Pathfinding clinically useful early diagnostics

Michael N Kammer, PhD

2024-09-07

Contents

1 Introduction

This document is part of a course from the IASLC World Conference on Lung Cancer 2024 in San Diego, presented by the Early Detection Research Network. It aims to guide users through the analysis of the added value of biomarkers in diagnostic models for cancer.

The code to reproduce this pdf is available at https://github.com/mnkammer/wclc-workshop

The included dataset is derived from "Integrated Biomarkers for the Management of Indeterminate Pulmonary Nodules," American Journal of Respiratory and Critical Care Medicine, 2021.

The data included is derived from 4 cohorts (described in the above paper), alloweing you to perform training/testing/validation.

2 Prologue: Understanding your data

Load in your data and define the variables and outcomes. After renaming and defining variables, check to ensure that the data is in the expected format. This is easily handled by looking at the "head" of the dataframe, using head(df), which displays the top 5 rows by default. If the data has too many variables such that the table would be too wide for the page, it can be helpful to transpose the table for this step, so that variables will be listed on each row, with the first five data points displayed to the right.

Table 1: Clinical Predictors

	CBM001	CBM002	CBM003	CBM004	CBM005	CBM006
Age	64	80	95	64	47	75
Male	1	0	1	1	0	1
Smoking	1	1	1	1	1	1
PKY	50	35	30	29	45	100
BMI	25.80	30.40	21.90	20.70	18.90	31.36
Prior.Cancer	0	1	0	0	1	0
Nodule.Diam.mm	6.0	12.0	9.0	25.0	25.0	23.3
Nodule.Spiculation	0	1	0	0	0	1
Nodule.Upper.Lobe	0	0	0	1	0	1

Table 2: Biomarkers to Evaluate

	CBM001	CBM002	CBM003	CBM004	CBM005	CBM006
Largest.Diameter	6.113	38.585	68.750	25.793	25.543	19.712
AntPost.Length	5.373	40.095	56.734	22.948	24.805	18.298
L3.Distance	4.604	16.131	27.059	12.950	19.001	13.264
Area.Density	1.089	1.332	1.623	1.089	1.242	1.234
Volume.Density	0.417	0.375	0.303	0.332	0.538	0.538
Flatness	0.668	0.466	0.428	0.609	0.701	0.691
Joint.Entropy	5.172	6.832	8.607	5.950	8.164	6.042
Sum.Entropy	3.919	5.049	5.988	4.576	5.923	4.669
Dep.Entropy	5.616	6.671	7.202	6.325	7.174	6.152
Strength	47.974	21.829	14.343	35.348	36.331	48.203

Table 3: Outcome Data

	CBM001	CBM002	CBM003	CBM004	CBM005	CBM006
TTD	NA	10	NA	NA	227	NA
Total.CT	0	3	NA	NA	3	NA
Total.PET	1	1	NA	NA	2	NA
Bronch	0	0	NA	NA	2	NA
EBUS.NAV	0	0	NA	NA	0	NA
Total.Inv	0	0	NA	NA	2	NA
Sx	0	0	NA	NA	1	NA
Sputum.Cytology	0	0	NA	NA	0	NA
Fungal	0	0	NA	NA	1	NA
TTNA	0	0	NA	NA	0	NA

2.1 Population Summary Table

Your manuscript's "Table 1." We advise showing the median and interquartile range for continuous variables. Clinical data is often non-normally distributed, causing mean and standard deviation to inaccurately represent the data.

It is imperative that the clinical data be split by case status. It can be helpful to include a third column showing the summary for the data in the entire cohort, but not necessary.

We actually advise against including p-values in this table, as they can mislead readers into thinking that the differences in clinical predictors are an important result of the study. However, some journals ask for them, and in case you wish to show them, they have been included in the code.

