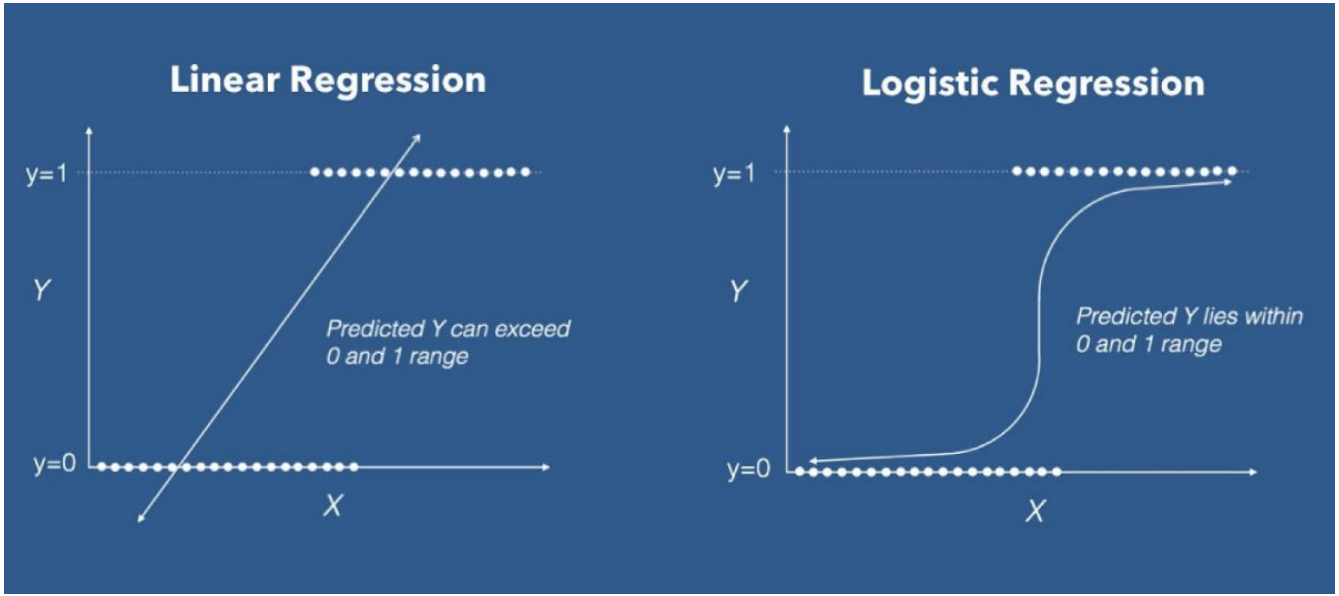
Logistic regression

Data Science for Biologists, Fall 2021
Dr. Spielman

Linear regression (linear model)	Logistic regression (logistic model)
Response is **numeric/continuous**	Response is **binary** (Yes/No, Sick/Healthy)
Model is a **straight line** without bounds	Model is a **logistic curve**, where 0 <= Y <= 1

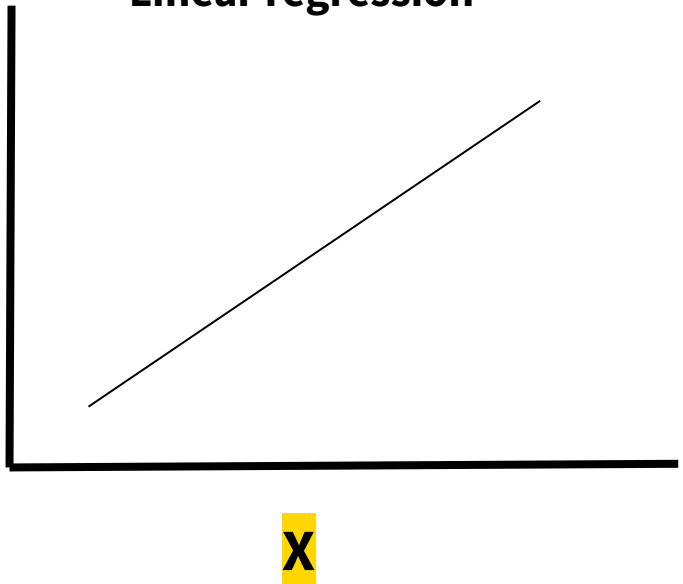


Logistic regression is just a *transformation* of linear regression

- "Binary" outcome can be thought of numerically: **0 and 1 are the response values**
- Logistic regression steps:
  - Perform linear regression! $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_n X_n + \epsilon$
  - Let's call that *t*: $t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_N X_N + \epsilon$
  - Get a new Y axis through *logistic transformation* : $Y = \frac{1}{1+e^{-t}}$

Visually...

