HOSEA Aim I – Report

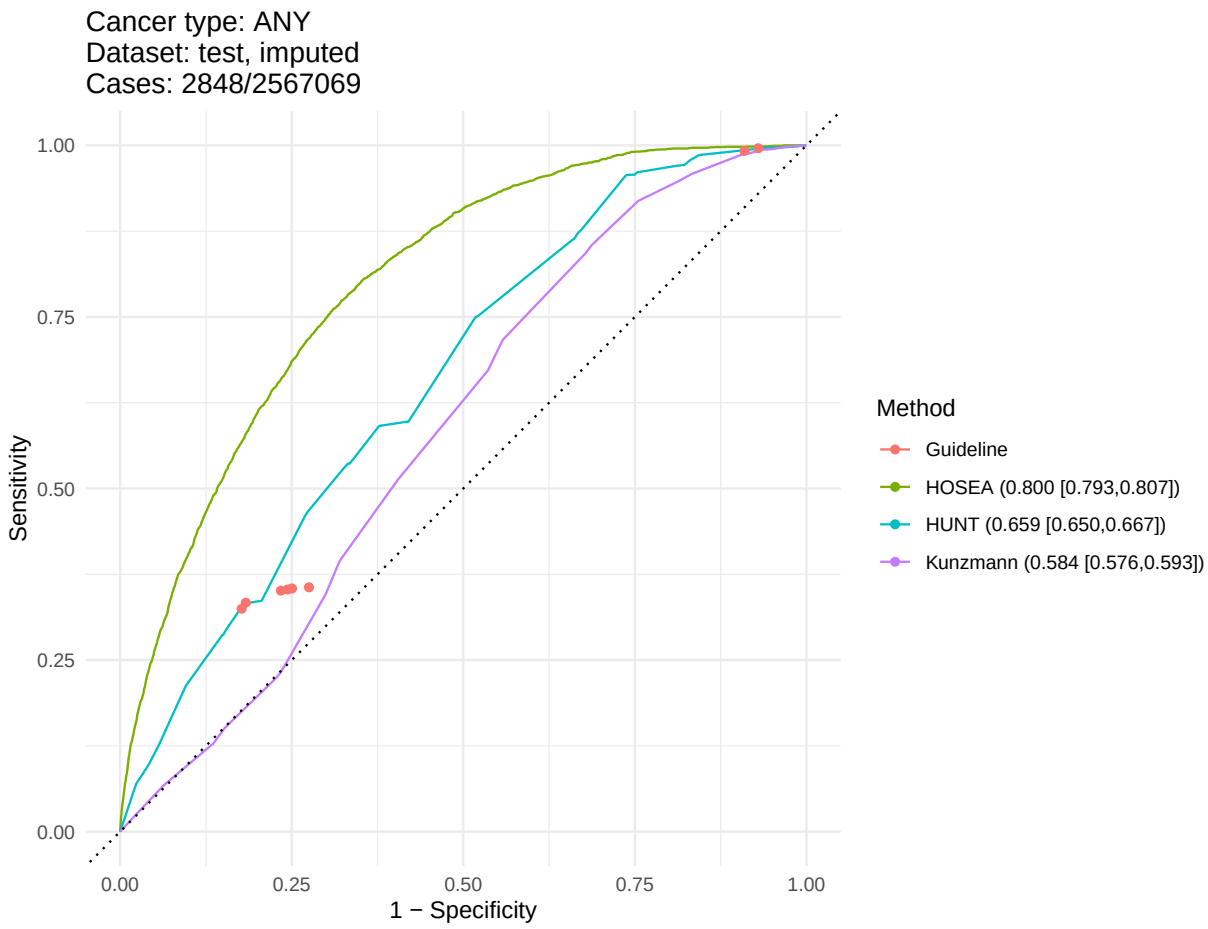
Simon Fontaine

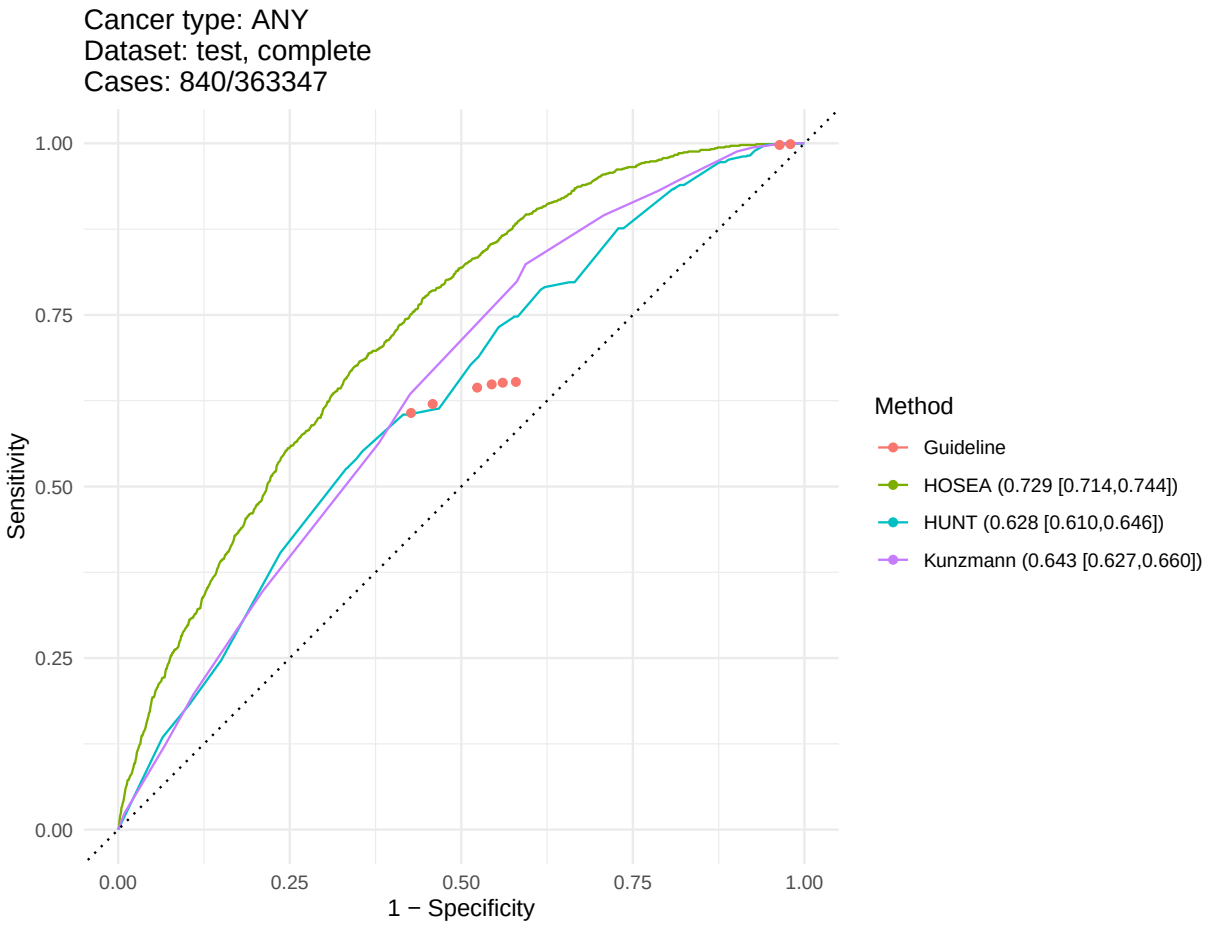
September 13, 2022

1 Current status

1.1 Comparison

1.1.1 ANY


Cancer type: ANY
Dataset: test, imputed
Cases: 2848/2567069
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Method
Guideline
HOSEA (0.800 [0.793,0.807])
HUNT (0.659 [0.650,0.667])
Kunzmann (0.584 [0.576,0.593])


Cancer type: ANY
Dataset: test, complete
Cases: 840/363347
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
Method
Guideline
HOSEA (0.729 [0.714,0.744])
HUNT (0.628 [0.610,0.646])
Kunzmann (0.643 [0.627,0.660])

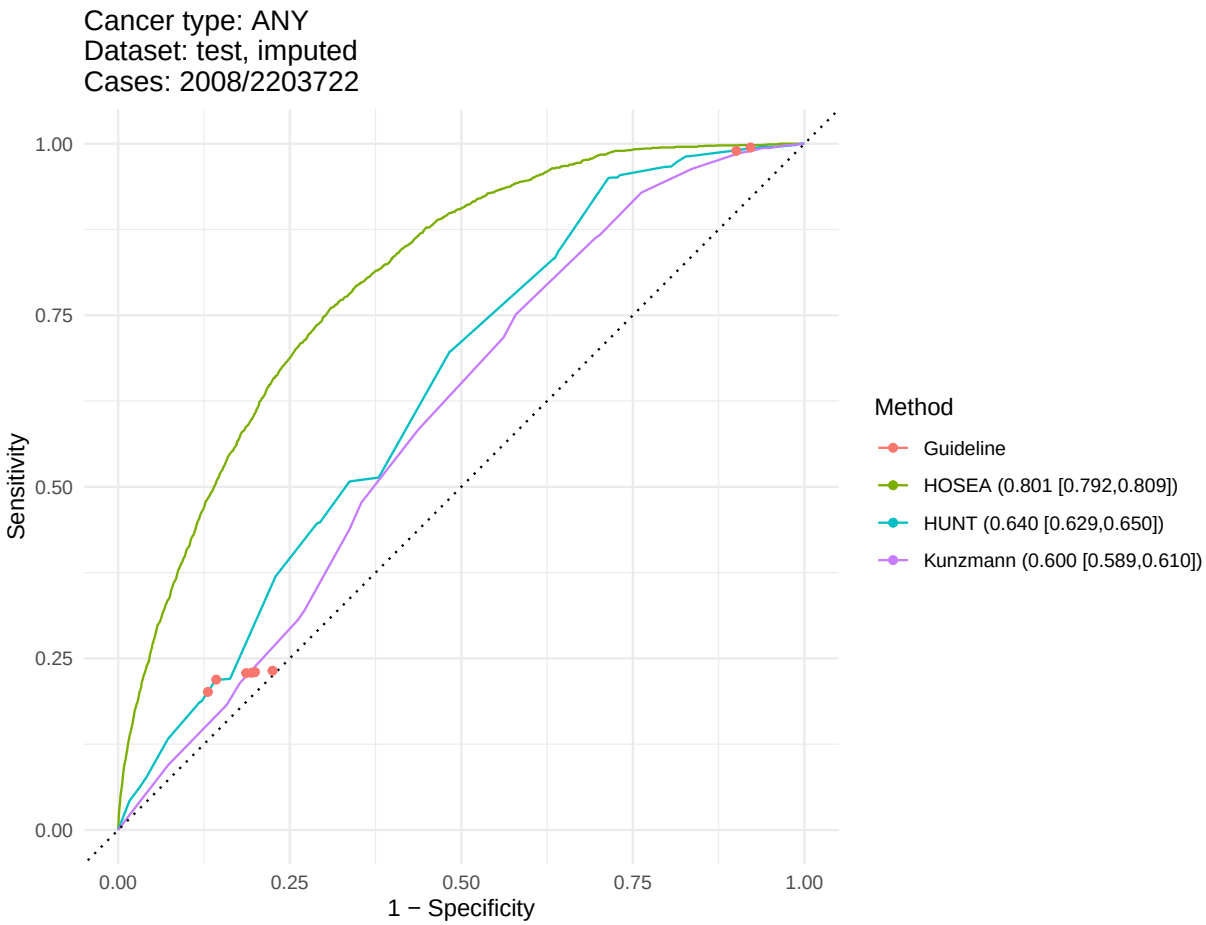
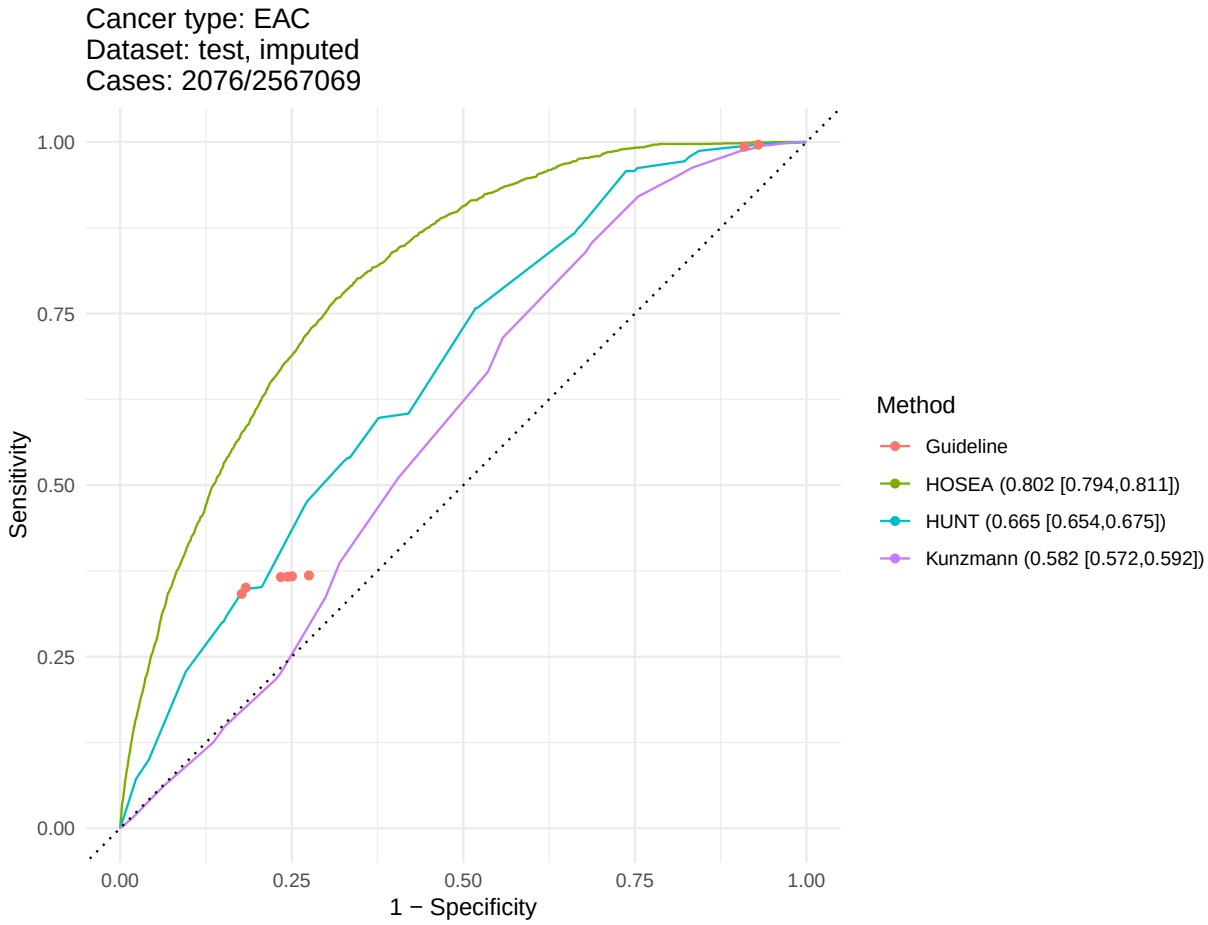
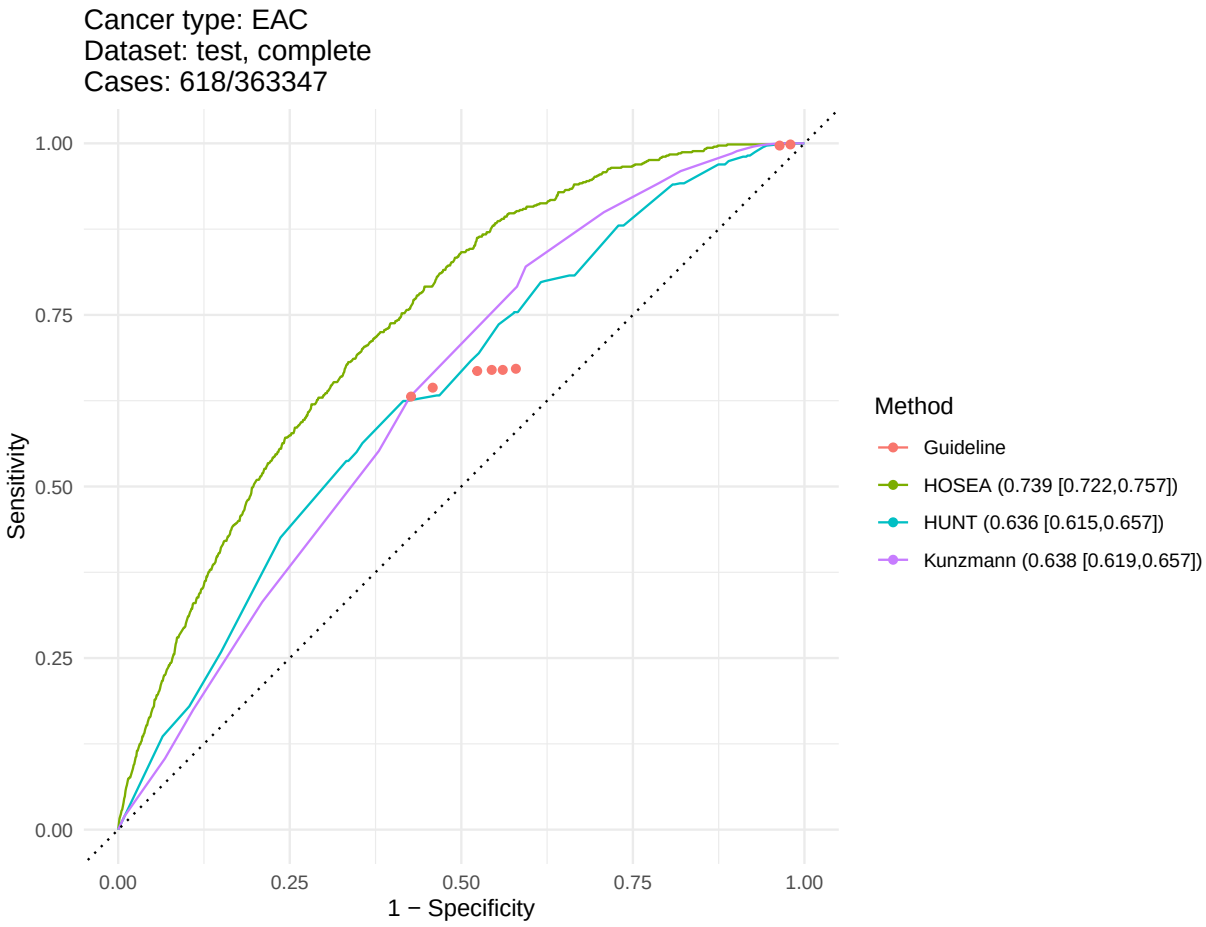


Figure 1: This should read "incomplete"...

1.1.2 EAC


Cancer type: EAC
Dataset: test, imputed
Cases: 2076/2567069
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Method
Guideline
HOSEA (0.802 [0.794,0.811])
HUNT (0.665 [0.654,0.675])
Kunzmann (0.582 [0.572,0.592])


Cancer type: EAC
Dataset: test, complete
Cases: 618/363347
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
Method
Guideline
HOSEA (0.739 [0.722,0.757])
HUNT (0.636 [0.615,0.657])
Kunzmann (0.638 [0.619,0.657])

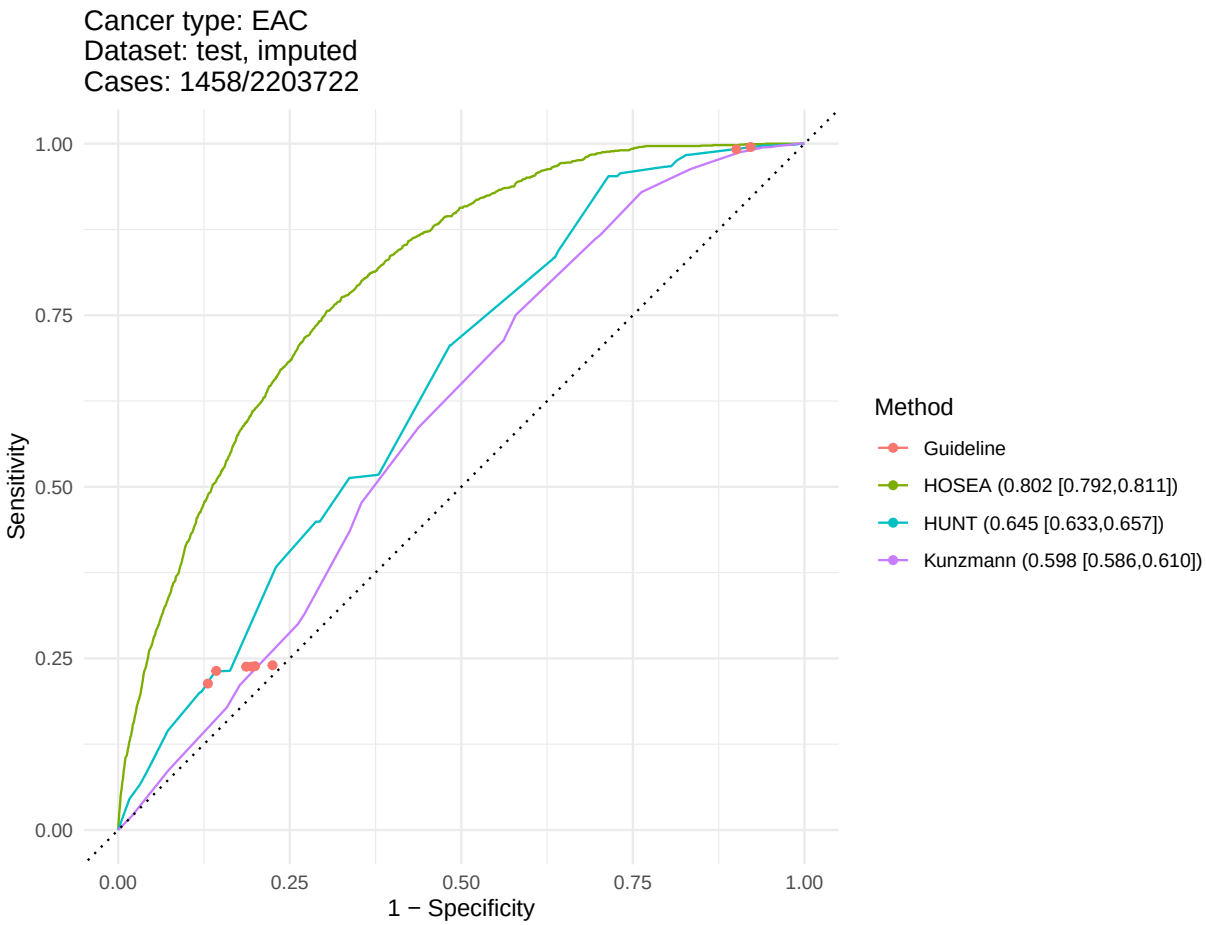
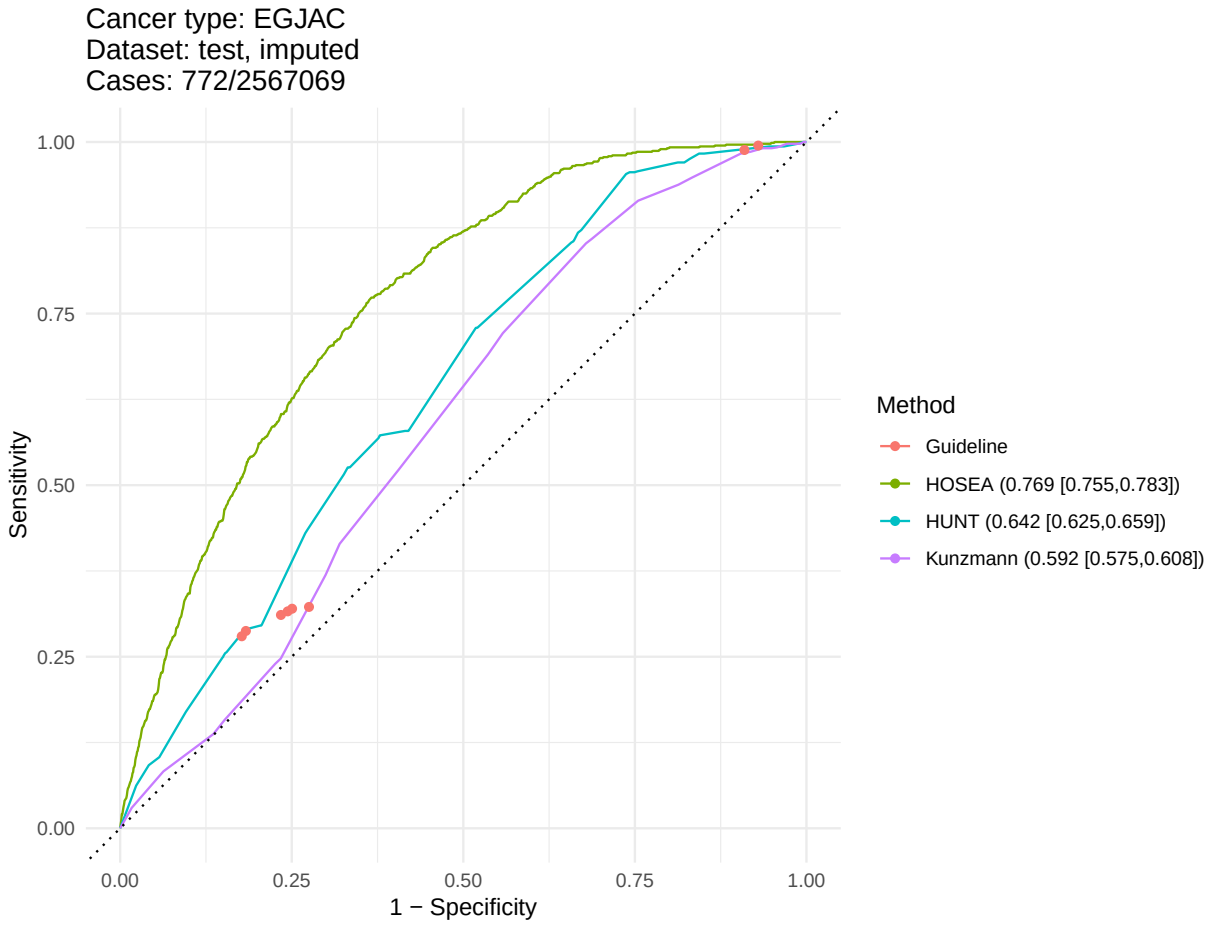
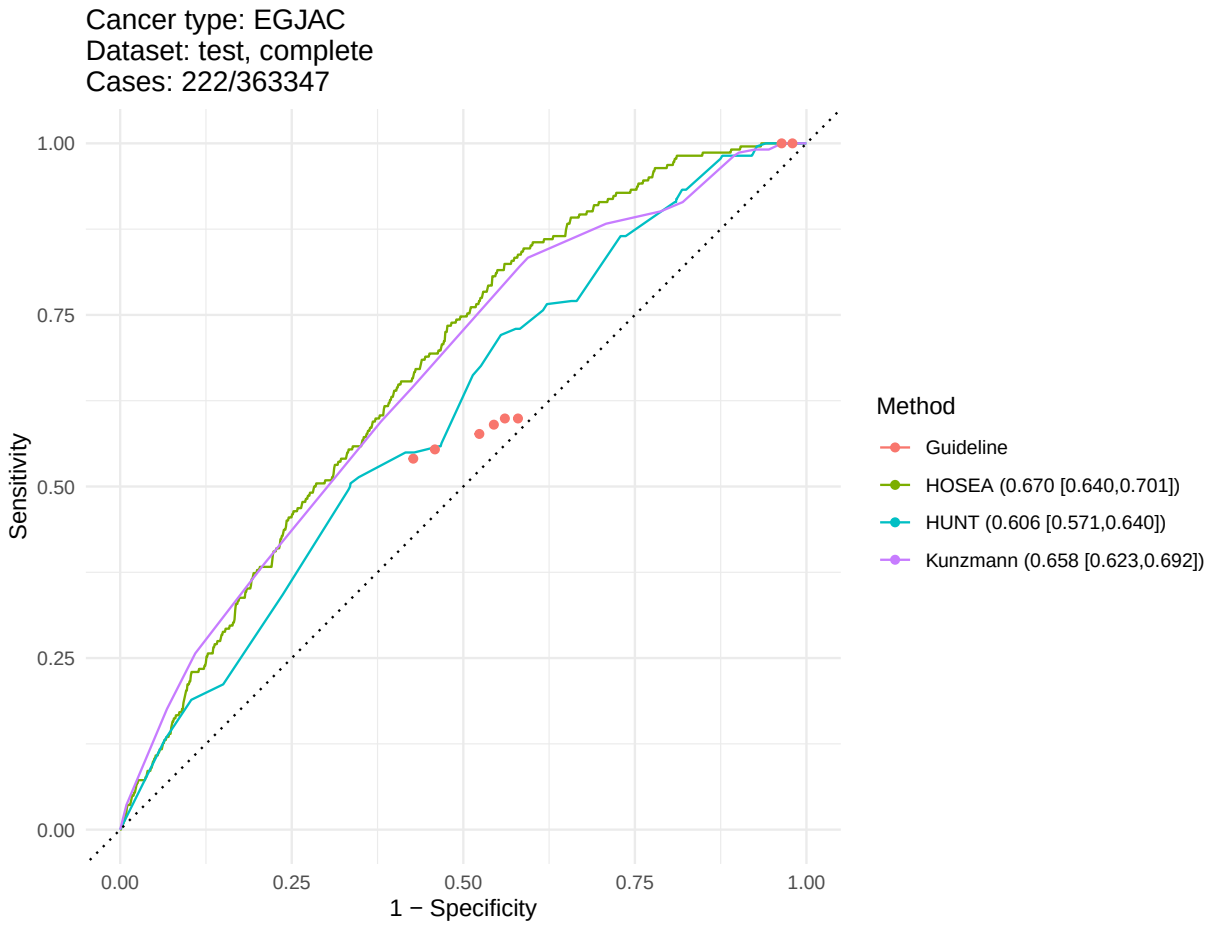


