Diagnostic Accuracy using AUC-ROC

Dan N

2024-03-01

Data

First 10 rows of our example data (flu_dat). We are interested in knowing at what cut point of a biomarker (in this case we are using a lymphocyte-to-monocyte ratio) can we accurately predict influenza infection.

subject	biomarker	infection	Flu_Type
1	0.3	Flu	A
2	0.3	Flu	A
3	0.4	Flu	A
4	0.4	Flu	A
5	0.4	Flu	B
6	0.5	Flu	A
7	0.5	Flu	A
8	0.5	Flu	A
9	0.5	Flu	A
10	0.5	Flu	A

Logistic Regression with Continuous Biomarker

We can use logistic regression with a continuous predictor variable to determine the best cut point. There are different methods, such as Youden Index or distance you can try. From the output of the logistic model, we can use the logit, intercept and slope to determine the cut point.

```
PROC LOGISTIC DATA = flu_dat;
MODEL infection (EVENT= 'Flu') = biomarker / OUTROC = ROCDATA;
ROC; ROCCONTRAST;
RUN;

DATA LM_threshold;  SET ROCDATA;
LOGIT = LOG(_PROB_/(1-_PROB_));
CUT_POINT = (LOGIT - 1.0172) / -0.5300;
YD = _SENSIT_ + (1 - _1MSPEC_) - 1;
RUN;

PROC SORT DATA = LM_threshold;
BY YD;
RUN;
```

SOURCE	***PROB***	***POS***	***NEG***	***FALPOS***	***FALNEG***	***SENSIT***	***1MSPEC***	**LOGIT**	**CUT_POINT**	**YD**
Model	0.4892834	121	98	74	30	0.8013245	0.4302326	-0.0428731	2.000138	0.3710919
Model	0.5025328	115	104	68	36	0.7615894	0.3953488	0.0101313	1.900130	0.3662406
Model	0.5157787	103	109	63	48	0.6821192	0.3662791	0.0631356	1.800121	0.3158401
Model	0.4760490	123	92	80	28	0.8145695	0.4651163	-0.0958775	2.100146	0.3494533
Model	0.5290024	97	111	61	54	0.6423841	0.3546512	0.1161400	1.700113	0.2877329

When sorted by largest YD value, the best cut point is 2.0. We can now model build using this cut point as well as try other cut point out and determine the best model.