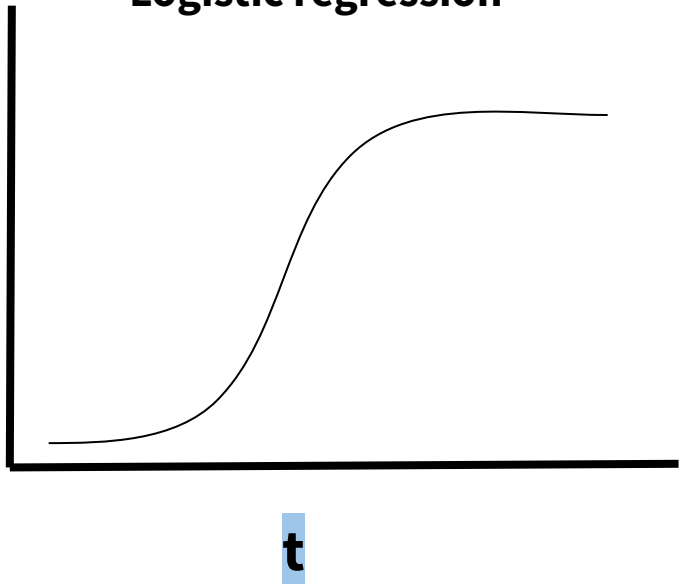
Linear regression



Y (or p)

$$Y = \frac{1}{1+e^{-t}}$$

Logistic regression



Why is this useful for *classification?*

$$Y = \frac{1}{1+e^{-t}}$$

- Our logistic Y-axis values only range between [0,1] (math!)
- **The Y-axis is a probability!!** Classifiers provide the **probability** of being in one group or the other.
- Requires a **threshold** for calling positive/negative: "anything about 75% is considered a positive result" for example.
    - Is your test result positive or negative? Is the tumor benign or malignant? **....Rarely have "yes" or "no" answers. Instead....**
    - "There is a 95% chance your tumor is benign."
    - "There is a 25% chance your test results mean you are positive."

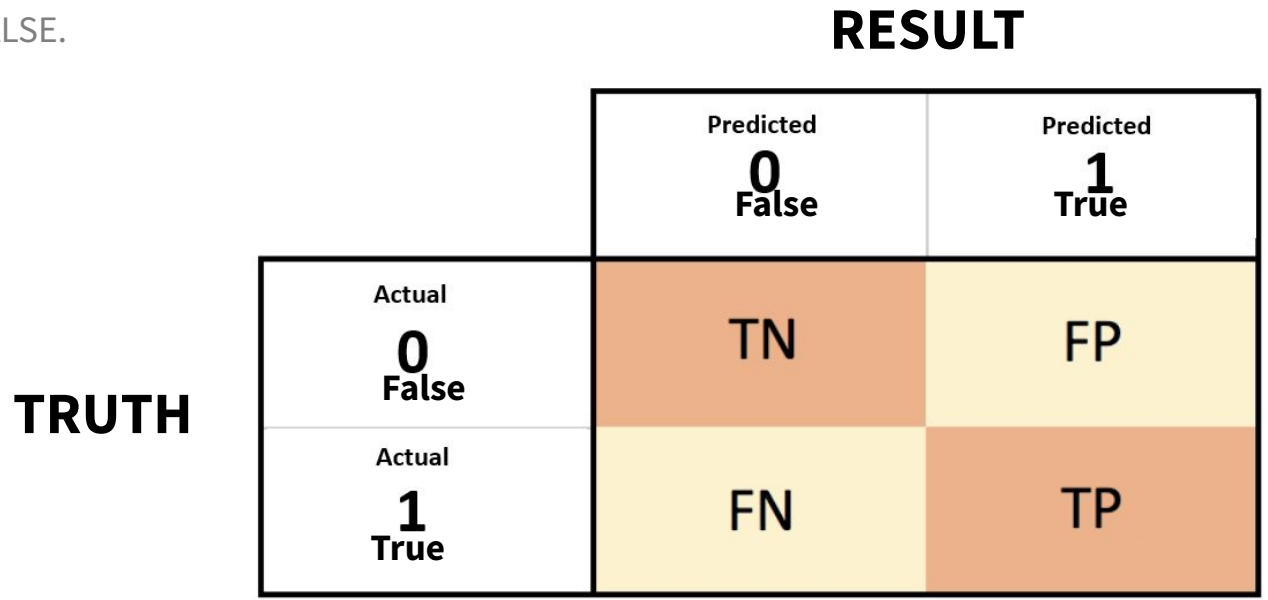
Evaluating performance of logistic regressions

