HOSEA Aim I – Report

Simon Fontaine

July 26, 2022

Calibration & threshold

Threshold	TPR	PPV	Det. prev.
10	99.51	0.14	79.20
15	99.19	0.14	76.68
20	99.05	0.15	74.69
25	98.74	0.15	72.82
30	98.46	0.15	70.91
35	97.93	0.16	68.92
40	97.26	0.16	66.88
45	96.77	0.17	64.81
50	95.82	0.17	62.74
55	94.66	0.17	60.67
60	93.86	0.18	58.59
65	92.80	0.18	56.51
70	91.54	0.19	54.44
75	89.93	0.19	52.41
80	88.77	0.20	50.40
85	87.40	0.20	48.42
90	86.03	0.21	46.47
95	84.35	0.21	44.57
100	82.91	0.22	42.70
105	80.84	0.22	40.87
110	79.26	0.23	39.08
115	77.89	0.23	37.35
120	76.06	0.24	35.67
125	74.34	0.24	34.03
130	72.27	0.25	32.44
135	69.95	0.25	30.91
140	68.30	0.26	29.46
145	66.34	0.26	28.06
150	64.41	0.27	26.71
155	62.97	0.27	25.44
160	61.21	0.28	24.21
165	59.39	0.29	23.05
170	57.63	0.29	21.94
175	55.70	0.30	20.88
180	53.88	0.30	19.89
185	52.61	0.31	18.93
190	50.97	0.31	18.02
195	49.49	0.32	17.17
200	47.81	0.32	16.36
210	45.03	0.34	14.88
220	41.84	0.34	13.53
230	39.35	0.35	12.34
240	36.96	0.36	11.25
250	34.85	0.38	10.28
260	32.82	0.39	9.40
270	31.52	0.41	8.62
280	29.52	0.41	7.91
290	27.87	0.43	7.26
300	26.04	0.43	6.67
325	22.01	0.45	5.43
350	18.67	0.46	4.46
375	16.15	0.49	3.66
400	14.04	0.51	3.03
425	12.57	0.55	2.52
450	10.74	0.57	2.10
475	9.23	0.58	1.76
500	8.11	0.61	1.47
600	4.32	0.64	0.75
700	2.81	0.79	0.39
800	1.76	0.90	0.22
900	1.02	0.92	0.12
1000	0.49	0.79	0.07
1100	0.28	0.74	0.04
1200	0.18	0.73	0.03
1300	0.07	0.50	0.02
1400	0.07	0.81	0.01
1500	0.07	1.18	0.01
1600	0.04	0.90	0.00
1700	0.04	1.33	0.00
1800	0.04	2.00	0.00
1900	0.04	2.94	0.00
2000	0.04	4.00	0.00

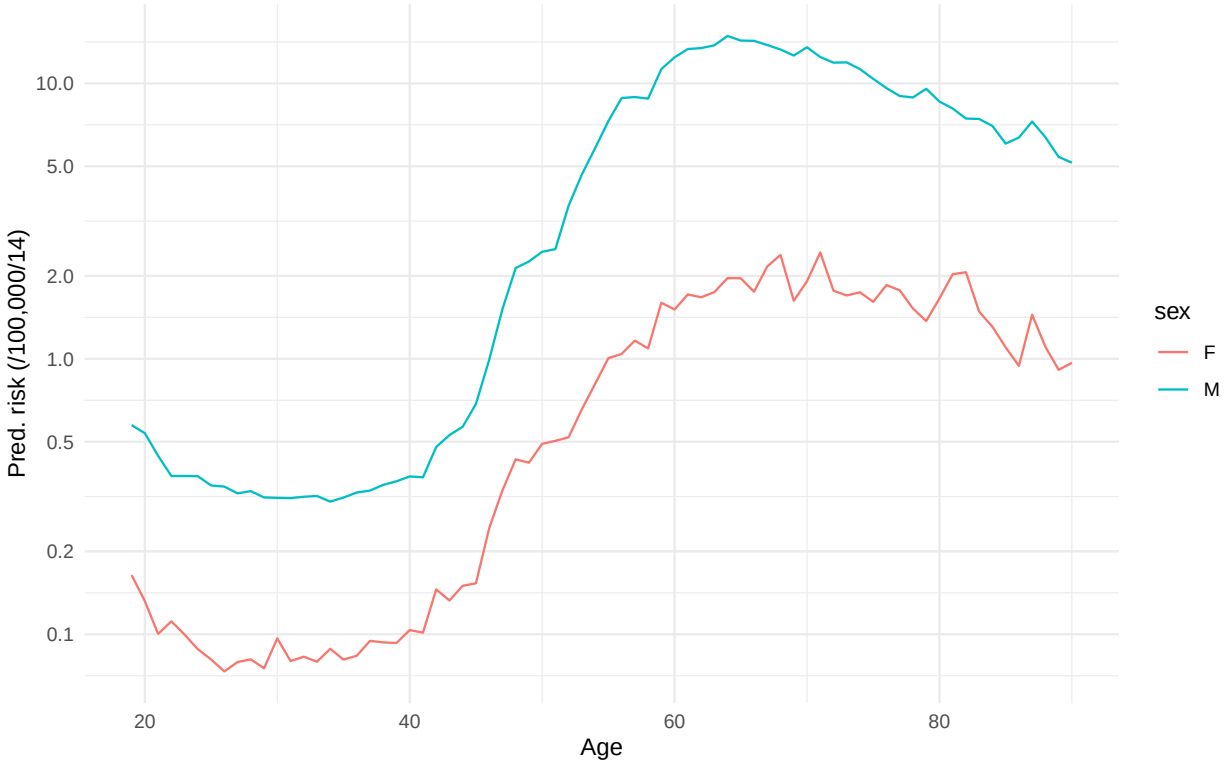
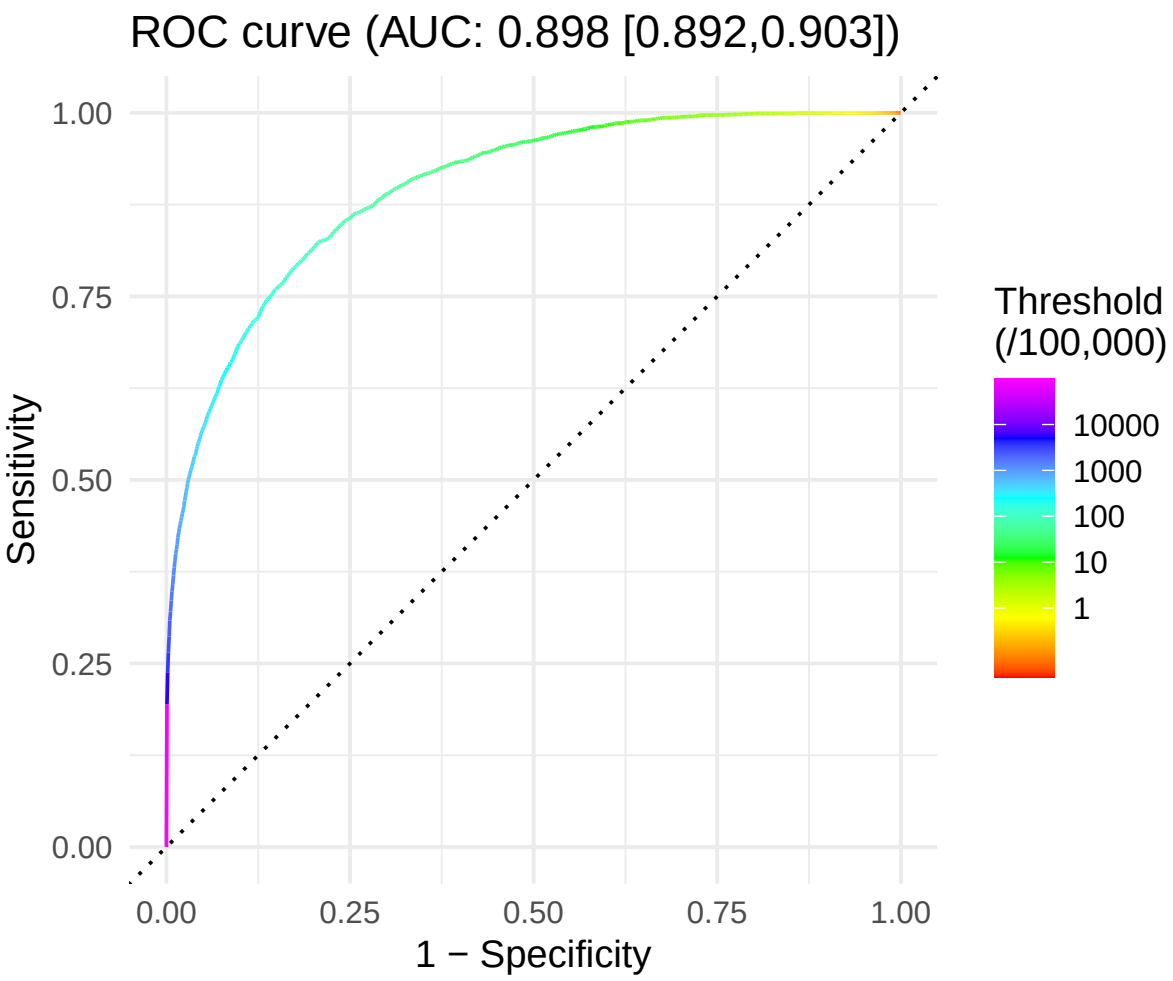
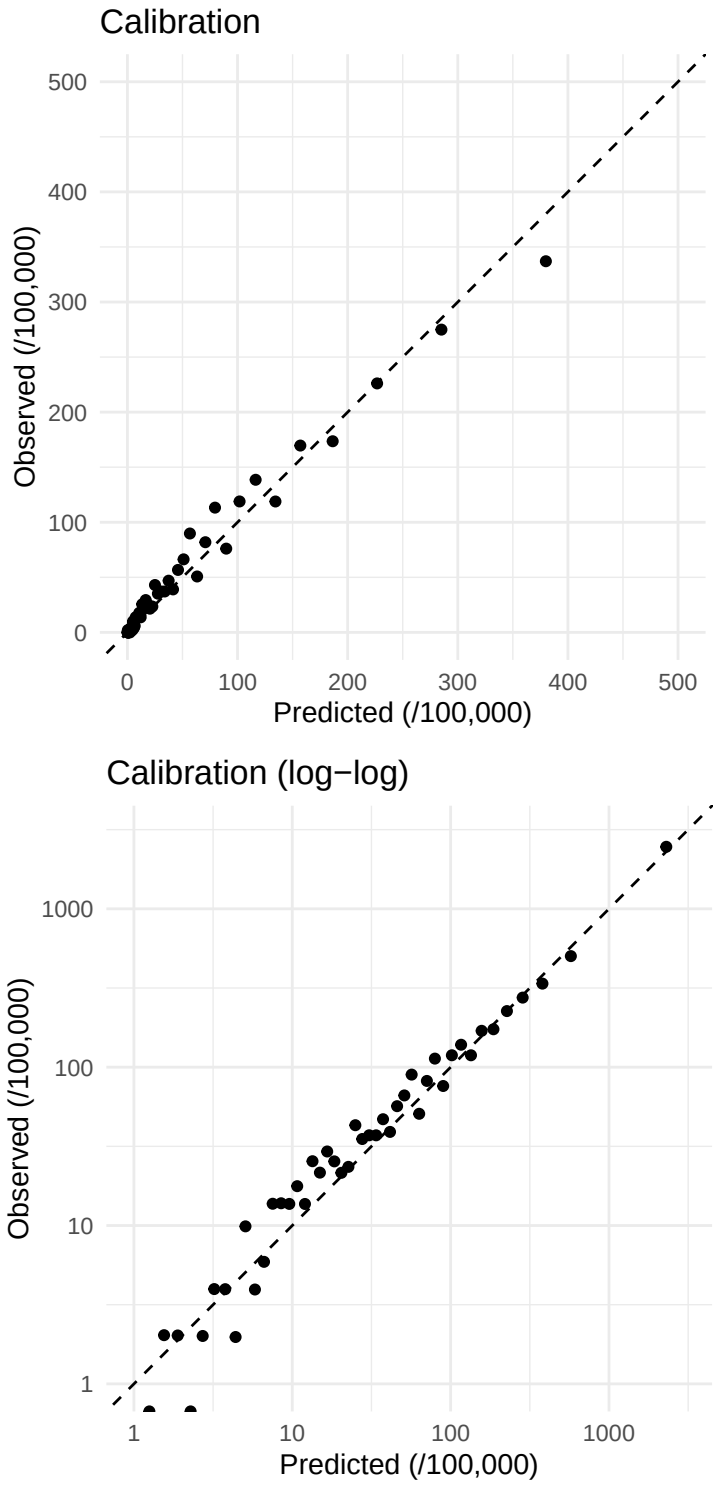


