HOSEA Aim I – ICD 10 Cohort Anlysis

Simon Fontaine

December 16, 2022

Contents

1 Pre-processing & filtering

A few details:

- Years 3 and 2
- Process all patients, filter after
- Impute missing values from training set (SRS)
- Use visitin4yrs to impute 0s in comorbidities

Inclusion criterion

- Not part of training or validation (both cases and controls)
- Exclude cases in the test set from being included again (both cases and control)
- A test control is allowed to become a case or a control with a new index date

Cohort	Nb. patients	Nb. controls	Nb. cases	Cases/100,000	% of cases
		ANY			
Test	2,567,059	2,564,221	2848	110.9	100.0
ICD 10	2,419,331	2,418,685	646	26.7	100.0
		EAC			
Test	2,567,059	2,564,993	2076	80.9	72.9
ICD 10	2,419,331	2,418,872	459	19.0	71.1
		EGJAC			
Test	2,567,059	2,566,297	772	30.1	27.1
ICD 10	2,419,331	2,419,144	187	7.7	28.9

Comments:

- We indeed have a much smaller case incidence rate, but this is mostly due to the smaller span (3(?) years instead of 14 years)
- EAC/EGJAC ratio seems relatively constant

2 Feature distribution

The ICD 10 cohort seems to be:

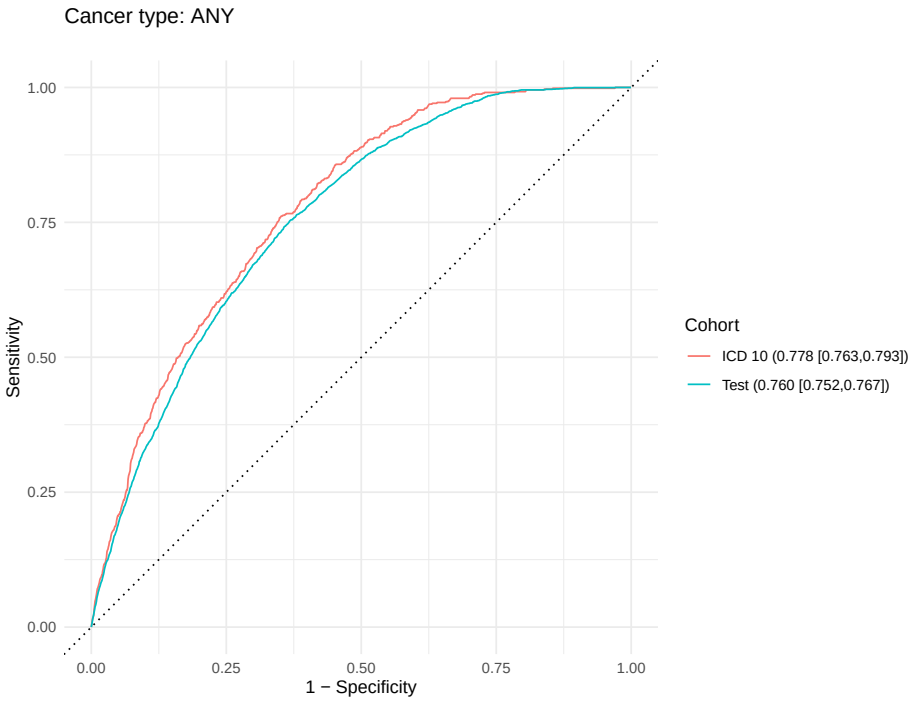
- Younger
- More female
- Fewer white (I only included Black, but the others show same trend)
- More obese (higher BMI & weight)
- Smoking slightly less
- More missing values? (hypothesis: more truncated patients since smaller span?)
- Labs & medication don't show major shifts