- R^2 and RMSE are some approaches to describing performance of linear regression, ***but not logistic regression!!***

- **We evaluate logistic regression with:**
  - **Confusion matrix quantities**
  - **Receiver Operating Characteristic (ROC) curves and AUC**
    - Unless data is highly imbalanced (e.g. 10000+ vs 2-), in which case there are other curves

Confusion matrix

- **First ask: is the result positive or negative?**
  - "Successes" are positive and "failures" are negative.
- **Then ask:** should we have gotten that result though?
  - If yes, TRUE. If not, FALSE.



Let's practice... result vs. truth

- Clinical trial results show a new arthritis drug reduces pain, but in reality it does not reduce pain.
- A person with HIV receives a positive test result for HIV.
- A person using illegal performance enhancing drugs passes a test clearing them of drug use.
- A study found a significant relationship between neck strain and jogging, when in reality there is no relationship.
- A healthy individual gets a negative cancer biopsy result.
- Someone with COVID-19 receives a negative COVID-19 test result.
- A study found that people who sleep 8 hours a night have less depression, and in reality sleeping 8 hours a night does reduce risk of depression.

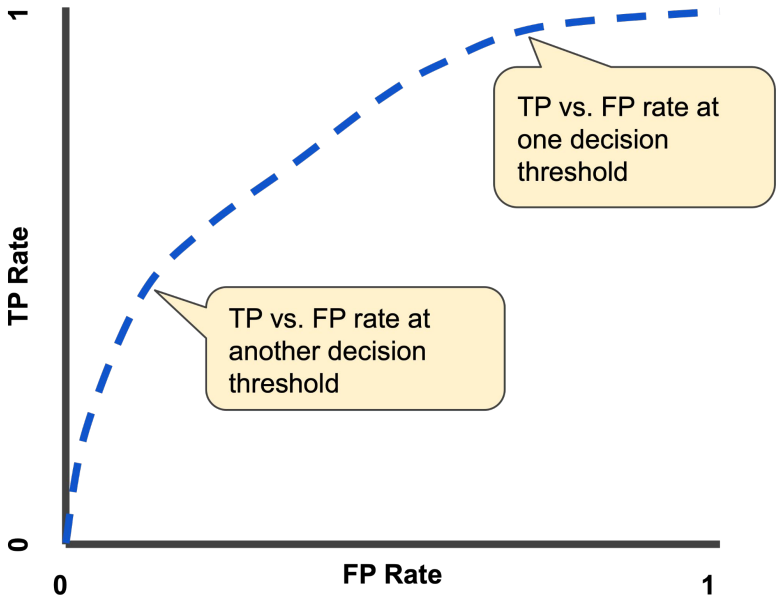
Confusion matrices help describe *performance of a classifier*

