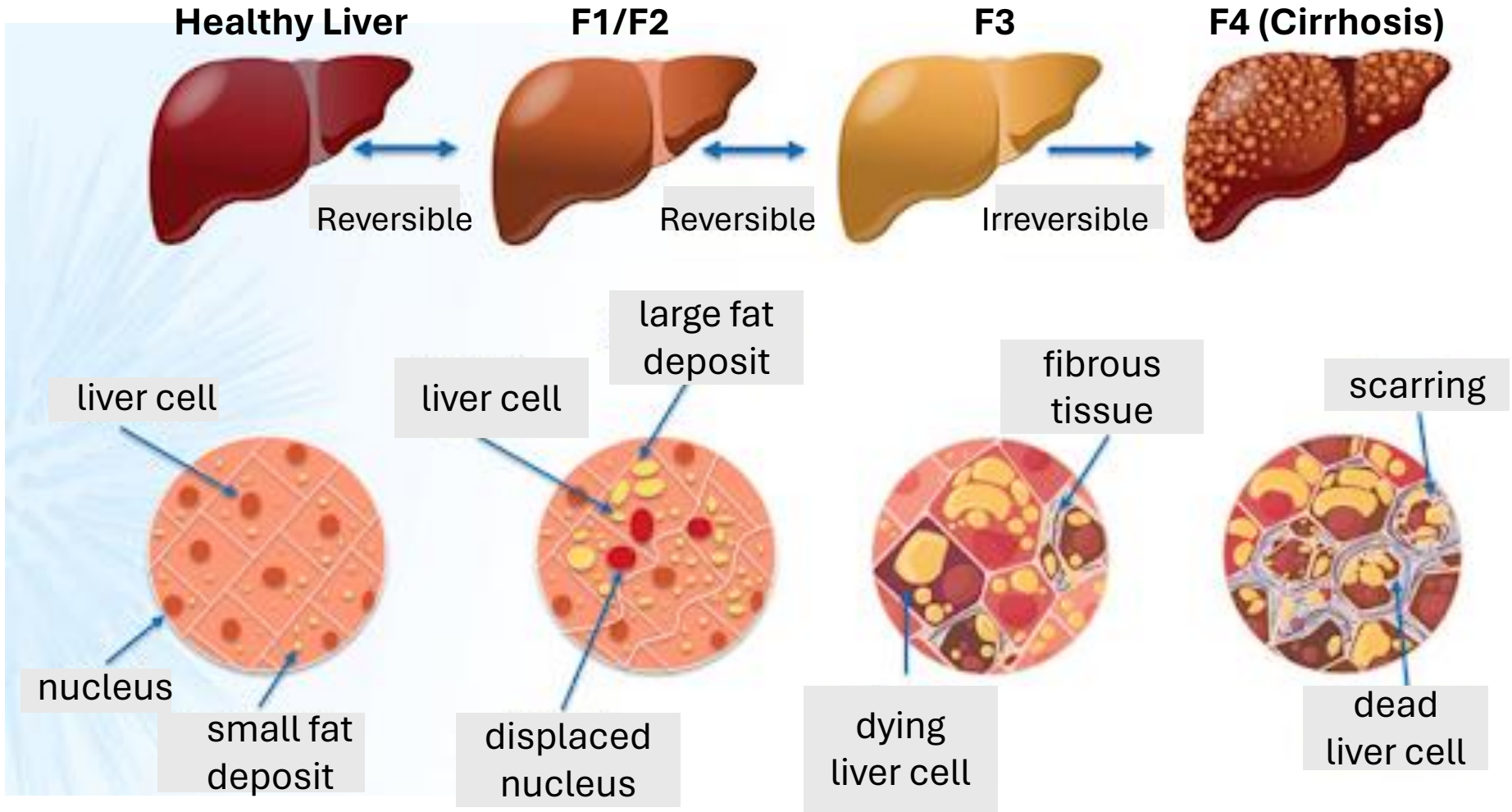
Prescreening for Fatty Liver Disease Using Blood Markers