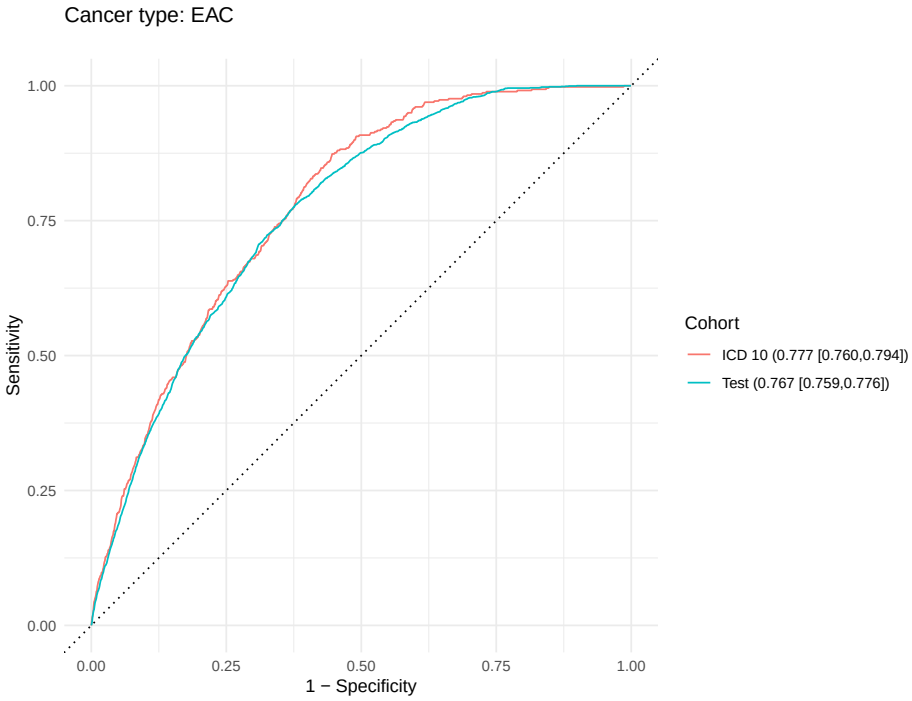
Before imputation **Variable**	**Cohort**	**% missing**	**Mean**
Age	Test [3-1]	0.6	59.58
	ICD 10	3.7	57.17
Sex	Test [3-1]	0.0	91.64
	ICD 10	0.0	86.56
Black	Test [3-1]	13.0	16.86
	ICD 10	12.9	18.99
BMI	Test [3-1]	34.2	29.30
	ICD 10	40.5	29.87
Weight	Test [3-1]	33.8	200.77
	ICD 10	40.1	204.67
Smoking (current)	Test [3-1]	43.0	43.46
	ICD 10	73.8	40.42
Smoking (former)	Test [3-1]	43.0	41.69
	ICD 10	73.8	41.72
COPD	Test [3-1]	15.3	14.34
	ICD 10	21.7	12.53
Gerd	Test [3-1]	15.3	16.85
	ICD 10	21.7	17.11
A few example lab variables			
WBC (Mean)	Test [3-1]	42.7	7.19
	ICD 10	45.9	7.13
Na (Mean)	Test [3-1]	40.4	139.28
	ICD 10	44.3	139.24
K (Mean)	Test [3-1]	40.3	4.27
	ICD 10	44.3	4.24