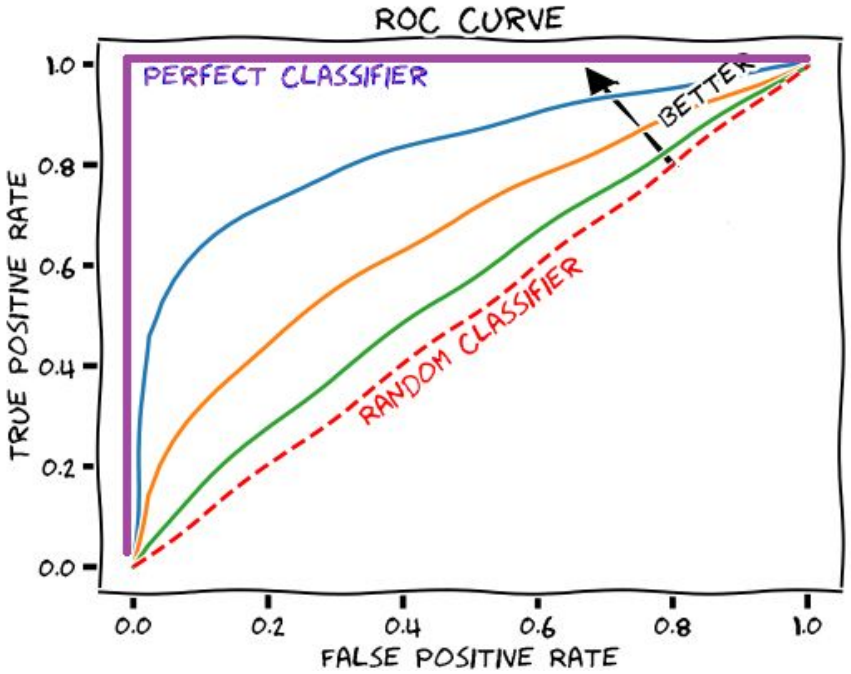
- https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
- https://towardsdatascience.com/taking-the-confusion-out-of-confusion-matrices-c1ce054b3d3e
- https://en.wikipedia.org/wiki/Confusion_matrix
- An abbreviated list of formulas
  - **True positive rate:** *TP / (TP + FN)*
    - **aka Sensitivity aka Recall**
  - **True negative rate:** *TN/(TN + FP)*
    - **aka Specificity**
  - **False positive rate:** *FP / (FP + TN)*
    - aka **"1 - Specificity"**
  - **False discovery rate:** *FP / (FP + TP)*
  - **Precision:** *TP / (TP + FP)*
    - **aka Positive Predictive Value**
  - **Accuracy:** *(TP + TN) / (TP + TN + FP + FN)*

Calculating confusion matrix metrics

- Choose a probability threshold for what you consider "positive" (success) vs "negative" (failure) model result
- Determine classifications of results based on this threshold, eg…
  - I choose at **80% threshold**
  - This data point had an **85%** chance of being a success. **It's a positive result under this threshold**
  - Another data point had a **75%** chance of being a success. **It's a negative result under this threshold**
- Plug and chug!!! (Back to R!)

