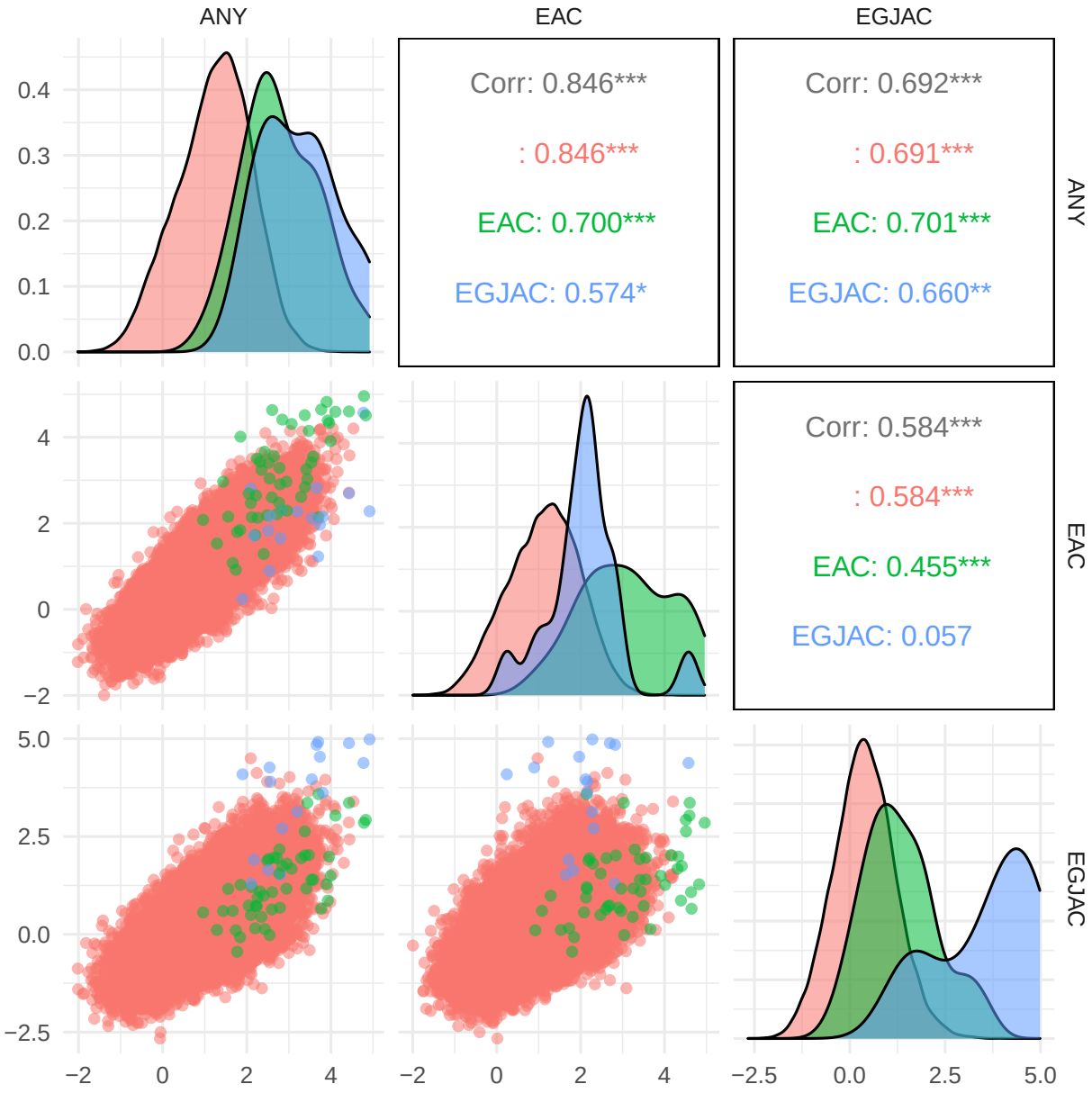


Figure 9: Scatter plots of (log) predicted risk by all three models stratified by case/control status (none, EAC, EGJAC). We have very strong correlation between EAC and ANY, but much less between ANY and EGJAC. There is no correlation between EAC and EGJAC for EGJAC patients. EGJAC assigns very high predicted risk to EGJAC patients compared to what ANY/EAC does to their respective targets.