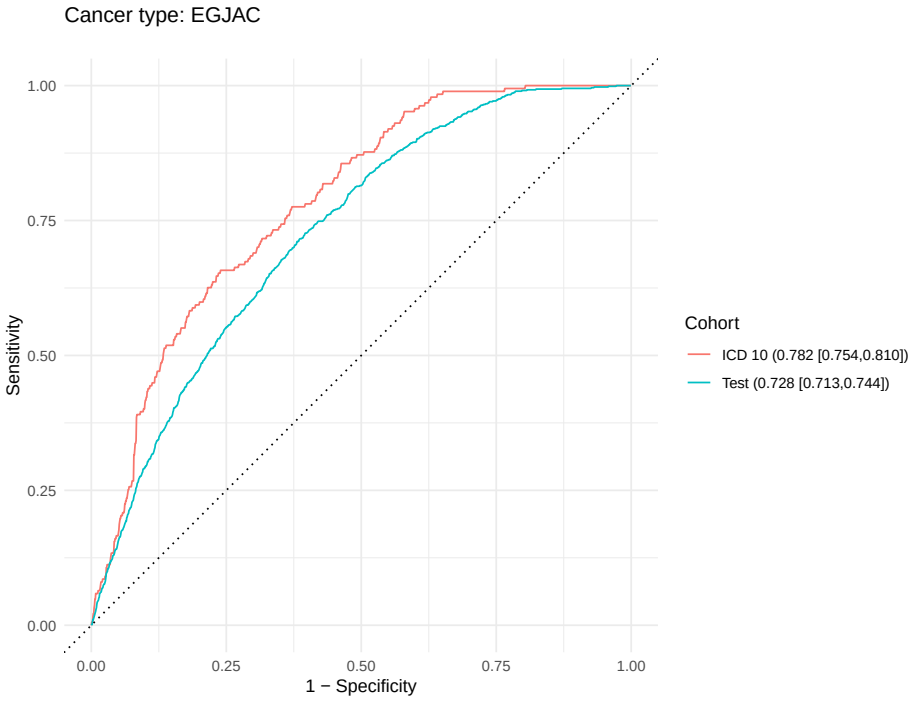
3 Discriminative performance comparison

Comments

- Slightly improved performance for ANY
- No improvement for EAC
- Large improvment for EGJAC



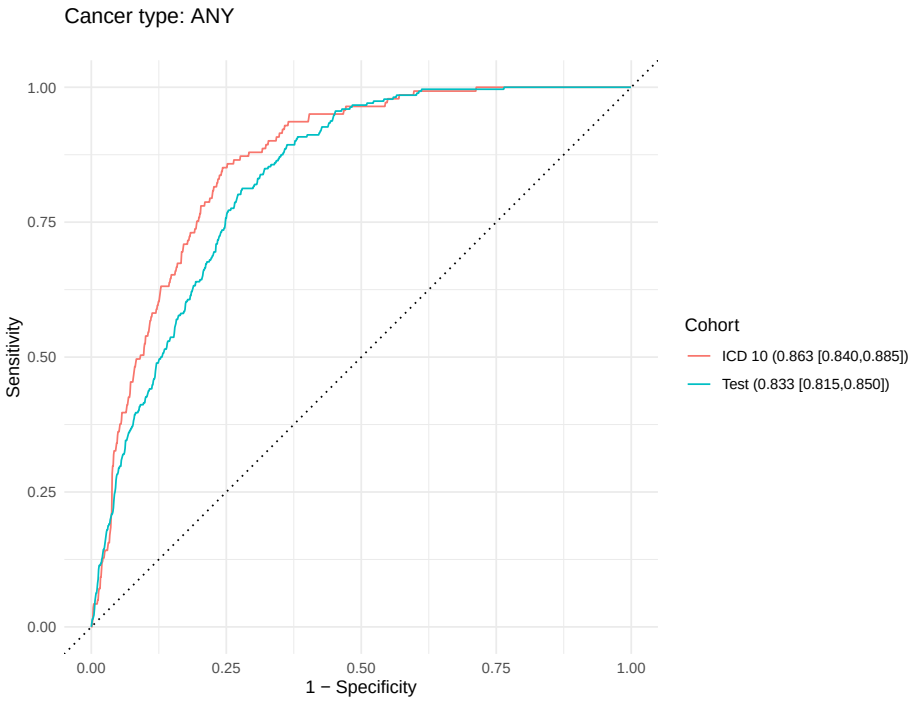

Cancer type: EAC
1.00
0.75
0.50
0.25
0.00
Sensitivity
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Cohort
ICD 10 (0.777 [0.760,0.794])
Test (0.767 [0.759,0.776])