Table 4: Summary of Variables with Statistical Tests

	Benign		Cancer		All			
Variable	Median	IQR	Median	IQR	Median	IQR	Missing	p
Age	66.0	(60.5 - 71)	68.0	(63 - 74)	67.0	(62 - 73)	0	0.004
Male	130	65.3%	162	63%	292	64%	0	0.684
Smoking	181	91%	241	93.8%	422	92.5%	0	0.339
PKY	40.0	(25 - 52.2)	44.0	(28.5 - 65)	42.0	(25 - 60)	0	0.035
BMI	27.2	(24.4 - 31.3)	26.6	(23.3 - 30.5)	27.0	(23.8 - 30.7)	8	0.093
Prior.Cancer	41	20.6%	75	29.2%	116	25.4%	0	0.048
Nodule.Diam.mm	12.0	(8.2 - 16)	19.0	(13 - 23.3)	15.4	(11 - 21.4)	0	0.000
Nodule.Spiculation	51	25.6%	99	38.5%	150	32.9%	0	0.005
Nodule.Upper.Lobe	104	52.3%	168	65.4%	272	59.6%	0	0.006
Largest.Diameter	12.9	(8.8 - 18.3)	20.3	(15.3 - 25.6)	17.1	(12 - 22.9)	0	0.000
AntPost.Length	11.3	(7.6 - 15.8)	17.6	(13.2 - 22.6)	14.8	(10.5 - 20.2)	0	0.000
L3.Distance	8.1	(5.4 - 10.3)	12.2	(9.1 - 15.5)	10.1	(7.1 - 13.9)	0	0.000
Area.Density	1.2	(1.1 - 1.2)	1.3	(1.2 - 1.3)	1.2	(1.1 - 1.3)	4	0.000
Volume.Density	0.4	(0.4 - 0.5)	0.4	(0.3 - 0.5)	0.4	(0.4 - 0.5)	4	0.014
Flatness	0.6	(0.5 - 0.7)	0.7	(0.6 - 0.7)	0.6	(0.5 - 0.7)	0	0.001
Joint.Entropy	6.9	(6.3 - 7.5)	7.3	(6.7 - 7.8)	7.2	(6.5 - 7.7)	0	0.000
Sum.Entropy	5.1	(4.7 - 5.5)	5.3	(4.9 - 5.6)	5.2	(4.8 - 5.6)	0	0.001
Dep.Entropy	6.3	(6 - 6.5)	6.4	(6.2 - 6.6)	6.4	(6.1 - 6.6)	1	0.000
Strength	39.5	(30 - 54.4)	32.6	(24.7 - 41.6)	35.3	(26.5 - 46.4)	12	0.000

Note:
Binary variables are represented by count (percentage), and continuous variables are represented by median (25th - 75th percentile). P-values are calculated using t-test for normally distributed variables, Mann-Whitney U test for non-normally distributed variables, and the Chi-Square test for binary variables

2.2 Dealing with missing data

3 Handling Missing Data

Introduction to Missing Data: Handling missing data is crucial for ensuring the validity of statistical analysis. Missing data can arise due to various reasons like errors in data collection, and ignoring them can lead to biased results. It is important to recognize the type of missing data: *Missing Completely at Random (MCAR)*, *Missing at Random (MAR)*, and *Missing Not at Random (MNAR)*.

Approaches to Handling Missing Data:

- **Deletion Methods:**
  - *Listwise Deletion:* Removes any cases with missing values, which can reduce sample size.
  - *Pairwise Deletion:* Uses all available data for each analysis, which can lead to varying sample sizes.
- **Imputation Methods:**
  - *Mean/Median Imputation:* Replaces missing values with the mean or median, potentially distorting data distribution.
  - *Last Observation Carried Forward (LOCF):* Uses the last observed value to fill in missing data points, common in longitudinal studies.
- **Multiple Imputation:** Creates multiple datasets with different imputed values, analyzes each separately, and combines the results.

Using the `mice` Package: The `mice` package (Multivariate Imputation by Chained Equations) provides a powerful method for handling missing data via multiple imputation.

3.1 Understanding the Biomarkers

- **Largest.Diameter**
  Category: Morphological
  Description: A measure of the largest planar diameter of the structure, which refers to the longest straight line that can fit entirely inside an XY-plane slice of the 3D structure, from edge to edge, without leaving the structure.

- **AntPost.Length**
  Category: Morphological
  Description: A measure of the anterior-posterior (front-to-back) distance of the region of interest (ROI).

- **L3.Distance**
  Category: Morphological
  Description: The length of the normal (L3) full principal axis, measured from edge to edge of the ROI, in millimeters. It is IBSI-consistent.

- **Area.Density**
  Category: Volume Density
  Description: IBSI-consistent surface area of the ROI over the surface area of the approximate enclosing ellipsoid (AEE).

- **Volume.Density**
  Category: Volume Density
  Description: IBSI-consistent volume fraction of the bounding box (AABB) occupied by the ROI.

- **Flatness**
  Category: Morphological
  Description: IBSI-consistent ratio of the least principal axis to the major principal axis, where a maximum value of 1 indicates a spherical shape.

- **Joint.Entropy**
  Category: Texture (GLCM)
  Description: IBSI-consistent joint entropy of the gray-level co-occurrence matrix (GLCM) of the unpadded ROI, binned for CT images, with aggregation by slice without merging.

- **Sum.Entropy**
  Category: Texture (GLCM)
  Description: IBSI-consistent sum entropy of the gray-level co-occurrence matrix (GLCM) of the unpadded ROI, binned for CT images, with aggregation by slice without merging.

- **Dep.Entropy**
  Category: Texture (NGLDM)
  Description: IBSI-consistent dependence entropy of the neighborhood gray-level dependence matrix (NGLDM) of the unpadded ROI, binned for CT images, with aggregation by slice and merging.