Andy Vong

Late Stages of NASH Deadly; Early Stages of NASH Reversible

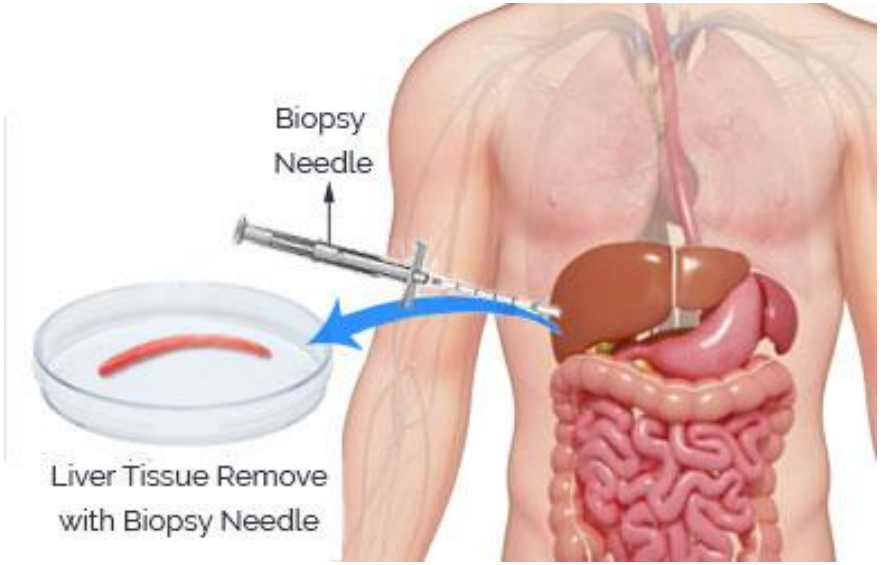


http://www.hcv-trials.com/nash/NAFLD-NASH.asp

Pre-screening for Fibrosis Staging Saves Costs and Improves Patient Care

Liver Biopsy (Gold Standard)

- Invasive



MRI-PDFF (Currently used by FDA for clinical trial enrollment)

- Costly
- Need specialized equipment

https://www.fda.gov/media/124105/download

Biomarkers Correlate with Fibrosis Staging; Blood Tests Detect Fibrosis Staging

ALBI

$$\log[\text{Bilirubin}] - \text{Albumin}$$

APRI

$$\frac{\text{AST}}{\text{Platelet}}$$

FIB-4

$$\frac{\text{Age} * \text{AST}}{\text{Platelet} * \sqrt{\text{ALT}}}$$

- Higher scores trend with fibrosis staging
- Bilirubin builds up in blood as liver is compromised
- Albumin levels decrease as liver is compromised
- Aspartate aminotransferase (AST) is liver enzyme
- Platelet count decreases with liver damage, cause is multifactorial
- Alanine aminotransferase (ALT) also liver enzyme