Logistic Regression with Dichotomous Biomarker

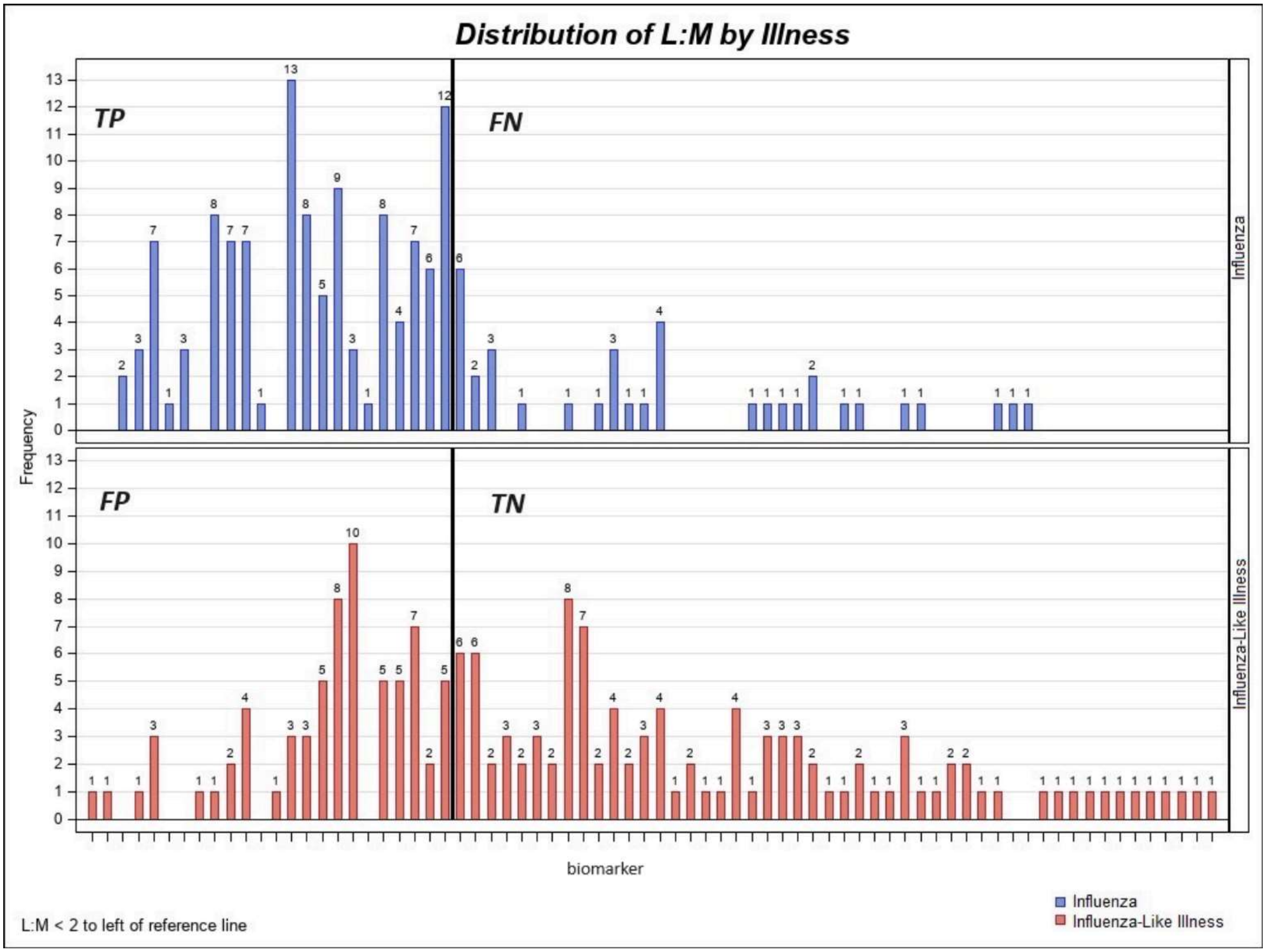
Lets create a dichotomous variable with the cut point set at 2.0 based on the output above. We can also graph the continuous distribution by infection type and see the overlap in biomarker between flu and flu-like groups.

```
DATA a; SET flu_dat;
IF biomarker < 2.0 then L_M_Ratio = "<2 L:M";
  ELSE L_M_Ratio = ">2 L:M";
RUN;

PROC FREQ DATA = a ORDER = formatted;
TABLES L_M_Ratio * infection / NOROW NOPERCENT;
RUN;
```

Frequency
Col Pct

Table of L_M_Ratio by Infection			
	Infection(Infection)		
L_M_Ratio(L/M Ratio)	Flu	Flu-Like	Total
<2 L:M	115 76.16	68 39.53	183
>2 L:M	36 23.84	104 60.47	140
Total	151	172	323