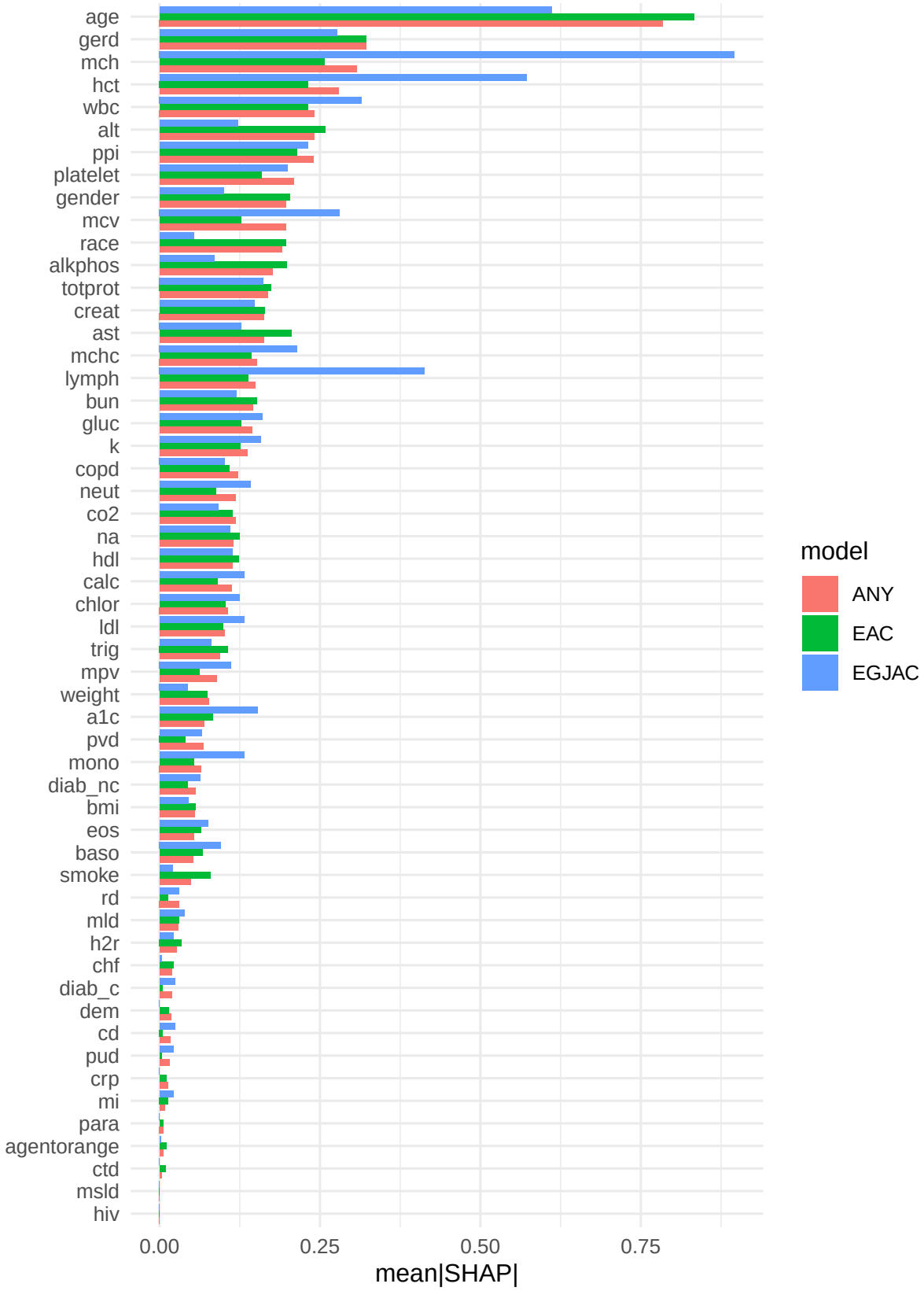


Figure 10: SHAP variable importance wihtin all three models. Notably, we see large differences `age`, `mch`, `hct`, `mcv`, `lymph`, `race` and a few others.