Purpose: Prescreen for Fatty Liver Disease Using Only Blood Test Results

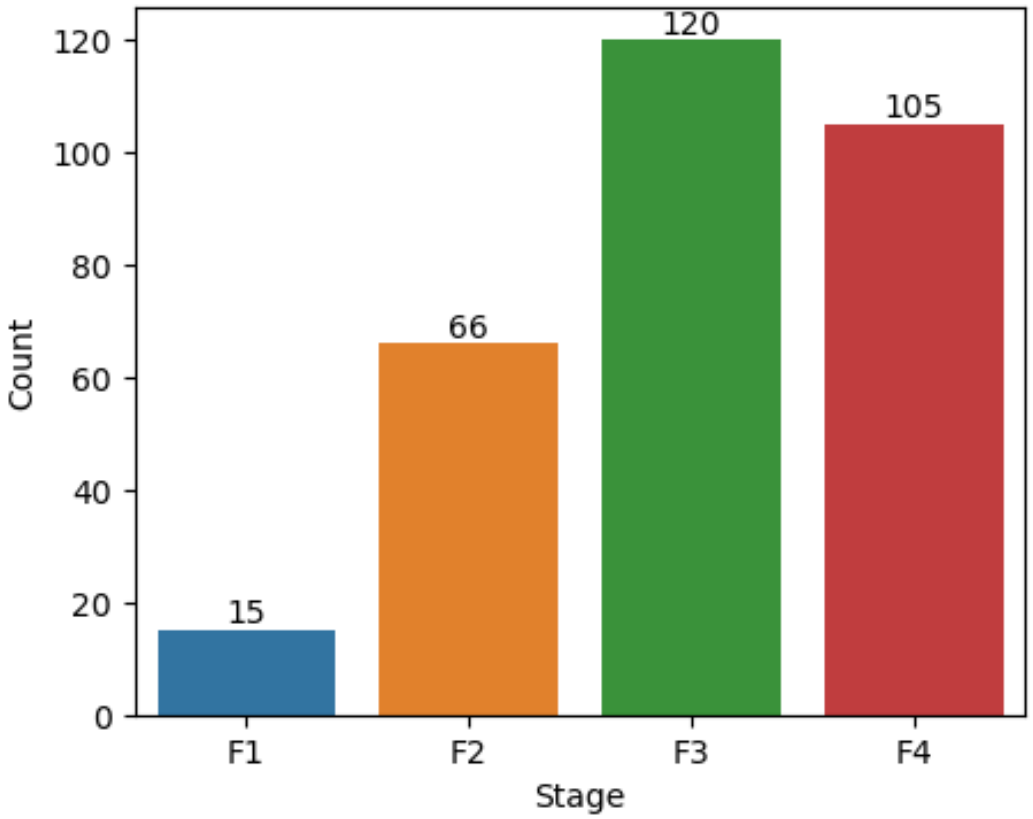
With focus on F3 due to limitations with data

Trained with 306 Patient Records Provided by Mayo Clinic

- Benchmarks
  - Random Guess
  - ALBI Score
  - APRI Score
  - FIB-4 Score
- Result:
  - Logistic Regression model outperforms benchmarks and other machine learning models

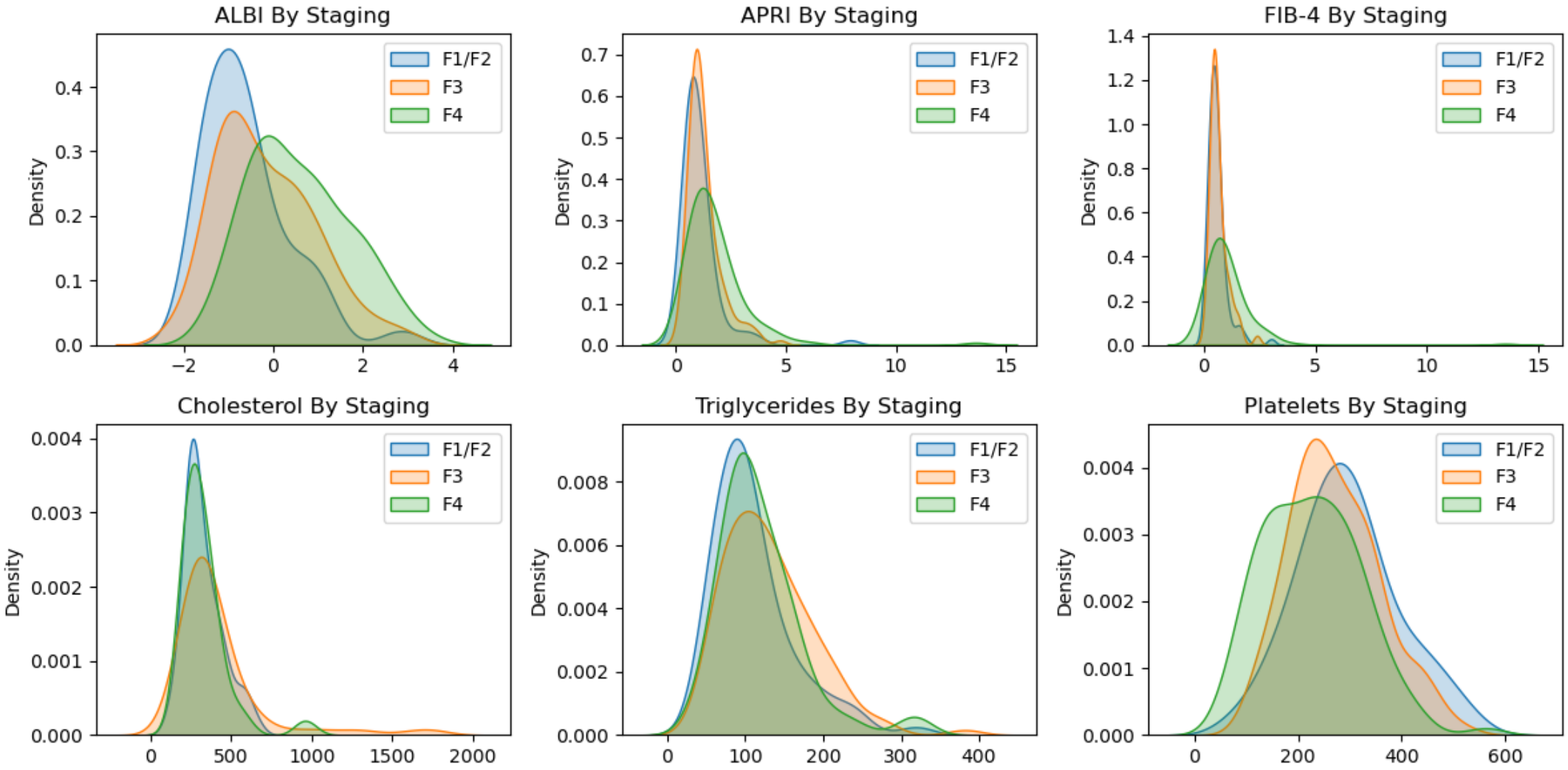
https://www.mayo.edu/research/documents/pbchtml/DOC-10027635