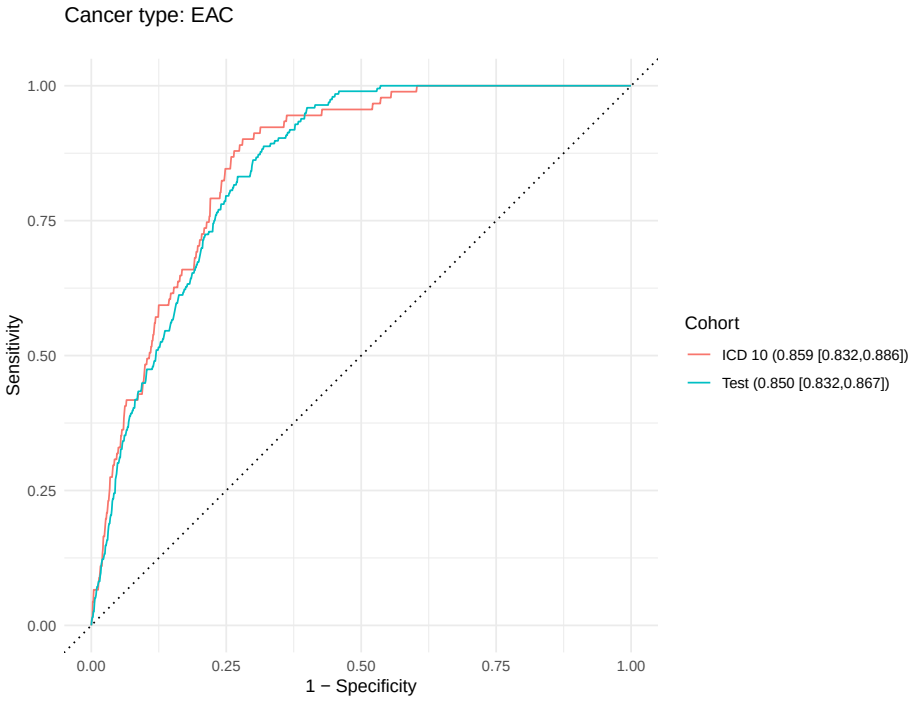

Cancer type: EGJAC
1.00
0.75
0.50
0.25
0.00
Sensitivity
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Cohort
ICD 10 (0.782 [0.754,0.810])
Test (0.728 [0.713,0.744])

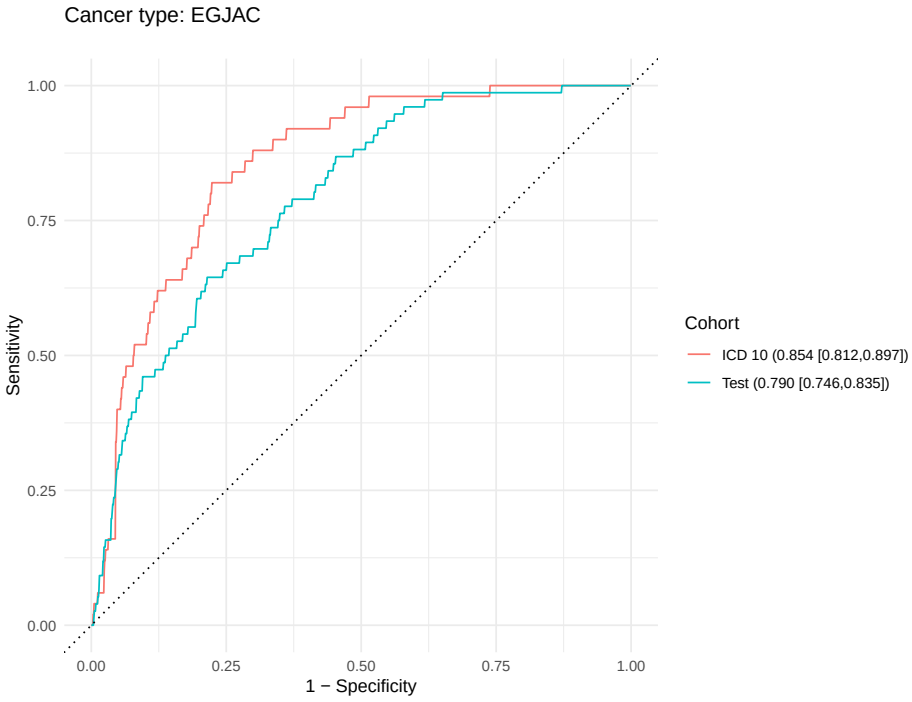
4 Discriminative performance comparison on a representative sample