Figure 2: This should read "incomplete"...

1.1.3 EGJAC


Cancer type: EGJAC
Dataset: test, imputed
Cases: 772/2567069
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Method
Guideline
HOSEA (0.769 [0.755,0.783])
HUNT (0.642 [0.625,0.659])
Kunzmann (0.592 [0.575,0.608])


Cancer type: EGJAC
Dataset: test, complete
Cases: 222/363347
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Method
Guideline
HOSEA (0.670 [0.640,0.701])
HUNT (0.606 [0.571,0.640])
Kunzmann (0.658 [0.623,0.692])

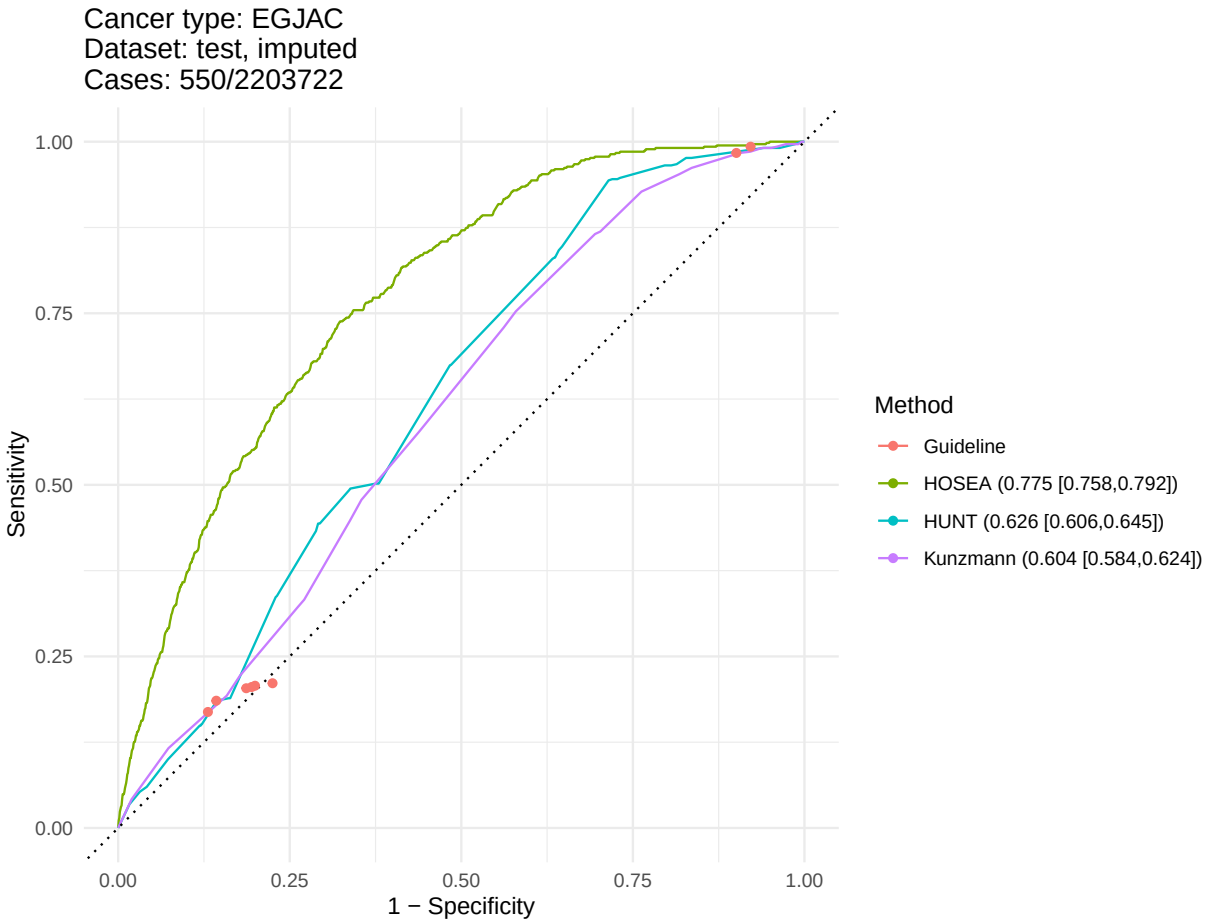
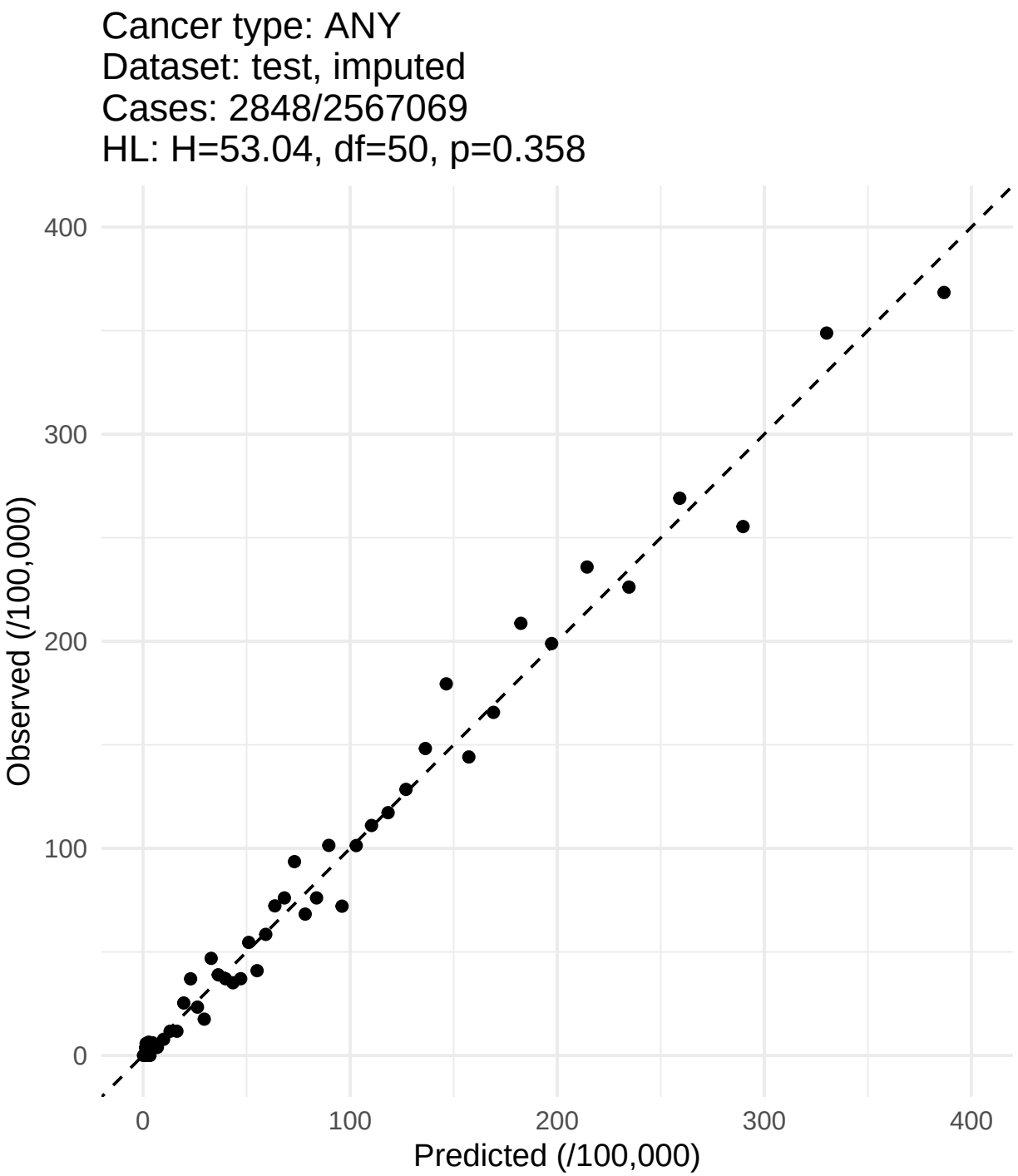
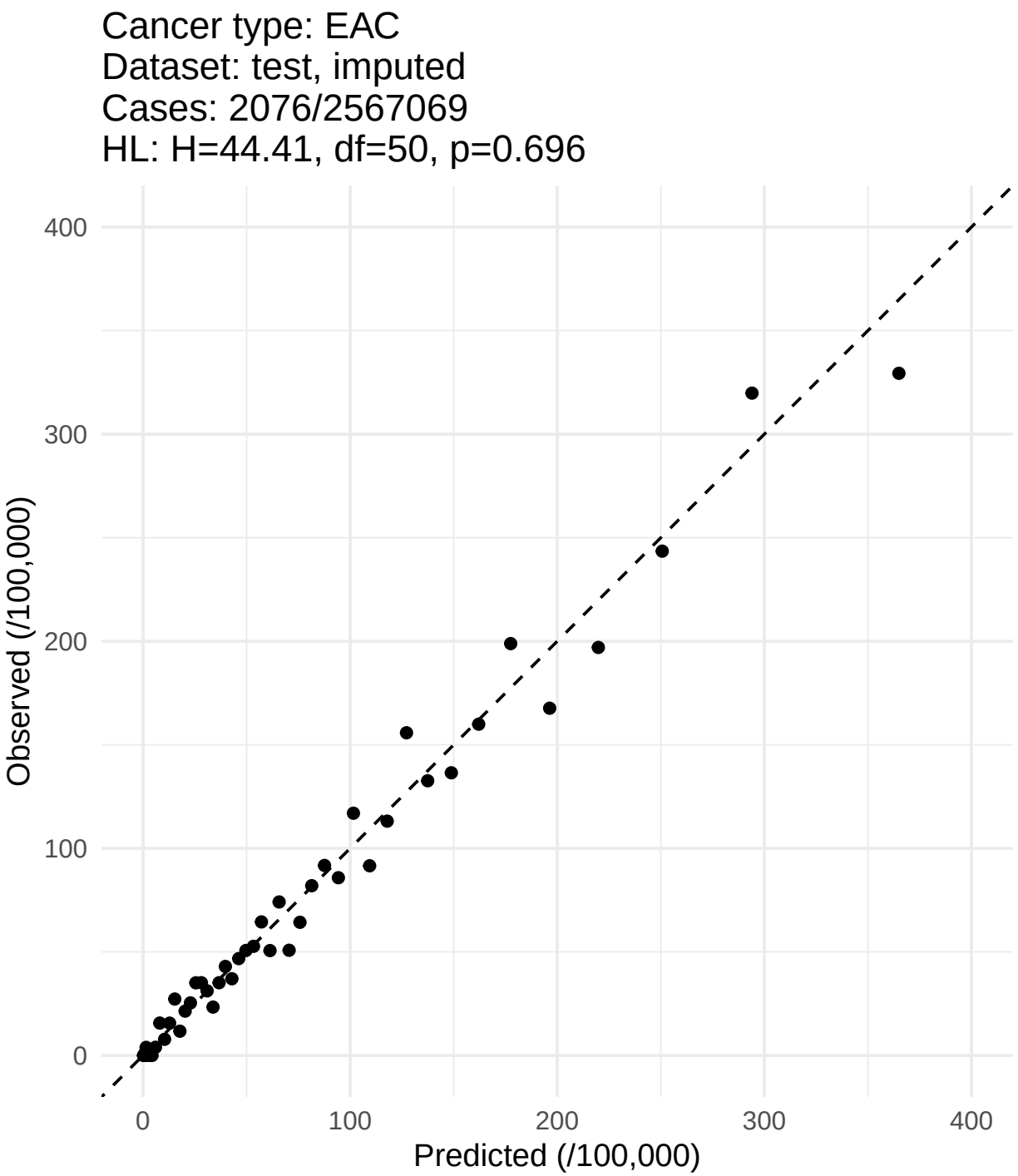
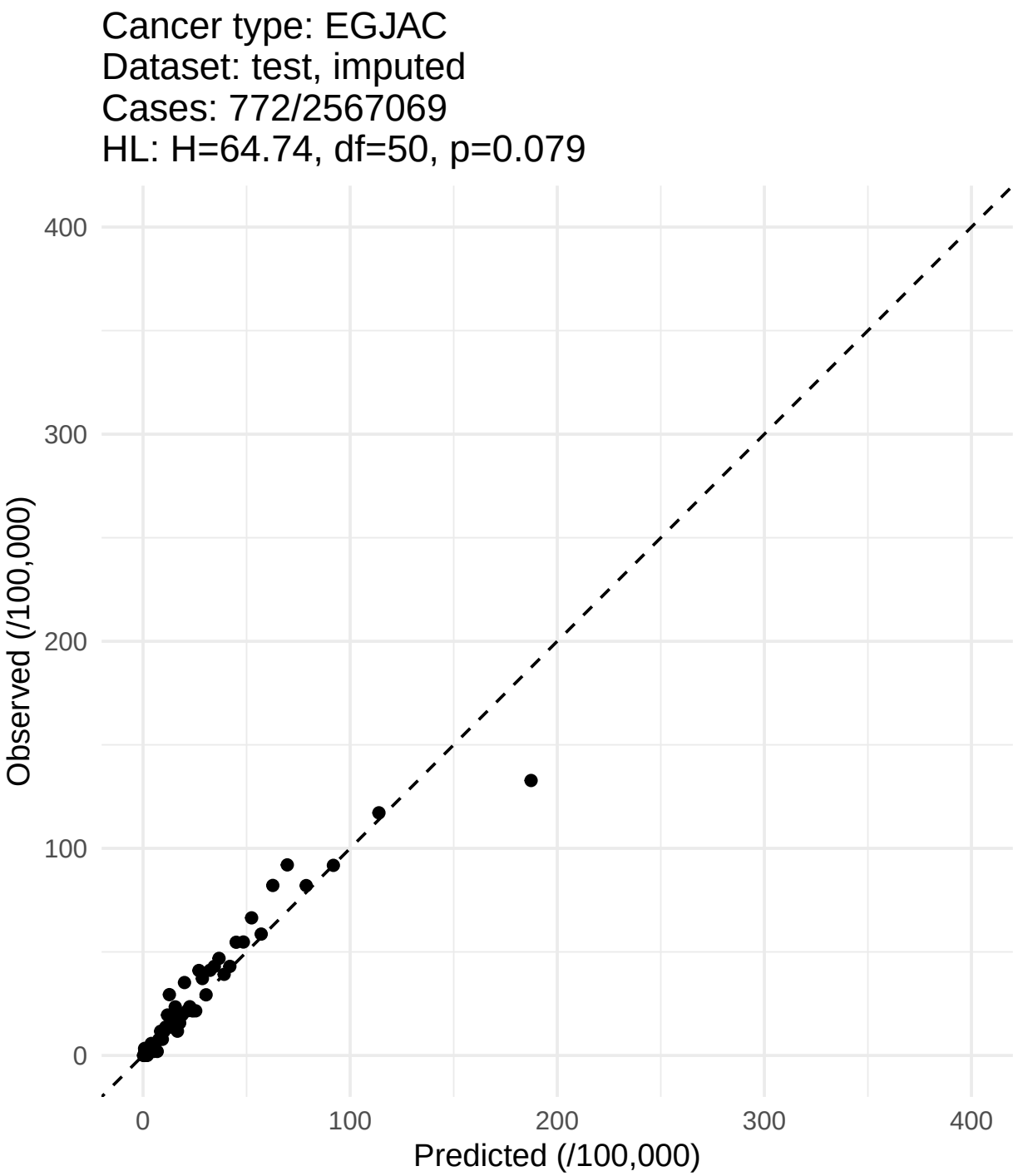


Figure 3: This should read "incomplete"...

1.2 Calibration


Cancer type: ANY
Dataset: test, imputed
Cases: 2848/2567069
HL: H=53.04, df=50, p=0.358
Observed (/100,000)
Predicted (/100,000)
0
100
200
300
400


Cancer type: EAC
Dataset: test, imputed
Cases: 2076/2567069
HL: H=44.41, df=50, p=0.696
Observed (/100,000)
Predicted (/100,000)
0
100
200
300
400


Cancer type: EGJAC
Dataset: test, imputed
Cases: 772/2567069
HL: H=64.74, df=50, p=0.079
Observed (/100,000)
Predicted (/100,000)
0
100
200
300
400

1.3 Threshold

ANY

Threshold	TPR	PPV	DetPrevalence
0	100.00	0.11	100.00
5	99.54	0.13	82.38
10	99.40	0.14	78.88
15	99.09	0.15	75.76
20	98.60	0.15	72.73
25	97.68	0.16	69.71
30	97.16	0.16	66.67
35	96.10	0.17	63.70
40	95.19	0.17	60.83
45	94.24	0.18	58.08
50	93.22	0.19	55.47
55	92.17	0.19	52.98
60	91.12	0.20	50.60
65	89.68	0.21	48.36
70	88.24	0.21	46.22
75	86.73	0.22	44.21
80	85.32	0.22	42.32
85	84.23	0.23	40.53
90	82.94	0.24	38.87
95	81.64	0.24	37.30
100	80.65	0.25	35.81
105	79.39	0.26	34.39
110	77.98	0.26	33.06
115	76.65	0.27	31.78
120	75.42	0.27	30.57
125	73.98	0.28	29.41
130	72.75	0.29	28.30
135	71.52	0.29	27.24
140	70.19	0.30	26.23
145	68.71	0.30	25.24
150	67.10	0.31	24.30
155	65.77	0.31	23.39
160	64.64	0.32	22.51
165	63.38	0.32	21.66
170	62.08	0.33	20.85
175	61.17	0.34	20.07
180	59.59	0.34	19.33
185	58.39	0.35	18.61
190	57.09	0.35	17.93
195	56.04	0.36	17.27
200	54.95	0.37	16.64
220	50.49	0.39	14.38
240	46.52	0.41	12.50
260	42.84	0.44	10.92
280	39.54	0.46	9.56
300	37.18	0.49	8.40
320	34.02	0.51	7.41
340	31.07	0.53	6.56
360	29.14	0.56	5.81
380	26.83	0.58	5.16
400	24.72	0.60	4.59
420	23.24	0.63	4.09
440	21.38	0.65	3.65
460	19.73	0.67	3.26
480	18.64	0.71	2.92
500	17.38	0.74	2.62
550	14.54	0.81	2.00
600	12.46	0.90	1.53
650	10.29	0.96	1.19
700	8.25	0.98	0.93
750	7.16	1.08	0.73
800	5.76	1.09	0.58
850	4.88	1.16	0.47
900	4.39	1.30	0.37
950	3.79	1.38	0.31
1000	3.41	1.52	0.25