Data Cannot Reliably Classify F2 and Below



- No healthy patients
- Insufficient data points to classify F1
- Group F1 and F2 to balance data
- Focus on classifying F3

Features Not Well Separable by Staging



ALBI Scoring Sets Baseline F1-score 0.45

Random		Predicted			
		F1/F2	**F3**	**F4**	
True	**F1/F2**	21.4	31.8	27.8	81
	F3	31.8	47.0	41.2	120
	F4	27.8	41.2	36.0	105
		81	120	105	
	Precision:	0.26	0.39	0.34	
	Recall:	0.26	0.39	0.34	
	F1-Score:	0.26	0.39	0.34	

ALBI		Predicted			
		F1/F2	**F3**	**F4**	
True	**F1/F2**	20.0	38.0	23.0	81
	F3	9.0	52.0	59.0	120
	F4	1.0	19.0	85.0	105
		30	109	167	
	Precision:	0.67	0.48	0.51	
	Recall:	0.25	0.43	0.81	
	F1-Score:	0.36	**0.45**	0.63	

Logistic Regression Model Outperforms Benchmarks and Random Forest

Random Forest		Predicted			
		F1/F2	**F3**	**F4**	
True	**F1/F2**	20.1	52.7	8.2	81
	F3	20.5	77.3	22.2	120
	F4	4.0	45.8	55.2	105
		44.6	175.8	85.6	
	Precision:	0.45	0.44	0.64	
	Recall:	0.25	0.64	0.53	
	F1-Score:	0.32	0.52	0.58	

Logistic Reg		Predicted			
		F1/F2	**F3**	**F4**	
True	**F1/F2**	10.8	65.3	4.9	81
	F3	6.8	96.5	16.7	120
	F4	1.4	57.5	46.1	105
		19	219.3	67.7	
	Precision:	0.57	**0.44**	0.68	
	Recall:	0.13	**0.80**	0.44	
	F1-Score:	0.22	**0.57**	0.53	

Random Forest		Predicted			
		F1/F2	F3	F4	
True	F1/F2	20.1	52.7	8.2	81
	F3	20.5	77.3	22.2	120
	F4	4.0	45.8	55.2	105
		44.6	175.8	85.6	
	Precision:	0.45	0.44	0.64	
	Recall:	0.25	0.64	0.53	
	F1-Score:	0.32	0.52	0.58	

Logistic Reg		Predicted			
		F1/F2	F3	F4	
True	F1/F2	10.8	65.3	4.9	81
	F3	6.8	96.5	16.7	120
	F4	1.4	57.5	46.1	105
		19	219.3	67.7	
	Precision:	0.57	**0.44**	0.68	
	Recall:	0.13	**0.80**	0.44	
	F1-Score:	0.22	**0.57**	0.53	