Comments

- Subsample men to 1:1 ratio (not adjusting to case incidence)
- All AUCs are increased as was previosuly observed for representative samples
- Similar comparison as above



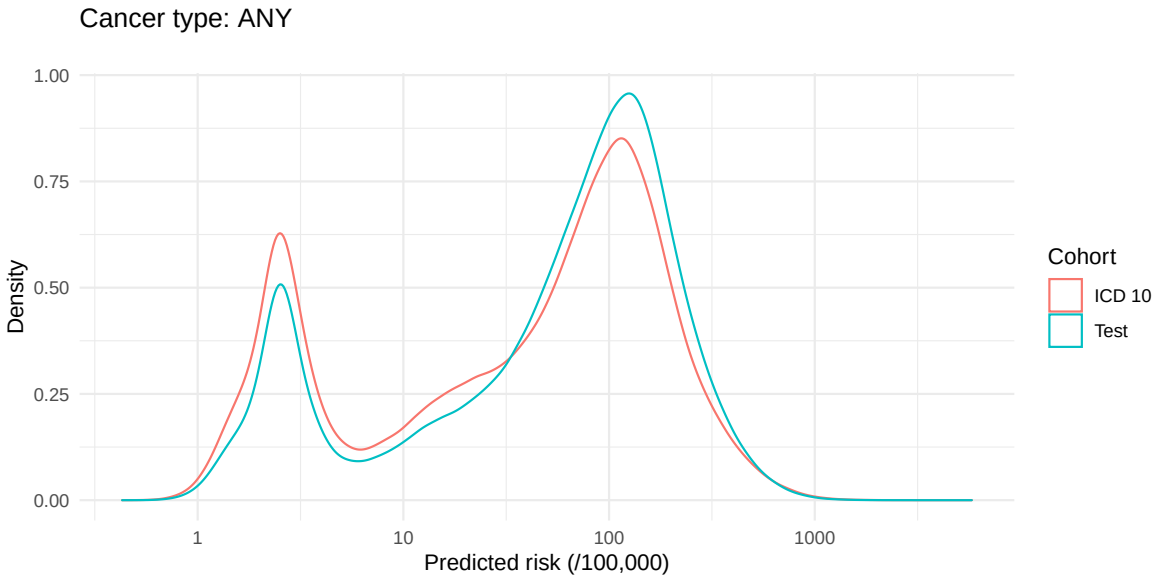

Cancer type: EAC
1.00
0.75
0.50
0.25
0.00
Sensitivity
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Cohort
ICD 10 (0.859 [0.832,0.886])
Test (0.850 [0.832,0.867])