Figure 1: The average annalized risk is 7.6/100,000: for males, it is 8.2/100,000 and, for females, it is 0.7/100,000.


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

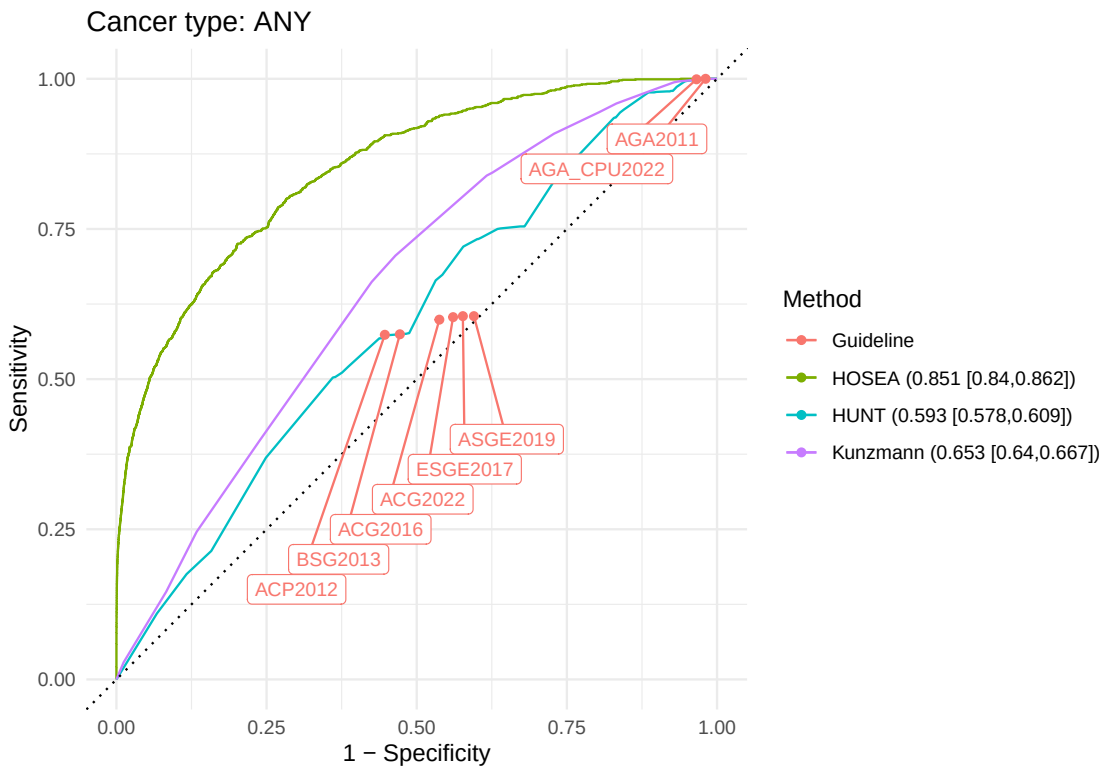


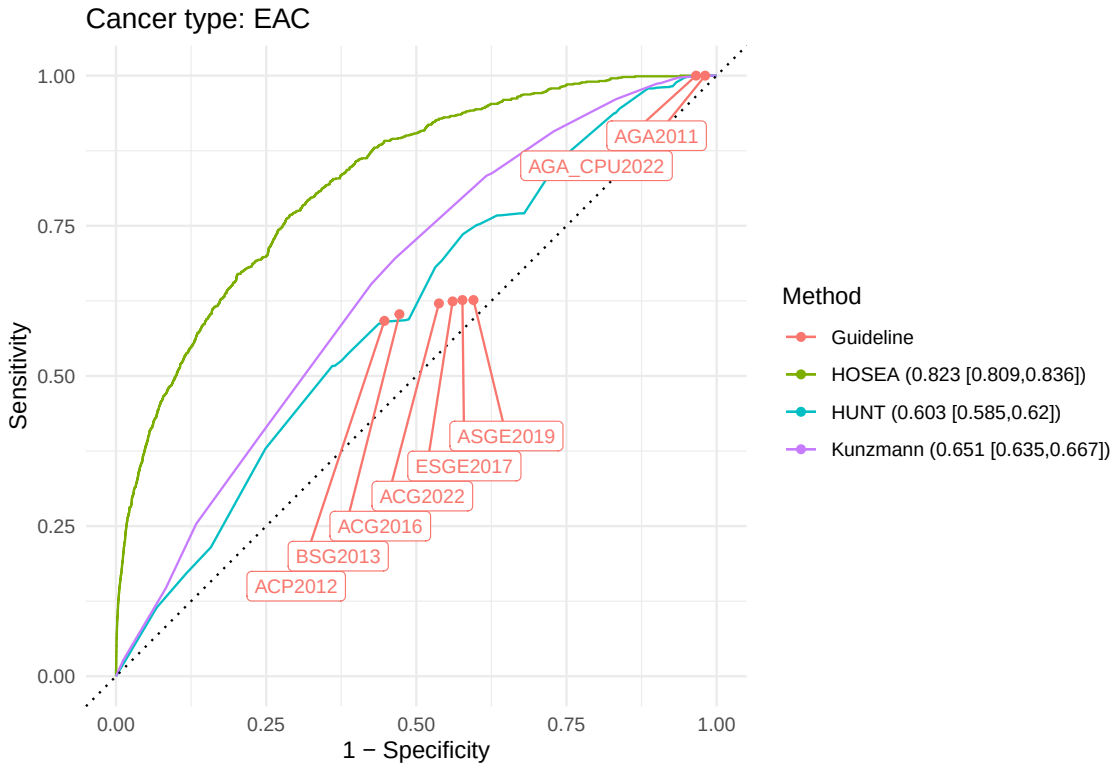
Each point represent 2% of the test data, split using predicted risk quantiles. The top plot is cropped on the right and top so we can focus on the more important region. HL test: "H=49.8507485754787, df=49, p=0.439294514740096"