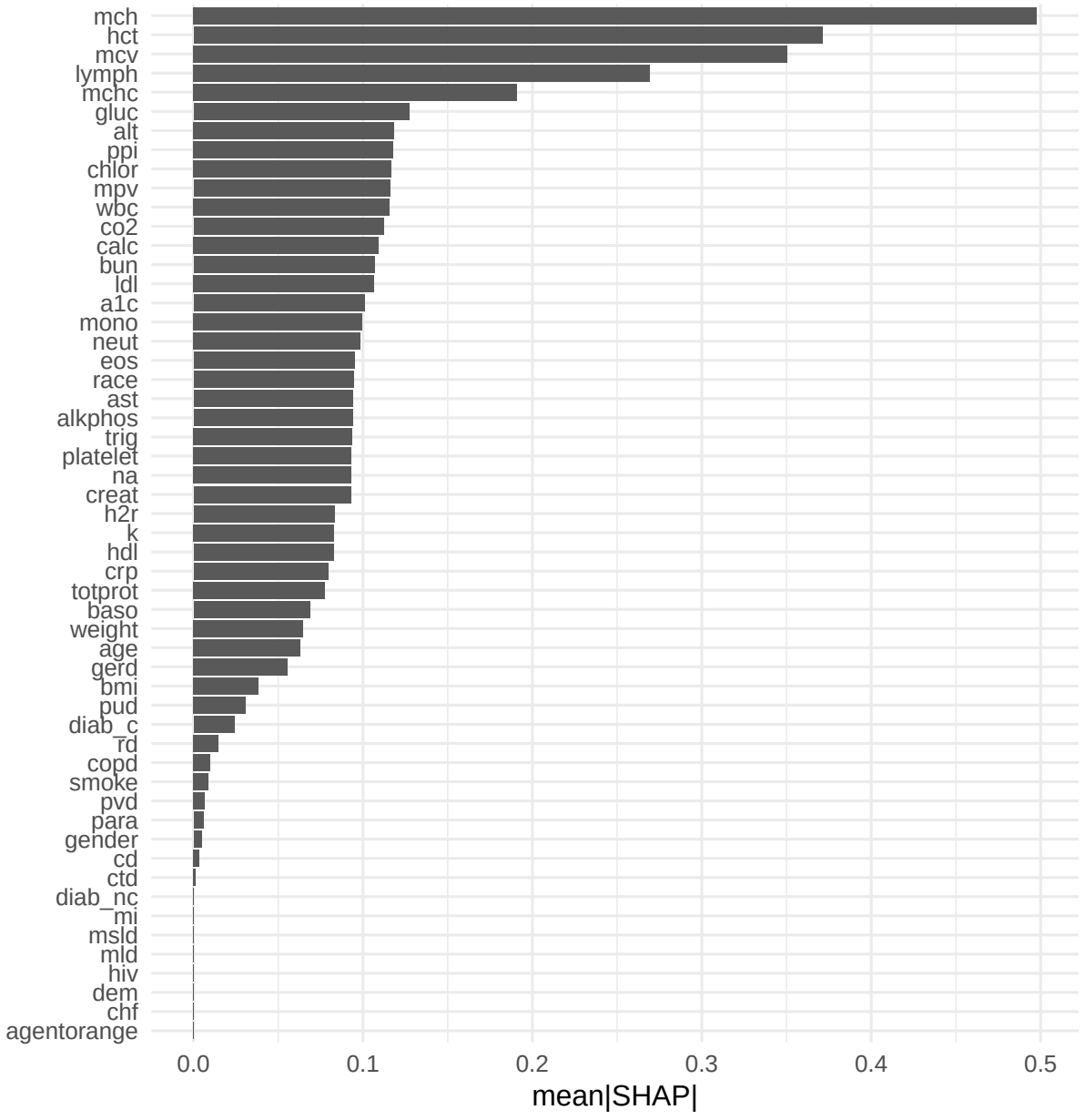


Figure 11: SHAP variable importance for a model differentiating EAC from EGJAC. We again find `mch`, `hct`, `mcv`, `lymph` and `mchc` at the top