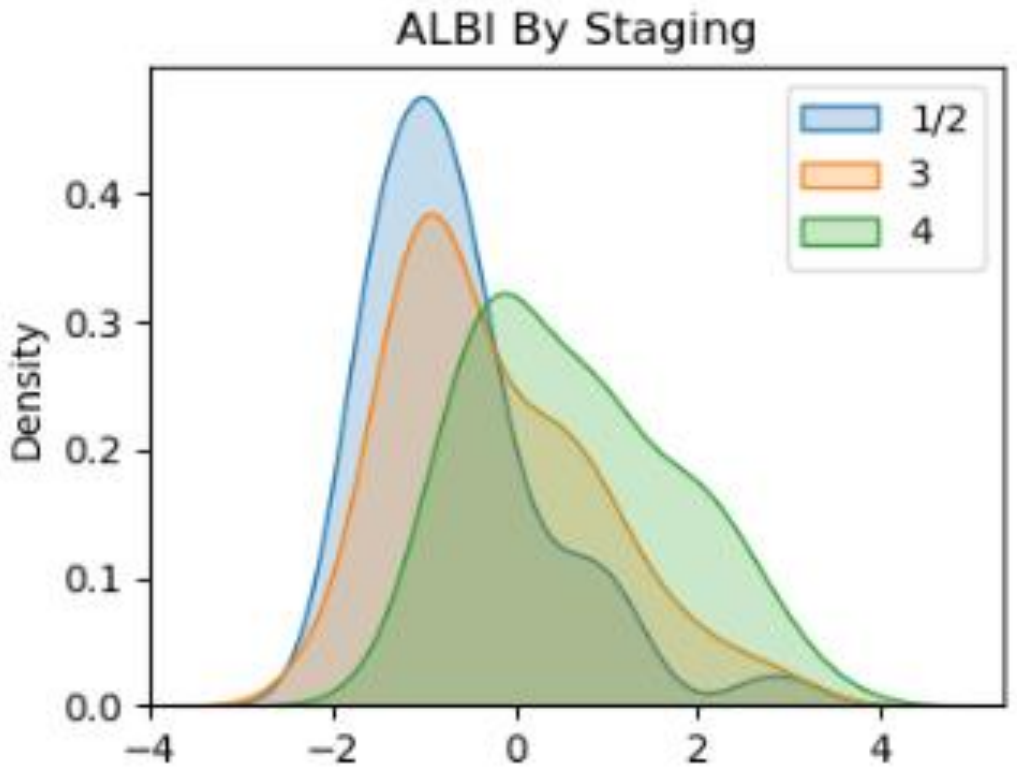
Random		Predicted			
		F1/F2	F3	F4	
True	F1/F2	21.4	31.8	27.8	81
	F3	31.8	47.0	41.2	120
	F4	27.8	41.2	36.0	105
		81	120	105	
	Precision:	0.26	0.39	0.34	
	Recall:	0.26	0.39	0.34	
	F1-Score:	0.26	0.39	0.34	

ALBI		Predicted			
		F1/F2	F3	F4	
True	F1/F2	20.0	38.0	23.0	81
	F3	9.0	52.0	59.0	120
	F4	1.0	19.0	85.0	105
		30	109	167	
	Precision:	0.67	0.48	0.51	
	Recall:	0.25	0.43	0.81	
	F1-Score:	0.36	**0.45**	0.63	

Summary

- Pre-screening for NASH with cheap, non-invasive tests can improve patient care and reduce costs
  - Treating F3 NASH reduce risk of cancer and liver-related deaths
- Data is limited
  - 306 patients not sufficient to simulate larger population
  - Lacked healthy population
  - Data rather homogeneous
- Machine learning models can outperform the best non-invasive diagnostic tests for fatty liver disease
- Additional data needed

ALBI Model



- Idea:
  - Choose threshold of ALBI value
  - Those with ALBI < threshold, assign as non-advanced or non-cirrhotic
  - Those with ALBI > threshold, assign as advanced or cirrhotic