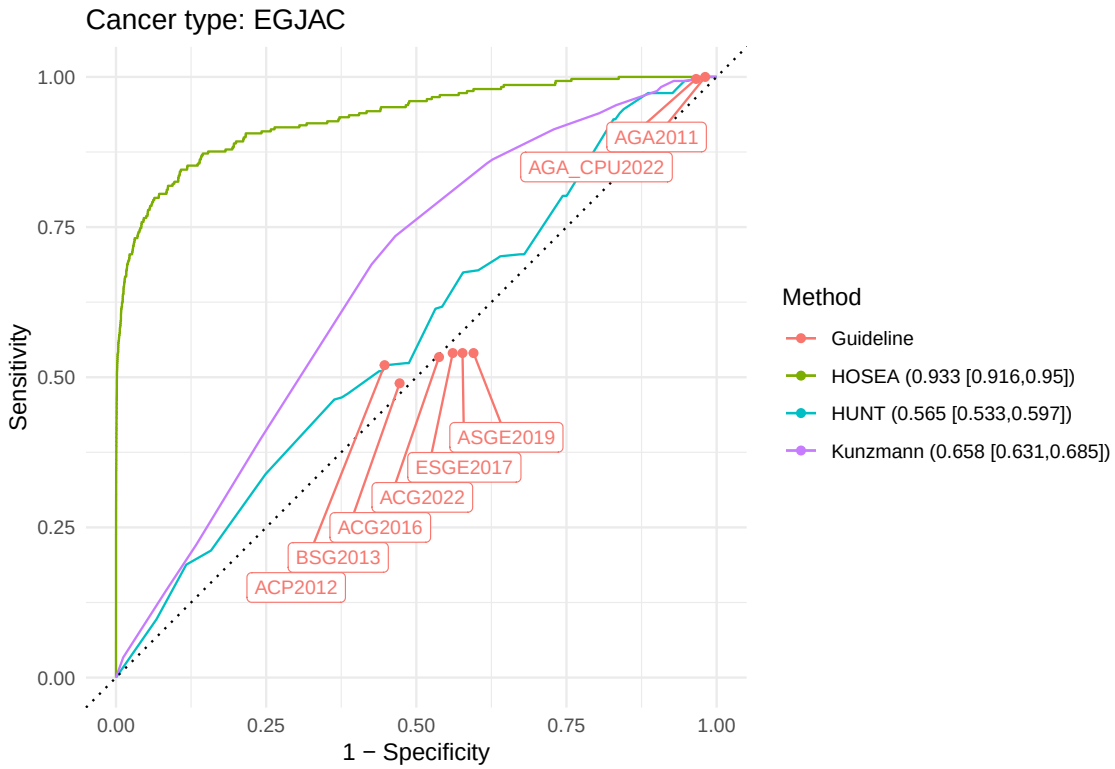
Comparison

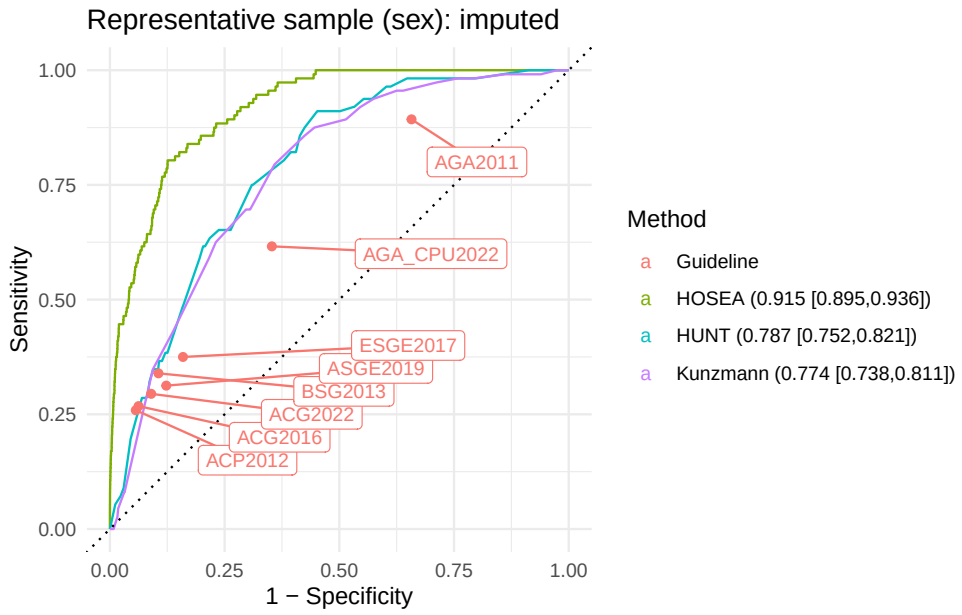
For these comparisons:

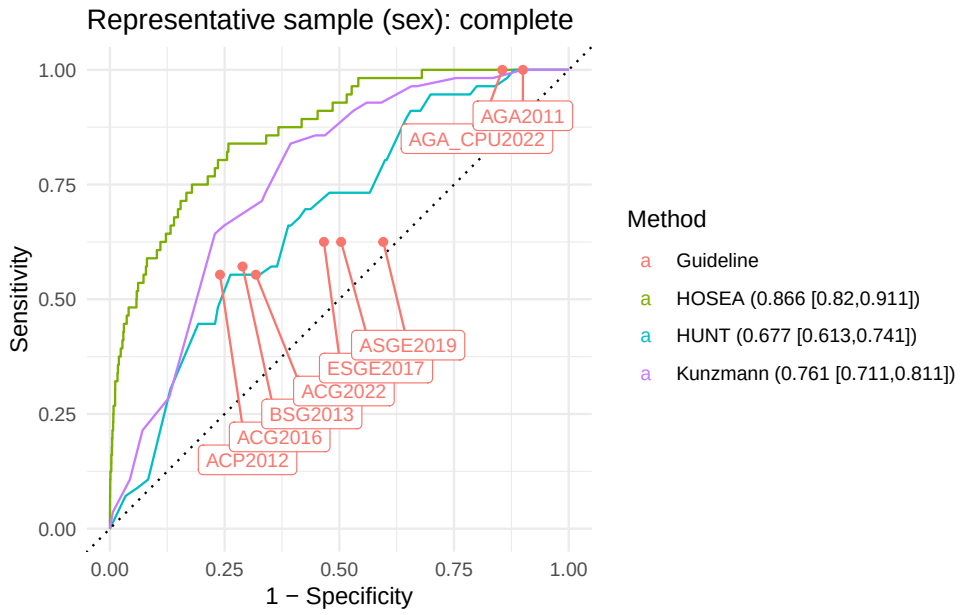
- Subset to test observations
- Filter out patients with missing information for HUNT/Kunzmann/Guidelines
- i.e., require age, BMI, race, smoking status, gerd, h2r/ppi
- 407K patients, 1192 cases (292/100,000)
- AUC + 95% CI using DeLong method




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])


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): 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])

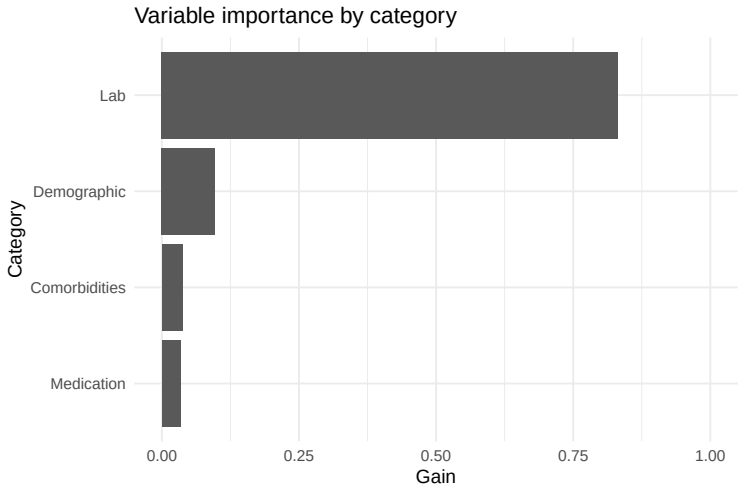

Representative sample (sex): complete
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
AGA2011
AGA_CPU2022
ASGE2019
ESGE2017
ACG2022
BSG2013
ACG2016
ACP2012
Method
a Guideline
a HOSEA (0.866 [0.82,0.911])
a HUNT (0.677 [0.613,0.741])
a Kunzmann (0.761 [0.711,0.811])

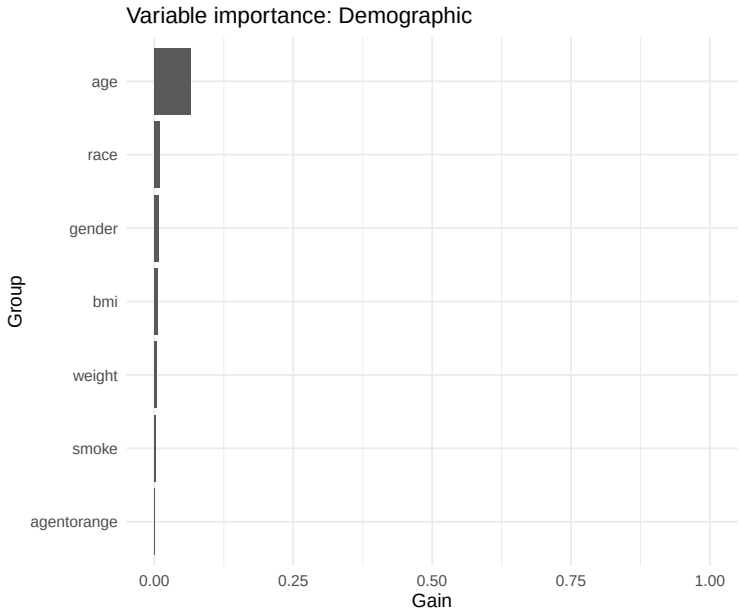
Gain Variable importance

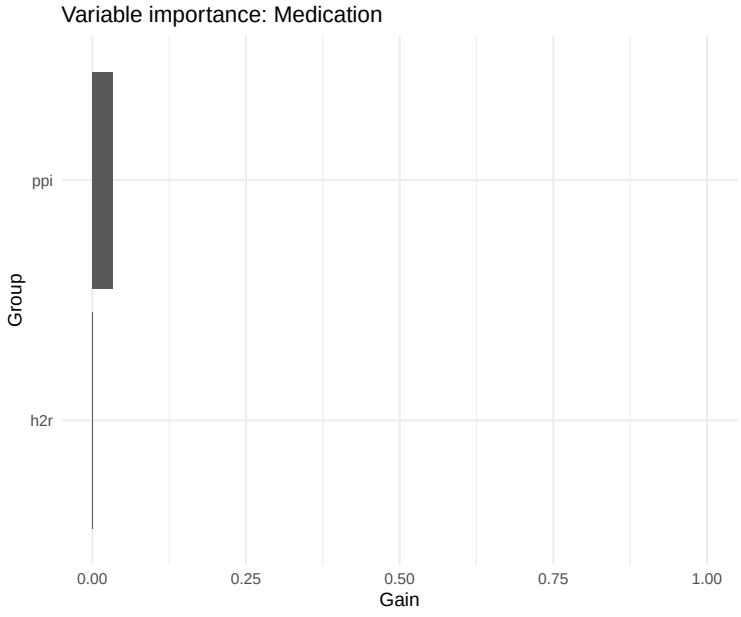
"Gain represents fractional contribution of each feature to the model based on the total gain of this feature's splits. Higher percentage means a more important predictive feature."