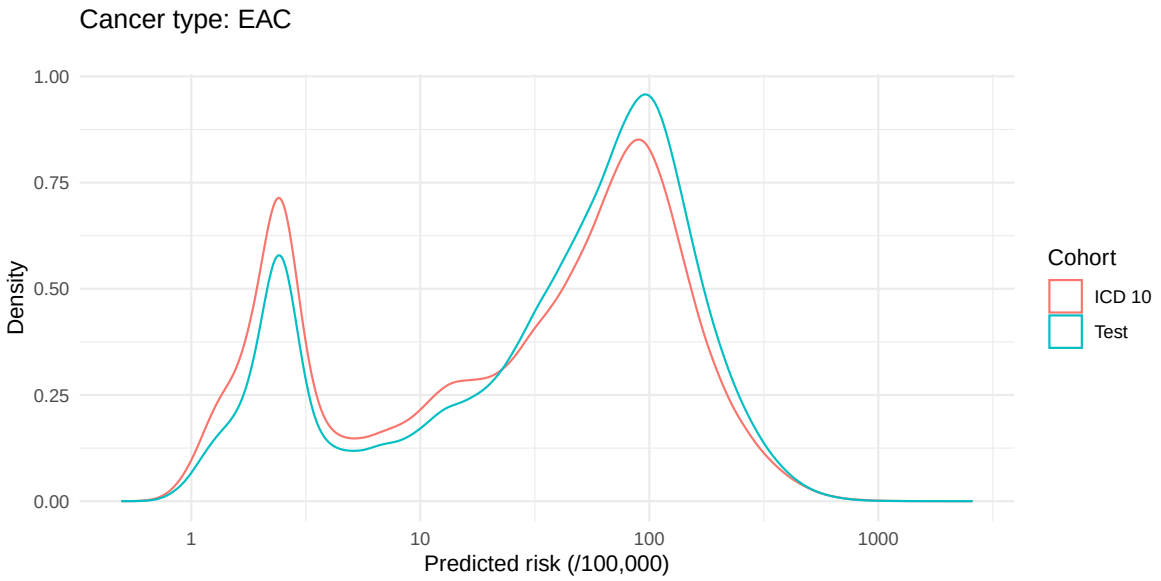

Cancer type: EGJAC
1.00
0.75
0.50
0.25
0.00
Sensitivity
0.00
0.25
0.50
0.75
1.00
1 − Specificity
Cohort
ICD 10 (0.854 [0.812,0.897])
Test (0.790 [0.746,0.835])

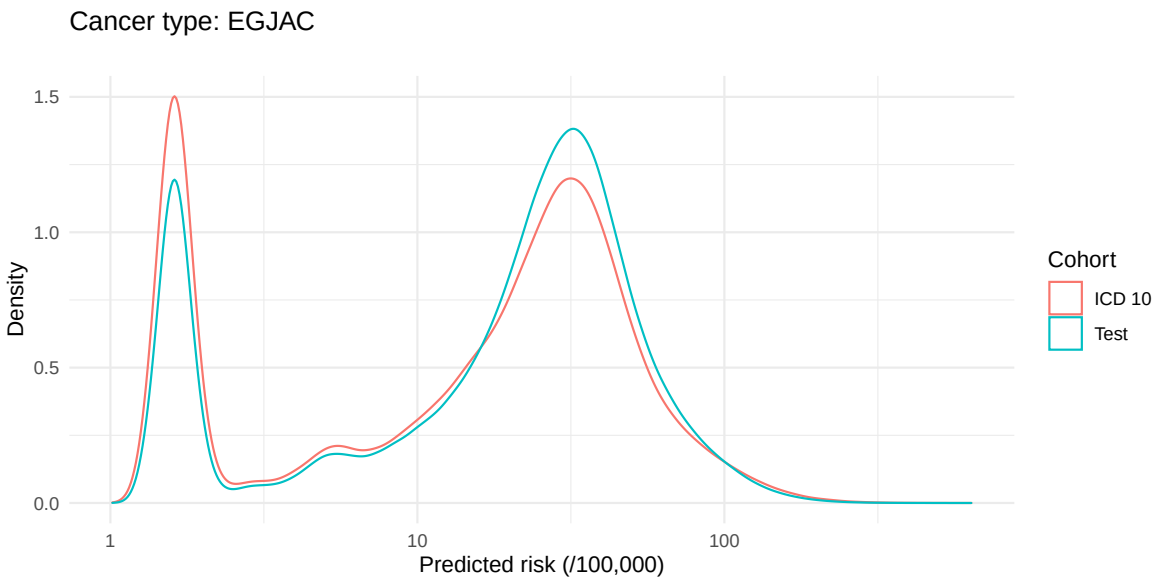
5 Risk distribution comparison

Comments

- All three show the exact same trend:
- Same two peaks, but the higher risk peak is smaller for ICD 10
- Possibly due to the distributional shift; i.e., fewer high-risk patients (fewer men, younger on average)






Cancer type: EGJAC
1.5
1.0
0.5
0.0
Density
1
10
100
Predicted risk (/100,000)
Cohort
ICD 10
Test