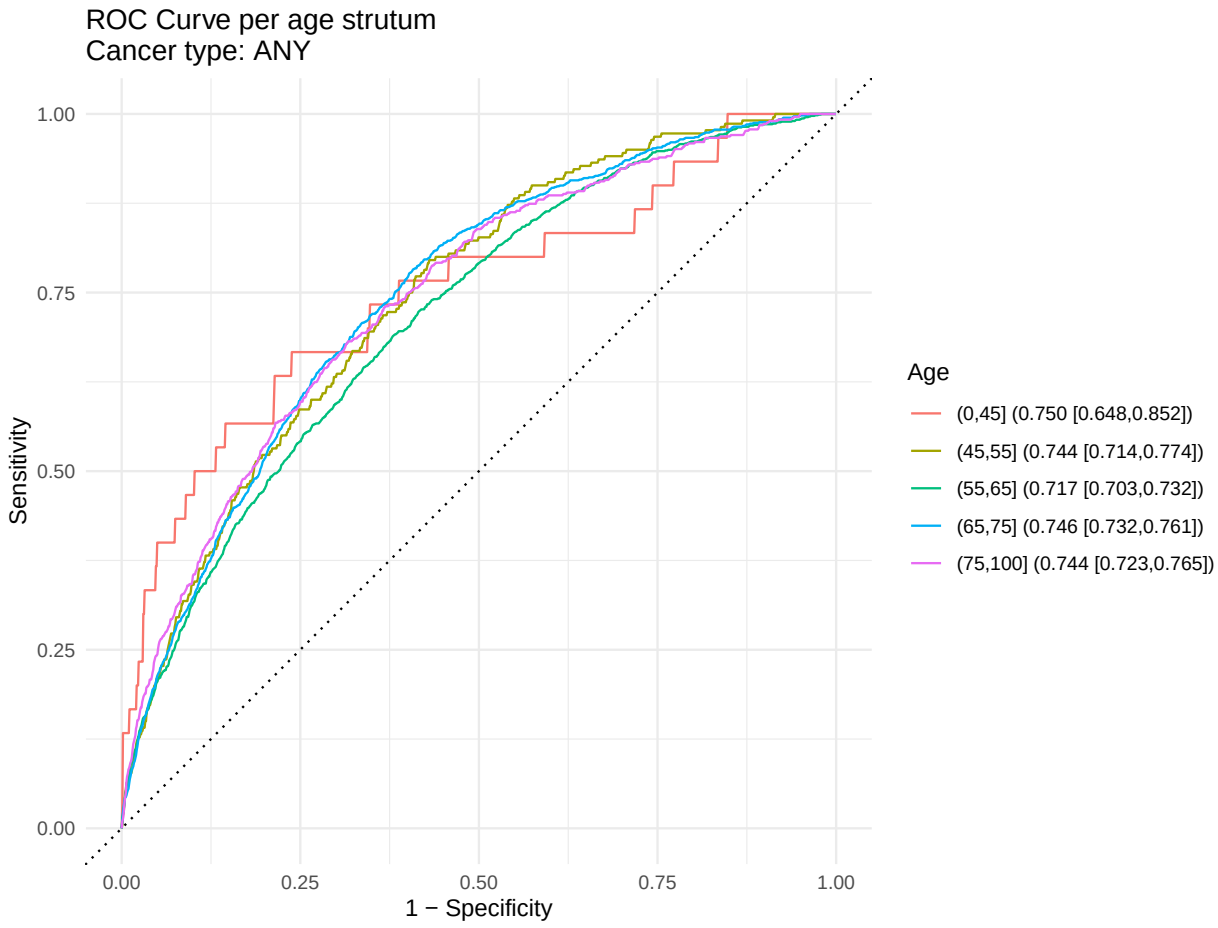
EAC

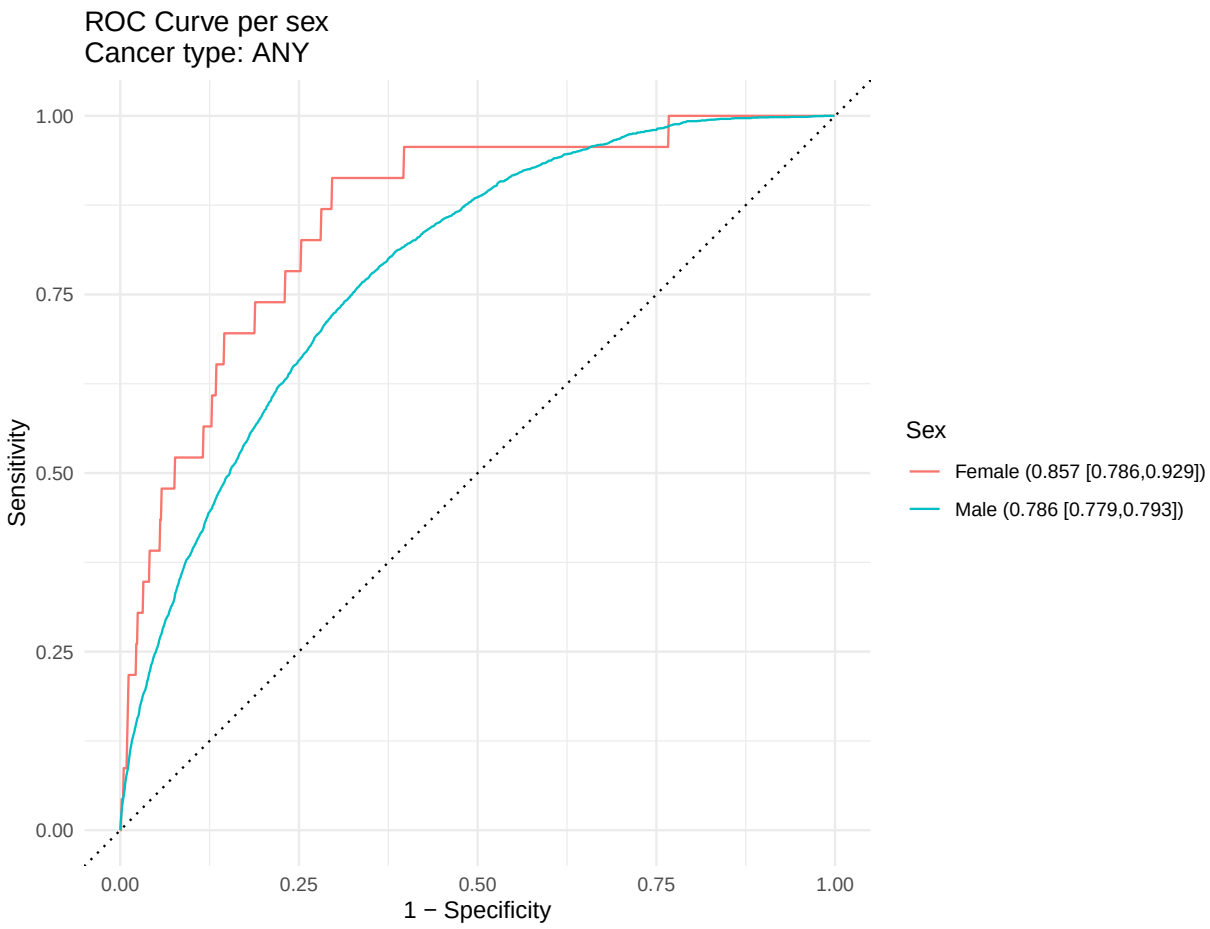
Threshold	TPR	PPV	DetPrevalence
0	100.00	0.08	100.00
5	99.71	0.10	79.96
10	99.18	0.11	75.25
15	98.55	0.11	71.17
20	97.59	0.12	67.18
25	96.15	0.12	63.30
30	94.75	0.13	59.59
35	93.55	0.13	56.09
40	92.05	0.14	52.76
45	90.37	0.15	49.65
50	88.87	0.15	46.76
55	86.95	0.16	44.05
60	84.97	0.17	41.56
65	83.33	0.17	39.26
70	81.79	0.18	37.16
75	80.39	0.18	35.24
80	78.85	0.19	33.45
85	77.31	0.20	31.80
90	75.53	0.20	30.23
95	73.70	0.21	28.77
100	72.11	0.21	27.39
105	70.42	0.22	26.08
110	68.64	0.22	24.82
115	67.29	0.23	23.63
120	65.70	0.24	22.50
125	64.07	0.24	21.43
130	62.19	0.25	20.41
135	60.16	0.25	19.43
140	58.67	0.26	18.50
145	57.42	0.26	17.62
150	56.02	0.27	16.78
155	54.62	0.28	16.00
160	53.42	0.28	15.27
165	51.73	0.29	14.57
170	50.43	0.29	13.90
175	49.37	0.30	13.27
180	47.54	0.30	12.69
185	45.86	0.31	12.13
190	44.89	0.31	11.61
195	43.93	0.32	11.10
200	42.82	0.33	10.62
220	38.97	0.35	8.96
240	35.50	0.38	7.59
260	32.23	0.40	6.47
280	28.37	0.42	5.52
300	25.72	0.44	4.74
320	23.03	0.46	4.08
340	20.86	0.48	3.52
360	19.08	0.51	3.04
380	17.15	0.53	2.64
400	15.90	0.56	2.29
420	14.60	0.59	1.99
440	12.96	0.60	1.73
460	11.85	0.63	1.52
480	10.69	0.65	1.33
500	9.78	0.68	1.16
550	7.66	0.73	0.85
600	5.92	0.77	0.62
650	4.53	0.79	0.47
700	3.85	0.90	0.35
750	3.32	1.01	0.27
800	2.46	0.98	0.20
850	1.83	0.95	0.16
900	1.49	1.01	0.12
950	1.25	1.08	0.09
1000	1.16	1.27	0.07

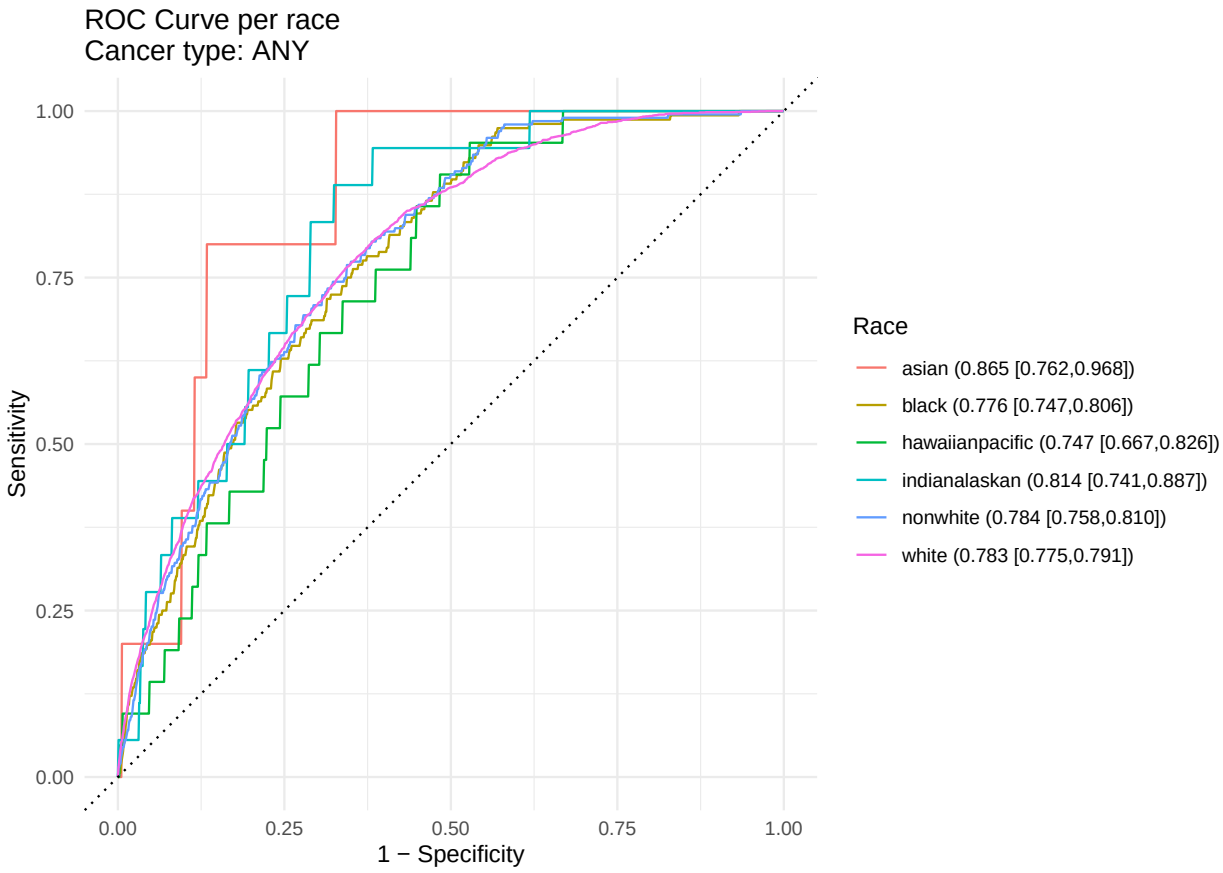
EGJAC

Threshold	TPR	PPV	DetPrevalence
0	100.00	0.03	100.00
5	98.58	0.04	76.82
10	96.11	0.04	65.06
15	89.25	0.05	53.93
20	83.68	0.06	44.85
25	77.72	0.06	37.42
30	70.85	0.07	31.35
35	64.64	0.07	26.37
40	58.55	0.08	22.29
45	53.89	0.09	18.91
50	48.32	0.09	16.12
55	43.26	0.09	13.78
60	38.99	0.10	11.84
65	34.97	0.10	10.23
70	30.83	0.10	8.87
75	27.85	0.11	7.71
80	25.26	0.11	6.73
85	22.54	0.11	5.90
90	19.56	0.11	5.19
95	18.01	0.12	4.58
100	16.84	0.12	4.06
105	15.41	0.13	3.60
110	14.51	0.14	3.20
115	12.82	0.14	2.86
120	11.40	0.13	2.55
125	10.23	0.13	2.28
130	8.94	0.13	2.05
135	8.16	0.13	1.84
140	7.38	0.13	1.66
145	6.87	0.14	1.50
150	6.35	0.14	1.36
155	5.96	0.15	1.23
160	5.83	0.16	1.12
165	5.05	0.15	1.02
170	4.79	0.16	0.92
175	4.27	0.15	0.84
180	4.27	0.17	0.77
185	4.15	0.18	0.70
190	4.02	0.19	0.64
195	3.63	0.19	0.59
200	3.37	0.19	0.54
220	2.46	0.19	0.39
240	2.07	0.22	0.28
260	1.55	0.22	0.21
280	1.17	0.22	0.16
300	0.78	0.20	0.12
320	0.78	0.26	0.09
340	0.39	0.17	0.07
360	0.39	0.21	0.05
380	0.39	0.27	0.04
400	0.39	0.34	0.03
420	0.13	0.14	0.03
440	0.13	0.18	0.02
460	0.13	0.22	0.02
480	0.13	0.28	0.01
500	0.13	0.34	0.01
550	0.13	0.53	0.01
600	0.00	0.00	0.00
650	0.00	0.00	0.00
700	0.00	0.00	0.00
750	0.00	0.00	0.00
800	0.00	0.00	0.00
850	0.00	0.00	0.00
900	0.00	0.00	0.00
950	0.00	0.00	0.00
1000	0.00	0.00	0.00

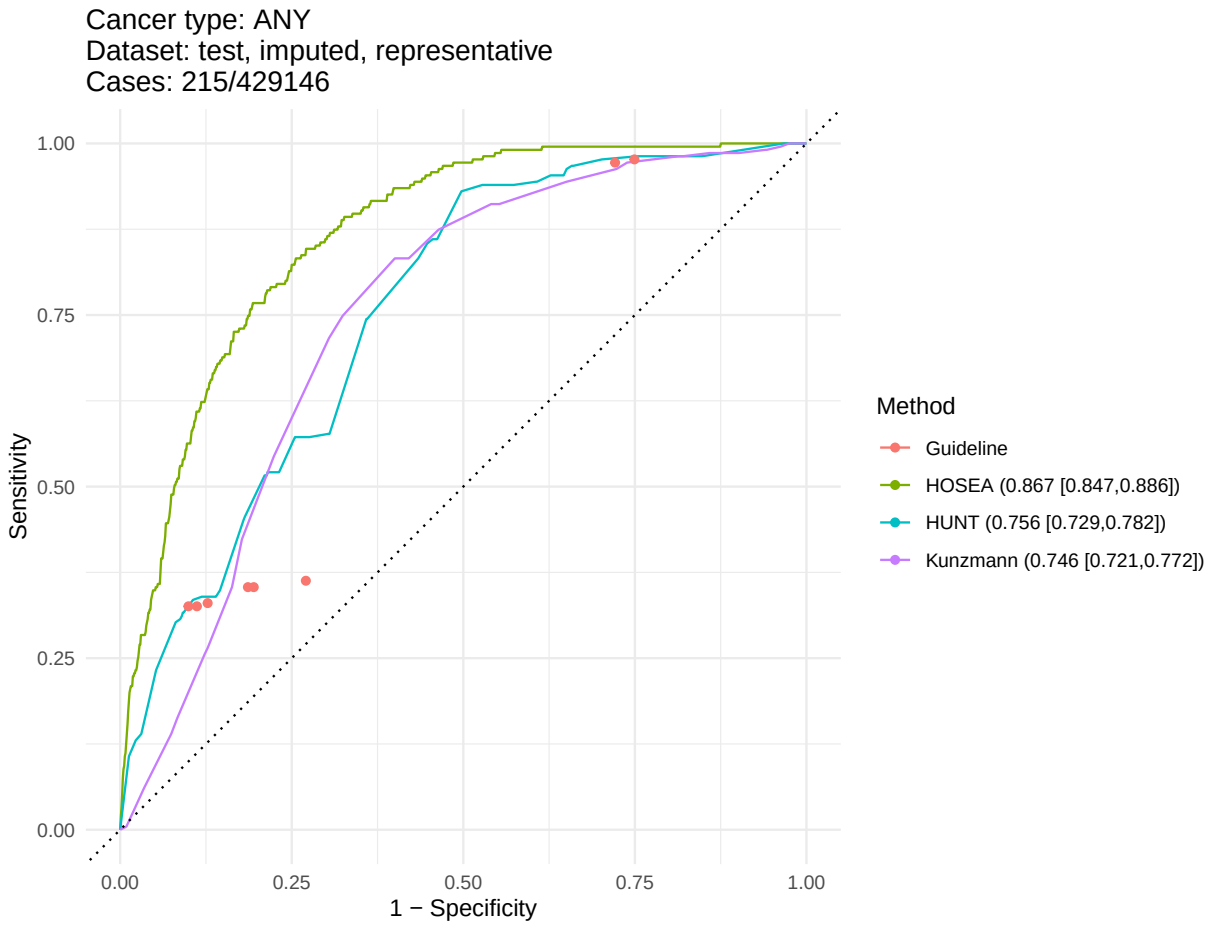
1.4 Identity

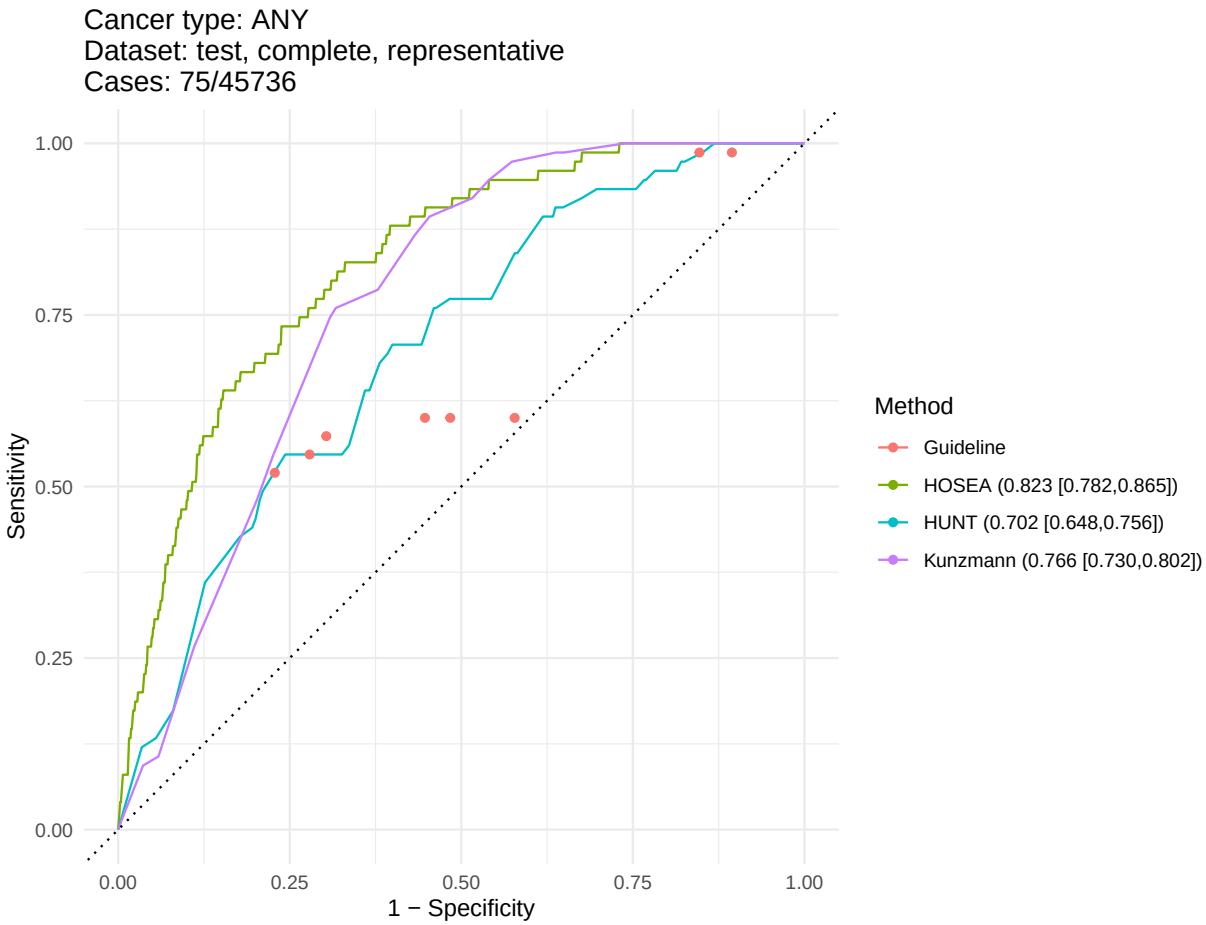

ROC Curve per age strutum
Cancer type: ANY
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Age
(0,45] (0.750 [0.648,0.852])
(45,55] (0.744 [0.714,0.774])
(55,65] (0.717 [0.703,0.732])
(65,75] (0.746 [0.732,0.761])
(75,100] (0.744 [0.723,0.765])


ROC Curve per sex
Cancer type: ANY
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Sex
Female (0.857 [0.786,0.929])
Male (0.786 [0.779,0.793])


ROC Curve per race
Cancer type: ANY
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
Race
asian (0.865 [0.762,0.968])
black (0.776 [0.747,0.806])
hawaiianpacific (0.747 [0.667,0.826])
indianalaskan (0.814 [0.741,0.887])
nonwhite (0.784 [0.758,0.810])
white (0.783 [0.775,0.791])

1.5 Representative samples




Cancer type: ANY
Dataset: test, complete, representative
Cases: 75/45736
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Method
Guideline
HOSEA (0.823 [0.782,0.865])
HUNT (0.702 [0.648,0.756])
Kunzmann (0.766 [0.730,0.802])