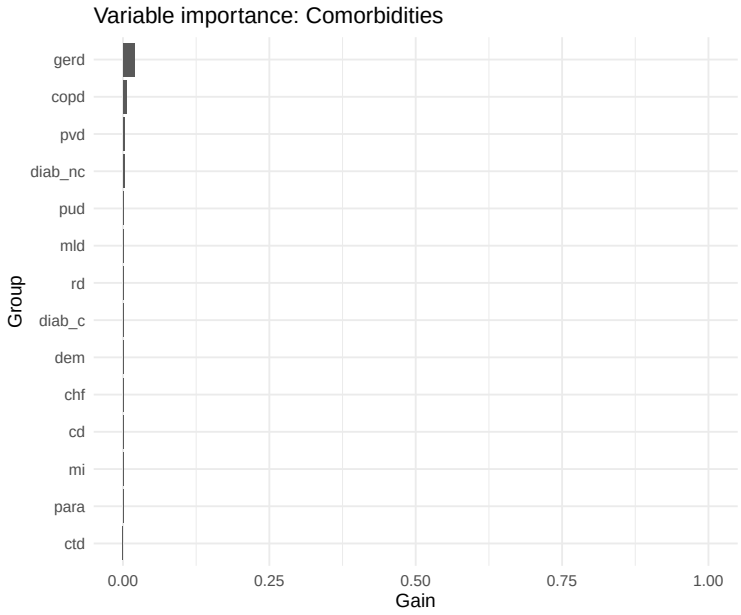
Some notes:

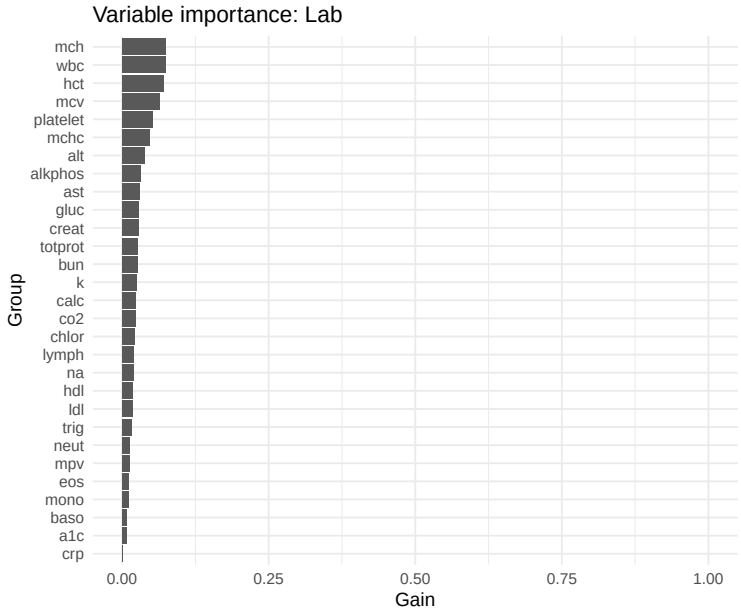
- These are additive, so we can compute the importance of a group of features.
- They sum up to 1
- We can only look at the relationship with the feature value (e.g., mostly positive, mostly negative) for single features




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

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

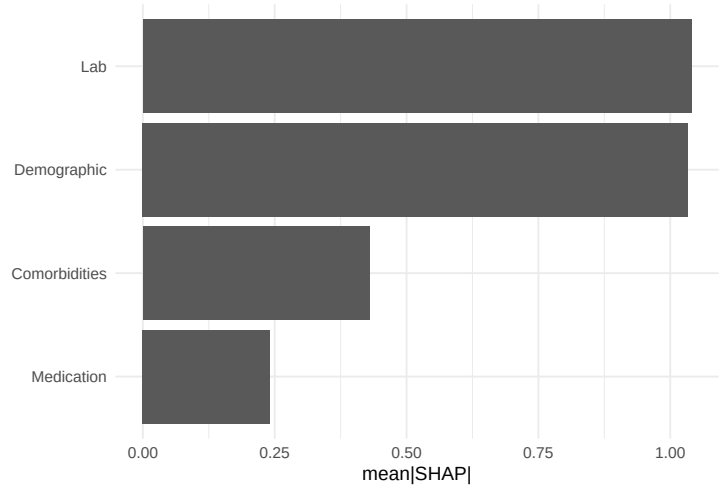
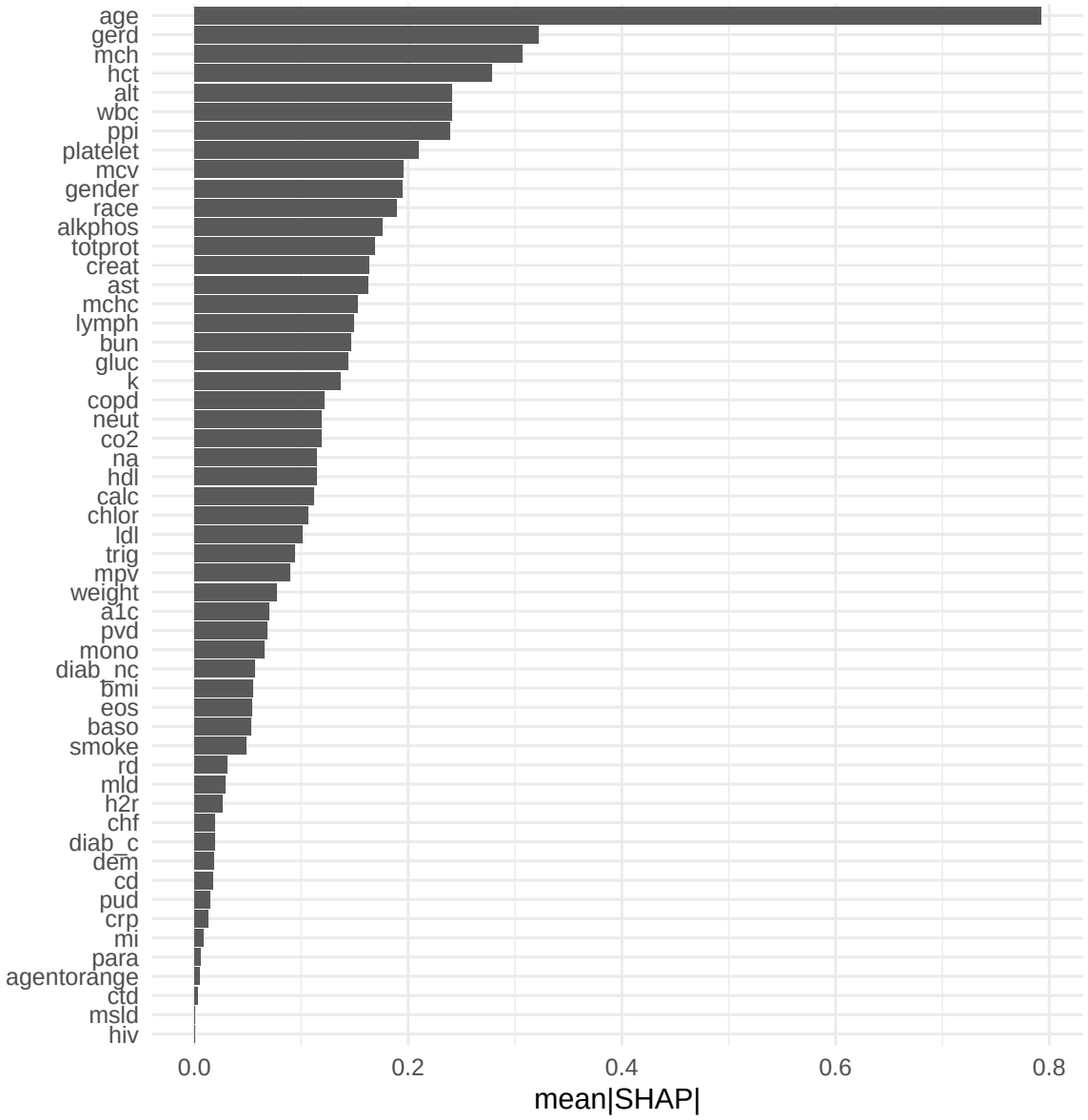


Figure 2: This has mean updated using a representative sample; previously, I was using a sample that over represented cases so age was much less important.