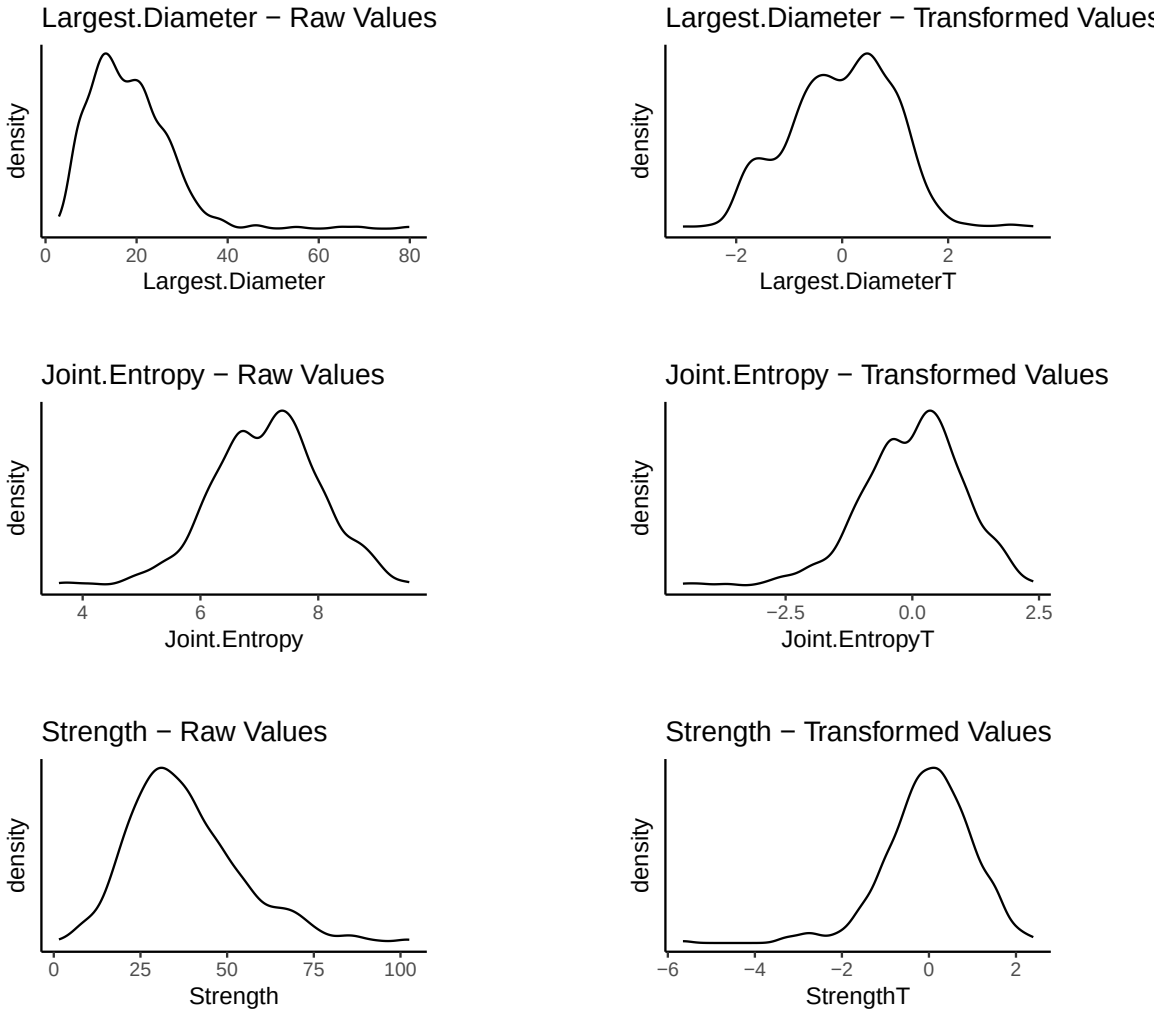
- **Strength**
  Category: Texture (NGTDM)
  Description: IBSI-consistent strength of the neighborhood gray-tone difference matrix (NGTDM) of the unpadded ROI, with metrics averaged across all matrices.

3.2 Normalize/Standardize the data

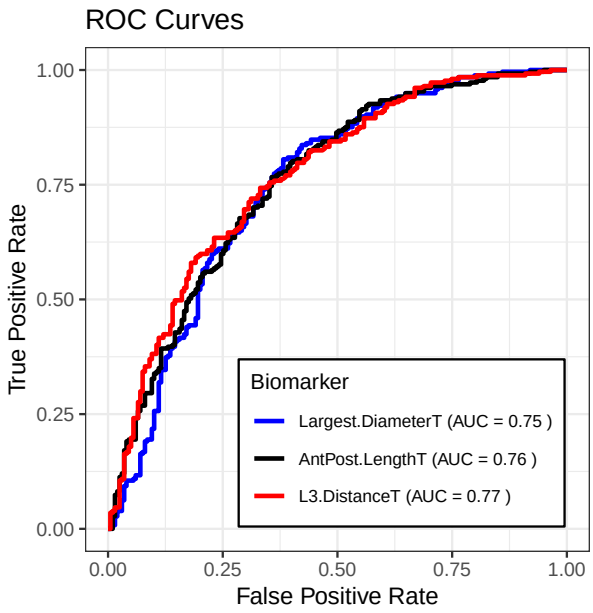
Normalizing of predictors is crucial when building logistic regression models, especially when the biomarkers have different scales. Without normalization, predictors with larger scales might disproportionately influence the model, making it harder for the model to converge and interpret the coefficients meaningfully.