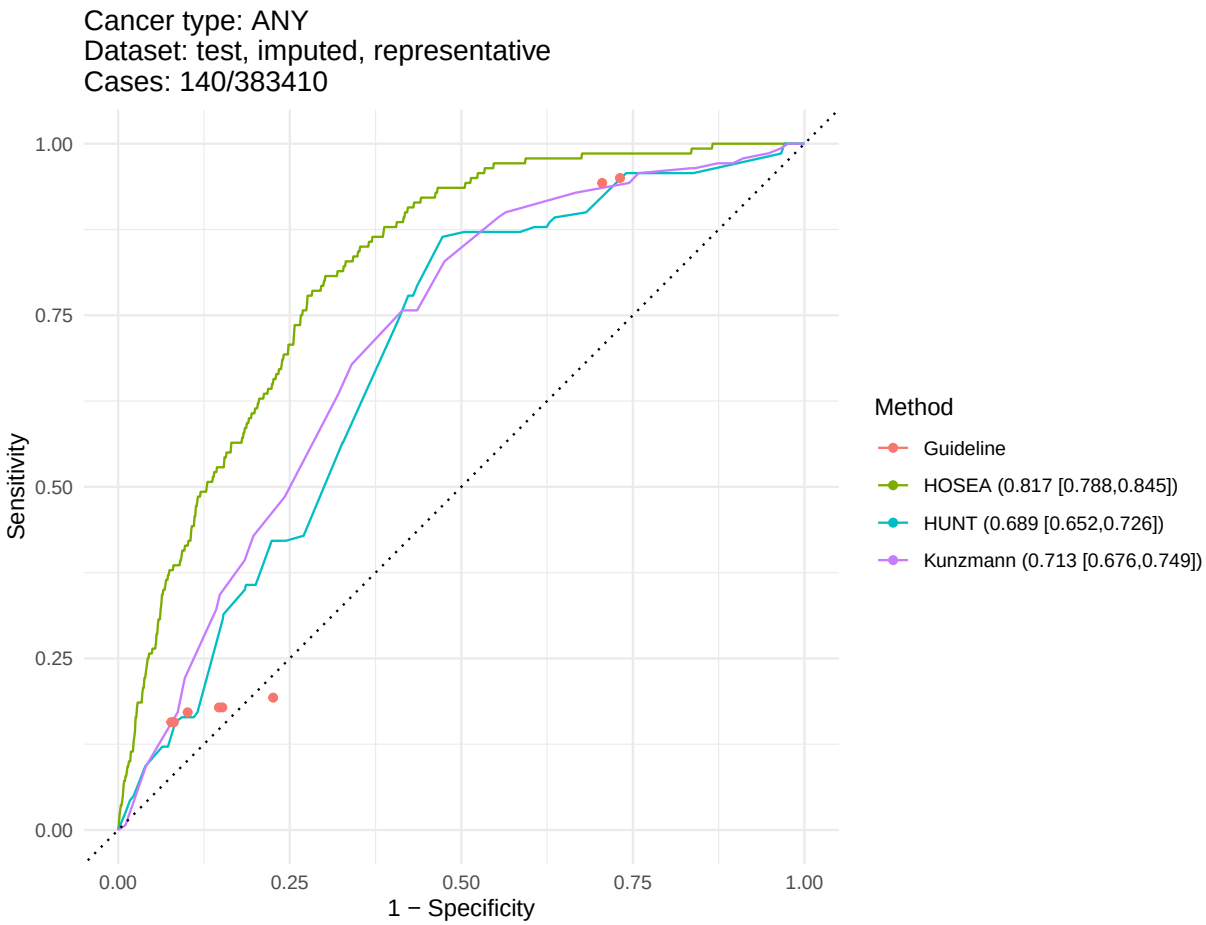
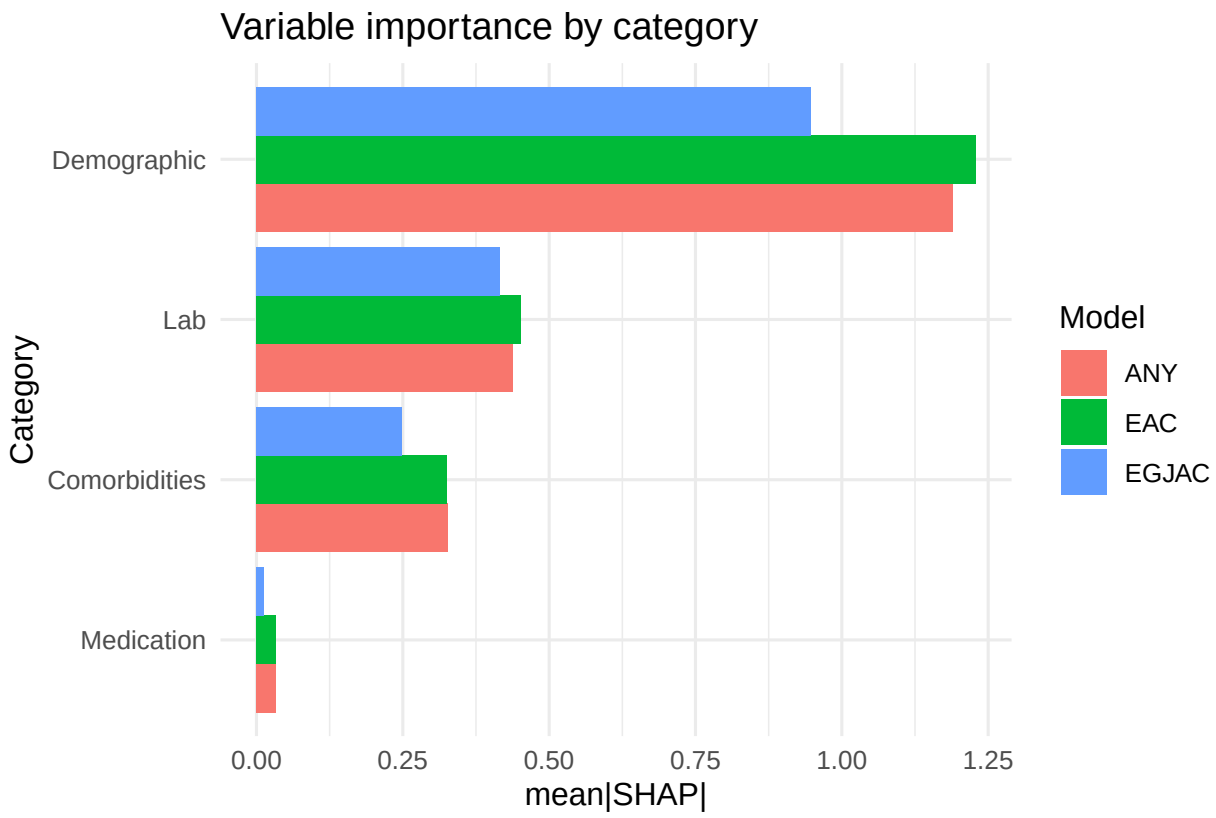


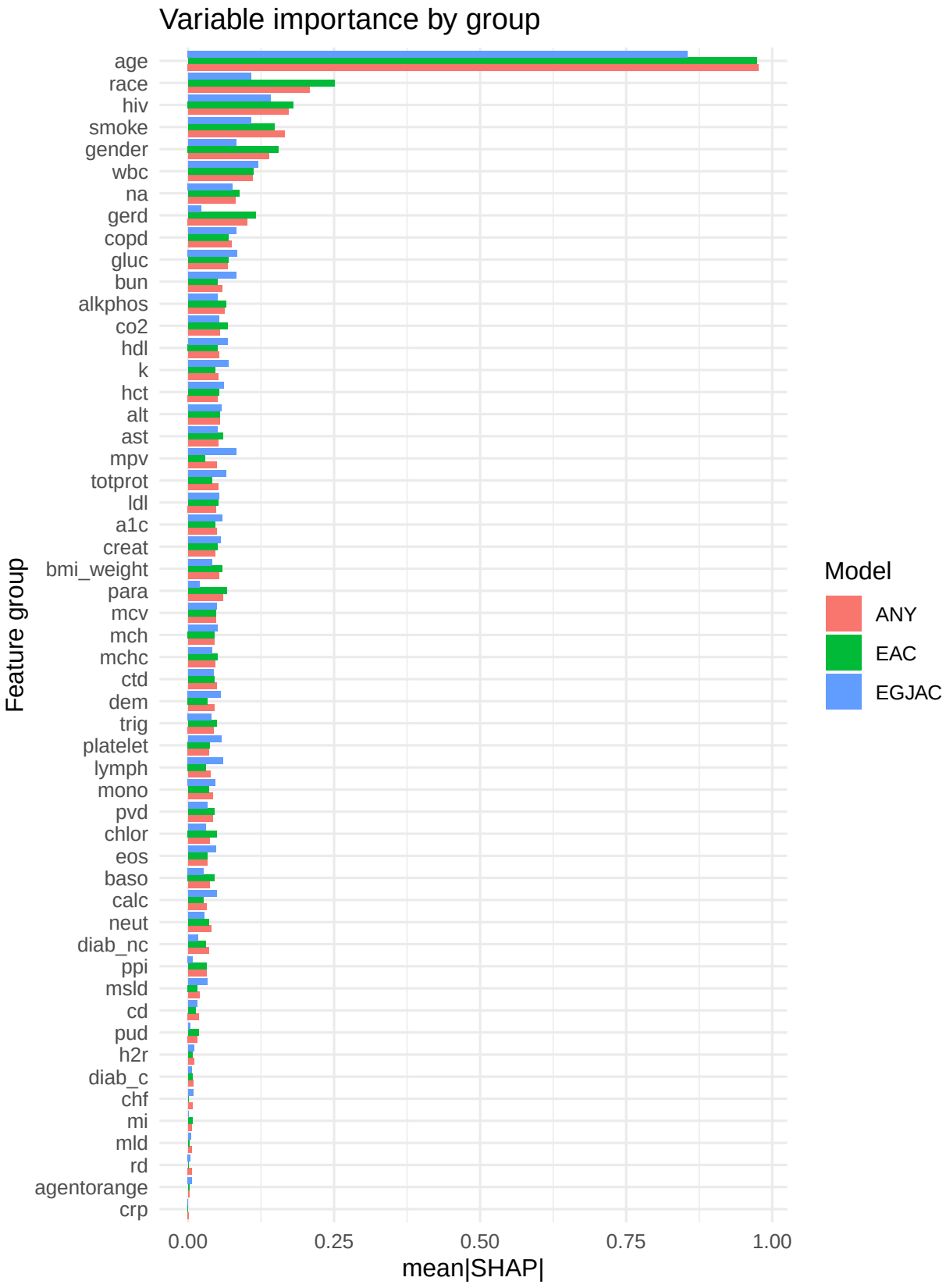
Figure 4: This should read "incomplete"...

1.6 Variable importance

A few observations:

- Age, gender, race, smoking, GERD all among the top 8 most important in terms of SHAP
- BMI/weight much farther
- HIV is now one of the most important
- Important labs: WBC, NA
- EAC: mostly similar to ANY (not surprising because of large prevalance)
- EGJAC: Age, race, smoking, gender, GERD are all less important




Variable importance by group
age
race
hiv
smoke
gender
wbc
na
gerd
copd
gluc
bun
alkphos
co2
hdl
k
hct
alt
ast
mpv
totprot
ldl
a1c
creat
bmi_weight
para
mcv
mch
mchc
ctd
dem
trig
platelet
lymph
mono
pvd
chlor
eos
baso
calc
neut
diab_nc
ppi
msld
cd
pud
h2r
diab_c
chf
mi
mld
rd
agentorange
crp
Feature group
0.00
0.25
0.50
0.75
1.00
mean|SHAP|
Model
ANY
EAC
EGJAC

Previous update

- I am currently investigating the difference in performance between the update of EGJAC lab results. In particular, there remained a few issues:
  - Why is there such a difference in the EAC model? For this outcome, only a few control (the EGJAC cases) changed slightly, so I do not expect significant changes
  - Why is there such a difference between complete data (wrt to HUNT/Kunzmann) and general data? Why are we doing so much worse for complete data?
  - Why is there a difference in the HUNT and Kunzmann results? These should not change since the data is exactly the same.
  - Why is the performance so much worse for older patients?
- Findings:
  - Tuning: I had to tweak tuning parameters because I saw some overfitting for EGJAC; I did a few tests and this doesn't seem to be the issue
  - Feature distribution: labs are mostly the same, differences are minor and cannot account for the various differences
  - I verified that the test set were identical
  - I get fewer "complete cases" now (407K patients, 1192 cases; 363K, 845 cases). No idea why yet. This surely explains the difference in HUNT/Kunzmann scores, but should still be understood.
  - I can almost reproduce the past results by using [4-0] data instead of [5-1] data. There remains some small gaps, but that could be from the imputation, small changes in processing, etc.
  - In terms of "complete data", we are now much closer: 394267 patients 1173 and cases. There remains a small gap, but that might just be small changes in processing. Indeed, I used the lates processing for [4-0], while an earlier version was used.
  - Remains to understand the shift in EAC and complete data.

I compared a few iterations:

- Original: the model currently in the package
- Post EGJAC: the new model with update EGJAC
- Pre EGJAC: Everything kept the same except for the EGJAC

I compared a few datasets:

- Post EGJAC: the latest dataset
- Pre EGJAC: same, but without updated EGJAC
- Shifted: using the post_egjac processing, but to [4-0] data instead of [5-1]
- c_xxxxxx: a few processing iterations from the past

	data	model	ANY	EAC	EGJAC
			MICE		
6	post_egjac	post_egjac	**0.800** [0.793,0.807]	**0.802** [0.794,0.811]	**0.769** [0.755,0.783]
			SRS		
4	post_egjac	original	0.737 [0.729,0.745]	0.742 [0.733,0.751]	0.646 [0.628,0.664]
5	post_egjac	pre_egjac	0.749 [0.742,0.756]	0.763 [0.754,0.771]	0.593 [0.574,0.612]
6	post_egjac	post_egjac	**0.769** [0.762,0.776]	**0.771** [0.763,0.78]	**0.741** [0.725,0.756]
1	pre_egjac	original	0.797 [0.789,0.805]	0.763 [0.754,0.773]	0.909 [0.895,0.922]
2	pre_egjac	pre_egjac	**0.816** [0.809,0.824]	**0.79** [0.781,0.798]	**0.941** [0.931,0.951]
3	pre_egjac	post_egjac	0.772 [0.765,0.779]	0.772 [0.763,0.78]	0.738 [0.723,0.753]
7	c_f6f1f465	original	0.806 [0.797,0.814]	0.778 [0.768,0.787]	0.905 [0.891,0.919]
8	c_f6f1f465	pre_egjac	0.819 [0.812,0.826]	0.792 [0.783,0.8]	0.941 [0.93,0.951]
9	c_f6f1f465	post_egjac	0.771 [0.764,0.778]	0.772 [0.764,0.781]	0.737 [0.722,0.753]
10	c_69cc9e5b	original	0.806 [0.797,0.814]	0.778 [0.768,0.787]	0.905 [0.891,0.919]
11	c_69cc9e5b	pre_egjac	0.819 [0.812,0.826]	0.792 [0.783,0.8]	0.941 [0.93,0.951]
12	c_69cc9e5b	post_egjac	0.771 [0.764,0.778]	0.772 [0.764,0.781]	0.737 [0.722,0.753]
1	shifted	original	**0.879** [0.872,0.885]	**0.86** [0.853,0.868]	**0.955** [0.945,0.965]
2	shifted	pre_egjac	0.849 [0.842,0.855]	0.821 [0.814,0.829]	0.969 [0.962,0.977]
3	shifted	post_egjac	0.805 [0.798,0.812]	0.807 [0.799,0.814]	0.773 [0.759,0.787]
	(figures)	(figures)	**0.898** [0.892,0.903]	**0.858** [0.842,0.873]	**0.949** [0.926,0.972]
	testing	post_egjac	0.781	0.763	0.741
	testing	pre_egjac	0.833	0.781	0.954
	testing	original	0.920	0.870	0.970

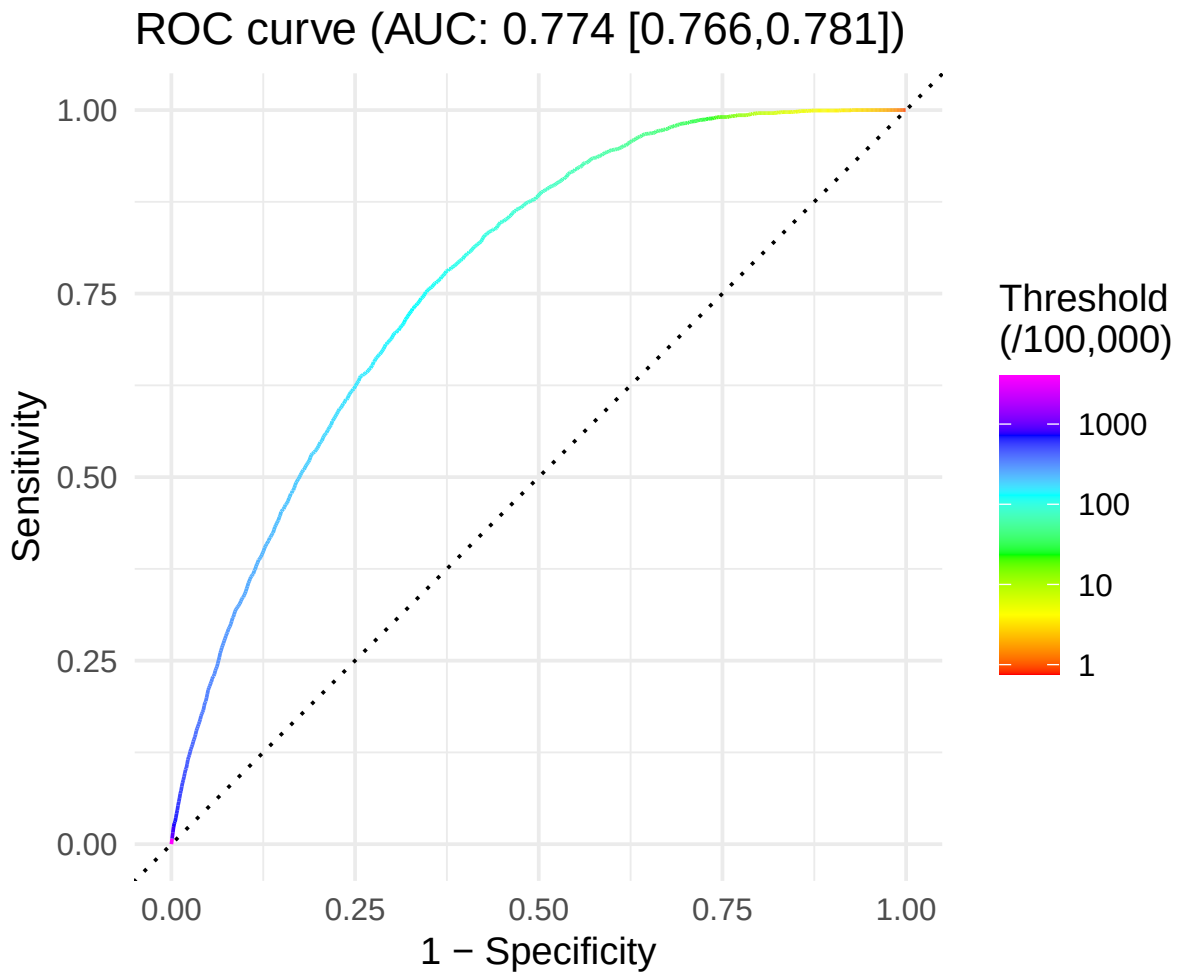
Table 1: This is for the full test set (25% of controls and cases). Apart from a small difference in ANY, EAC and EGJAC are almost identical between the figures and "shifted original". this is the only situation even in the same ballpark.

	data	model	ANY	EAC	EGJAC
6	post_egjac	original	0.691 [0.675,0.707]	0.68 [0.662,0.698]	0.607 [0.571,0.644]
7	post_egjac	pre_egjac	0.704 [0.689,0.72]	0.727 [0.71,0.743]	0.61 [0.573,0.647]
8	post_egjac	post_egjac	**0.714** [0.699,0.728]	**0.736** [0.72,0.753]	**0.681** [0.649,0.713]
9	post_egjac	Kunzmann	0.637 [0.62,0.654]	0.635 [0.616,0.653]	0.64 [0.604,0.676]
10	post_egjac	HUNT	0.597 [0.579,0.616]	0.615 [0.594,0.636]	0.625 [0.589,0.66]
1	pre_egjac	original	0.811 [0.796,0.827]	0.745 [0.727,0.763]	0.991 [0.987,0.996]
2	pre_egjac	pre_egjac	**0.814** [0.799,0.828]	**0.773** [0.757,0.789]	**0.996** [0.993,0.998]
3	pre_egjac	post_egjac	0.712 [0.697,0.727]	0.74 [0.723,0.756]	0.672 [0.641,0.704]
4	pre_egjac	Kunzmann	0.637 [0.62,0.654]	0.635 [0.616,0.653]	0.64 [0.604,0.676]
5	pre_egjac	HUNT	0.597 [0.579,0.616]	0.615 [0.594,0.636]	0.625 [0.589,0.66]
11	c_f6f1f465	original	0.828 [0.813,0.843]	0.779 [0.762,0.797]	0.989 [0.983,0.995]
12	c_f6f1f465	pre_egjac	0.813 [0.798,0.828]	0.771 [0.754,0.787]	0.993 [0.988,0.998]
13	c_f6f1f465	post_egjac	0.714 [0.699,0.729]	0.732 [0.715,0.748]	0.672 [0.64,0.704]
14	c_f6f1f465	Kunzmann	0.637 [0.62,0.654]	0.635 [0.616,0.653]	0.64 [0.604,0.676]
15	c_f6f1f465	HUNT	0.597 [0.579,0.616]	0.615 [0.594,0.636]	0.625 [0.589,0.66]
16	c_69cc9e5b	original	0.828 [0.813,0.843]	0.779 [0.762,0.797]	0.989 [0.983,0.995]
17	c_69cc9e5b	pre_egjac	0.813 [0.798,0.828]	0.771 [0.754,0.787]	0.993 [0.988,0.998]
18	c_69cc9e5b	post_egjac	0.714 [0.699,0.729]	0.732 [0.715,0.748]	0.672 [0.64,0.704]
19	c_69cc9e5b	Kunzmann	0.637 [0.62,0.654]	0.635 [0.616,0.653]	0.64 [0.604,0.676]
20	c_69cc9e5b	HUNT	0.597 [0.579,0.616]	0.615 [0.594,0.636]	0.625 [0.589,0.66]
1	shifted	original	**0.856** [0.845,0.868]	**0.813** [0.8,0.827]	**0.994** [0.99,0.998]
2	shifted	pre_egjac	0.82 [0.808,0.833]	0.769 [0.755,0.783]	0.993 [0.989,0.998]
3	shifted	post_egjac	0.73 [0.718,0.742]	0.732 [0.719,0.746]	0.678 [0.653,0.704]
4	shifted	Kunzmann	**0.649** [0.635,0.663]	**0.646** [0.631,0.662]	**0.65** [0.621,0.678]
5	shifted	HUNT	**0.593** [0.578,0.609]	**0.602** [0.584,0.62]	**0.584** [0.553,0.615]
	(figures)	(XGBoost)	**0.851** [0.84,0.862]	**0.823** [0.809,0.836]	**0.933** [0.916,0.95]
	(figures)	Kunzmann	**0.653** [0.64,0.667]	**0.651** [0.635,0.667]	**0.658** [0.631,0.685]
	(figures)	HUNT	**0.593** [0.578,0.609]	**0.603** [0.585,0.62]	**0.565** [0.533,0.597]

Table 2: Restricted to "complete cases" wrt to HUNT/Kunzmann. "shifted original" seems the closest to "figures XGBoost". Also, "shifted" Kunzmann and HUNT are much closer to those under "figures".

	data	seed	original	pre_egjac	post_egjac
11	pre_egjac	1	0.765 [0.756,0.774]	0.788 [0.779,0.796]	0.77 [0.762,0.779]
12	pre_egjac	2	0.768 [0.758,0.777]	0.788 [0.78,0.797]	0.772 [0.764,0.78]
13	pre_egjac	3	0.764 [0.755,0.774]	0.788 [0.78,0.797]	0.77 [0.762,0.779]
14	pre_egjac	4	0.765 [0.756,0.775]	0.785 [0.777,0.793]	0.77 [0.762,0.779]
15	pre_egjac	5	0.773 [0.764,0.782]	0.79 [0.782,0.798]	0.771 [0.763,0.779]
16	pre_egjac	6	0.768 [0.758,0.777]	0.787 [0.778,0.795]	0.769 [0.76,0.777]
17	pre_egjac	7	0.769 [0.76,0.779]	0.79 [0.782,0.798]	0.771 [0.763,0.779]
18	pre_egjac	8	0.77 [0.76,0.779]	0.788 [0.78,0.797]	0.769 [0.761,0.778]
19	pre_egjac	9	0.764 [0.754,0.773]	0.791 [0.782,0.799]	0.771 [0.763,0.78]
20	pre_egjac	10	0.766 [0.757,0.775]	0.79 [0.782,0.798]	0.771 [0.762,0.779]
1	post_egjac	1	0.739 [0.73,0.748]	0.76 [0.752,0.768]	0.769 [0.761,0.778]
2	post_egjac	2	0.738 [0.728,0.747]	0.76 [0.752,0.769]	0.769 [0.761,0.777]
3	post_egjac	3	0.735 [0.725,0.744]	0.76 [0.751,0.768]	0.77 [0.762,0.778]
4	post_egjac	4	0.738 [0.728,0.747]	0.764 [0.756,0.772]	0.767 [0.759,0.776]
5	post_egjac	5	0.744 [0.735,0.753]	0.763 [0.755,0.771]	0.771 [0.763,0.779]
6	post_egjac	6	0.742 [0.733,0.751]	0.763 [0.754,0.771]	0.768 [0.759,0.776]
7	post_egjac	7	0.738 [0.729,0.748]	0.763 [0.755,0.772]	0.769 [0.761,0.778]
8	post_egjac	8	0.739 [0.73,0.749]	0.761 [0.753,0.77]	0.769 [0.761,0.777]
9	post_egjac	9	0.734 [0.725,0.744]	0.765 [0.757,0.773]	0.772 [0.764,0.78]
10	post_egjac	10	0.736 [0.727,0.745]	0.761 [0.753,0.77]	0.769 [0.76,0.777]
21	shifted	1	0.862 [0.855,0.869]	0.823 [0.815,0.831]	0.807 [0.8,0.815]
22	shifted	2	0.857 [0.849,0.864]	0.821 [0.813,0.829]	0.807 [0.799,0.815]
23	shifted	3	0.858 [0.851,0.865]	0.819 [0.811,0.826]	0.804 [0.796,0.812]
24	shifted	4	0.856 [0.849,0.864]	0.822 [0.815,0.83]	0.805 [0.797,0.813]
25	shifted	5	0.861 [0.854,0.868]	0.82 [0.813,0.828]	0.806 [0.799,0.814]
26	shifted	6	0.858 [0.851,0.866]	0.819 [0.811,0.827]	0.806 [0.798,0.814]
27	shifted	7	0.86 [0.853,0.867]	0.819 [0.811,0.827]	0.807 [0.799,0.815]
28	shifted	8	0.86 [0.853,0.867]	0.821 [0.813,0.828]	0.806 [0.798,0.814]
29	shifted	9	0.856 [0.848,0.863]	0.822 [0.814,0.83]	0.807 [0.8,0.815]
30	shifted	10	0.862 [0.854,0.869]	0.822 [0.814,0.829]	0.804 [0.797,0.812]

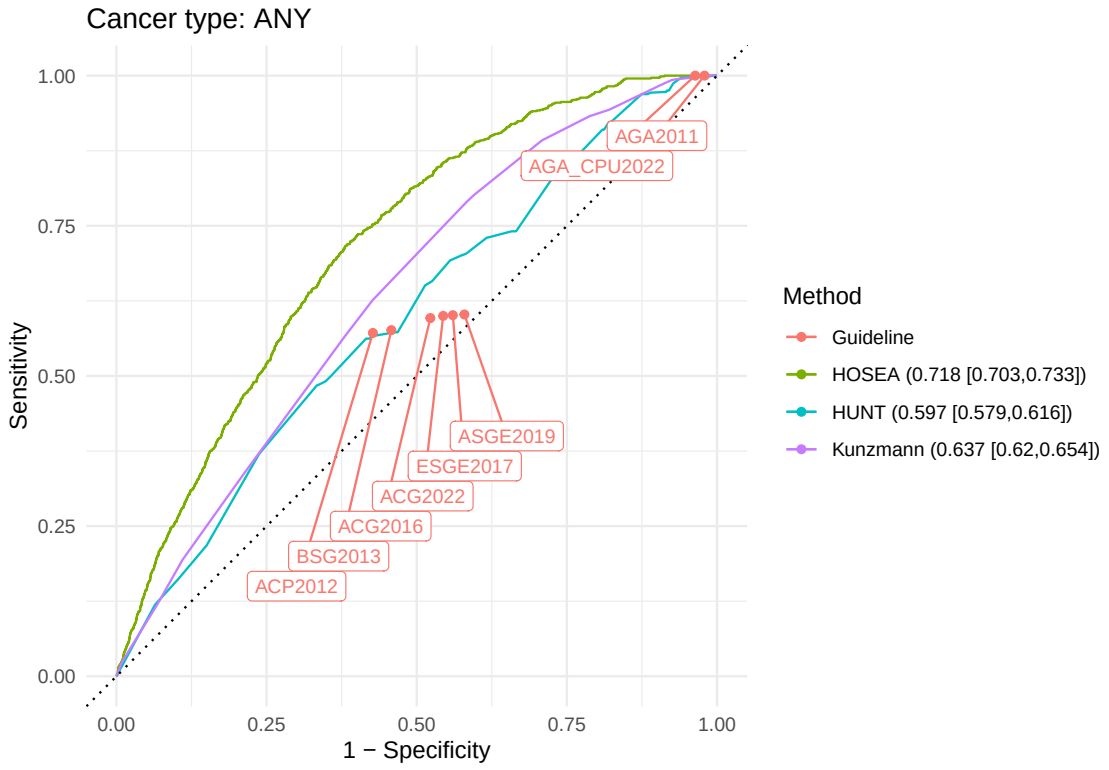
Table 3: EAC. There does seem to be a difference between pre- and post-egjac update.