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|

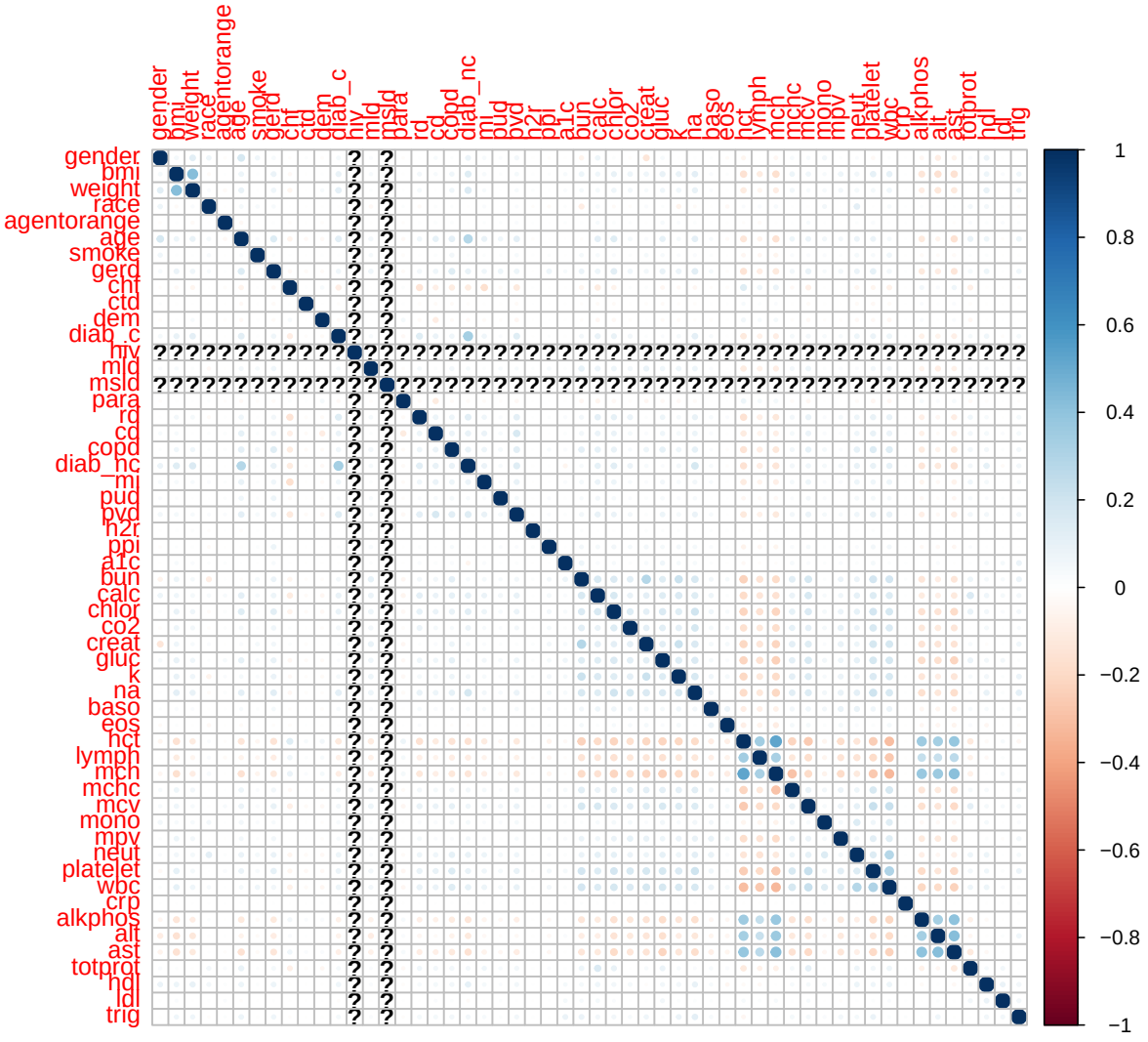
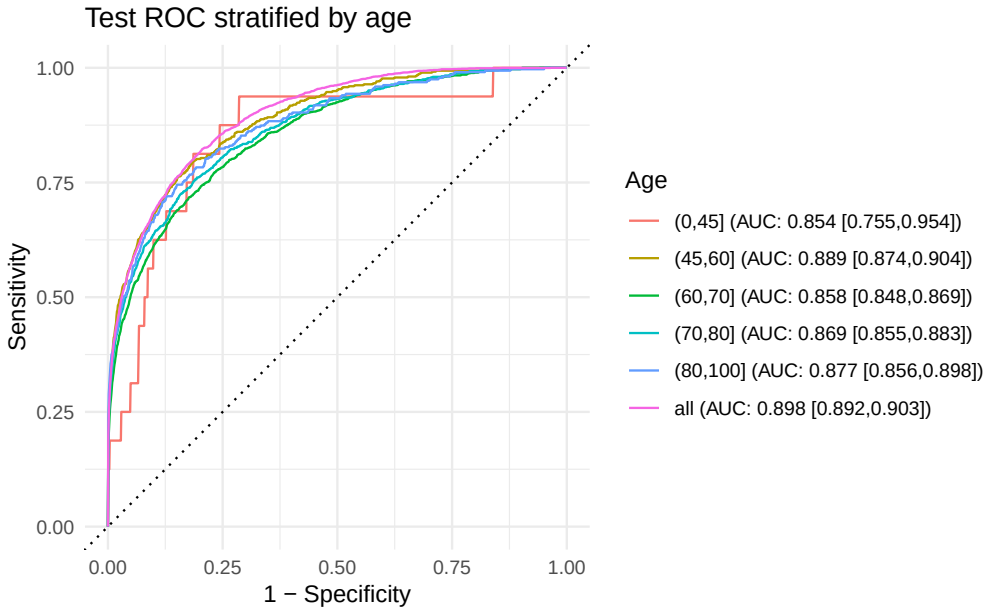


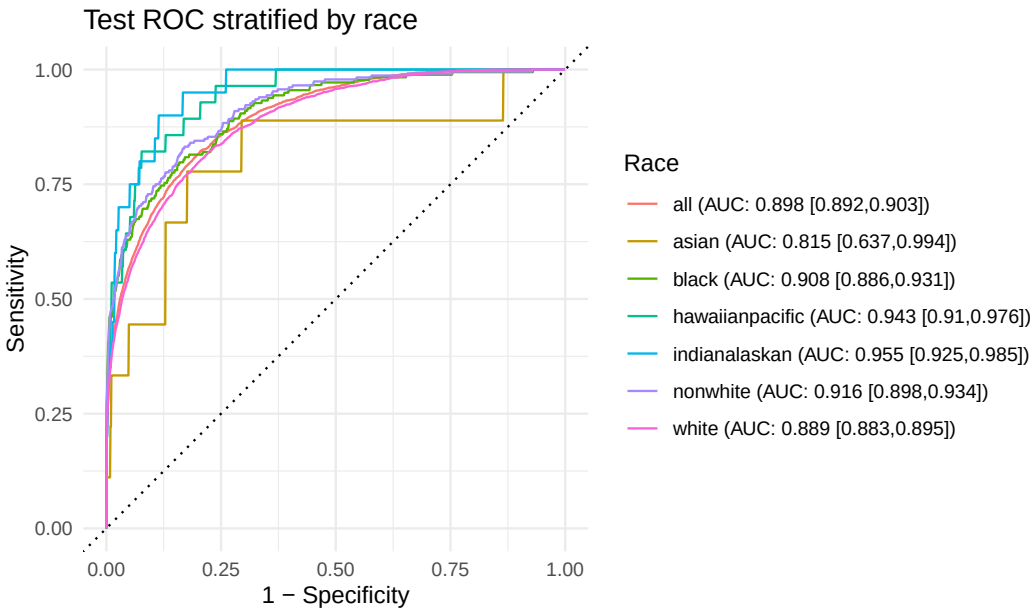
Figure 3: Correlation between group-level Shapley values. Largest absolute correlation in the following table. If two Shapley values are highly correlated, that means the underlying variables contribute essentially in the same way to the prediction and are therefore redundant.

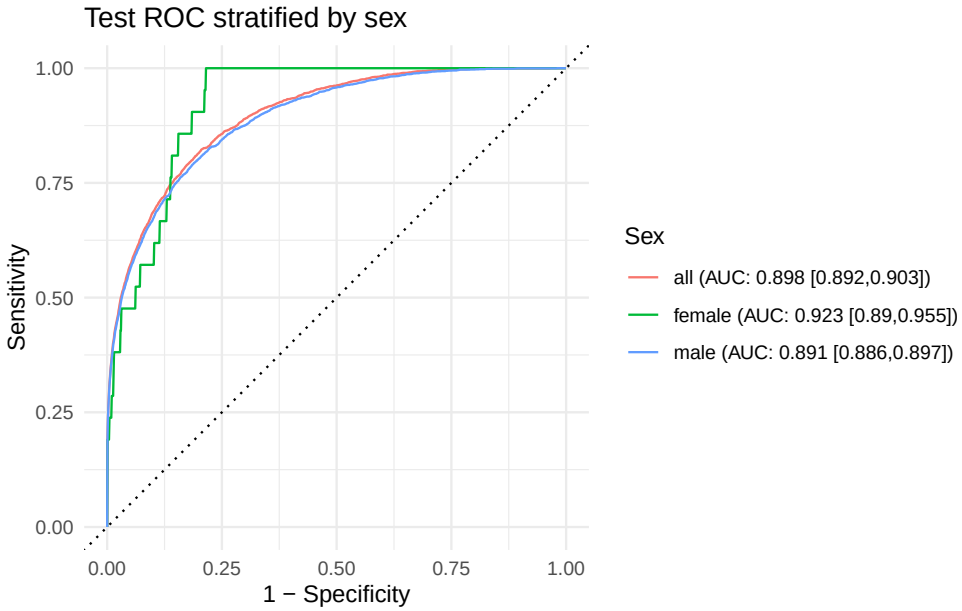
Feature pair		SHAP correlation
diab_c	diab_nc	0.31
bun	creat	0.28
bun	k	0.24
mch	mcv	0.23
bun	na	0.19
gender	creat	-0.18
diab_nc	gluc	0.18
chlor	co2	0.17
gluc	k	0.17
neut	wbc	0.17
gluc	na	0.17
creat	na	0.16
age	diab_nc	0.16
chlor	gluc	0.15
bmi_weight	diab_nc	0.15
alt	ast	0.15
k	na	0.14
gluc	mch	0.14
gender	age	0.14
cd	pvd	0.14