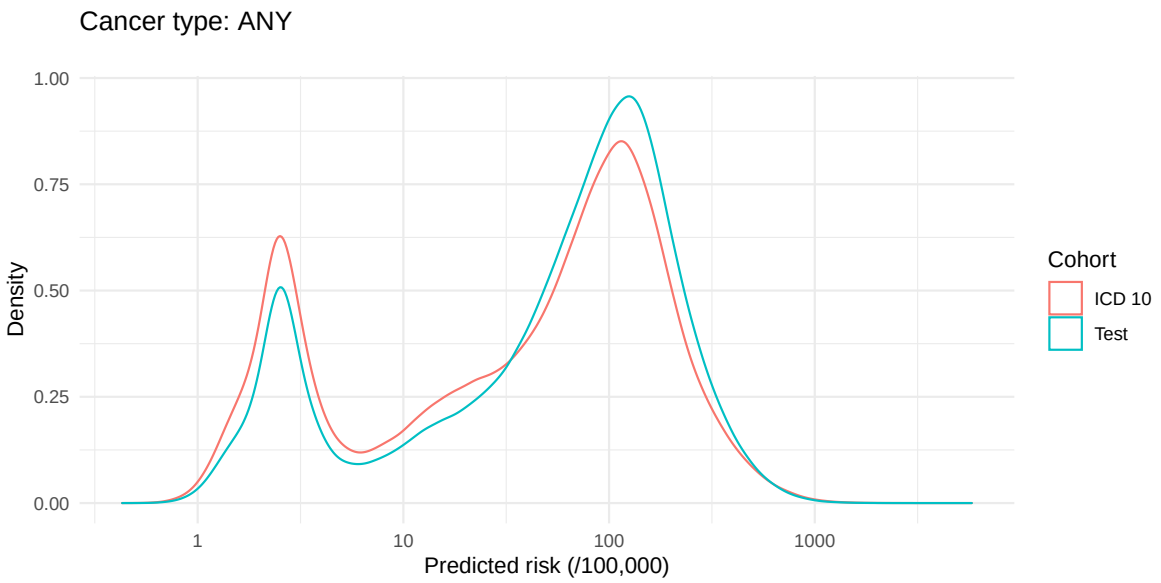
To understand the disparity better, we stratify by age and sex:

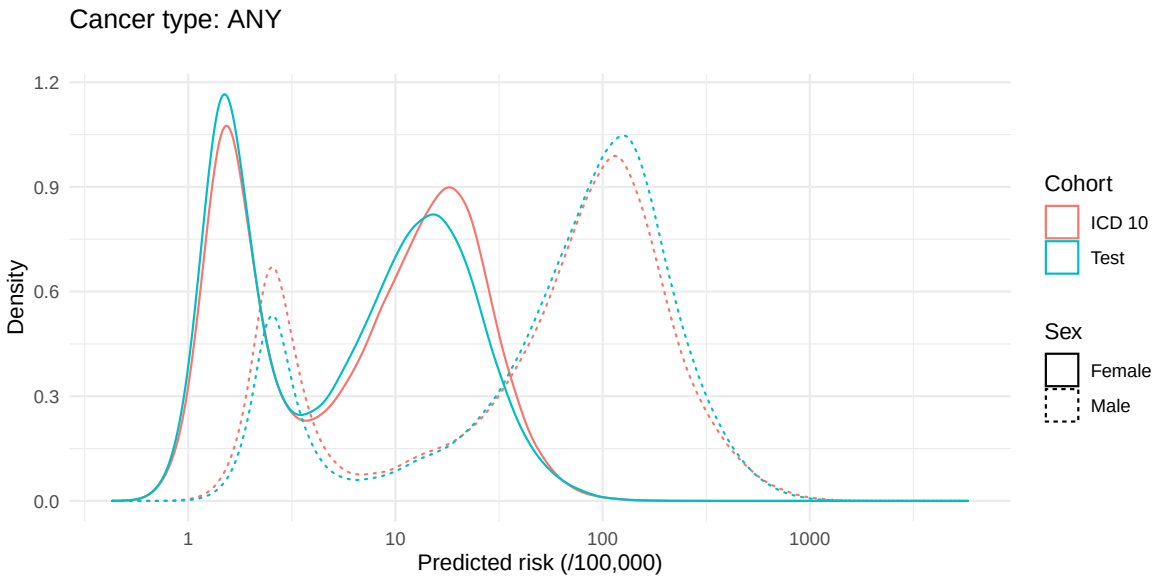
	Sex		
Cohort	Female	Male	**Ratio M:F**
Age 0-35			
ICD 10	89844	311775	3.5:1
Test	64313	268890	4.2:1
Age 35-50			
ICD 10	96179	344122	3.6:1
Test	63560	311452	4.9:1
Age 50-100			
ICD 10	154829	1457904	9.4:1
Test	86699	1772155	20.5:1