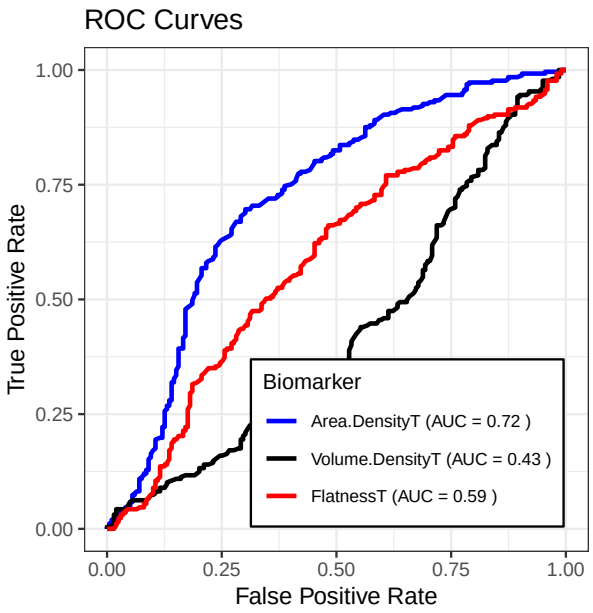
For example, in cases where one biomarker is measured in millimeters and another in a unitless ratio, their scales could vary significantly. By transforming and normalizing biomarkers to a standard normal distribution, we ensure that each biomarker contributes equally to the model, improving convergence and interpretability.

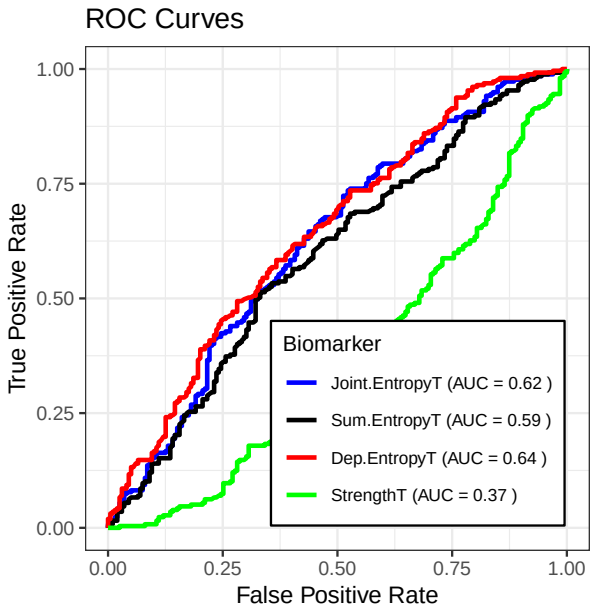
Normalization helps in: - **Stabilizing learning**: Logistic regression can converge faster and more reliably when the features are on similar scales. - **Interpretability**: Coefficients in the model become more comparable since the features are standardized. - **Reducing bias**: It helps avoid biasing the model towards features with larger ranges.