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

Identity groups



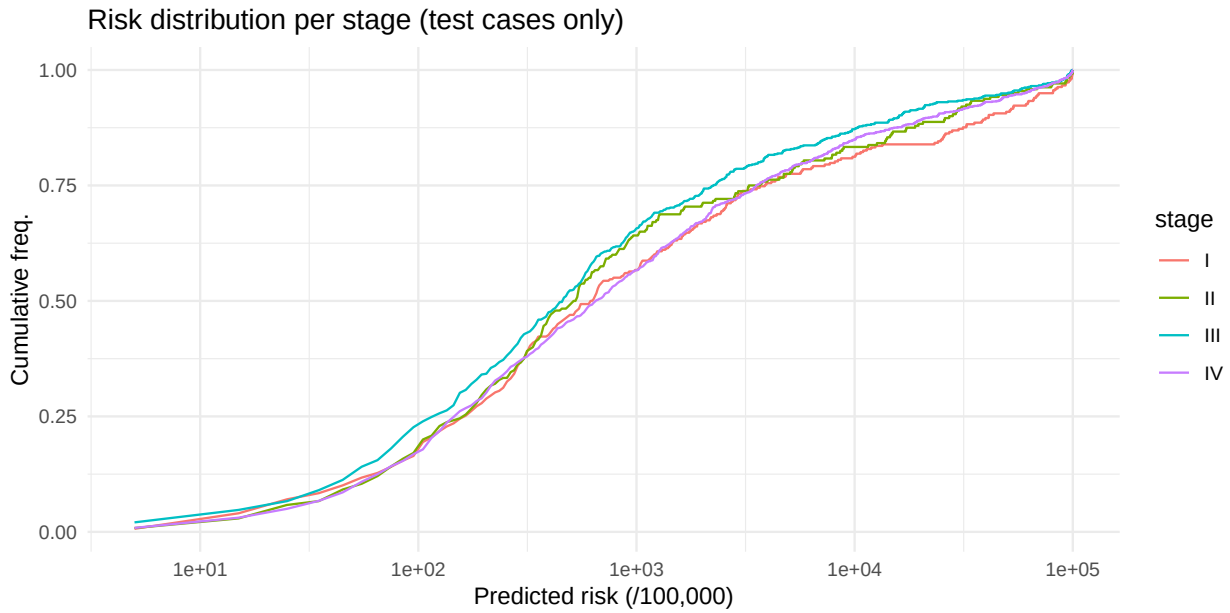



Test ROC stratified by sex
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Sex
all (AUC: 0.898 [0.892,0.903])
female (AUC: 0.923 [0.89,0.955])
male (AUC: 0.891 [0.886,0.897])

Cancer stage

Stage	Test. AUC	Nb. cases
Any	0.773 [0.765,0.78]	2810
I	0.795 [0.773,0.817]	298
II	0.798 [0.774,0.822]	240
III	0.772 [0.757,0.787]	631
IV	0.767 [0.755,0.778]	1041
I+	0.775 [0.767,0.783]	2254
II+	0.772 [0.764,0.78]	1956
III+	0.768 [0.759,0.777]	1716
IV+	0.767 [0.755,0.778]	1041

Table 2: Interestingly, it seems slightly harder to predict late stages?



Years prior

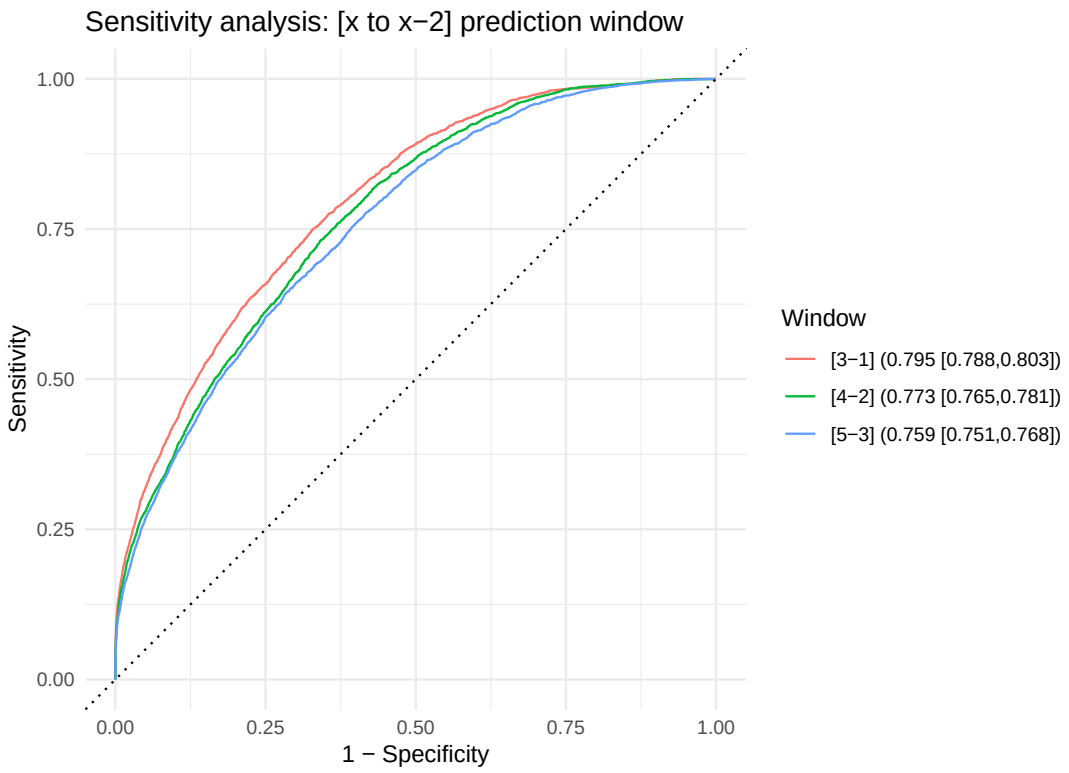


Figure 4: In this experiment, all three test set utilizes the same amount of data (2 years), but we predict farther in the future (1yr, 2yrs or 3yrs). As expected, we do worse and worse as we predict further in time, but only by a small margin. This seem to indicate are predictions are somewhat valid beyond the 1yr window we used.

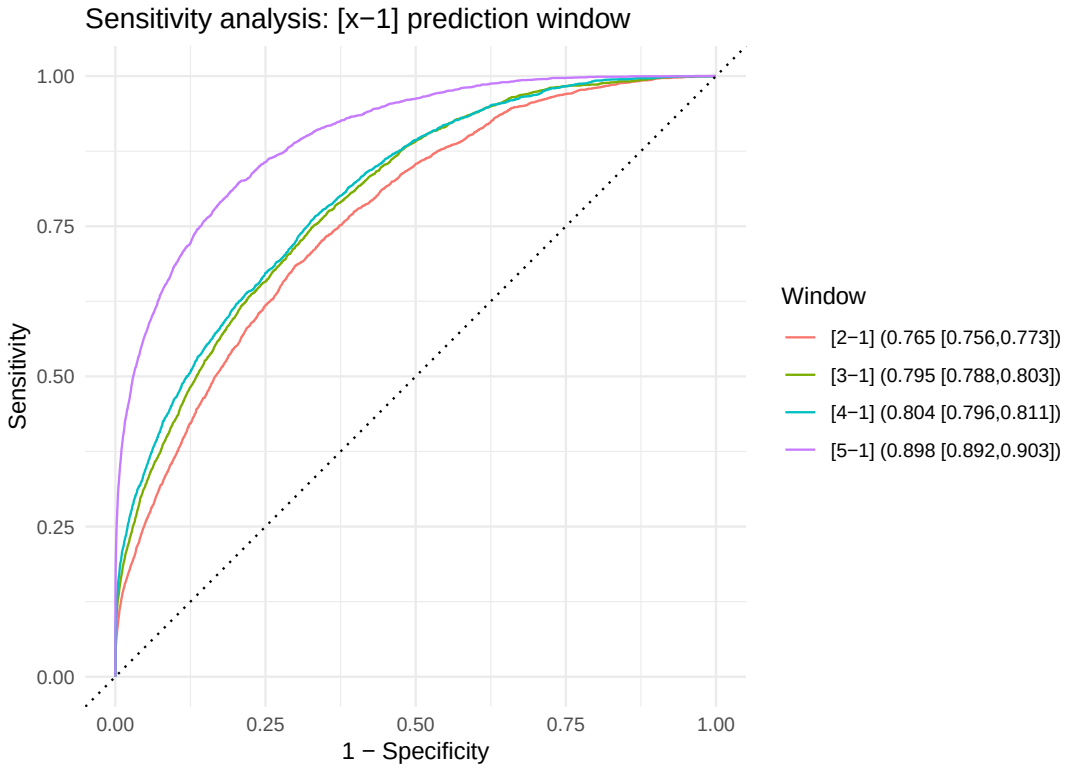


Figure 5: In this experiment, we decrease the amount of data (from 4yrs to 1yr) in the test set, but keep the prediction horizon to 1yr. As expected, performance decreases as the number of years decrease (we have less and less data). However, we notice a large jump in performance from 3yrs to 4yrs. A 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	0.001
gerd	0.232	0.357	0.269	0.005
baso_max	1.017	0.328	1.791	0.033

Table 3: Features with different means between EAC and EGJAC (at 5% level, BH). It appears that race and GERD are less important for EGJAC. We also find that baso could be different between the two.

Prop. Control	Prop. EAC	Prop. EGJAC	pvalue.adj

Table 4: Features with different proportions of non-NAs between EAC and EGJAC (at 5% level, BH). There are None!