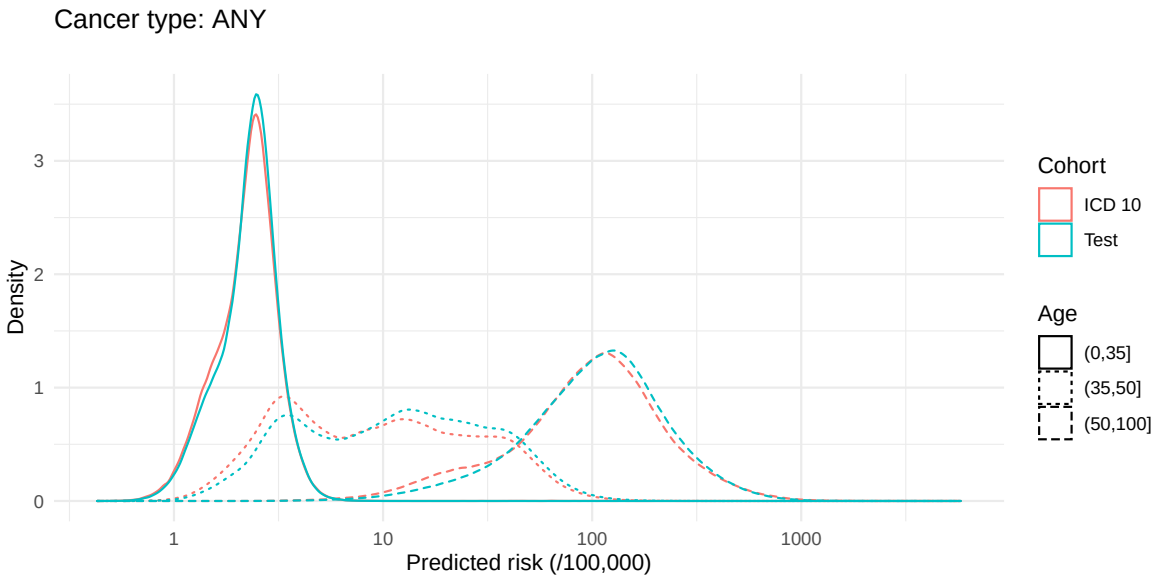
Average age		
	Sex	
Cohort	Male	Female
ICD 10	58.7	47.9
Test	60.8	46.2

Some observations:

- Females increased their average age; males decreased
- The biggest shift in proportion of females in the older patients (say, >50 yo)
- Females have an increase in the second "bump"; Males rather have a decrease. This seem to match with the change in ratios: the largest change in ratios is in the older, more at risk, patients
- Stratifying by age seem to completely remove the change in risk distribution. Thus, the conclusion seems to be that the shift in risk distribution is mostly explained by the shift in age: in particular, since males are by far the most frequent, the decrease in average in men explains most of the shift, even though females have the opposite effect


Cancer type: ANY
1.00
0.75
0.50
0.25
0.00
Density
1
10
100
1000
Predicted risk (/100,000)
Cohort
ICD 10
Test


Cancer type: ANY
1.2
0.9
0.6
0.3
0.0
Density
1
10
100
1000
Predicted risk (/100,000)
Cohort
ICD 10
Test
Sex
Female
Male


Cancer type: ANY
Density
0
1
2
3
1
10
100
1000
Predicted risk (/100,000)
Cohort
ICD 10
Test
Age
(0,35]
(35,50]
(50,100]