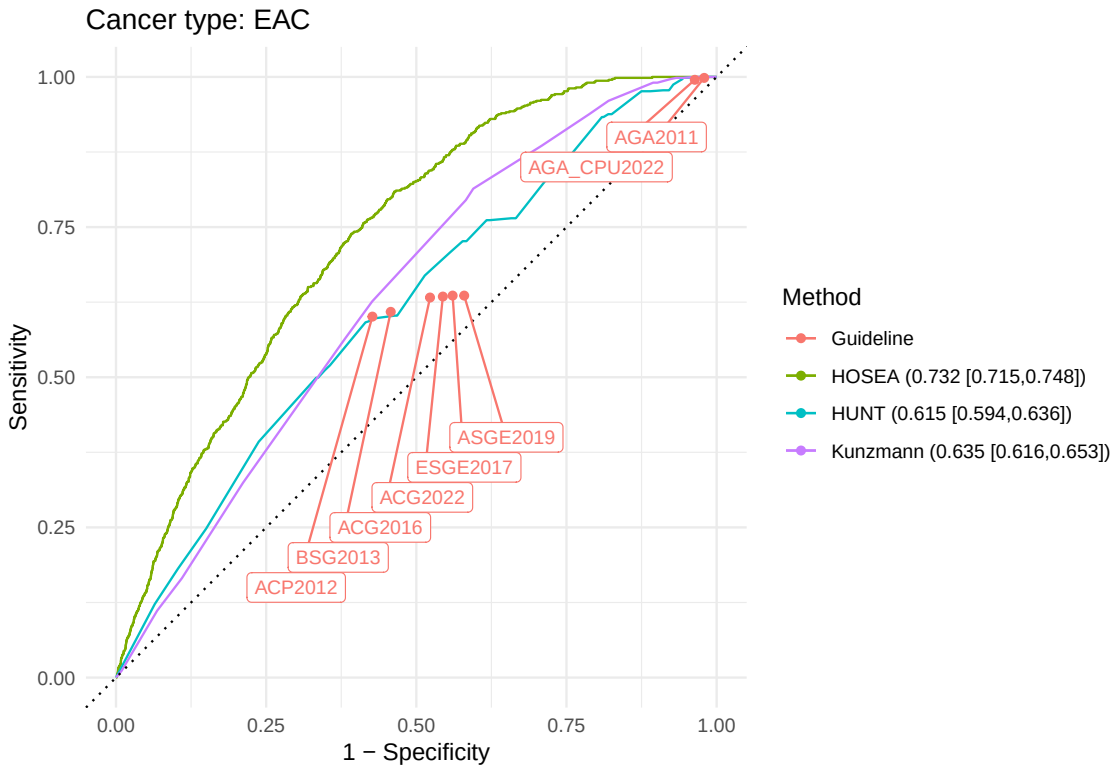


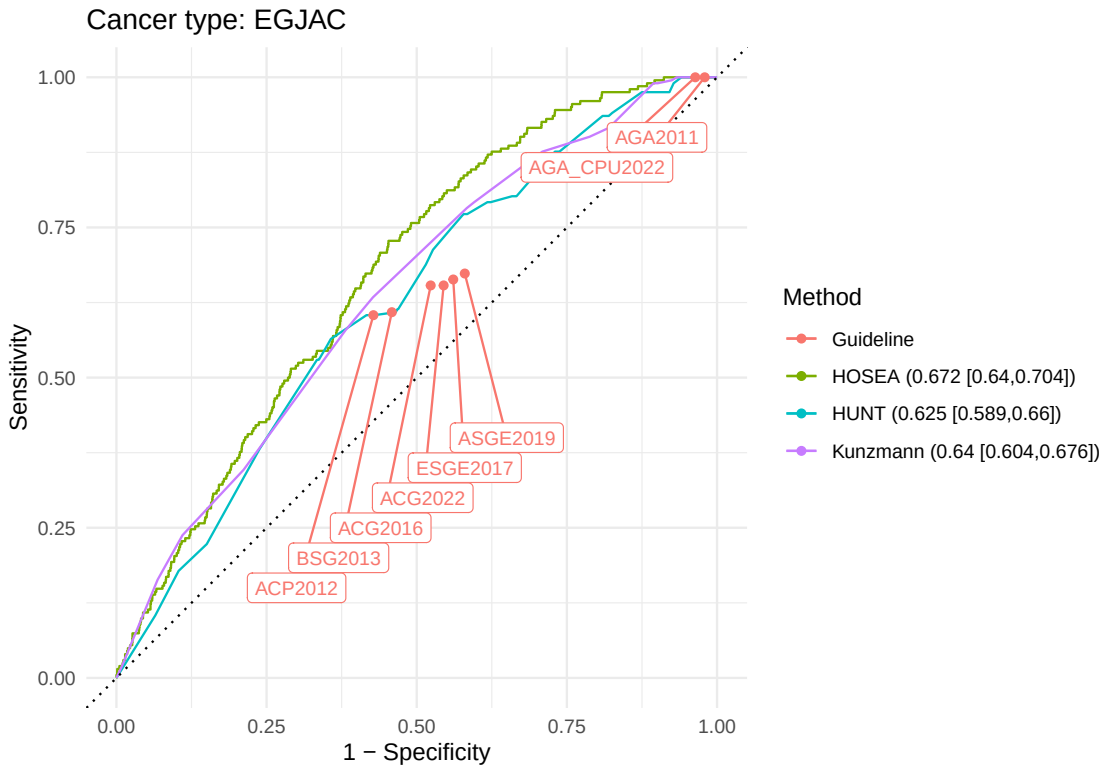
Comparison

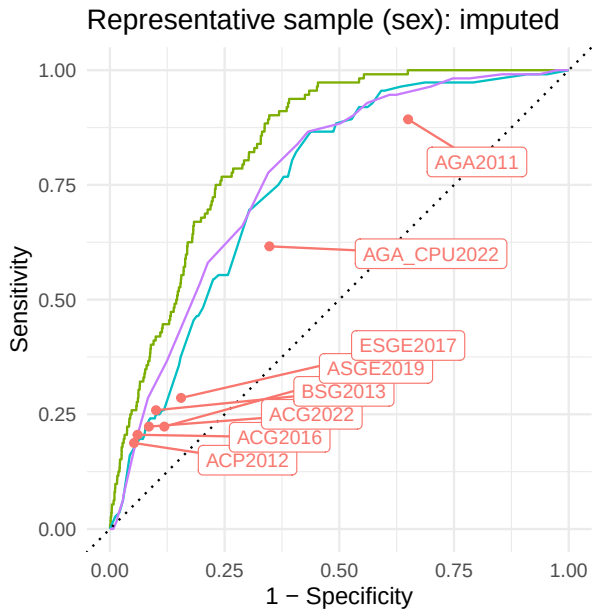
For these comparisons:

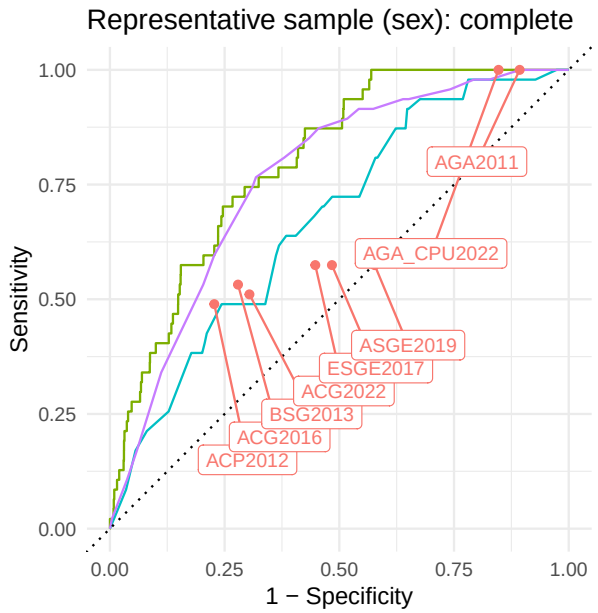
- Subset to test observations
- Filter out patients with missing information for HUNT/Kunzmann/Guidelines
- i.e., require age, BMI, race, smoking status, gerd, h2r/ppi
- 407K patients, 1192 cases (292/100,000)
- AUC + 95% CI using DeLong method
- repeat for all three models/outcome


Cancer type: ANY
1.00
0.75
0.50
0.25
0.00
Sensitivity
0.00
0.25
0.50
0.75
1.00
1 − Specificity
AGA2011
AGA_CPU2022
ASGE2019
ESGE2017
ACG2022
ACG2016
BSG2013
ACP2012
Method
Guideline
HOSEA (0.718 [0.703,0.733])
HUNT (0.597 [0.579,0.616])
Kunzmann (0.637 [0.62,0.654])


Cancer type: EAC
1.00
0.75
0.50
0.25
0.00
Sensitivity
0.00
0.25
0.50
0.75
1.00
1 − Specificity
AGA2011
AGA_CPU2022
ASGE2019
ESGE2017
ACG2022
ACG2016
BSG2013
ACP2012
Method
Guideline
HOSEA (0.732 [0.715,0.748])
HUNT (0.615 [0.594,0.636])
Kunzmann (0.635 [0.616,0.653])


Cancer type: EGJAC
1.00
0.75
0.50
0.25
0.00
Sensitivity
0.00
0.25
0.50
0.75
1.00
1 − Specificity
AGA2011
AGA_CPU2022
ASGE2019
ESGE2017
ACG2022
ACG2016
BSG2013
ACP2012
Method
Guideline
HOSEA (0.672 [0.64,0.704])
HUNT (0.625 [0.589,0.66])
Kunzmann (0.64 [0.604,0.676])

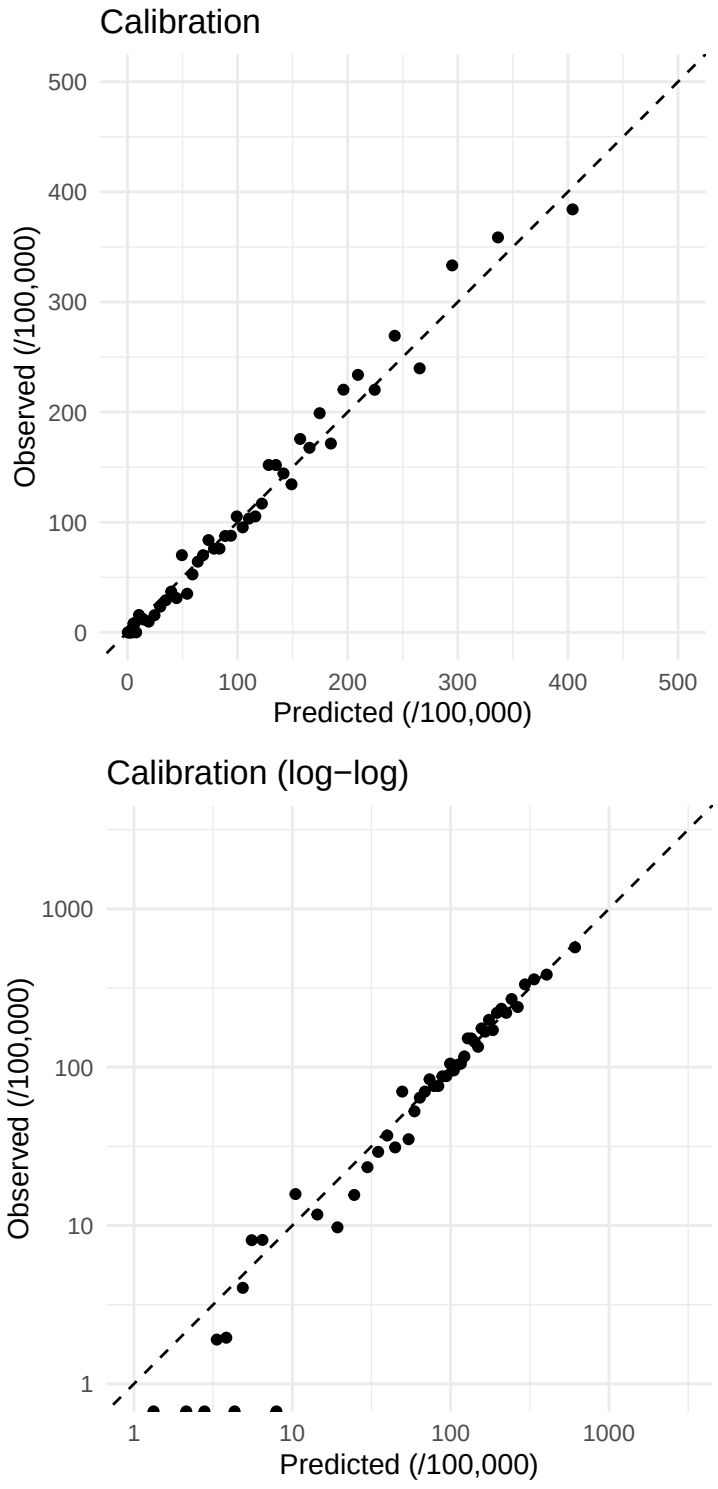

Representative sample (sex): imputed
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
AGA2011
AGA_CPU2022
ESGE2017
ASGE2019
BSG2013
ACG2022
ACG2016
ACP2012
Method
a Guideline
a HOSEA (0.835 [0.809,0.861])
a HUNT (0.75 [0.713,0.786])
a Kunzmann (0.767 [0.731,0.803])


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 (0.799 [0.748,0.849])
a HUNT (0.673 [0.602,0.743])
a Kunzmann (0.764 [0.706,0.822])

Calibration & threshold

Threshold	TPR	PPV	Det. prev.
5	99.86	0.13	86.24
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

Threshold	TPR	PPV	Det. prev.
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
3000	0.00	0.00	0.00
4000	0.00		0.00
5000	0.00		0.00
6000	0.00		0.00
7000	0.00		0.00
8000	0.00		0.00
10000	0.00		0.00



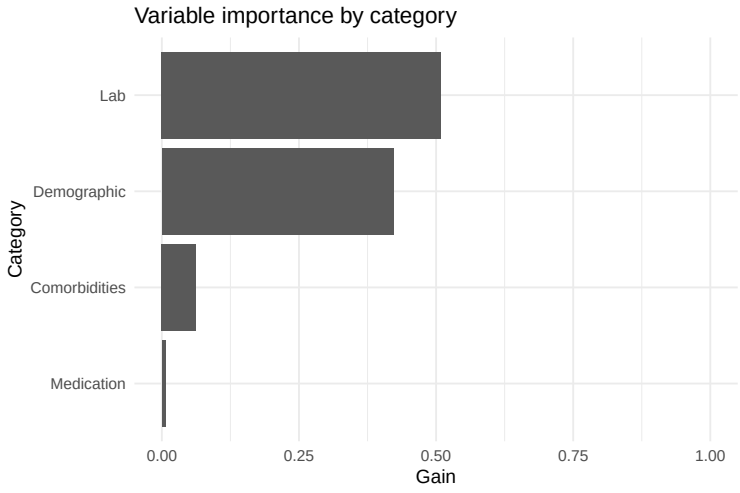
Each point represent 2% of the test data, split using predicted risk quantiles. The top plot is cropped on the right and top so we can focus on the more important region. HL test: H=49.851, df=49, p=0.439

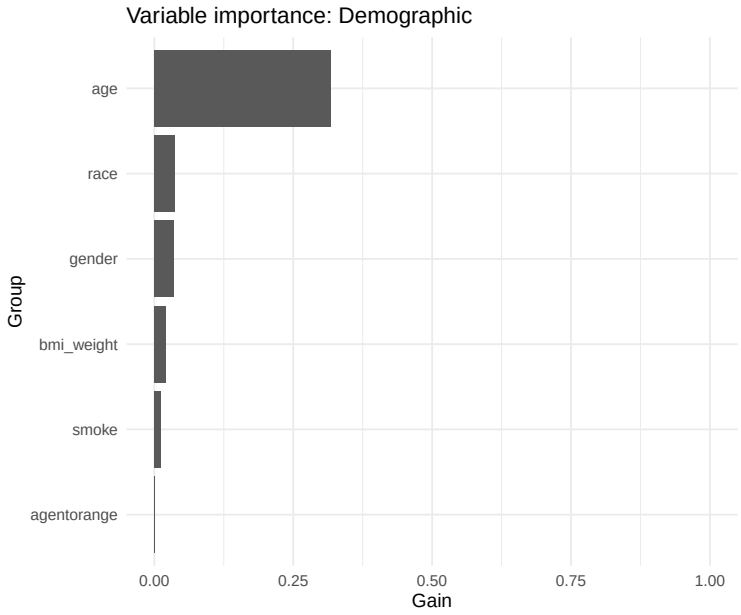
Gain Variable importance

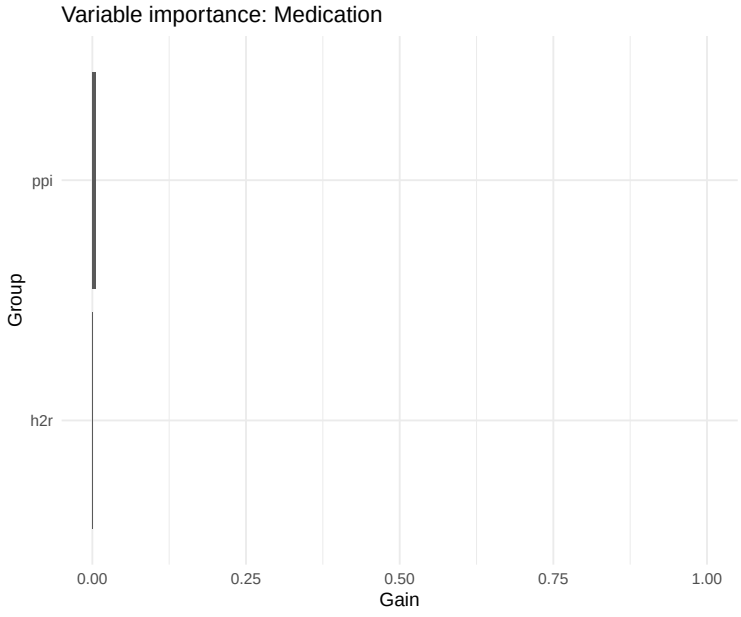
"Gain represents fractional contribution of each feature to the model based on the total gain of this feature's splits. Higher percentage means a more important predictive feature."

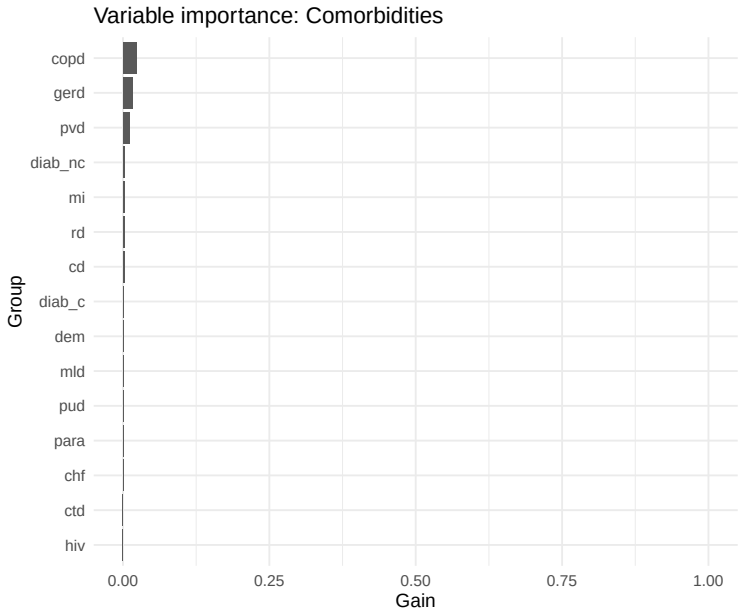
Some notes:

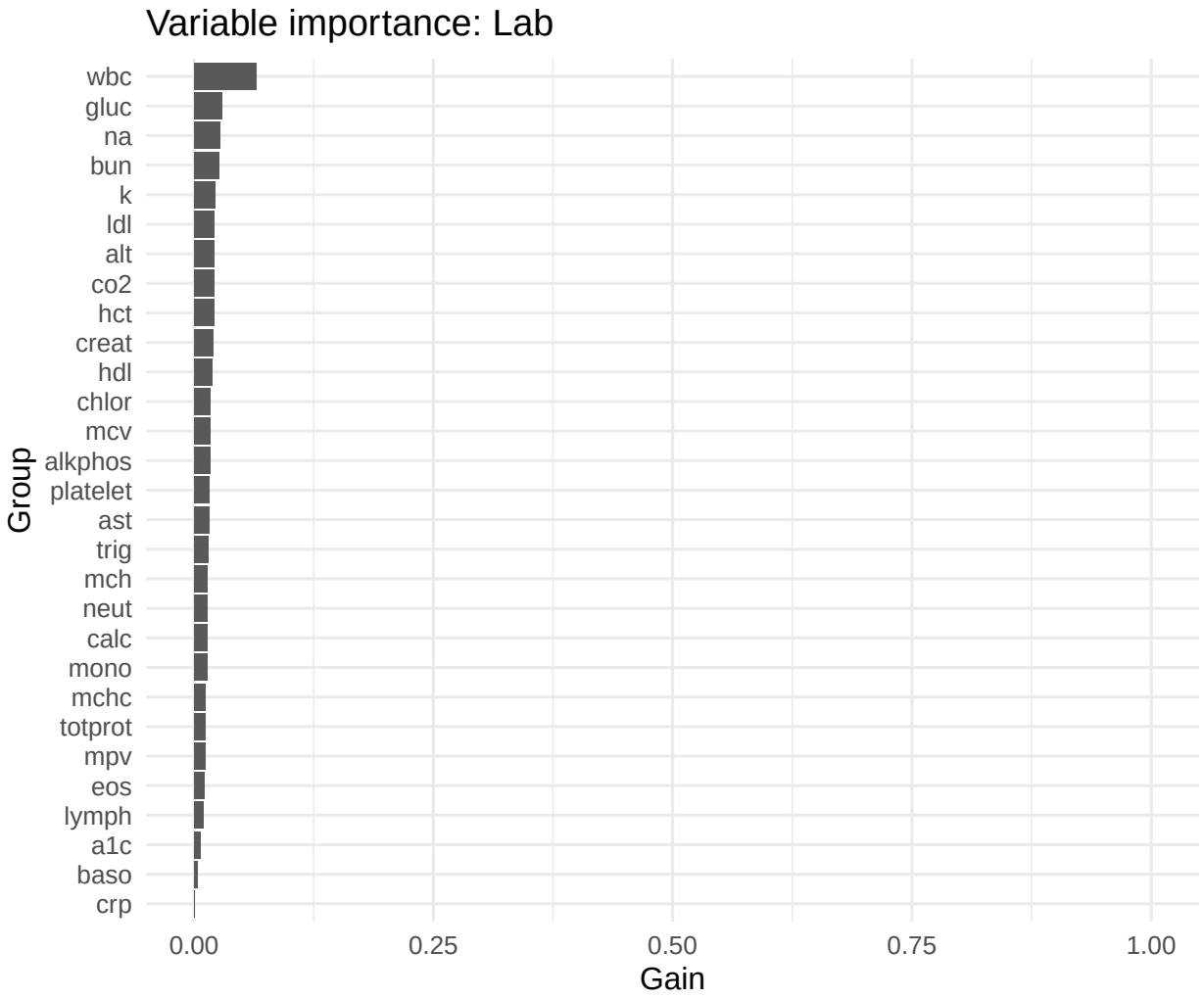
- These are additive, so we can compute the importance of a group of features.
- They sum up to 1
- We can only look at the relationship with the feature value (e.g., mostly positive, mostly negative) for single features