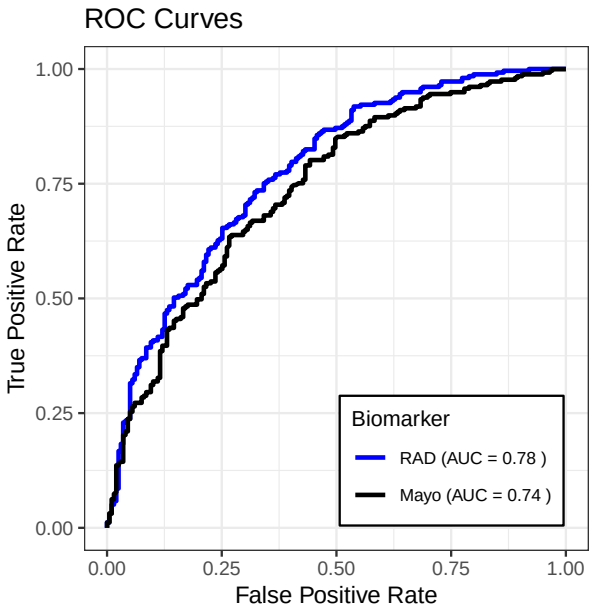


4 Building the Biomarker model

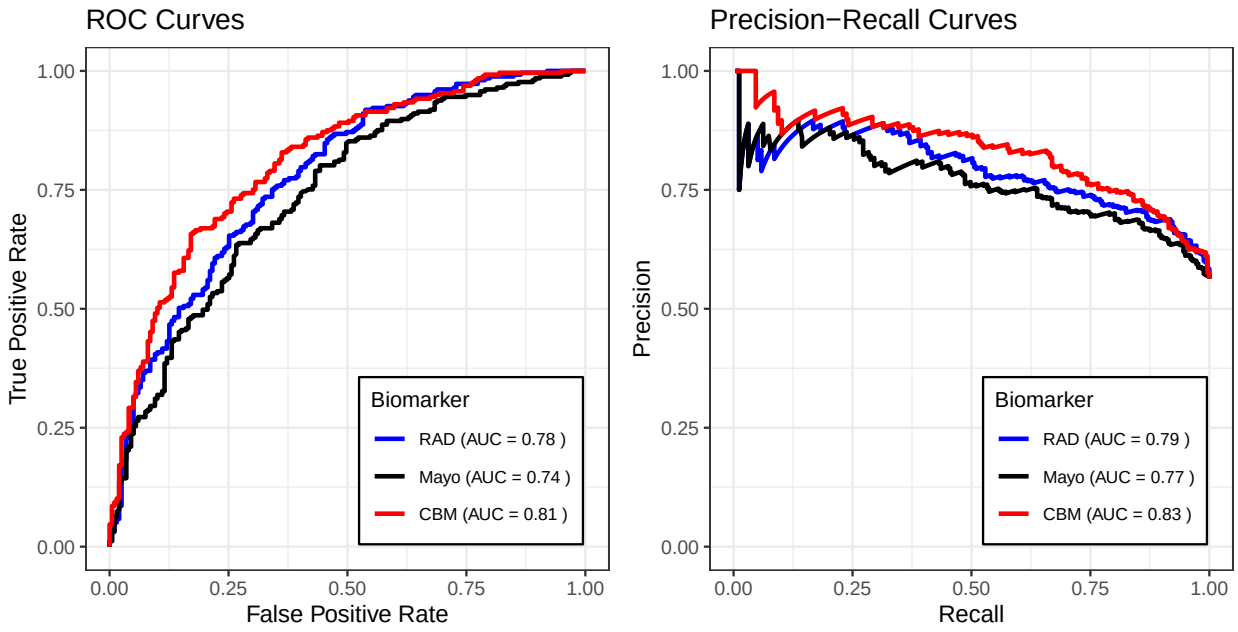








4.1 The combined model (CBM)



4.2 Significant improvement?

4.2.1 Likelihood Ratio Test (LRT)

The Likelihood Ratio Test (LRT) is a statistical method used to compare two nested models: a simpler model (a subset of a more complex model) and the more complex model (which includes additional parameters). The LRT helps determine whether the added parameters in the complex model significantly improve the fit of the model to the data.

In the context of diagnostic models for cancer, the LRT allows us to assess whether adding new biomarkers to an existing model, such as the Mayo model, provides a statistically significant improvement in predicting cancer outcomes. Specifically, we compare the Mayo model, which includes clinical predictors, with a more comprehensive model that incorporates biomarkers. This comparison can help determine whether the biomarkers add meaningful predictive value.

4.2.1.1 Mathematical Framework The LRT is based on the likelihood function, which measures how well the model explains the observed data. The test statistic is calculated as follows:

$$\text{LR Statistic} = -2\left(\log L_{\text{simpler}} - \log L_{\text{complex}}\right)$$

Where: - L_{simpler} is the likelihood of the data under the simpler model. - L_{complex} is the likelihood of the data under the more complex model.

This test statistic follows a chi-square distribution (χ^2) with degrees of freedom equal to the difference in the number of parameters between the simpler and complex models. The null hypothesis is that the simpler model fits the data as well as the complex model, implying that the additional parameters (in this case, the biomarkers) are not necessary. If the test statistic is large enough to exceed a critical value from the χ^2 distribution, we reject the null hypothesis and conclude that the added biomarkers provide significant improvement.

4.2.1.2 When to Use the LRT

- **Nested Models**: The LRT is applicable when one model is nested within another. That is, the simpler model must be a special case of the more complex model (i.e., the complex model contains all the parameters of the simpler model, plus additional ones).
- **Evaluating Biomarkers**: By comparing a model with only clinical variables (e.g., the Mayo model) to a combined model that includes both clinical variables and biomarkers, the LRT can assess whether the biomarkers significantly enhance the predictive accuracy of the model.

4.2.1.3 Application to this Analysis In this document, we compare: - **Mayo Model**: A model that includes only clinical predictors. - **CBM Model**: A combined model that includes both clinical predictors from the Mayo model and additional biomarkers.

The LRT allows us to formally test whether the addition of biomarkers to the Mayo model leads to a statistically significant improvement in predicting cancer outcomes.

By calculating the LR statistic and comparing it to the chi-square distribution, we can quantify whether the combined biomarker model (CBM) significantly outperforms the clinical-only Mayo model. This test helps validate the added value of biomarkers in the diagnostic model.

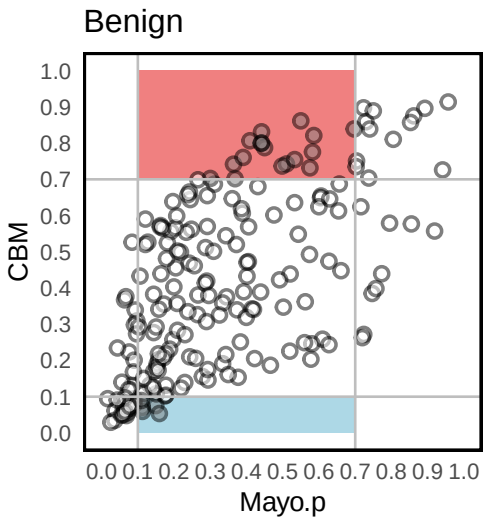
Likelihood ratio test for the CBM vs Mayo