ALBI Model

```
array([[20, 38, 23],
       [ 9, 52, 59],
       [ 1, 19, 85]], dtype=int64)
```

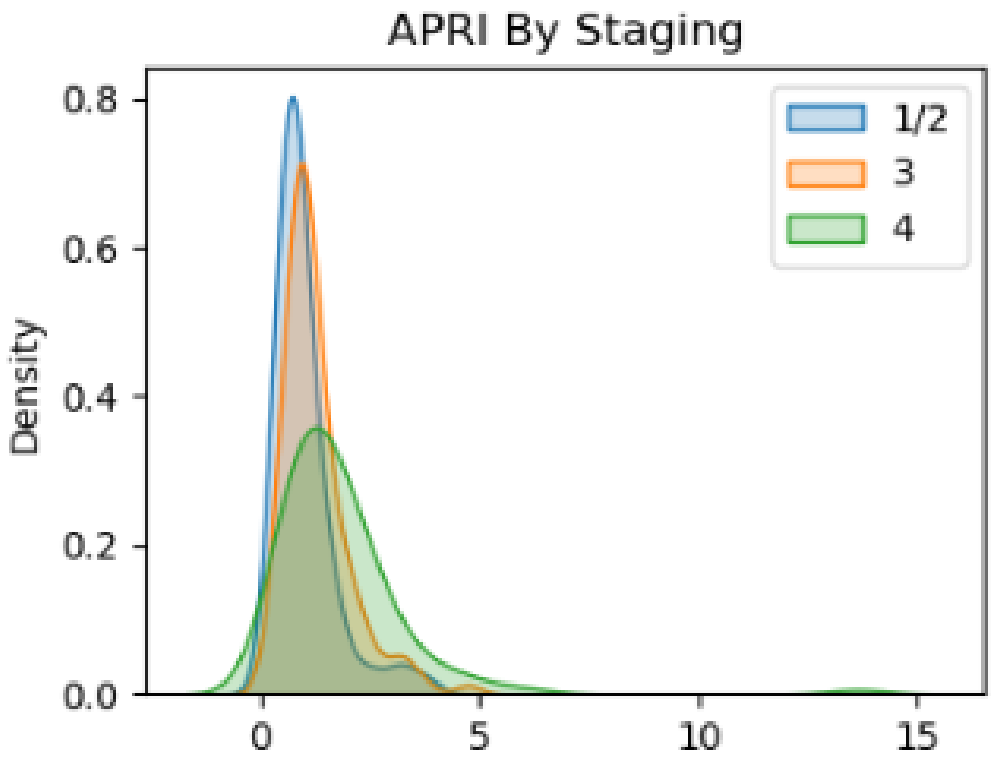
		Advanced Fibrosis (F3)	Cirrhosis (F4)
ALBI Score	AUC	0.705	0.733
	Best cutoff	-1.180	-0.300
	Sensitivity	0.956	0.810
	Specificity	0.247	0.592
	Accuracy	0.768	0.667
	Precision	0.779	0.509
	Neg Precision	0.667	0.856
	F1-Score	0.858	0.625
	F1-Precision	0.718	0.638

- Thresholds were chosen such that they maximize F1-score (thus balancing positive prediction rate and positive occurrence)
- For predicting advanced fibrosis, false negative rate is about the same as random guessing (1 in 3 of negative predictions are incorrectly labeled)
- For predicting cirrhosis, false negative rate is much better (1 in 7 of negative predictions are incorrectly labeled)
  - At same time, predict negative at much higher rate (48.7% for cirrhosis vs 9.8% for advanced fibrosis)
  - Meaning, much better false negative rate and NOT by making less false predictions

```
array([[0.06535948, 0.19934641],
       [0.03267974, 0.70261438]])
```

```
array([[0.38888889, 0.26797386],
       [0.06535948, 0.27777778]])
```

APRI Model



- Idea:
  - Choose threshold of APRI value
  - Those with APRI < threshold, assign as non-advanced or non-cirrhotic
  - Those with APRI > threshold, assign as advanced or cirrhotic

APRI Model

```
array([[ 1, 60, 20],
       [ 0, 71, 49],
       [ 0, 38, 67]], dtype=int64)
```