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


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

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

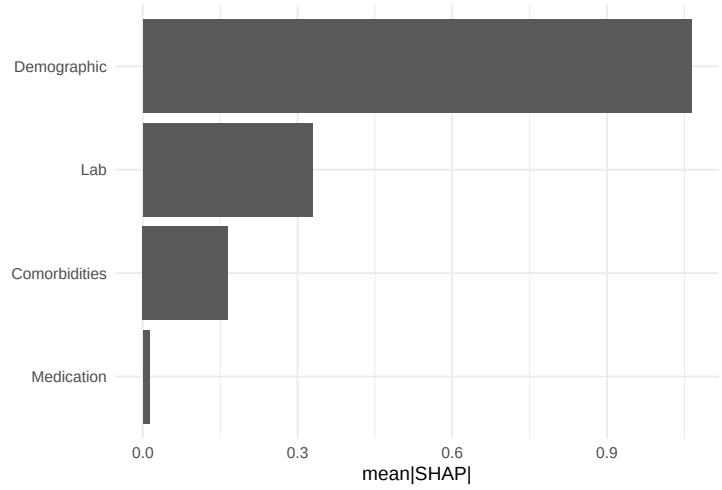
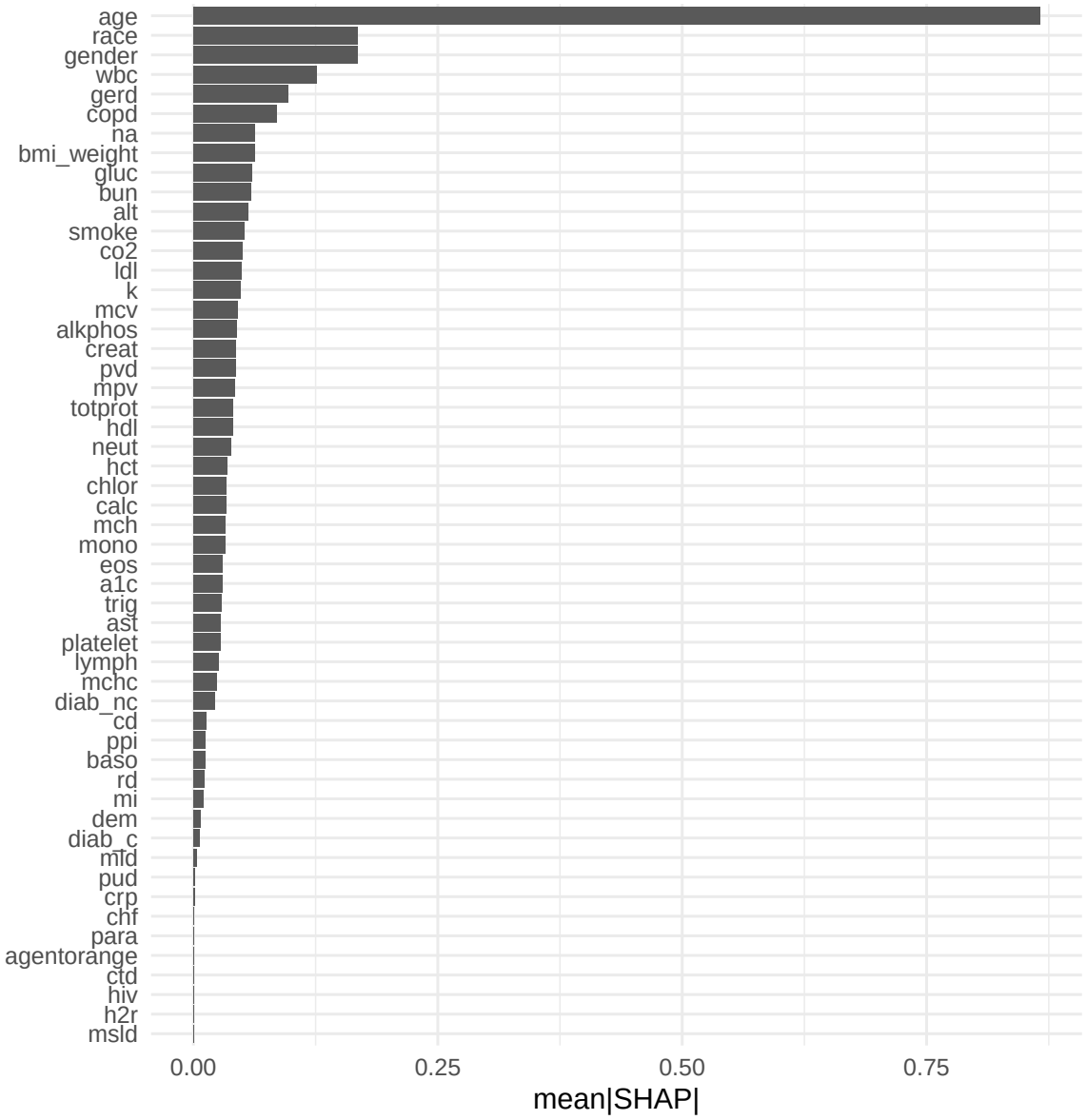


Figure 5: This has mean updated using a representative sample; previously, I was using a sample that over represented cases so age was much less important.


age
race
gender
wbc
gerd
copd
na
bmi_weight
gluc
bun
alt
smoke
co2
ldl
k
mcv
alkphos
creat
pvd
mpv
totprot
hdl
neut
hct
chlor
calc
mch
mono
eos
a1c
trig
ast
platelet
lymph
mchc
diab_nc
cd
ppi
baso
rd
mi
dem
diab_c
mld
pud
crp
chf
para
agentorange
ctd
hiv
h2r
msld
0.00
0.25
0.50
0.75
mean|SHAP|

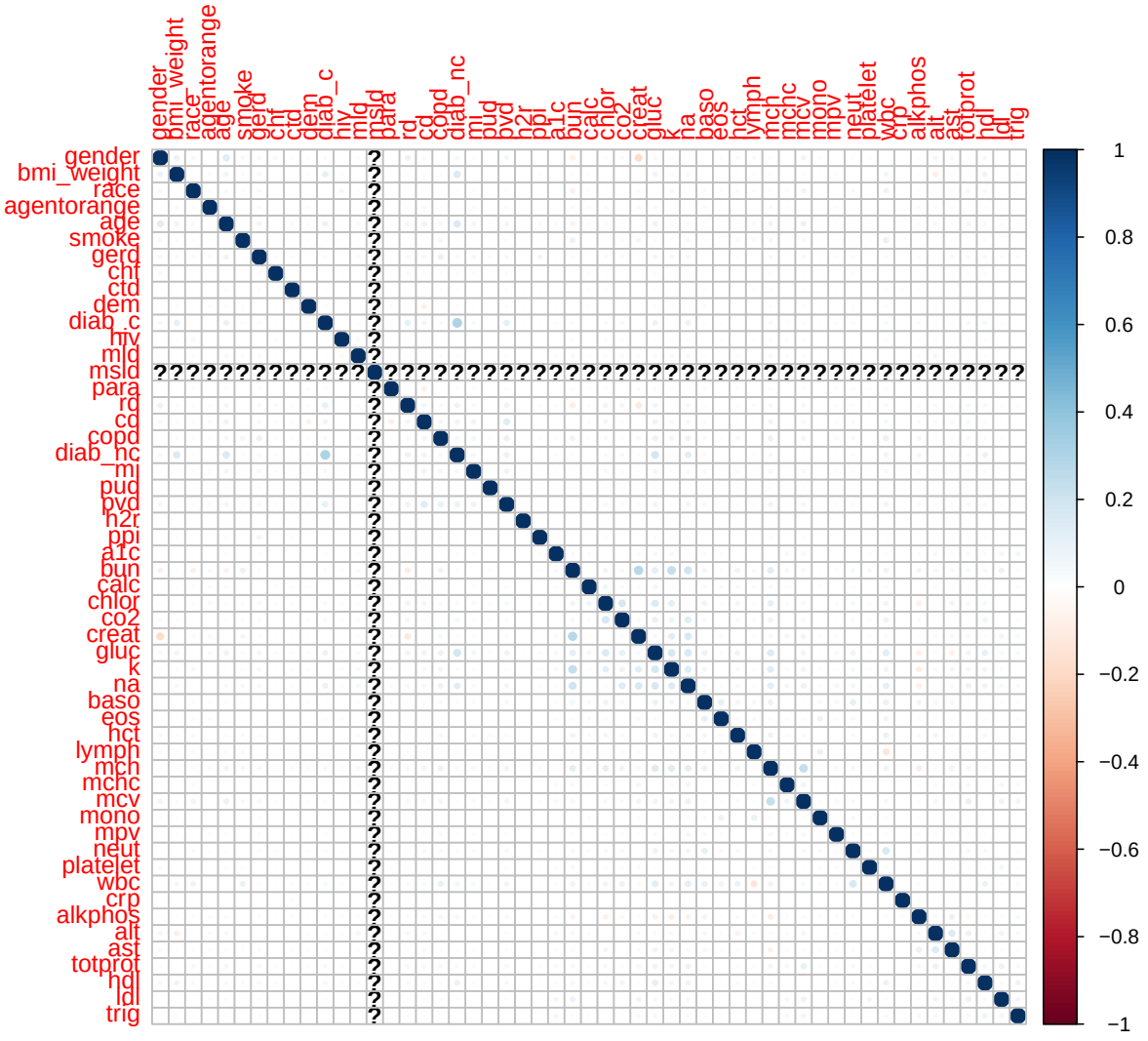
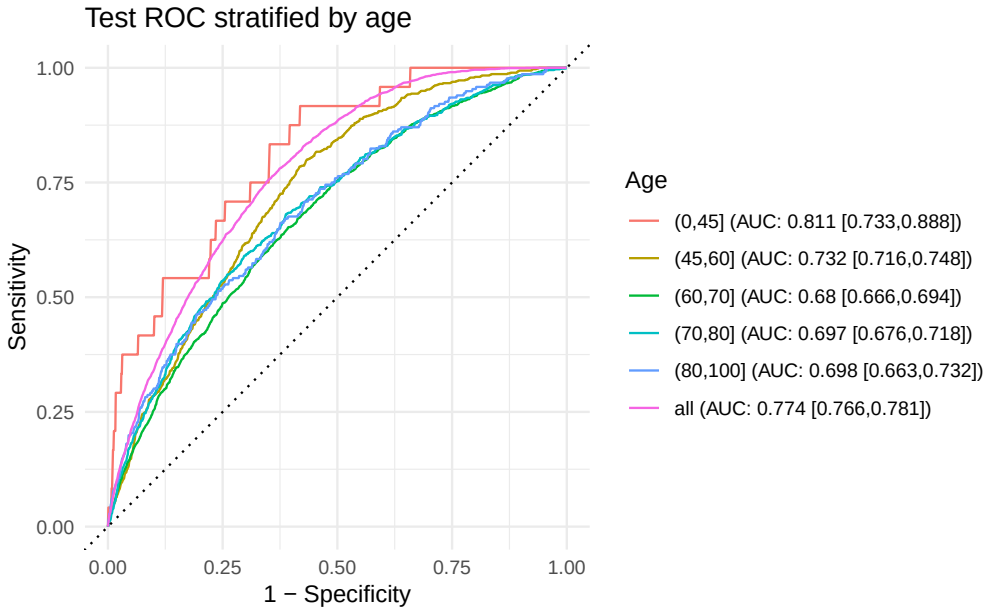


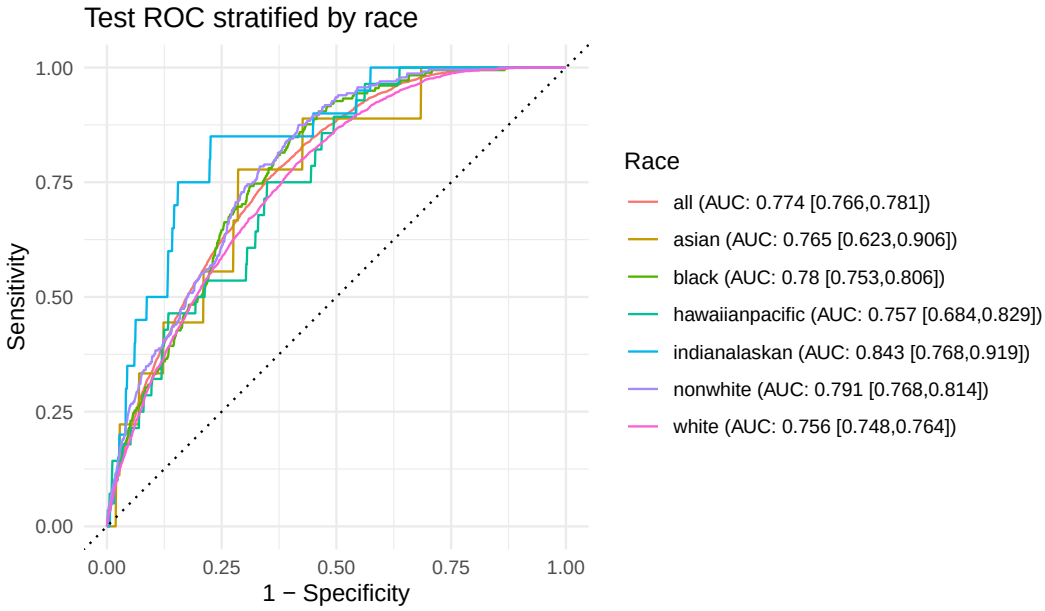
Figure 6: Correlation between group-level Shapley values. Largest absolute correlation in the following table. If two Shapley values are highly correlated, that means the underlying variables contribute essentially in the same way to the prediction and are therefore redundant.

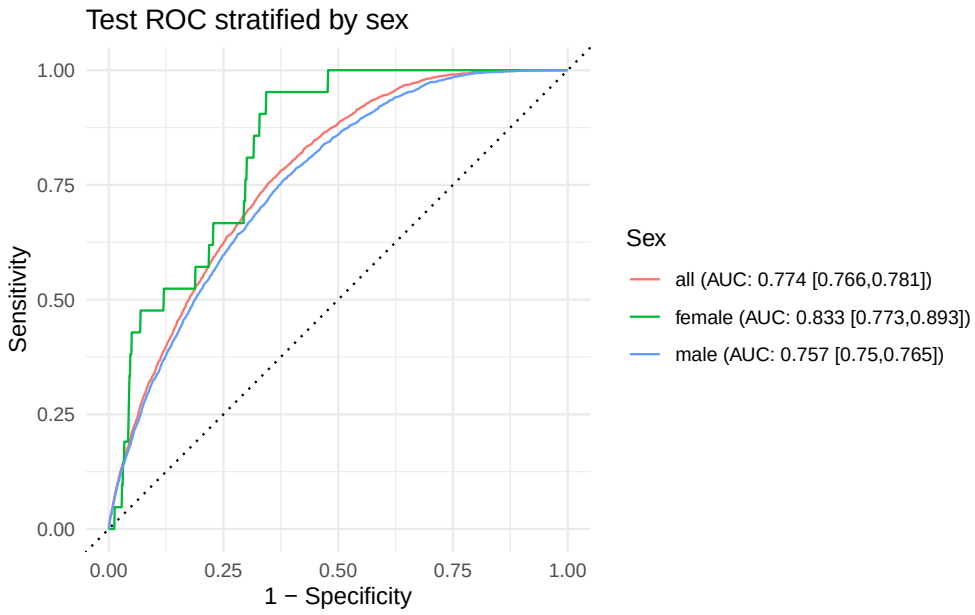
Feature pair		SHAP correlation
diab_c	diab_nc	0.31
bun	creat	0.28
bun	k	0.24
mch	mcv	0.23
bun	na	0.19
gender	creat	-0.18
diab_nc	gluc	0.18
chlor	co2	0.17
gluc	k	0.17
neut	wbc	0.17
gluc	na	0.17
creat	na	0.16
age	diab_nc	0.16
chlor	gluc	0.15
bmi_weight	diab_nc	0.15
alt	ast	0.15
k	na	0.14
gluc	mch	0.14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

Table 4: Since the largest Shapley correaltion is fairly low, we can be reassured we do not have redundant variables. Is the top one resulting from patients haveing diab_nc first and then diab_c later?

Identity groups



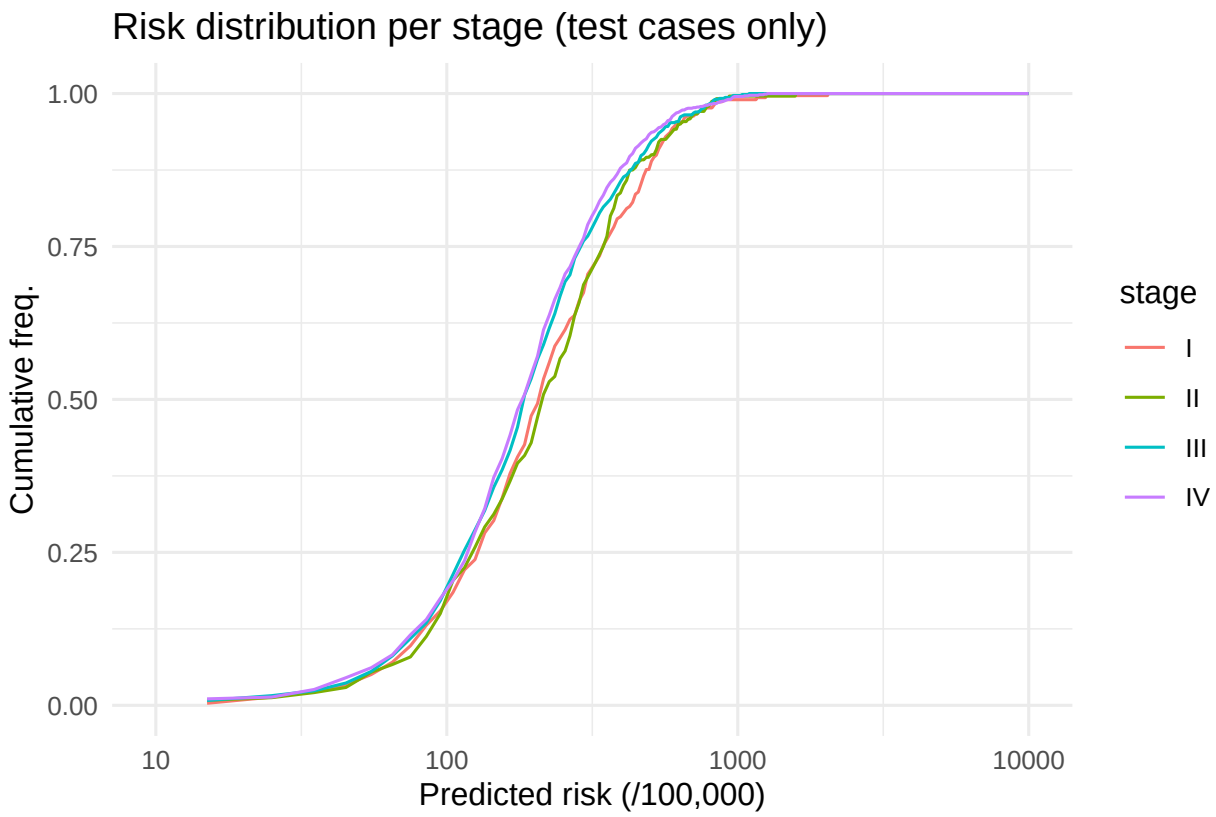



Test ROC stratified by sex
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
Sex
all (AUC: 0.774 [0.766,0.781])
female (AUC: 0.833 [0.773,0.893])
male (AUC: 0.757 [0.75,0.765])

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


Risk distribution per stage (test cases only)
Cumulative freq.
1.00
0.75
0.50
0.25
0.00
10
100
1000
10000
Predicted risk (/100,000)
stage
I
II
III
IV

Years prior

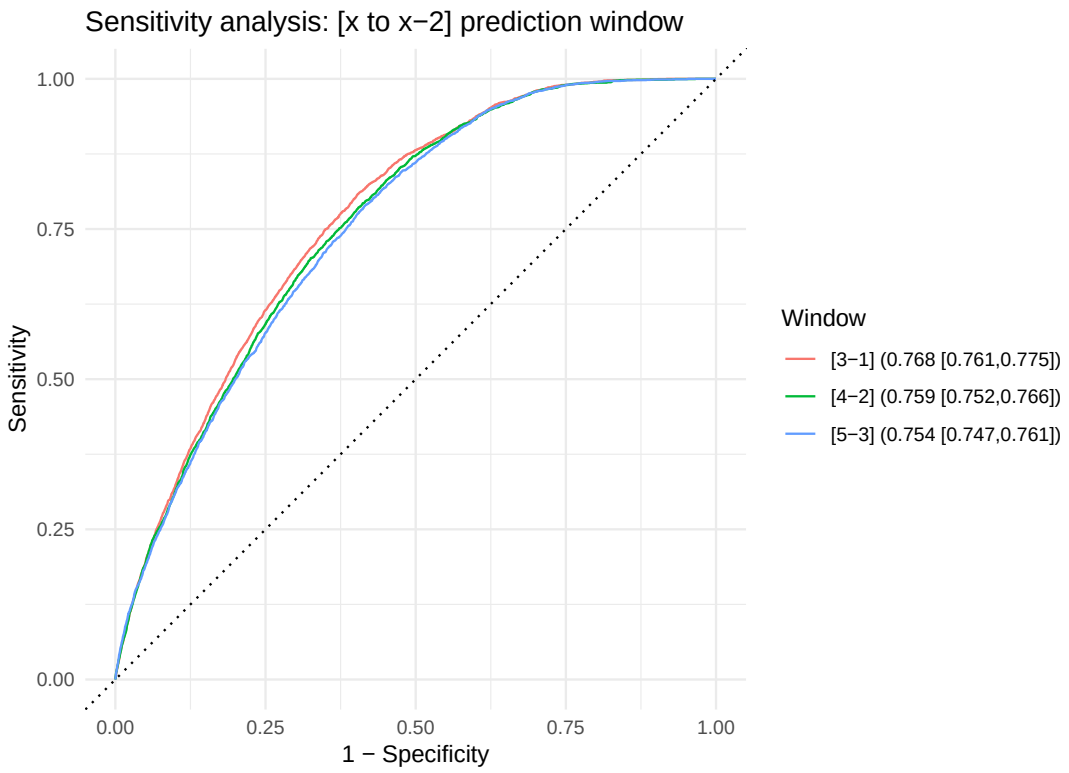


Figure 7: In this experiment, all three test set utilizes the same amount of data (2 years), but we predict farther in the future (1yr, 2yrs or 3yrs). As expected, we do worse and worse as we predict further in time, but only by a small margin. This seem to indicate our predictions are somewhat valid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

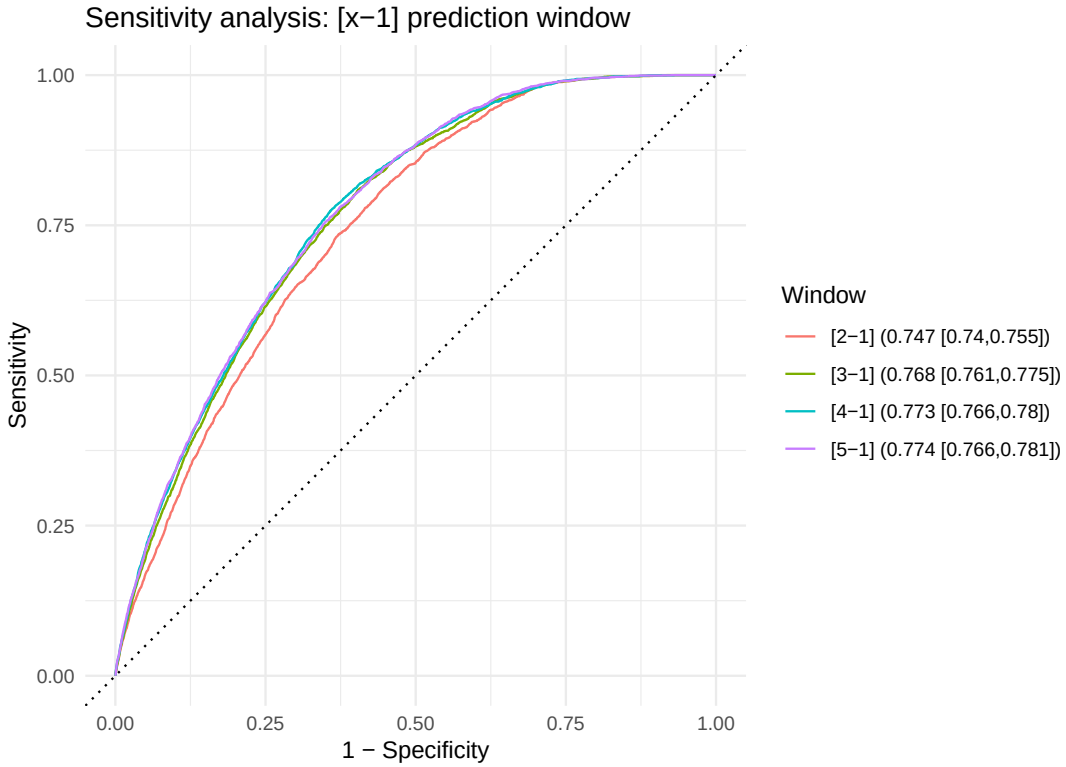


Figure 8: In this experiment, we decrease the amount of data (from 4yrs to 1yr) in the test set, but keep the prediction horizon to 1yr. Performance only seem to decrease when dropping from 2 to 1 year of data.