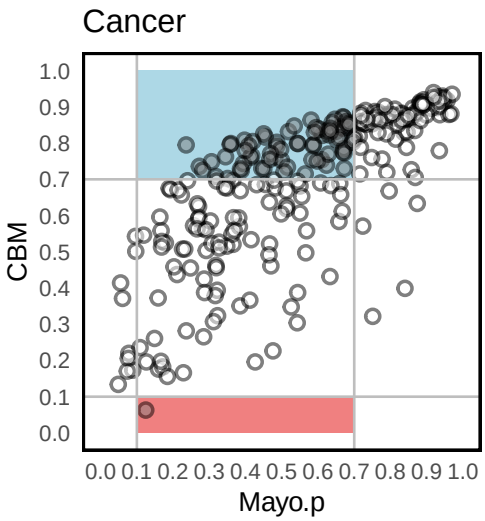
```
##
## Model 1: y ~ rcs(Mayo, 3) + rcs(RAD, 3) + rcs(BMI, 3)
## Model 2: y ~ Mayo
##
##   L.R. Chisq         d.f.            P
## 6.057610e+01 5.000000e+00 9.239831e-12
```

Likelihood ratio test for the CBM vs Radiomics

```
##
## Model 1: y ~ rcs(Mayo, 3) + rcs(RAD, 3) + rcs(BMI, 3)
## Model 2: y ~ (Strength) + (Flatness) + (AntPost.Length) + (L3.Distance) +
##     rcs(Area.Density, 3) + rcs(Volume.Density, 3) + (Dep.Entropy) +
##     (Sum.Entropy)
##
##   L.R. Chisq         d.f.            P
## 2.461636e+01 4.000000e+00 6.008203e-05
```

4.3 Reclassification





Benign

	Low	Medium	High	Total
Total	35	142	22	199
High	0	17	13	30
Medium	15	120	9	144
Low	20	5	0	25

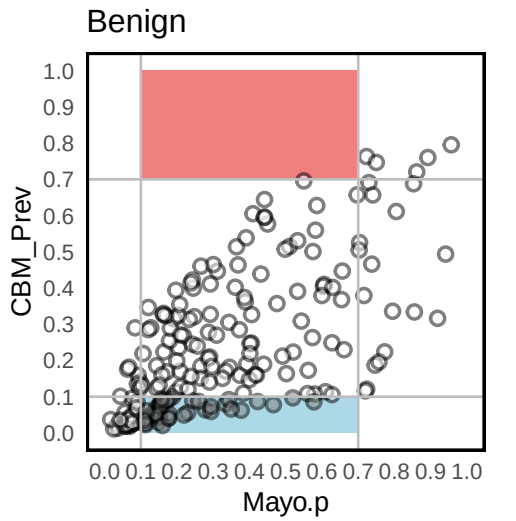
Cancer

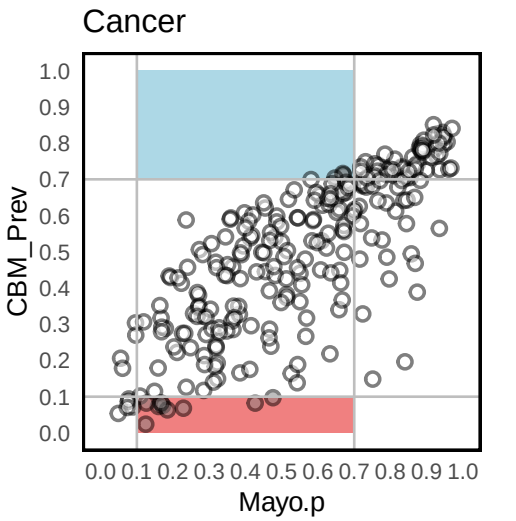
	Low	Medium	High	Total
Total	9	165	83	257
High	0	71	78	149
Medium	9	93	5	107
Low	0	1	0	1

For Benign: Observed = -0.0845, Expected = 0.0486, Net = -0.1332

For Cancer: Observed = 0.4242, Expected = 0.1779, Net = 0.2464

4.4 Reclassification Adjusted for Prevalence





Benign

	Low	Medium	High	Total
Total	35	142	22	199
High	0	0	5	5
Medium	8	105	17	130
Low	27	37	0	64