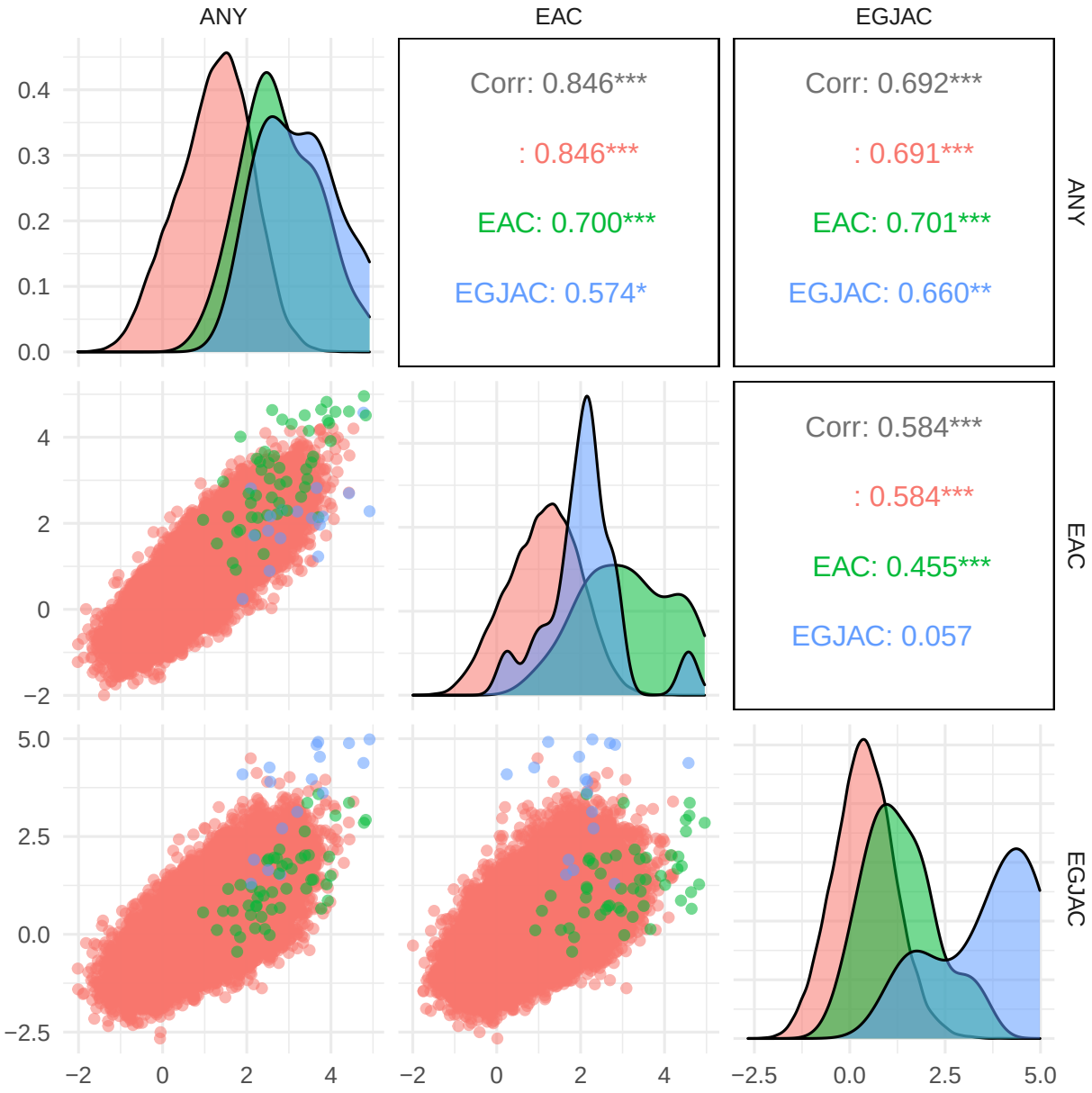


Figure 6: Scatter plots of (log) predicted risk by all three models stratified by case/control status (none, EAC, EGJAC). We have very strong correlation between EAC and ANY, but much less between ANY and EGJAC. There is no correlation between EAC and EGJAC for EGJAC patients. EGJAC assigns very high predicted risk to EGJAC patients compared to what ANY/EAC does to their respective targets.

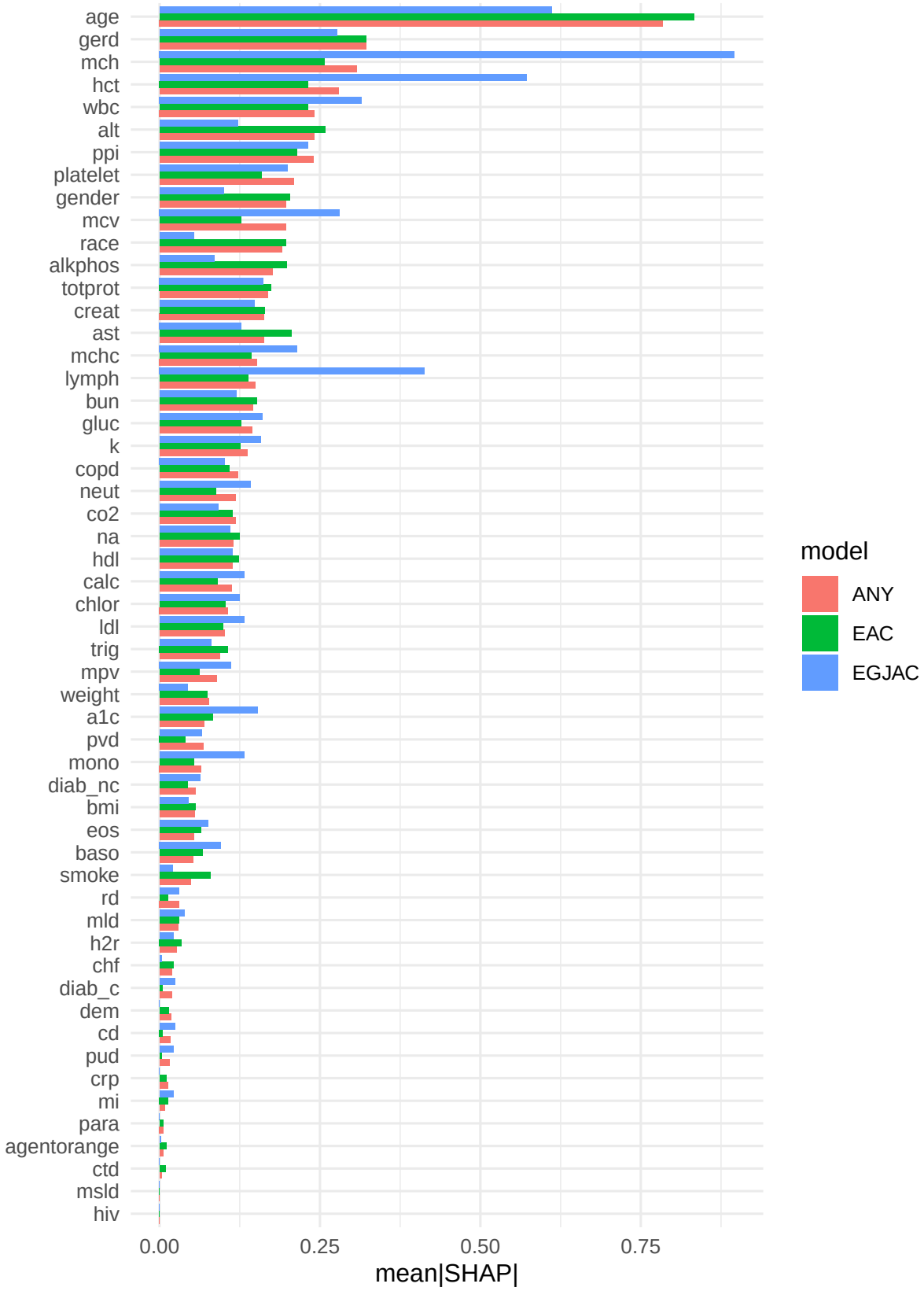


Figure 7: SHAP variable importance wihtin all three models. Notably, we see large differences `age`, `mch`, `hct`, `mcv`, `lymph`, `race` and a few others.