EAC v EGJAC

- This is now looking a lot better now
- Almost no difference in features
- No difference in missing values (though there still is between Cases and Controls)
-

	Training on ...		
Testing on ...	ANY	EAC	EGJAC
ANY	0.774 [0.766,0.781]	0.768 [0.761,0.776]	0.745 [0.737,0.752]
EAC	0.778 [0.770,0.786]	0.775 [0.767,0.784]	0.747 [0.738,0.755]
EGJAC	0.761 [0.746,0.776]	0.747 [0.732,0.763]	0.738 [0.723,0.754]

Table 6: It seems that testing on EAC yields the best result all the time interestingly.

	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 7: 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.

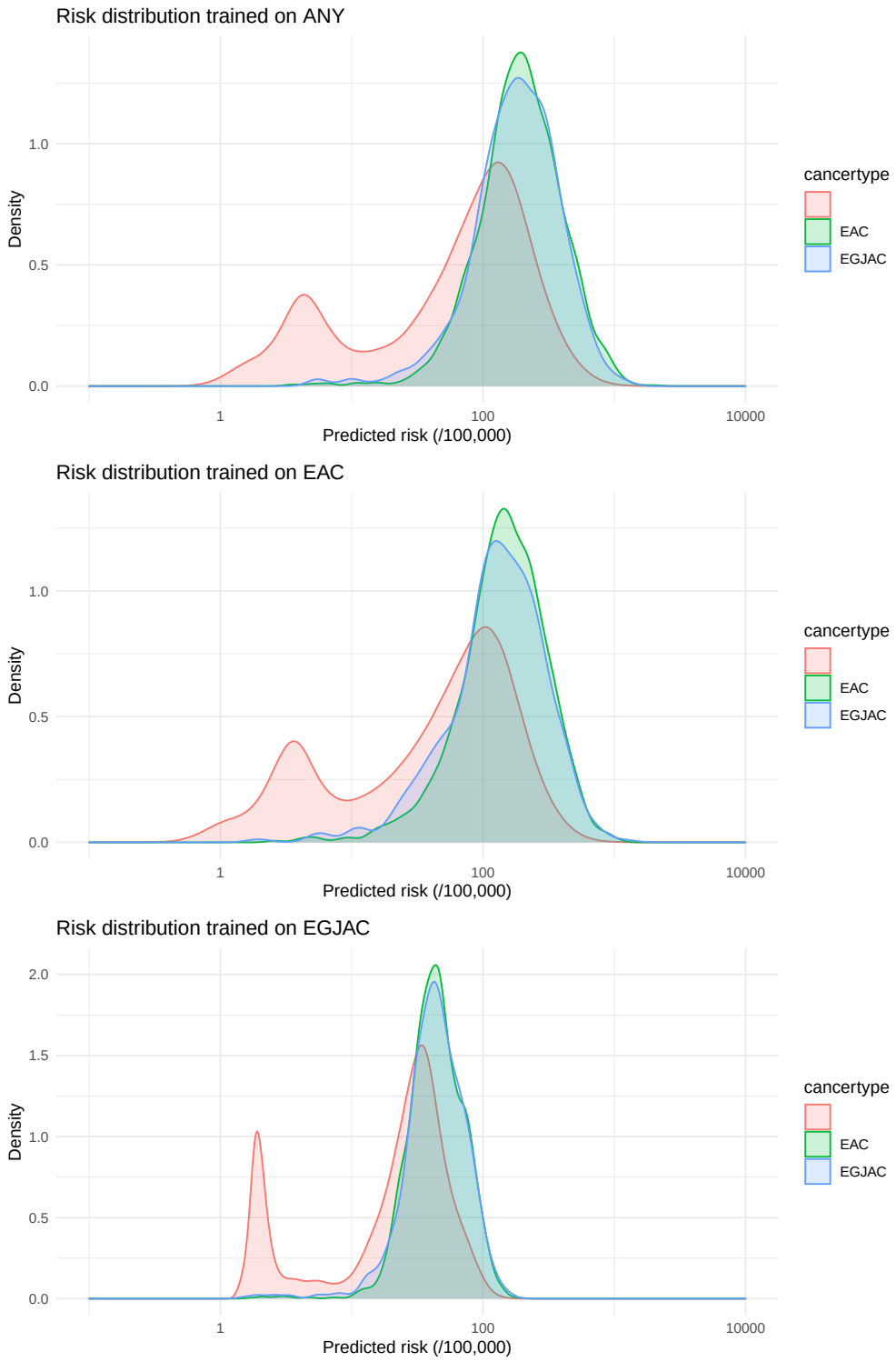


Figure 9: Predicted risk don't go as large as before, there are essentially none above 1,000/100,000. For EGJAC, there is almost none beyond 150/100,000. For EGJAC, the spread is smaller, which indicates less certainty by the model, and this transpires in the lower AUC. For ANY and EAC, the risk distribution for EAC and EGJAC seem very similar. We also note two modes within the cases: risk between 1 and 10 and between 50 and 500.

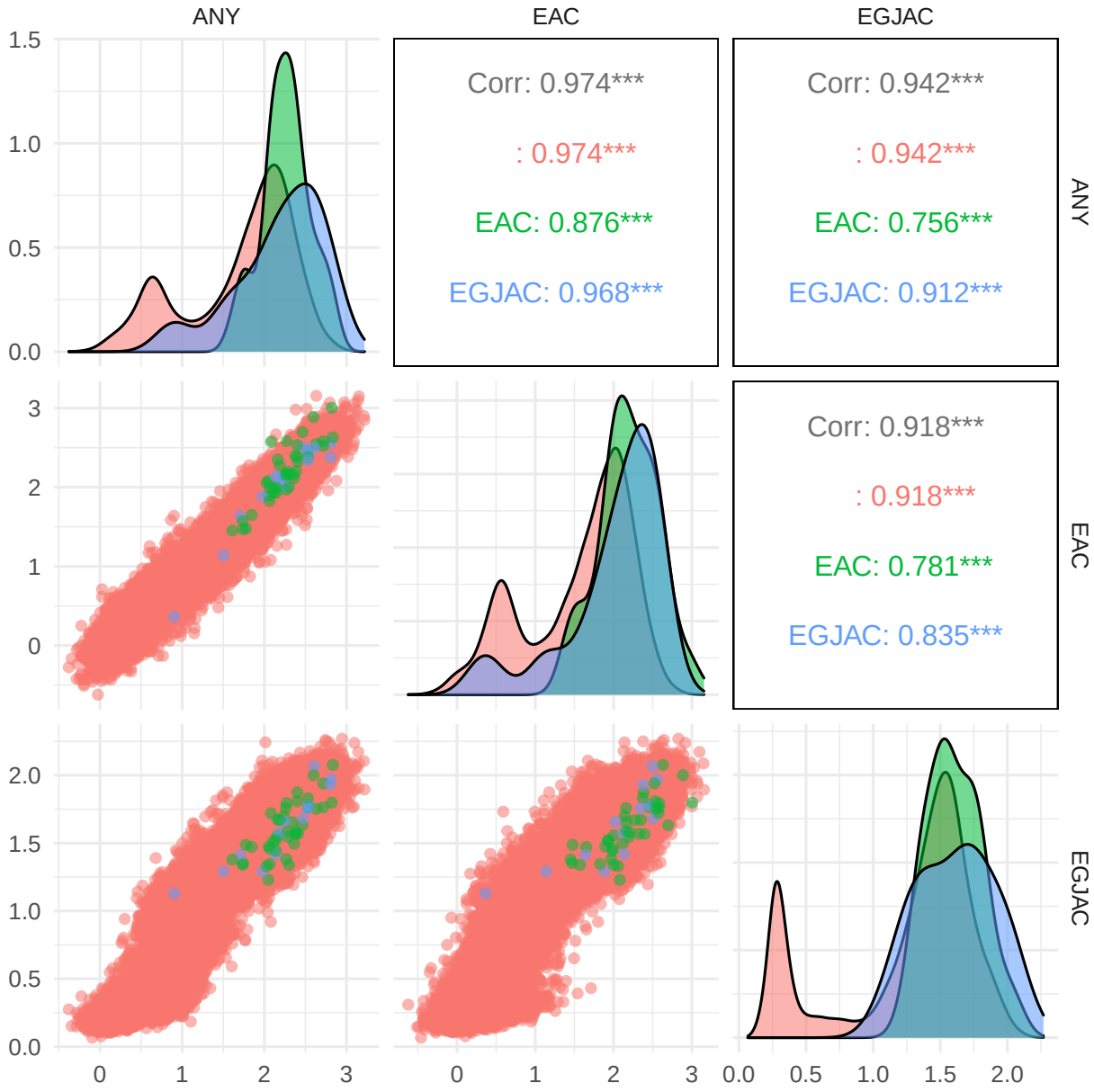


Figure 10: Scatter plots of (log) predicted risk by all three models stratified by case/control status (none, EAC, EGJAC). We have almost perfect correlation between ANY and EAC risks within Controls and EGJAC, while the correlation drops for EAC, indicating that the EAC-only model learns something somewhat different. There is better agreement on controls and EGJAC between EAC and EGJAC models than between ANY and EGJAC models. Within EAC, the highest correlation is between ANY and EAC; within EGJAC, the highest correlation is also between ANY and EGJAC, which seems to indicate our EGJAC is doing worse.

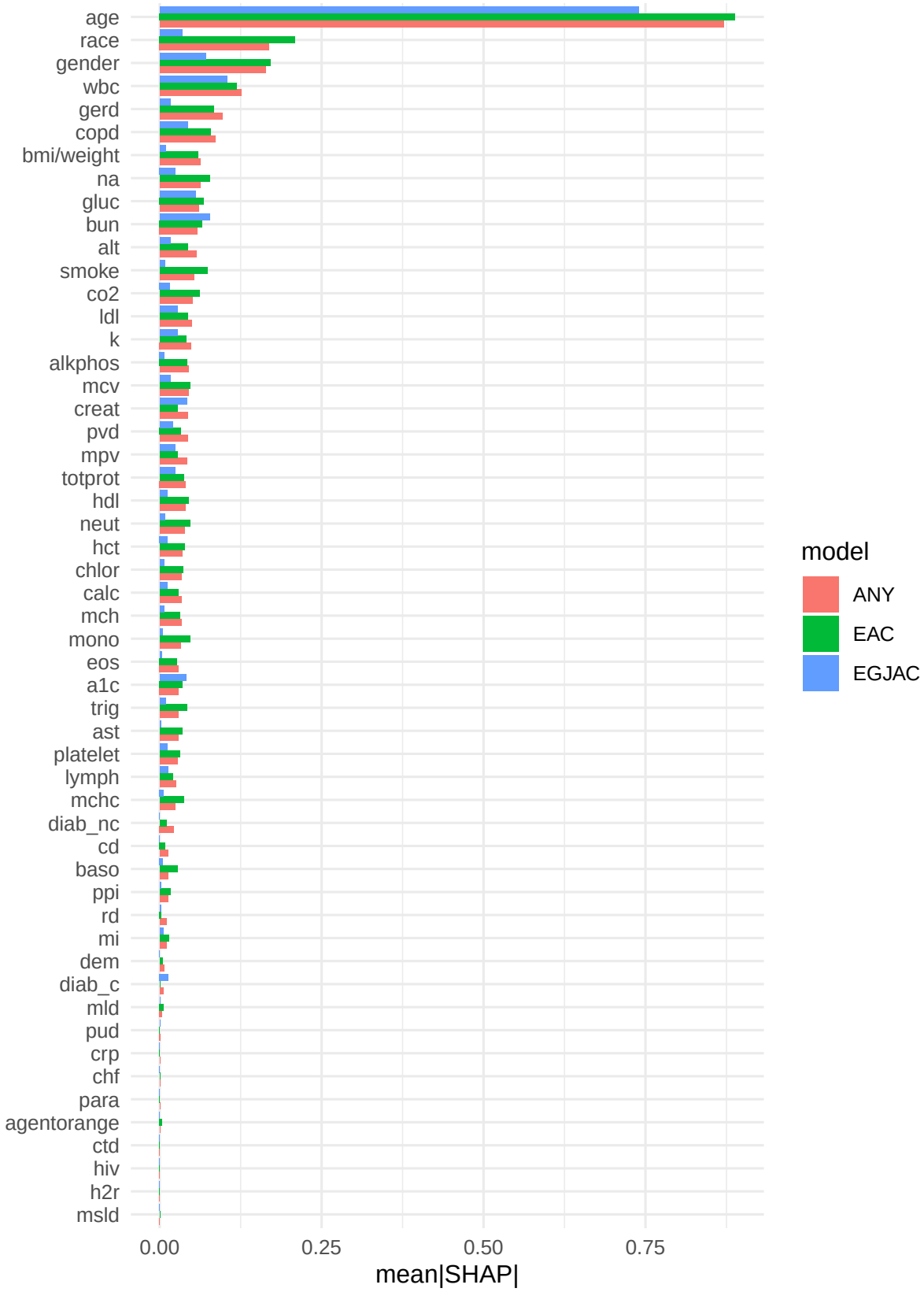


Figure 11: SHAP variable importance wihtin all three models. Notably, we see large differences `age`, `race`, `gender`, `gerd`, `bmi/weight`, `smoke`, and a few others.

Prop. Control	Prop. EAC	Prop. EGJAC	P-value (adj.)

Table 8: Features with different proportions of non-NAs between EAC and EGJAC (at 5% level, BH). There are None!

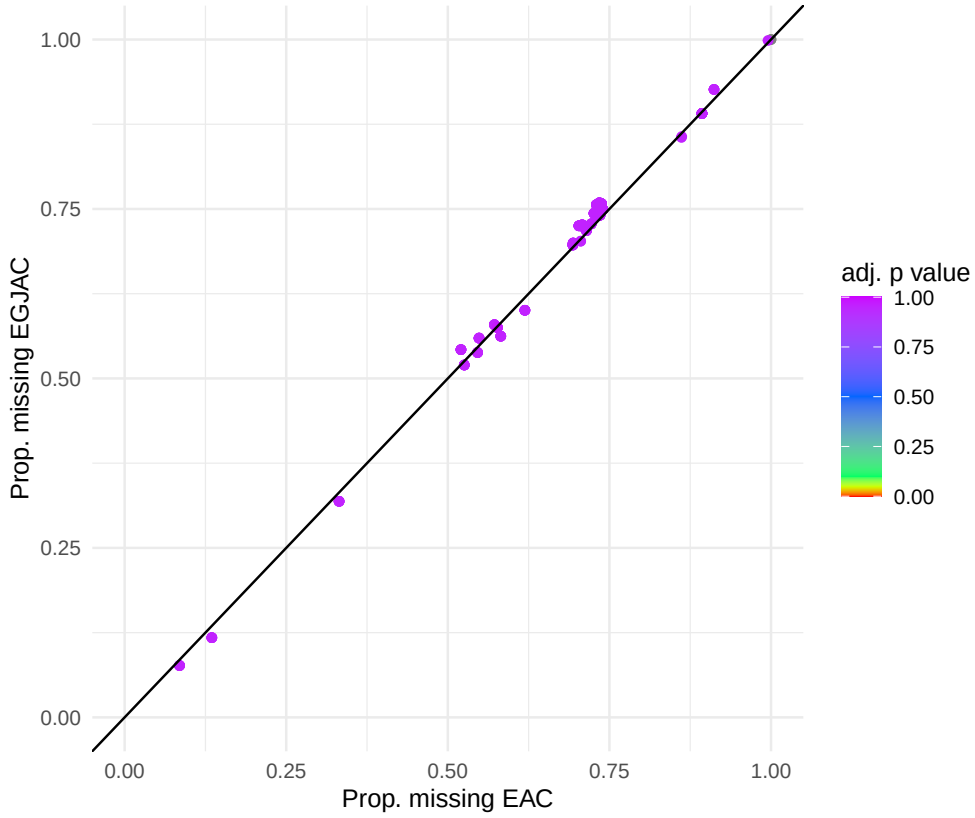


Figure 12: No issues with missing proportions!

EAC Thrashold

Threshold	TPR	PPV	Det. prev.
5	99.65	0.13	82.06
10	99.05	0.14	75.94
15	98.41	0.15	72.68
20	97.60	0.15	69.72
25	96.65	0.16	66.87
30	95.52	0.16	64.09
35	94.17	0.17	61.37
40	92.69	0.17	58.70
45	91.35	0.18	56.08
50	89.62	0.18	53.50
55	87.96	0.19	51.01
60	86.30	0.20	48.54
65	84.36	0.20	46.15
70	82.94	0.21	43.83
75	80.86	0.21	41.57
80	79.24	0.22	39.37
85	77.01	0.23	37.23
90	74.89	0.23	35.18
95	72.21	0.24	33.19
100	69.95	0.25	31.29
105	67.55	0.25	29.46
110	65.22	0.26	27.73
115	62.99	0.27	26.06
120	60.49	0.27	24.50
125	58.40	0.28	23.02
130	56.00	0.29	21.64
135	53.67	0.29	20.31
140	51.69	0.30	19.07
145	49.68	0.31	17.91
150	47.88	0.31	16.81
155	45.87	0.32	15.80
160	44.53	0.33	14.83
165	42.16	0.33	13.94
170	40.22	0.34	13.10
175	38.95	0.35	12.33
180	37.75	0.36	11.60
185	36.48	0.37	10.93
190	35.45	0.38	10.30
195	34.25	0.39	9.70
200	32.27	0.39	9.14

Threshold	TPR	PPV	Det. prev.
210	29.94	0.41	8.12
220	27.72	0.42	7.24
230	25.53	0.44	6.47
240	23.52	0.45	5.79
250	21.15	0.45	5.18
260	19.56	0.46	4.65
270	18.26	0.48	4.18
280	17.20	0.50	3.76
290	15.75	0.51	3.39
300	14.30	0.52	3.06
325	12.04	0.56	2.38
350	9.96	0.59	1.87
375	8.02	0.60	1.47
400	6.64	0.63	1.17
425	5.58	0.66	0.93
450	4.27	0.63	0.75
475	3.71	0.68	0.60
500	2.75	0.63	0.49
600	1.20	0.61	0.22
700	0.85	0.91	0.10
800	0.42	0.91	0.05
900	0.18	0.70	0.03
1000	0.11	0.73	0.02
1100	0.04	0.42	0.01
1200	0.04	0.61	0.01
1300	0.00	0.00	0.00
1400	0.00	0.00	0.00
1500	0.00	0.00	0.00
1600	0.00	0.00	0.00
1700	0.00	0.00	0.00
1800	0.00	0.00	0.00
1900	0.00	0.00	0.00
2000	0.00	0.00	0.00
3000	0.00	0.00	0.00
4000	0.00	0.00	0.00
5000	0.00	0.00	0.00
6000	0.00	0.00	0.00
7000	0.00	0.00	0.00
8000	0.00	0.00	0.00
10000	0.00	0.00	0.00

EGJAC Thrashold

Threshold	TPR	PPV	Det. prev.
5	99.39	0.13	80.54
10	99.04	0.14	77.35
15	97.83	0.15	72.34
20	95.26	0.16	64.80
25	89.02	0.18	55.36
30	80.54	0.20	44.87
35	69.53	0.22	34.55
40	58.13	0.25	25.80
45	48.11	0.27	19.27
50	38.28	0.29	14.62
55	31.68	0.31	11.26
60	26.05	0.33	8.76
65	21.70	0.35	6.81
70	17.64	0.37	5.26
75	13.86	0.38	4.02
80	10.91	0.39	3.02
85	7.95	0.38	2.26
90	6.41	0.42	1.66
95	4.78	0.43	1.20
100	3.53	0.44	0.87
105	2.46	0.43	0.62
110	1.96	0.49	0.44
115	1.43	0.51	0.31
120	1.00	0.50	0.22
125	0.86	0.60	0.16
130	0.57	0.56	0.11
135	0.46	0.65	0.08
140	0.36	0.70	0.06
145	0.21	0.62	0.04
150	0.14	0.56	0.03
155	0.11	0.60	0.02
160	0.11	0.87	0.01
165	0.07	0.81	0.01
170	0.04	0.62	0.01
175	0.04	1.03	0.00
180	0.04	1.59	0.00
185	0.04	2.50	0.00
190	0.00	0.00	0.00
195	0.00	0.00	0.00
200	0.00	0.00	0.00

Threshold	TPR	PPV	Det. prev.
210	0.00	0.00	0.00
220	0.00	0.00	0.00
230	0.00		0.00
240	0.00		0.00
250	0.00		0.00
260	0.00		0.00
270	0.00		0.00
280	0.00		0.00
290	0.00		0.00
300	0.00		0.00
325	0.00		0.00
350	0.00		0.00
375	0.00		0.00
400	0.00		0.00
425	0.00		0.00
450	0.00		0.00
475	0.00		0.00
500	0.00		0.00
600	0.00		0.00
700	0.00		0.00
800	0.00		0.00
900	0.00		0.00
1000	0.00		0.00
1100	0.00		0.00
1200	0.00		0.00
1300	0.00		0.00
1400	0.00		0.00
1500	0.00		0.00
1600	0.00		0.00
1700	0.00		0.00
1800	0.00		0.00
1900	0.00		0.00
2000	0.00		0.00
3000	0.00		0.00
4000	0.00		0.00
5000	0.00		0.00
6000	0.00		0.00
7000	0.00		0.00
8000	0.00		0.00
10000	0.00		0.00