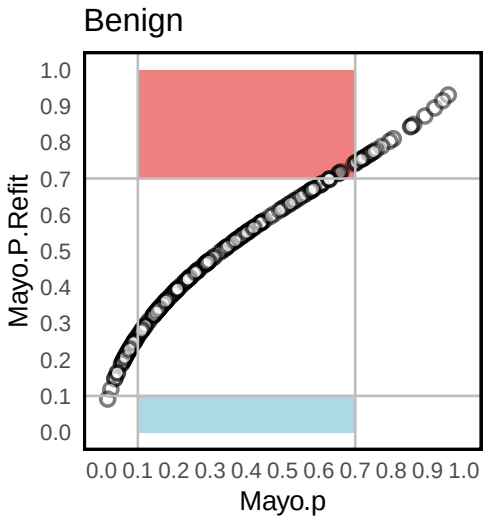
Cancer

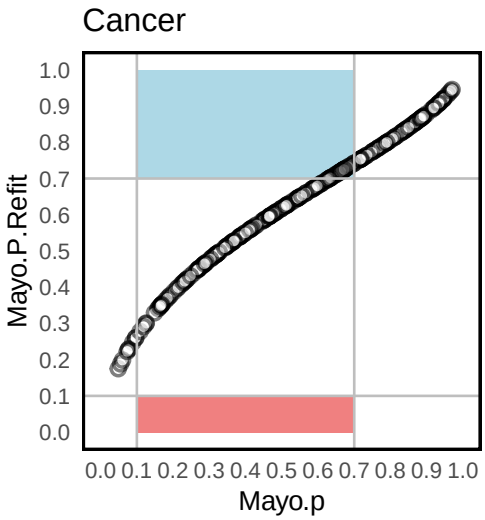
	Low	Medium	High	Total
Total	9	165	83	257
High	0	3	52	55
Medium	4	153	31	188
Low	5	9	0	14

For Benign: Observed = 0.2606, Expected = -0.139, Net = 0.3995

For Cancer: Observed = -0.0364, Expected = -0.0875, Net = 0.0512

4.5 Did we reclassify just by adjusting the pretest probability for our current prevalence?





Benign

	Low	Medium	High	Total
Total	35	142	22	199
High	0	4	22	26
Medium	34	138	0	172
Low	1	0	0	1

Cancer

	Low	Medium	High	Total
Total	9	165	83	257
High	0	27	83	110
Medium	9	138	0	147
Low	0	0	0	0

For Benign: Observed = -0.0282, Expected = 0.1187, Net = -0.1469

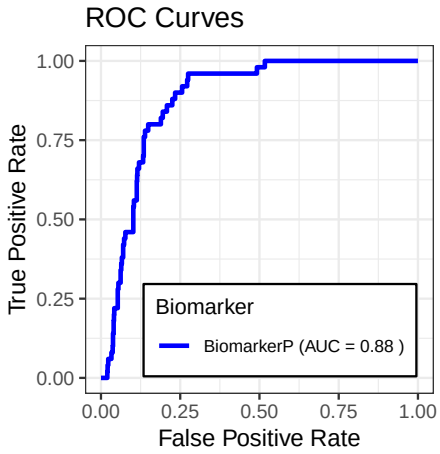
For Cancer: Observed = 0.1636, Expected = 0.0984, Net = 0.0653

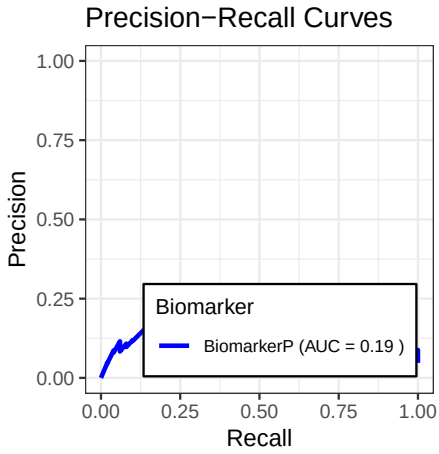
5 Appendix: Helpful Demonstratations

5.1 Precision Recall

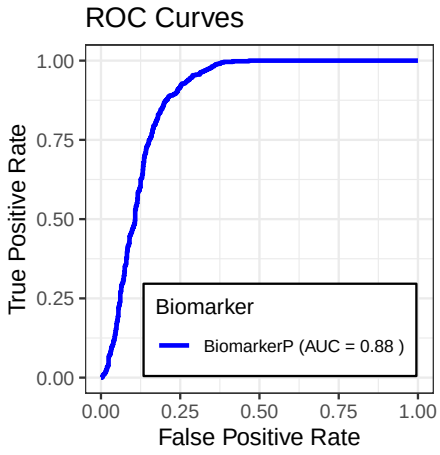
In a case:control imbalanced population, a "good" ROC curve can be a result of simply having way more controls than cases, as long as the controls, on average, have a lower biomarker level that the cases, on average.

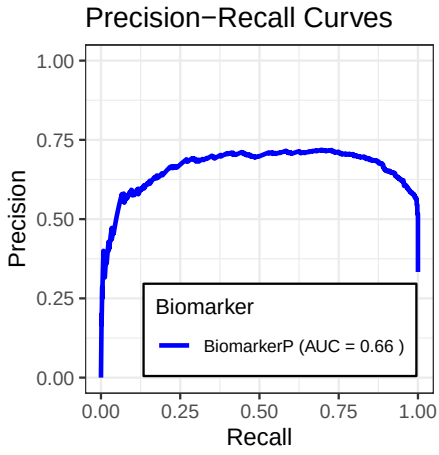
For example, with 50 cases and 1000 controls:



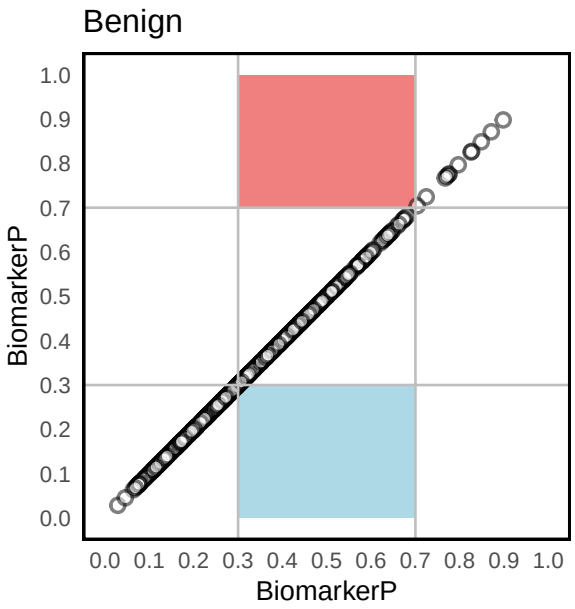


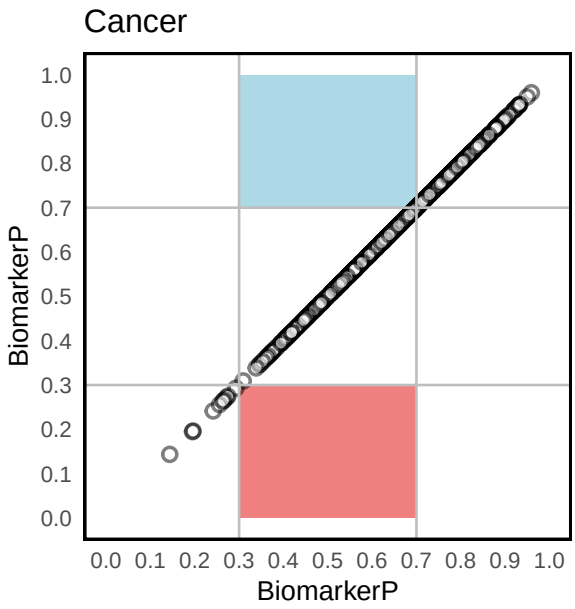
However, a biomarker with the exact same distribution in a population with 500 cases and 1000 controls:





5.2 Expected vs Observed Reclassification



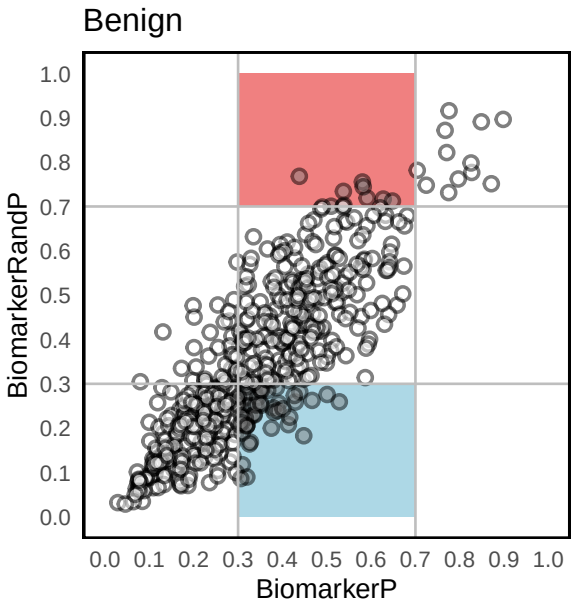


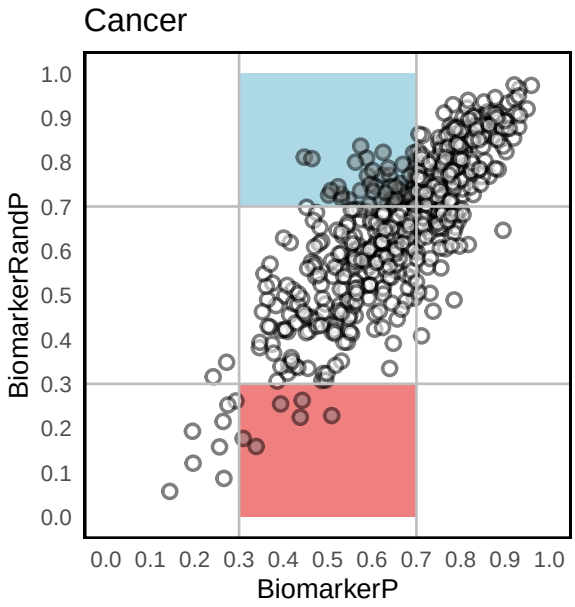
Benign

	Low	Medium	High	Total
Total	214	274	12	500
High	0	0	12	12
Medium	0	274	0	274
Low	214	0	0	214

Cancer

	Low	Medium	High	Total
Total	10	274	216	500
High	0	0	216	216
Medium	0	274	0	274
Low	10	0	0	10





Benign

	Low	Medium	High	Total
Total	214	274	12	500
High	0	9	12	21
Medium	31	216	0	247
Low	183	49	0	232

Cancer

	Low	Medium	High	Total
Total	10	274	216	500
High	0	47	163	210
Medium	2	221	53	276
Low	8	6	0	14