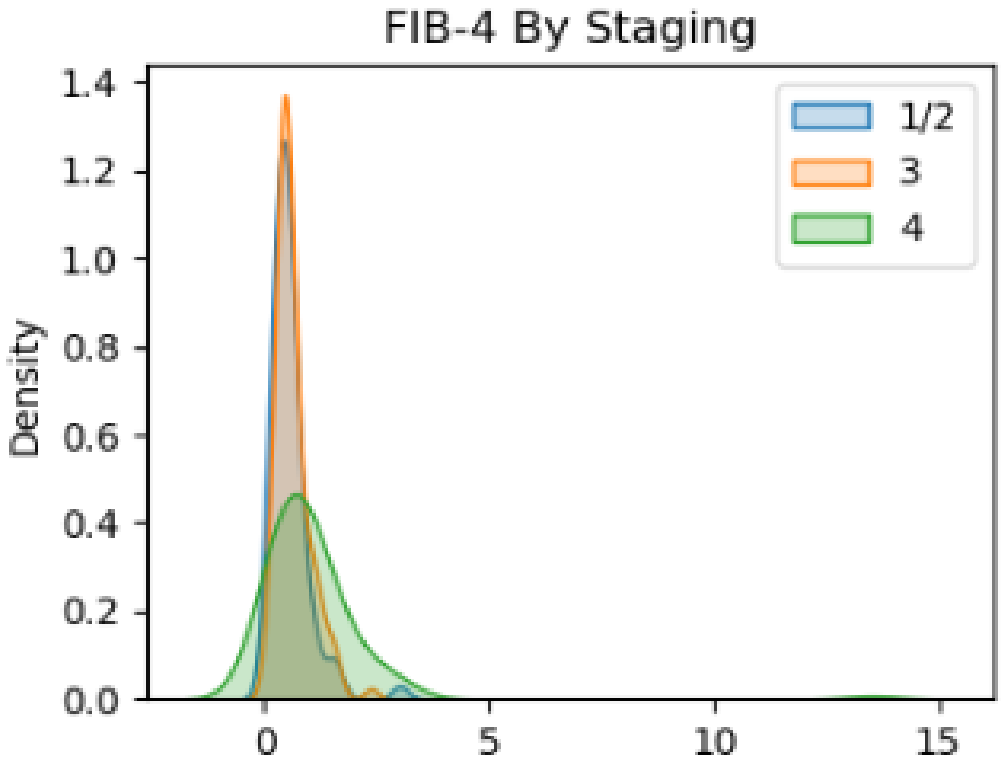
		Advanced Fibrosis (F3)	Cirrhosis (F4)
APRI Score	AUC	0.673	0.660
	Best cutoff	-1.060	-0.22
	Sensitivity	1.000	0.638
	Specificity	0.012	0.657
	Accuracy	0.739	0.65
	Precision	0.738	0.493
	Neg Precision	1.000	0.776
	F1-Score	0.849	0.556
	F1-Precision	0.849	0.603

- Thresholds were chosen such that they maximize F1-score (thus balancing positive prediction rate and positive occurrence)
- For predicting advanced fibrosis, predict negative only 1 time, but does so correctly
  - Assign all patients as having advanced fibrosis, precision is no better than random guessing
- For predicting cirrhosis, predict negative 55.5% of the time with 2 in 9 being incorrectly labeled
  - Compared to random guessing, predict negative less frequently, but still catch same number of true negatives
    - Improves purely on false negatives
  - Even though predicts positive more often, also significantly improves quality of positive prediction (49.3% are true positives, whereas would be 34.3% for random guessing). Also catches more true positives (86% more true positives)

```
array([[0.00326797, 0.26143791],
       [0.        , 0.73529412]])
```

```
array([[0.43137255, 0.2254902 ],
       [0.12418301, 0.21895425]])
```

FIB-4 Model



- Idea:
  - Choose threshold of FIB-4 value
  - Those with FIB-4 < threshold, assign as non-advanced or non-cirrhotic
  - Those with FIB-4 > threshold, assign as advanced or cirrhotic