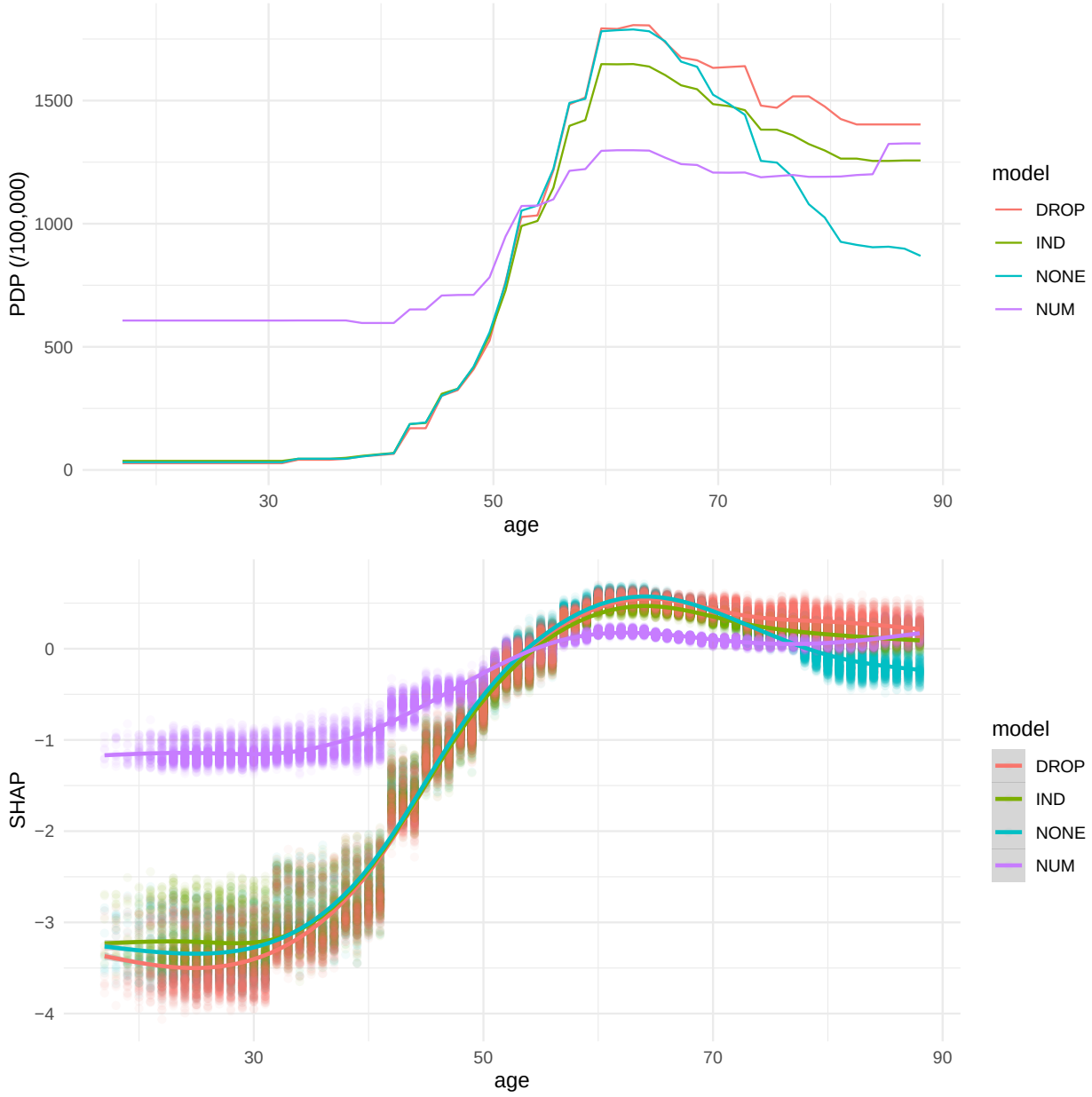
Data comparison

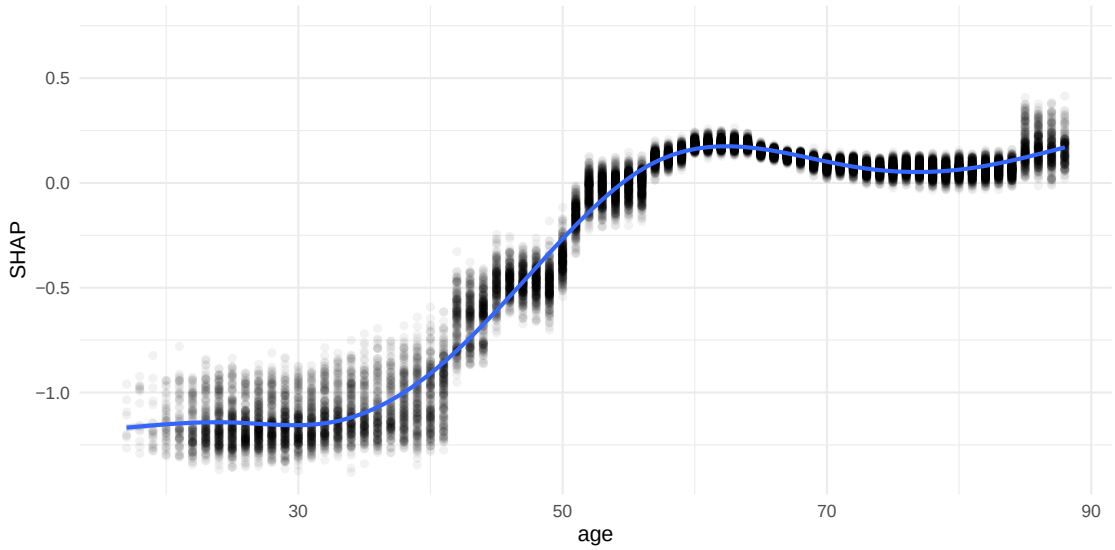
variable	cancertype	old	new	diff	Is diff. (%)	NA-NA (%)	Obs-NA(%)	NA-Obs.(%)
alkphos	EAC	86.02	86.14	0.00	0.43	42.04	0.07	10.94
	EGJAC	85.55	85.83	0.01	0.47	44.41	0.06	11.99
	All	85.89	86.06	0.00	0.44	42.70	0.06	11.23
alt	EAC	32.52	32.56	0.00	0.52	36.34	0.08	11.59
	EGJAC	31.86	31.52	-0.01	0.36	38.57	0.08	12.91
	All	32.34	32.28	-0.00	0.48	36.96	0.08	11.96
ast	EAC	30.00	30.00	-0.01	0.36	37.63	0.08	11.36
	EGJAC	29.56	29.54	-0.00	0.39	40.38	0.07	12.67
	All	29.88	29.88	-0.01	0.37	38.39	0.08	11.73
totprot	EAC	7.01	7.01	-0.00	0.62	52.92	0.00	0.00
	EGJAC	7.03	7.03	0.00	0.53	55.71	0.00	0.00
	All	7.02	7.02	0.00	0.59	53.70	0.00	0.00
baso	EAC	0.22	0.22	0.00	0.00	50.35	0.00	0.00
	EGJAC	0.47	0.47	0.00	0.00	49.07	0.00	0.00
	All	0.29	0.29	0.00	0.00	49.99	0.00	0.00
eos	EAC	0.85	0.85	0.00	0.00	47.63	0.00	0.00
	EGJAC	1.30	1.30	0.00	0.00	47.80	0.00	0.00
	All	0.97	0.97	0.00	0.00	47.68	0.00	0.00
hct	EAC	39.84	39.84	-0.00	1.96	20.70	0.21	0.36
	EGJAC	40.80	39.22	0.03	1.88	19.74	1.46	78.37
	All	39.84	39.67	0.00	1.96	20.43	0.56	22.08
hgb	EAC	13.54	13.54	0.00	0.00	20.65	0.00	0.00
	EGJAC	13.30	13.30	0.00	0.00	22.33	0.00	0.00
	All	13.47	13.47	0.00	0.00	21.12	0.00	0.00
lymph	EAC	3.45	3.43	-0.00	5.99	46.95	0.17	0.28
	EGJAC	2.52	3.76	-0.12	0.60	45.68	1.06	52.98
	All	3.44	3.52	-0.01	5.98	46.60	0.41	14.95
mch	EAC	30.62	30.63	0.00	2.83	20.17	0.20	0.39
	EGJAC	31.06	30.45	-0.00	1.14	19.94	1.45	78.19
	All	30.63	30.58	0.00	2.83	20.10	0.55	22.05

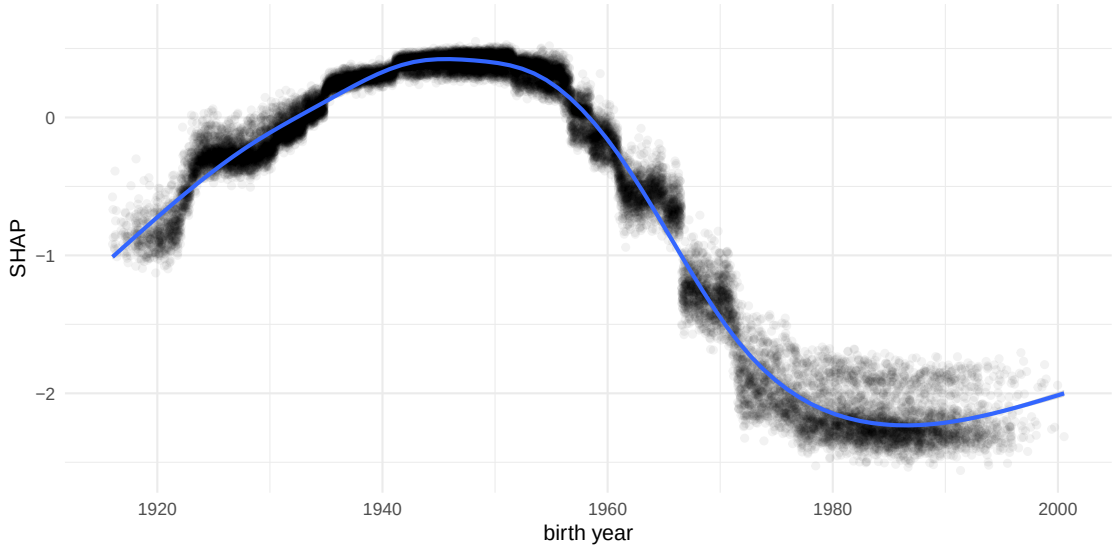
variable	cancertype	old	new	diff	Is diff. (%)	NA-NA (%)	Obs-NA(%)	NA-Obs.(%)
mchc	EAC	33.65	33.65	0.00	0.00	22.65	0.00	0.00
	EGJAC	33.51	33.51	0.00	0.00	24.23	0.00	0.00
	All	33.61	33.61	0.00	0.00	23.09	0.00	0.00
mcv	EAC	90.03	90.03	0.00	0.00	23.02	0.00	0.00
	EGJAC	89.68	89.68	0.00	0.00	24.24	0.00	0.00
	All	89.94	89.94	0.00	0.00	23.36	0.00	0.00
mono	EAC	1.38	1.38	0.00	0.00	50.90	0.00	0.00
	EGJAC	1.44	1.44	0.00	0.00	51.51	0.00	0.00
	All	1.40	1.40	0.00	0.00	51.07	0.00	0.00
mpv	EAC	9.05	9.05	0.00	0.00	38.72	0.00	0.00
	EGJAC	9.19	9.19	0.00	0.00	40.84	0.00	0.00
	All	9.09	9.09	0.00	0.00	39.31	0.00	0.00
neut	EAC	8.54	8.48	-0.06	4.65	52.02	0.00	0.00
	EGJAC	8.55	8.60	0.06	4.94	53.24	0.00	0.00
	All	8.54	8.52	-0.02	4.73	52.36	0.00	0.00
platelet	EAC	236.13	236.13	0.00	0.00	19.80	0.00	0.00
	EGJAC	235.36	235.36	0.00	0.00	21.08	0.00	0.00
	All	235.92	235.92	0.00	0.00	20.16	0.00	0.00
rbc	EAC	4.44	4.44	0.00	0.00	19.34	0.00	0.00
	EGJAC	4.38	4.38	0.00	0.00	20.62	0.00	0.00
	All	4.42	4.42	0.00	0.00	19.69	0.00	0.00
rdw	EAC	15.04	15.04	0.00	0.00	21.32	0.00	0.00
	EGJAC	14.99	14.99	0.00	0.00	23.05	0.00	0.00
	All	15.03	15.03	0.00	0.00	21.80	0.00	0.00
wbc	EAC	8.23	8.23	0.00	0.00	19.75	0.00	0.00
	EGJAC	8.13	8.13	0.00	0.00	20.72	0.00	0.00
	All	8.20	8.20	0.00	0.00	20.02	0.00	0.00

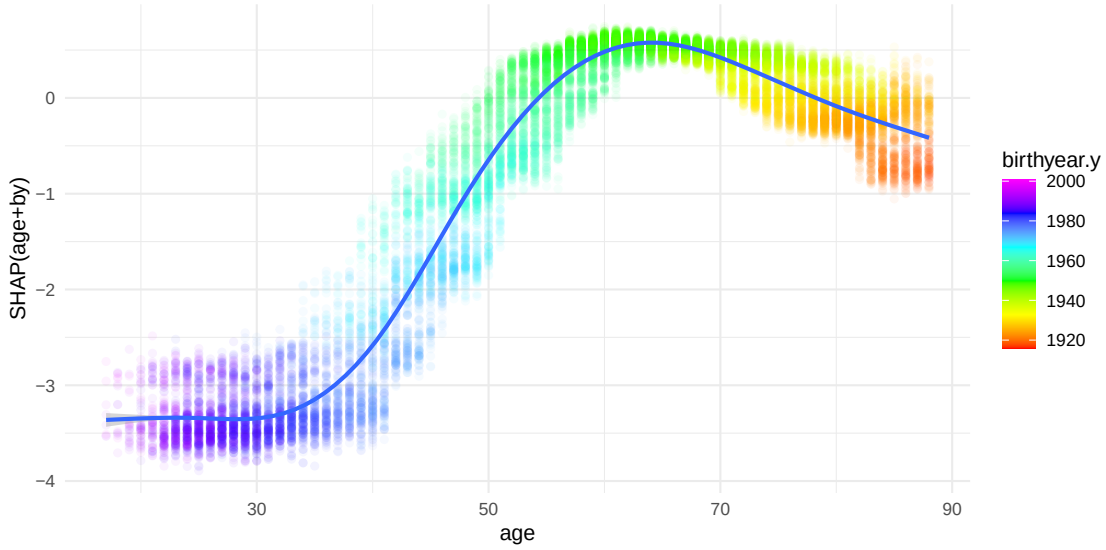
Table 9: Lots of values change for `hct`, `lymph`, `mch` and `neut`. Change of missing status frequent for `alkphos`, `alt`, and `ast`, but equally for both types; very frequent for `hct`, `lymph`, `mch` only in EGJAC.

Birth cohort effect




0.5
0.0
−0.5
−1.0
SHAP
30
50
70
90
age

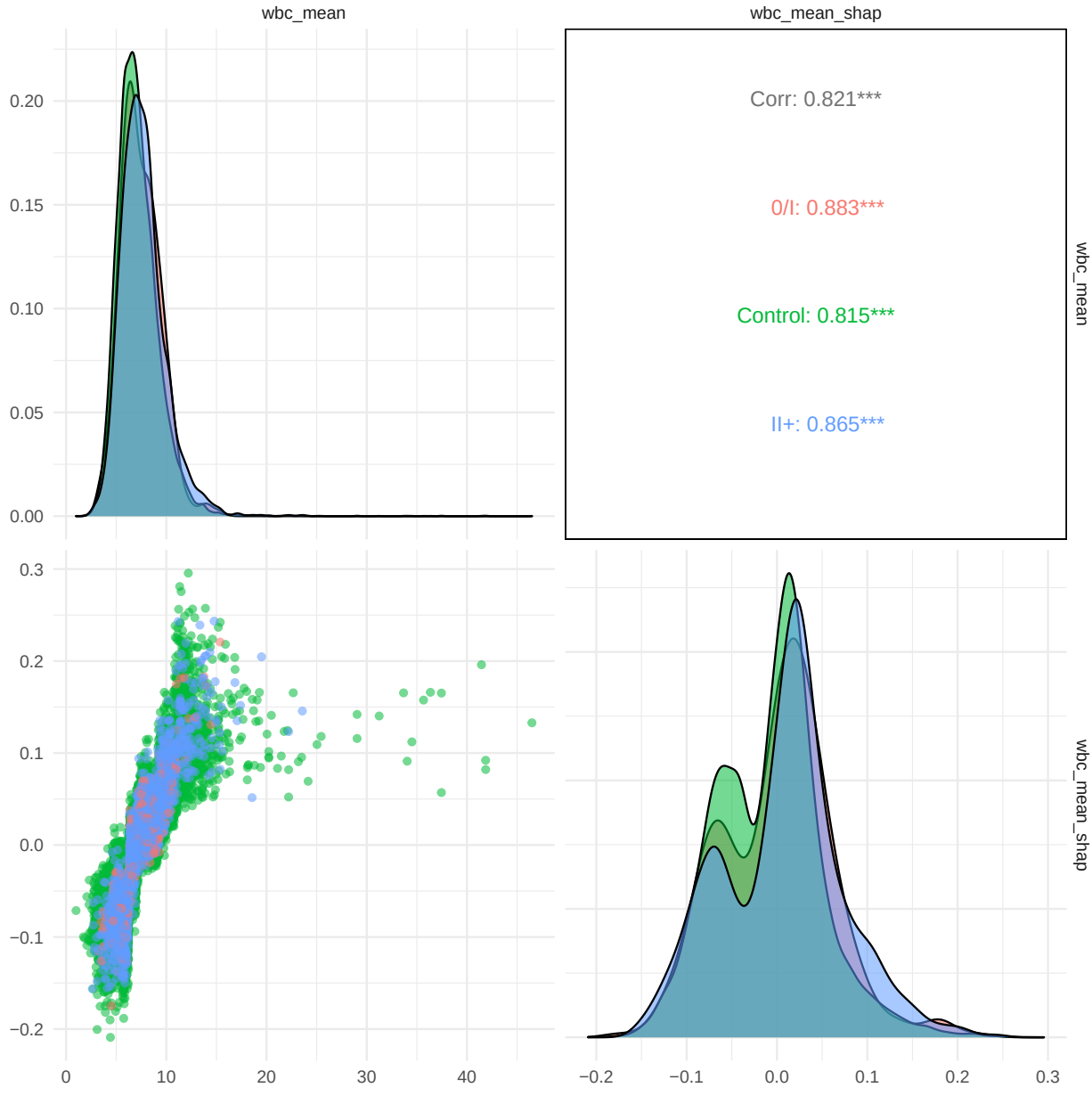

0
−1
−2
SHAP
1920
1940
1960
1980
2000
birth year

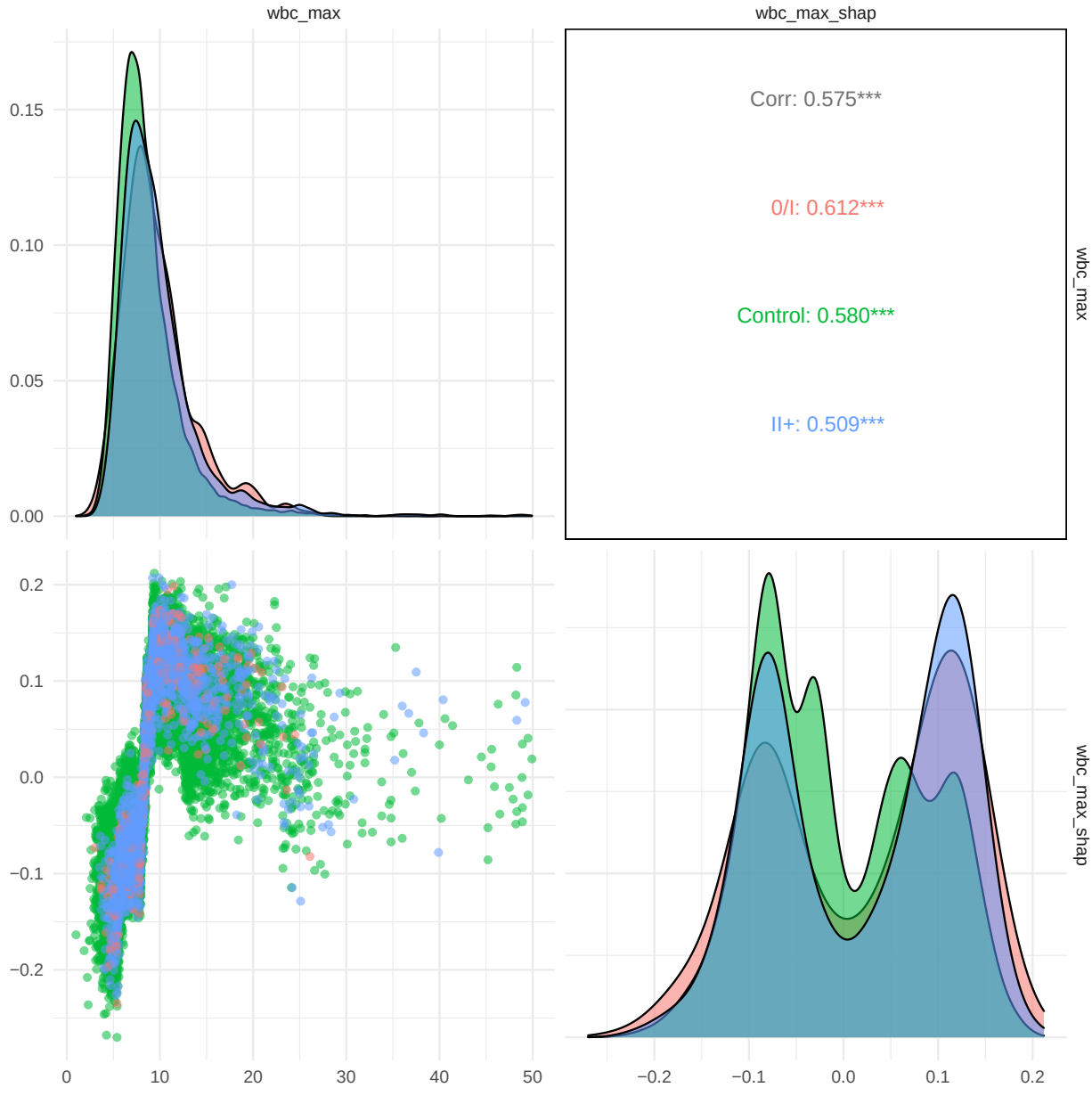

0
−1
−2
−3
−4
SHAP(age+by)
30
50
70
90
age
birthyear.y
2000
1980
1960
1940
1920

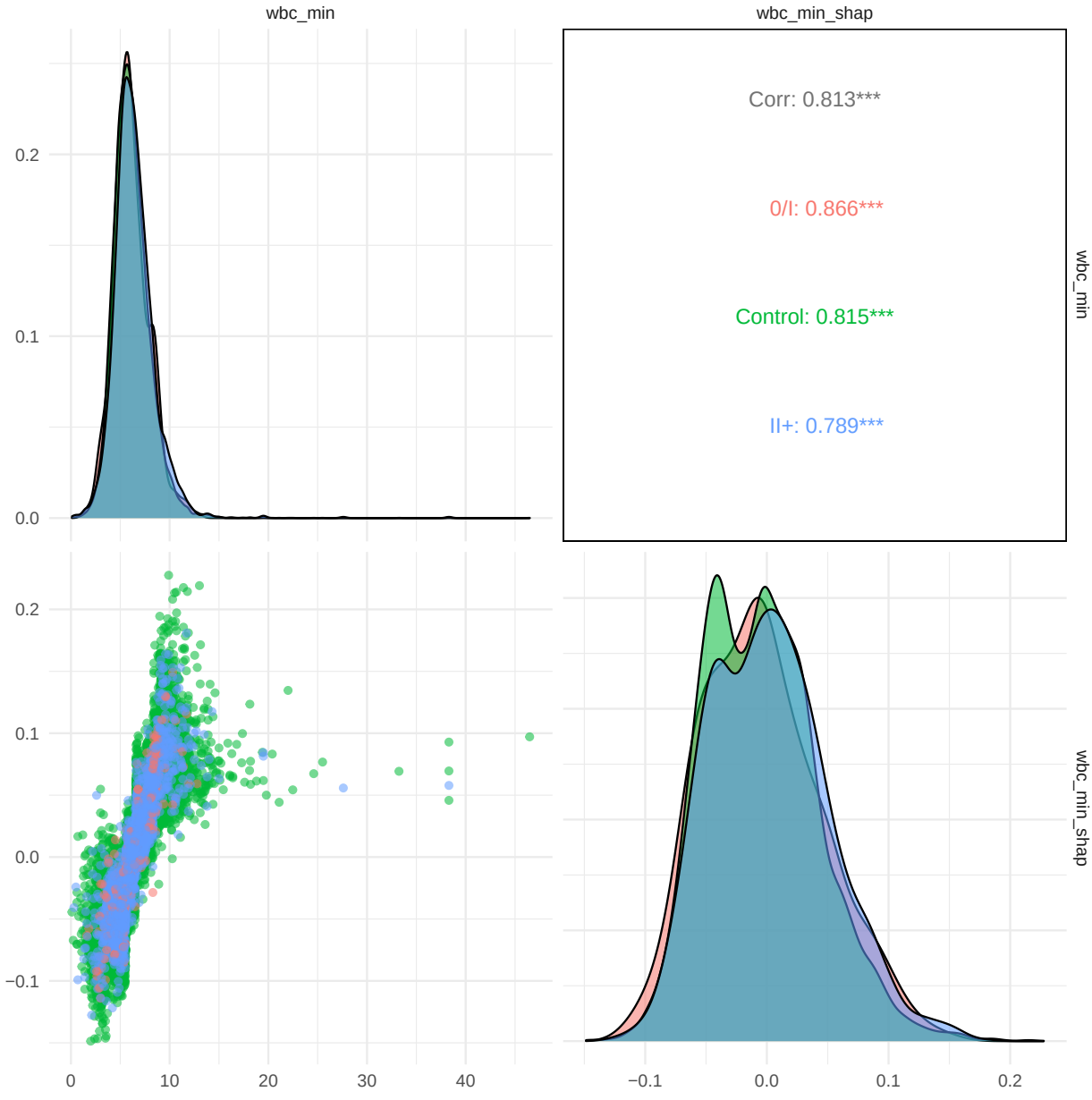
Are we simply diagnosing cancer?

Since WBC is the main lab result influencing predicted risk, we investigate its relationship to cancer stage.

- I decided to compare stage O/1 to II/III/IV
- The main conclusion is that there is a gradient of effect following stage: the higher the stage, the higher the effect of WBC
- I'm not sure this is the best approach to study this yet, but, so far, it doesn't seem too bad


wbc_mean
wbc_mean_shap
Corr: 0.821***
0/I: 0.883***
Control: 0.815***
II+: 0.865***
wbc_mean
wbc_mean_shap
0.20
0.15
0.10
0.05
0.00
0.3
0.2
0.1
0.0
−0.1
−0.2
0
10
20
30
40
−0.2
−0.1
0.0
0.1
0.2
0.3


wbc_max
wbc_max_shap
Corr: 0.575***
0/I: 0.612***
Control: 0.580***
II+: 0.509***
wbc_max
wbc_max_shap
0.15
0.10
0.05
0.00
0.2
0.1
0.0
−0.1
−0.2
0
10
20
30
40
50
−0.2
−0.1
0.0
0.1
0.2


wbc_min
wbc_min_shap
Corr: 0.813***
0/I: 0.866***
Control: 0.815***
II+: 0.789***
wbc_min
wbc_min_shap
0.2
0.1
0.0
−0.1
0
10
20
30
40
−0.1
0.0
0.1
0.2