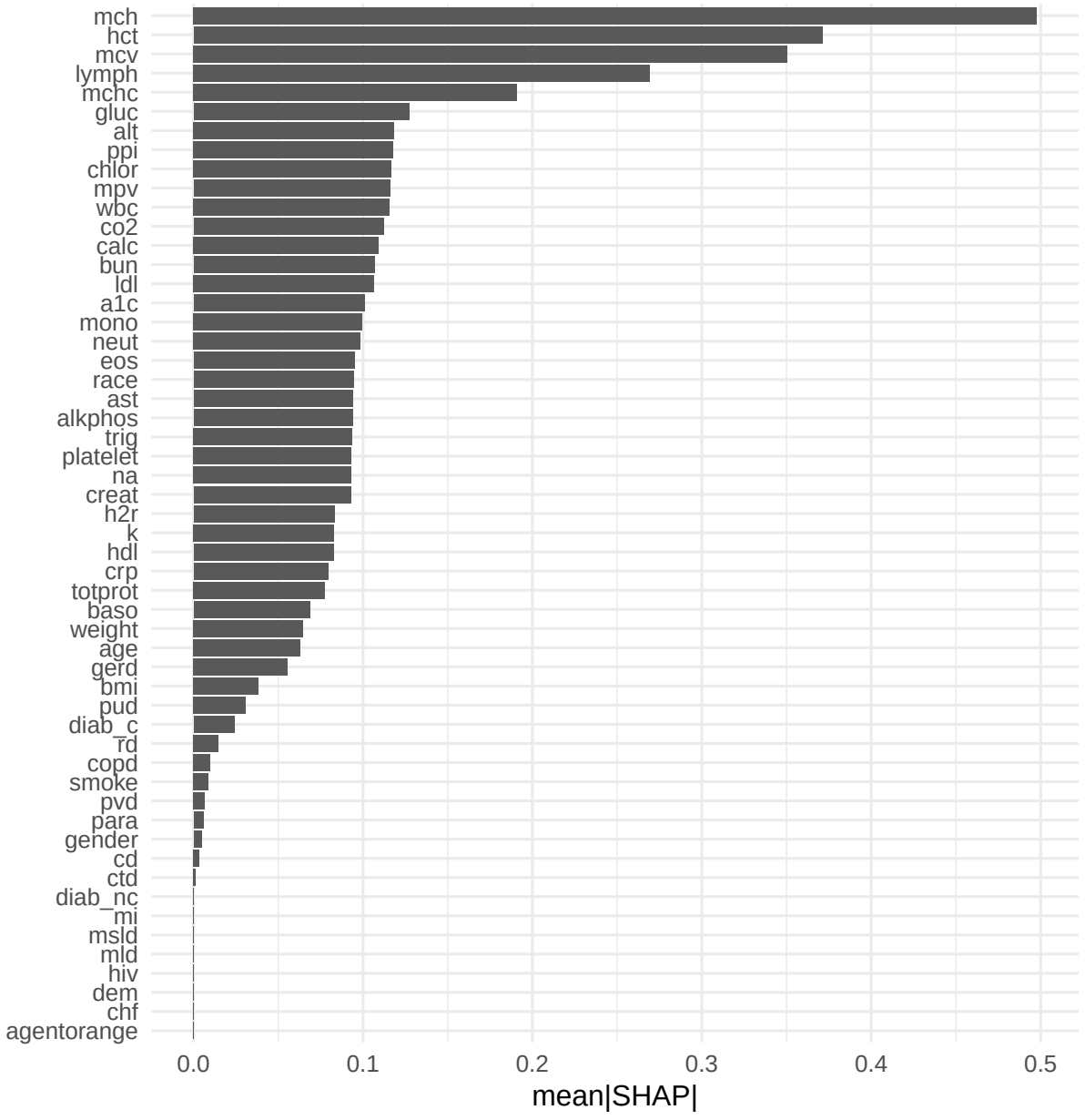


Figure 8: SHAP variable importance for a model differentiating EAC from EGJAC. We again find `mch`, `hct`, `mcv`, `lymph` and `mchc` at the top

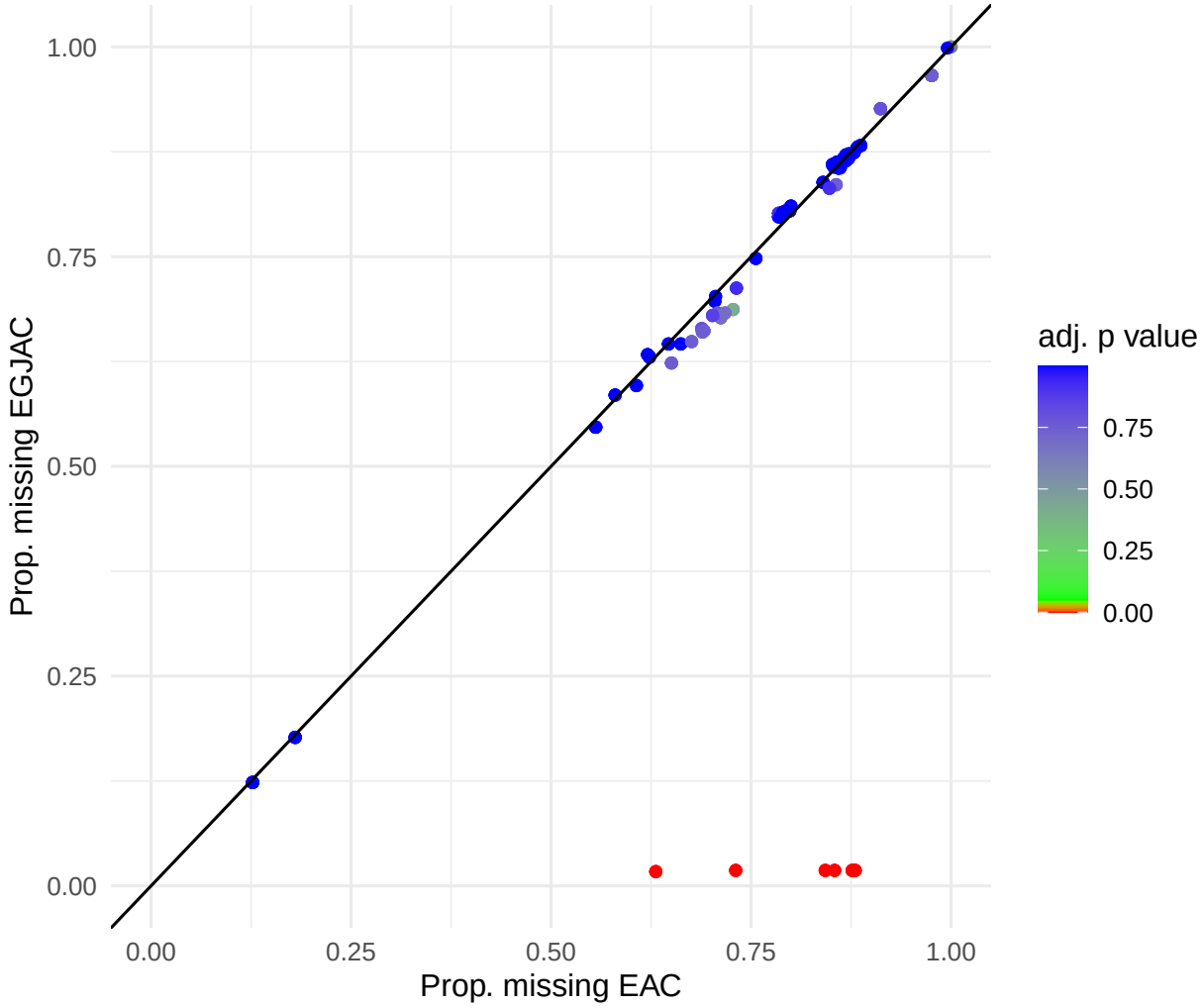
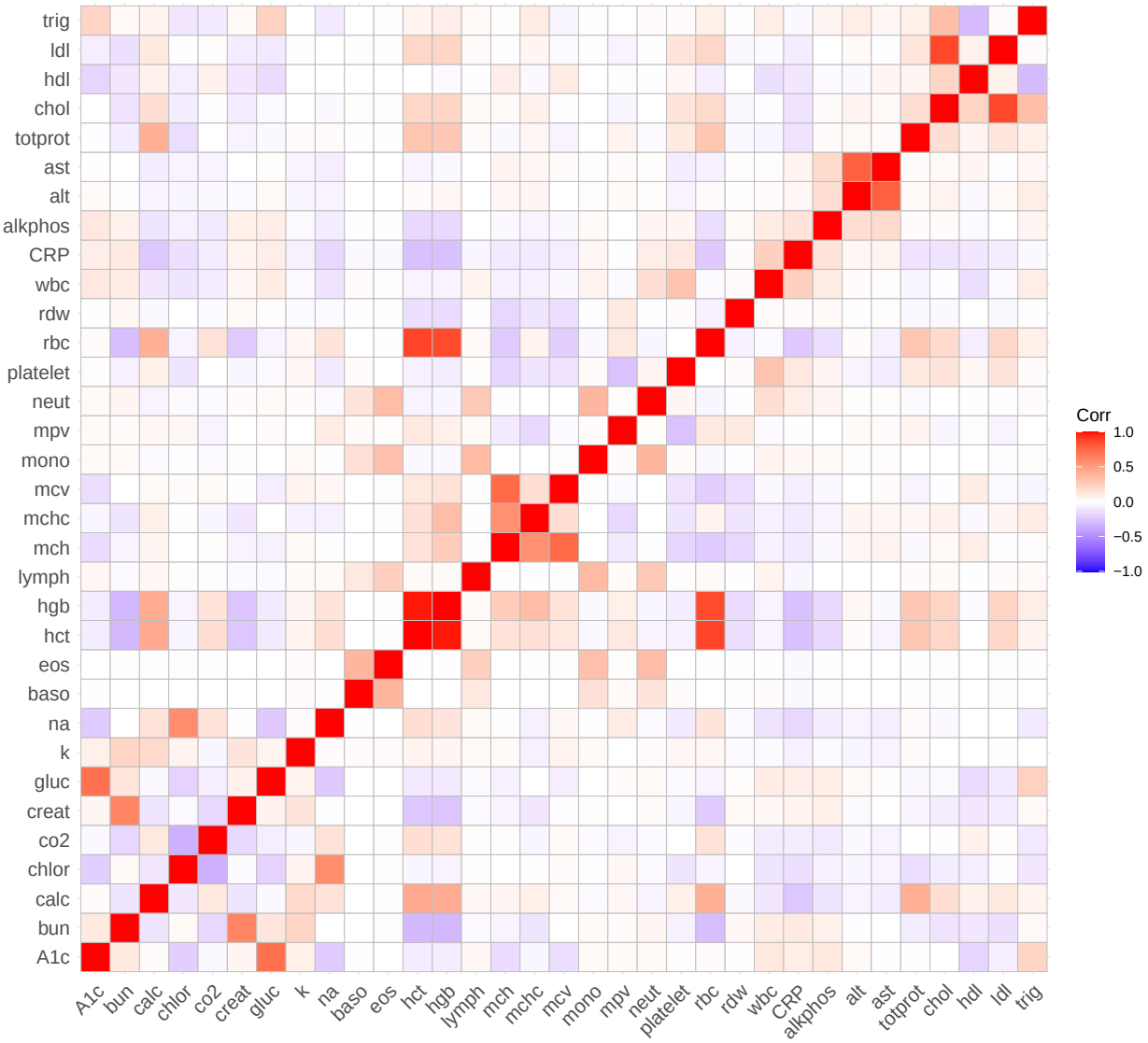


Figure 9: Labels are wrong and should read "non-missing." This shows it really is just there three that have different missing proportions.

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 5: 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