FIB-4 Model

```
array([[ 0, 59, 22],
       [ 0, 79, 41],
       [ 0, 43, 62]], dtype=int64)
```

		Advanced Fibrosis (F3)	Cirrhosis (F4)
FIB-4 Score	AUC	0.624	0.659
	Best cutoff	-2.000	-0.16
	Sensitivity	1.000	0.590
	Specificity	0.000	0.687
	Accuracy	0.735	0.654
	Precision	0.735	0.496
	Neg Precision	0.000	0.762
	F1-Score	0.660	0.539
	F1-Precision	0.000	0.601

- Thresholds were chosen such that they maximize F1-score (thus balancing positive prediction rate and positive occurrence)
- For predicting advanced fibrosis, it predicts everyone as having advanced fibrosis
- For cirrhosis, predict negative 59.1% of the time (compared to 65.7% for guessing)
  - Also of the negative predictions, only 23.8% are false negatives, where as for guessing that would be 34.3%
  - Even though predicts positive more often, also significantly improves quality of positive prediction (49.6% are true positives, whereas would be 34.3% for random guessing). Also catches more true positives (72% more true positives

```
array([[0.        , 0.26470588],
       [0.        , 0.73529412]])
```

```
array([[0.45098039, 0.20588235],
       [0.14052288, 0.20261438]])
```

RF

```
array([[20.1, 52.7,  8.2],
       [20.5, 77.3, 22.2],
       [ 4. , 45.8, 55.2]])
```

- Advanced fibrosis, then cirrhosis

```
7, 0.18300645],
7, 0.6682129 ]])
```

	recall	specificity	precision	neg_precision	f1-score	accuracy	support_pos	support_neg
mean	0.91	0.314	0.786	0.565	0.842	0.75	45.552	16.448
95.0% CI	[0.819, 0.979]	[0.118, 0.538]	[0.679, 0.889]	[0.25, 0.875]	[0.766, 0.904]	[0.645, 0.839]	[39.0, 52.0]	[10.0, 23.0]

```
516, 0.09054839],
839, 0.19185806]])
```

	recall	specificity	precision	neg_precision	f1-score	accuracy	support_pos	support_neg
mean	0.564	0.863	0.684	0.791	0.609	0.758	21.254	40.746
95.0% CI	[0.348, 0.789]	[0.75, 0.952]	[0.474, 0.889]	[0.667, 0.907]	[0.432, 0.758]	[0.661, 0.855]	[15.0, 28.0]	[34.0, 47.0]

Log Regression

```
array([[10.8, 65.3,  4.9],
       [ 6.8, 96.5, 16.7],
       [ 1.4, 57.5, 46.1]])
```

```
0.07347097, 0.1906129 ],
0.07696774, 0.65894839]])
```

	recall	specificity	precision	neg_precision	f1-score	accuracy	support_pos	support_neg
mean	0.896	0.286	0.777	0.512	0.83	0.732	45.627	16.373
95.0% CI	[0.773, 0.98]	[0.067, 0.536]	[0.661, 0.887]	[0.2, 0.875]	[0.753, 0.9]	[0.629, 0.839]	[39.0, 52.0]	[10.0, 23.0]

```
0.57569677, 0.08158065],
0.15047742, 0.19224516]])
```

	recall	specificity	precision	neg_precision	f1-score	accuracy	support_pos	support_neg
mean	0.564	0.877	0.707	0.794	0.619	0.768	21.249	40.751
95.0% CI	[0.346, 0.769]	[0.759, 0.974]	[0.5, 0.923]	[0.673, 0.905]	[0.437, 0.765]	[0.677, 0.855]	[15.0, 28.0]	[34.0, 47.0]

Adv Fibrosis

- RF and log regression nearly identical results, but RF slightly better (may not be statistically significant against larger population) at differentiating TP vs FP and TN vs FN
- ALBI model may be over predicting positive cases (90% positive prediction), but at least might lower FN rate than either RF or LR

Fibrosis

- Again, RF and LR are nearly identical, but this time LR is slightly better at distinguishing between TP vs FP and TN vs FN
- While ALBI model has lower FN rate, both RF and LR are more confident in N predictions (precision of negatives is higher)
- Both RF and LR also have higher precision