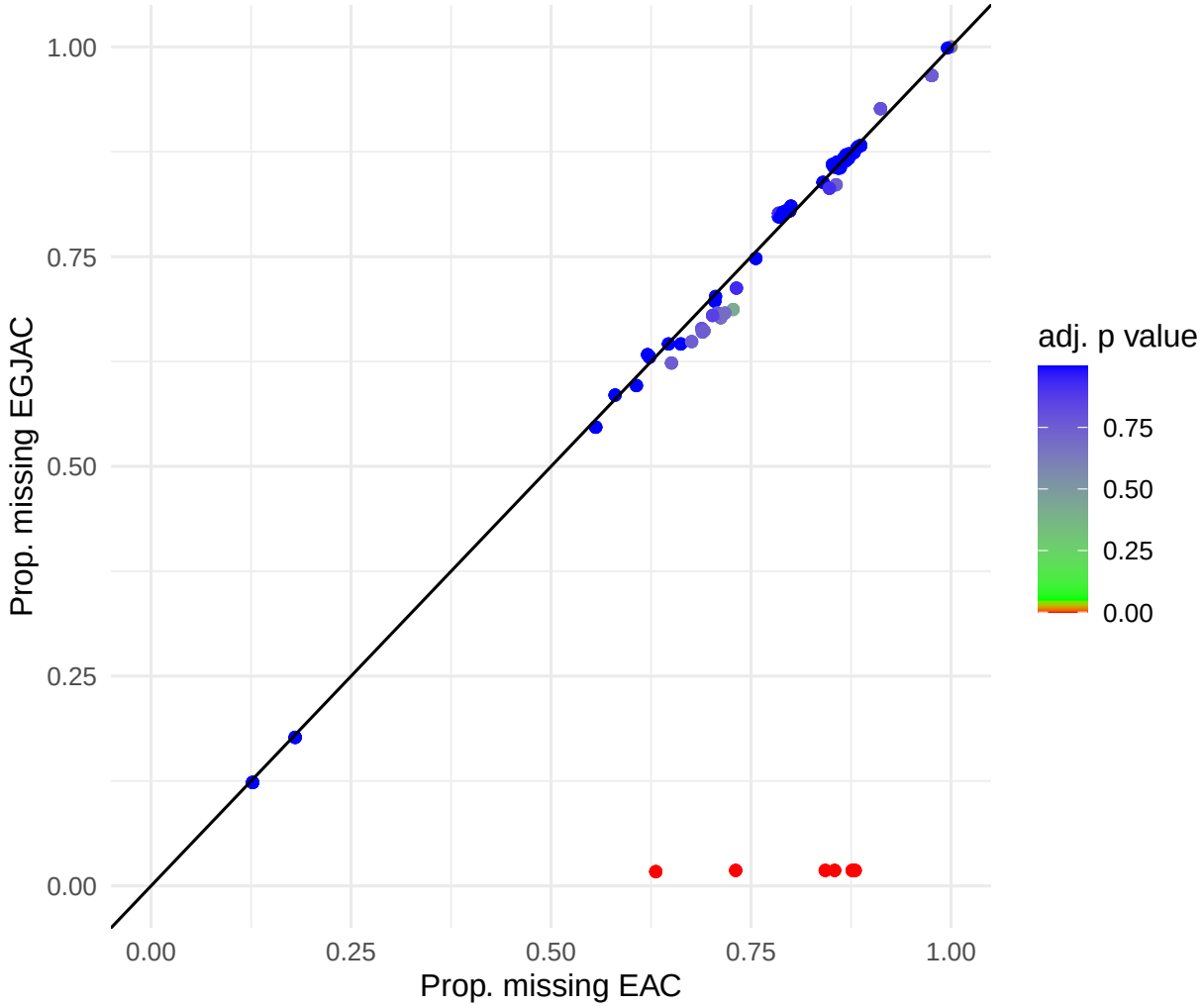
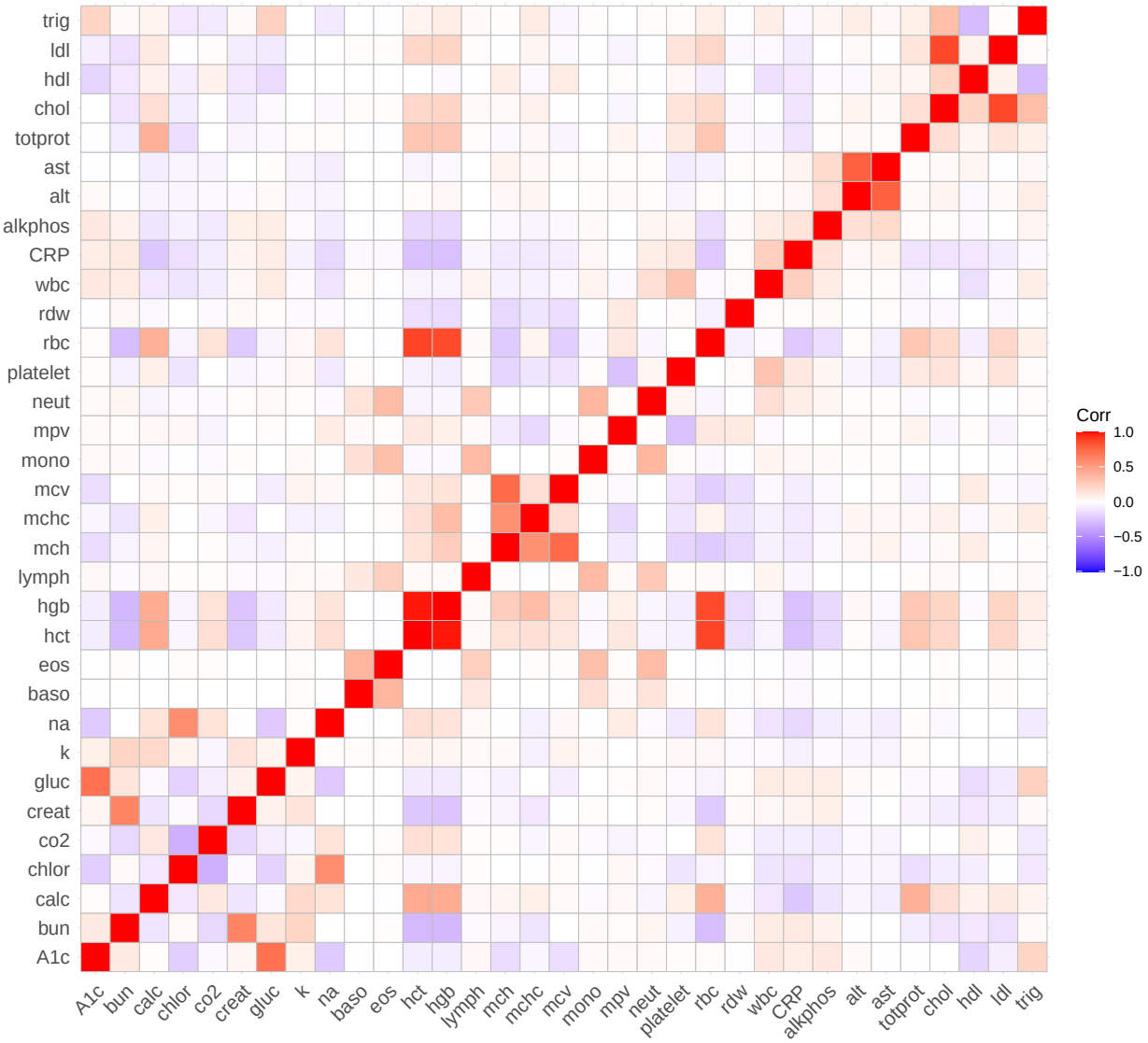