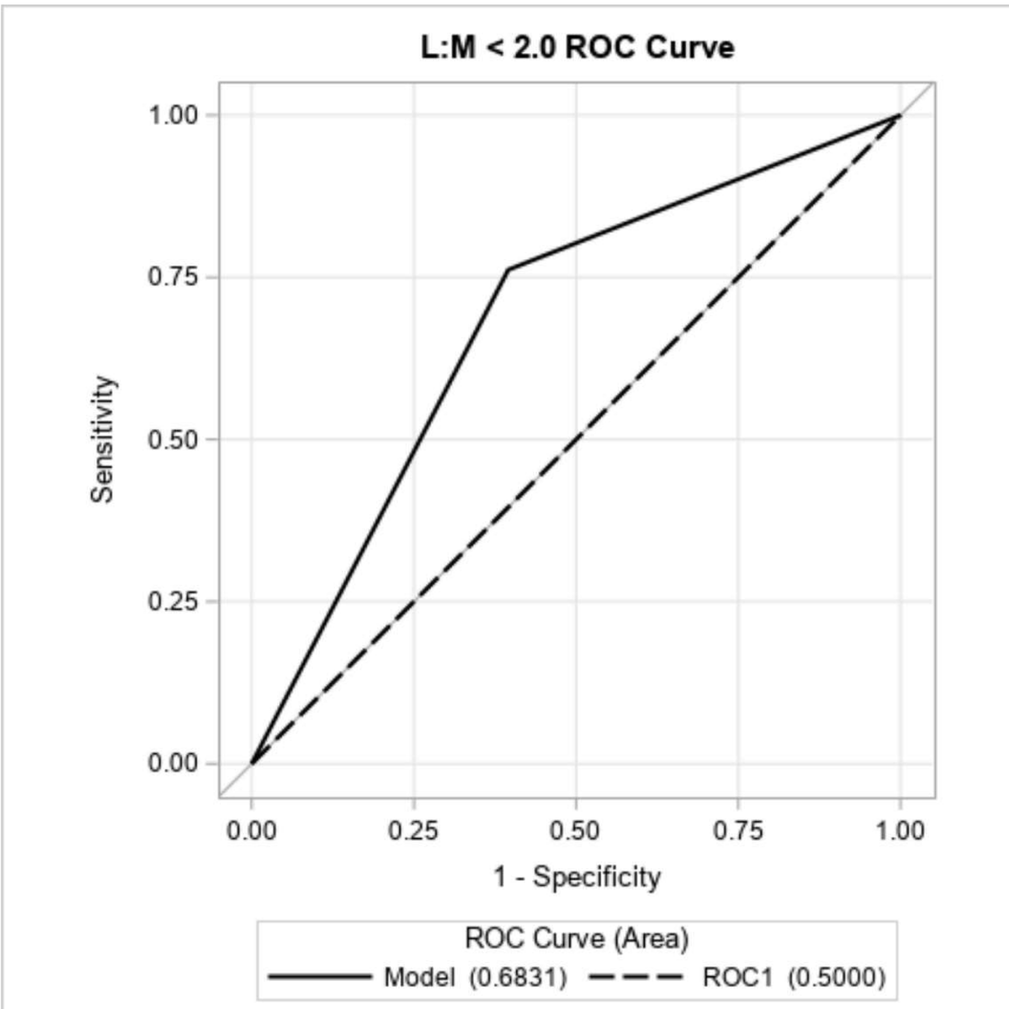
Using the column percentages, we see the sensitivity is 76.2% and the specificity is 60.5%. Also note the how the distribution of the continuous biomarker overlaps each infection group. This hints that the area under the curve might not be as high in value as we hope. But lets run the logistic model with our new variable and output the ROC curve and see what we get.

```
ODS GRAPHICS ON;
ODS LISTING STYLE = statistical SGE = on;
PROC LOGISTIC DATA = a PLOTS(ONLY) = ROC;
CLASS L_M_Ratio (PARAM = ref REF = ">2 L:M");
MODEL infection(EVENT = 'Flu') = L_M_Ratio;
ROC; ROCCONTRAST;
RUN;
ODS LISTING SGE = off;
ODS GRAPHICS OFF;
```