Figure 12: Labels are wrong and should read "non-missing." This shows it really is just there three that have different missing proportions.

Old stuff to update from here

Feature selection

Correlation:



	var1	var2	correlation
2	mchc	mch	0.569
3	chlor	na	0.578
6	creat	bun	0.617
8	gluc	A1c	0.707
10	mcv	mch	0.742
11	alt	ast	0.782
13	hgb	rbc	0.857
15	chol	ldl	0.875
18	rbc	hct	0.884
20	hgb	hct	0.976

Table 4: HGB and HCT are indeed very highly correlated

On chol/ldl/hdl: Missingness patterns does not help

Pattern (chol,hdl,ldl)	000	001	010	011	100	101	110	111
Prop (%)	59.5	0.7	0.5	0.6	0.2	0.0	0.2	38.3

Using performance:

- Drop various features:
  - (colonoscopy, fobt)
  - two of (hct, hgb, rct)
  - one or two of (chol, ldl, hdl)
- Refit with 1M controls & evaluate
- Proposed is dropping (colonoscopy, fobt, hbg, rct, chol)
- Conclusions:
  - Not much difference
  - Probably best to avoid dropping hct

Drop	Valid. AUC	Test AUC
None	0.832	0.827
Colonoscopy, fobt	0.830	0.826
rct, hgb	0.831	0.826
rbc, hct	0.824	0.821
hct, hgb	0.824	0.826
chol	0.834	0.831
ldl	0.831	0.829
hdl	0.832	0.825
hdl, ldl	0.831	0.830
(proposed)	0.827	0.827