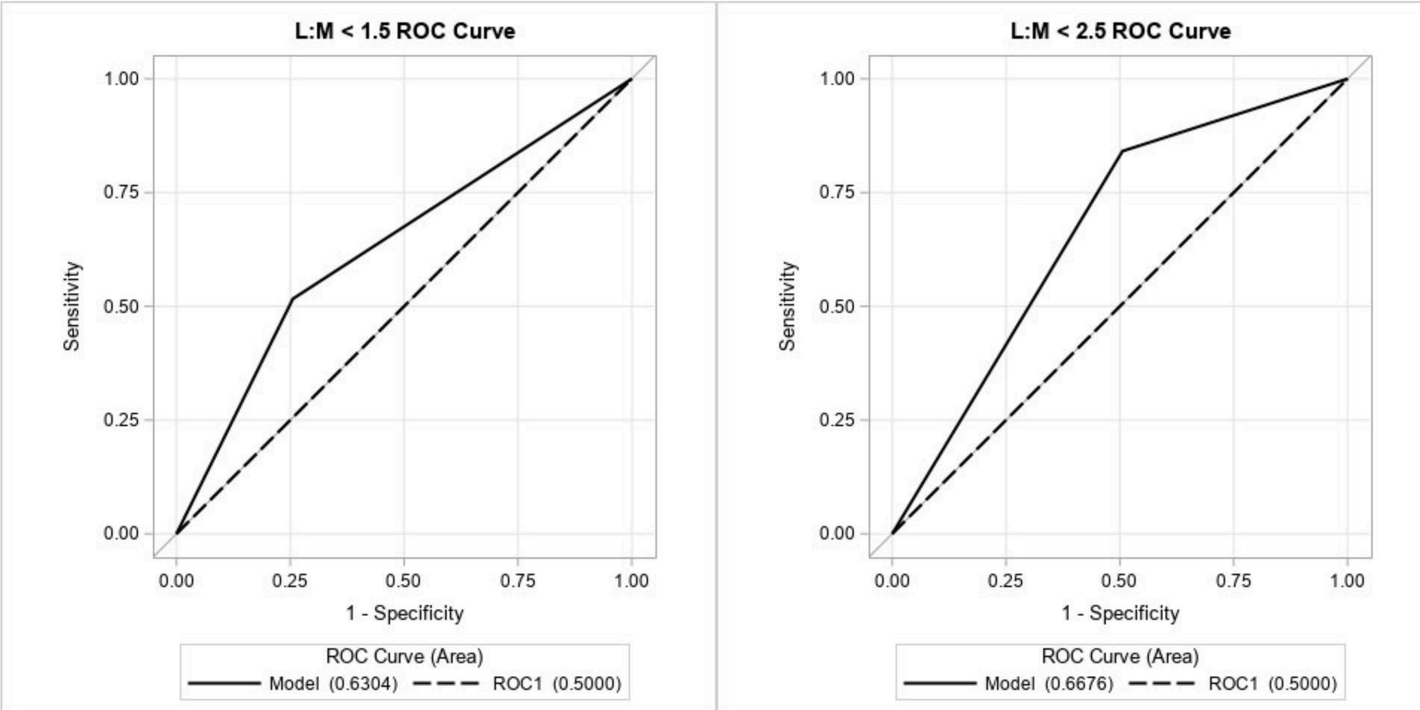
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.0609	0.1934	30.0977	<.0001
ratio	<2 L:M	1	1.5863	0.2466	41.3907	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
ratio <2 L:M vs >2 L:M	4.886	3.013	7.921

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	46.0	Somers' D	0.366
Percent Discordant	9.4	Gamma	0.660
Percent Tied	44.5	Tau-a	0.183
Pairs	25972	c	0.683



The area under the curve for this model with a biomarker cut point of 2.0 is 0.683. While not ideal, given that we have a false positive rate of ~40%, this cut point outperforms the chance model (AUC = 0.50) as well as other tested thresholds.



If our cut point of the continuous biomarker is set at 1.5, the AUC = .630. And if the cut point is set at 2.5, the AUC = 0.667. Both of these models show an unfavorable tradeoff between the true positive rate and the false positive rate.