ROC curves: When 1 threshold isn't enough

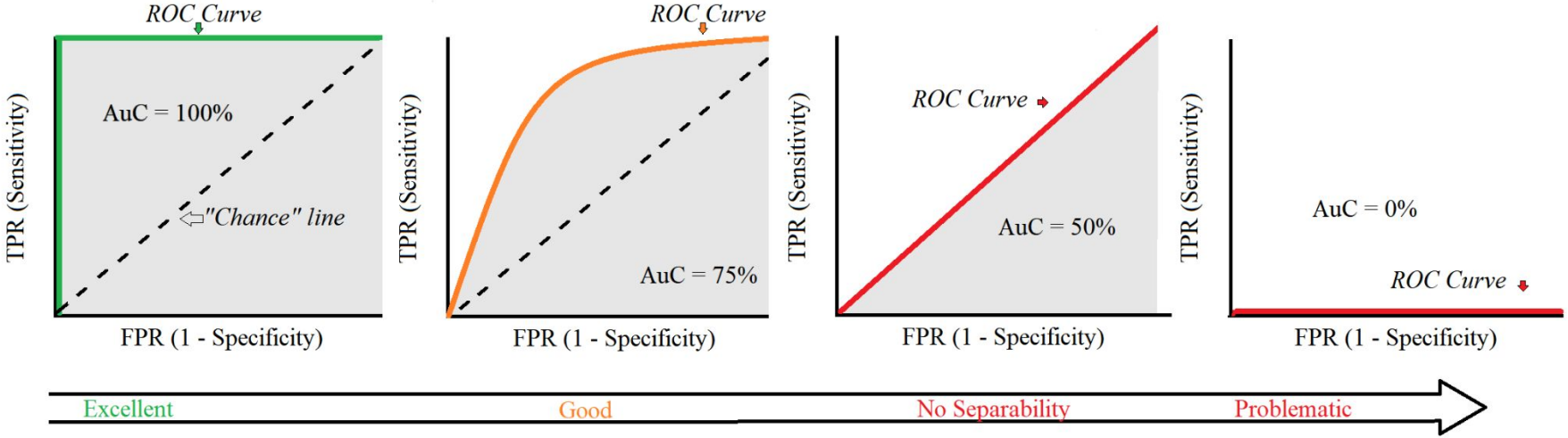




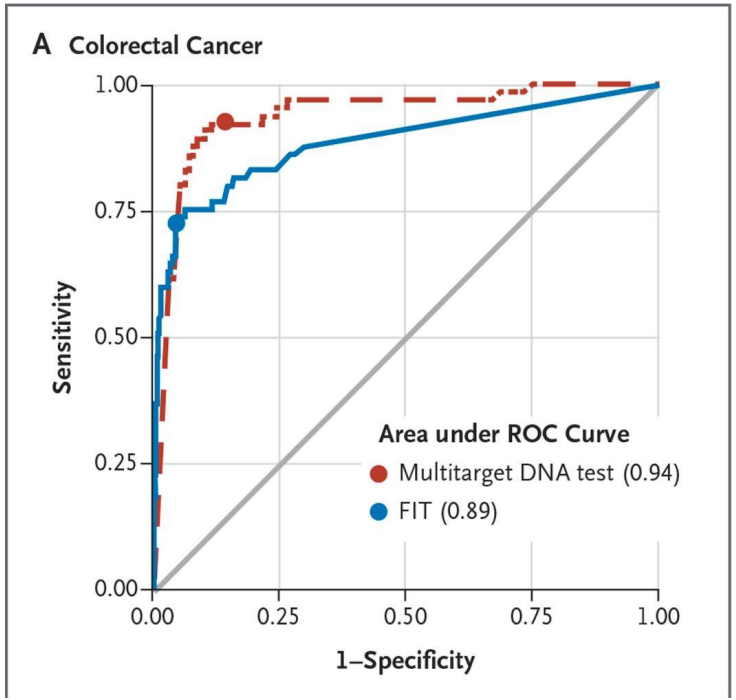
https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/

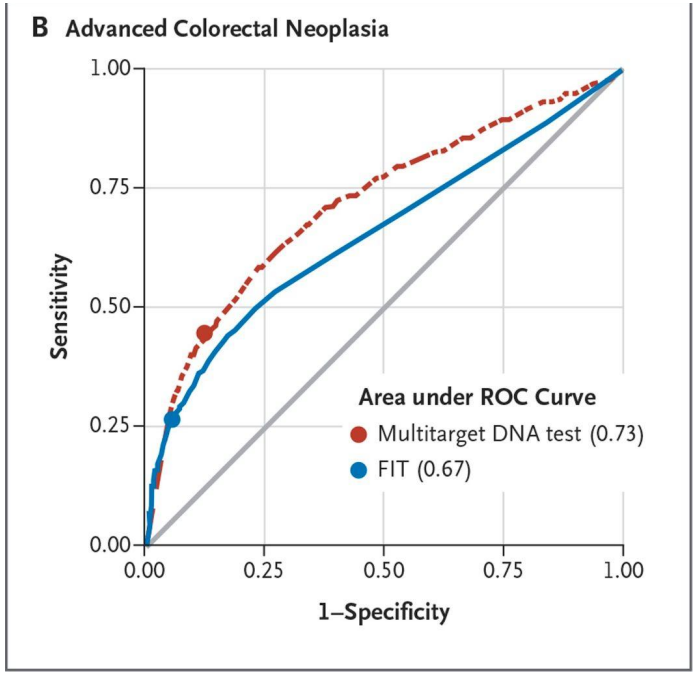
The Area Under the Curve (AUC) classifies overall performance across thresholds

- AUC ranges from 0-1.
  - 1 = perfect classifier! The mode is *always right.*
  - 0.5 = classifier is no better than *random chance*.
  - <0.5 usually indicates you have some kind of funky thing going on in your code.



ROC curves out in the wild





Imperiale et al. N Engl J Med 2014;

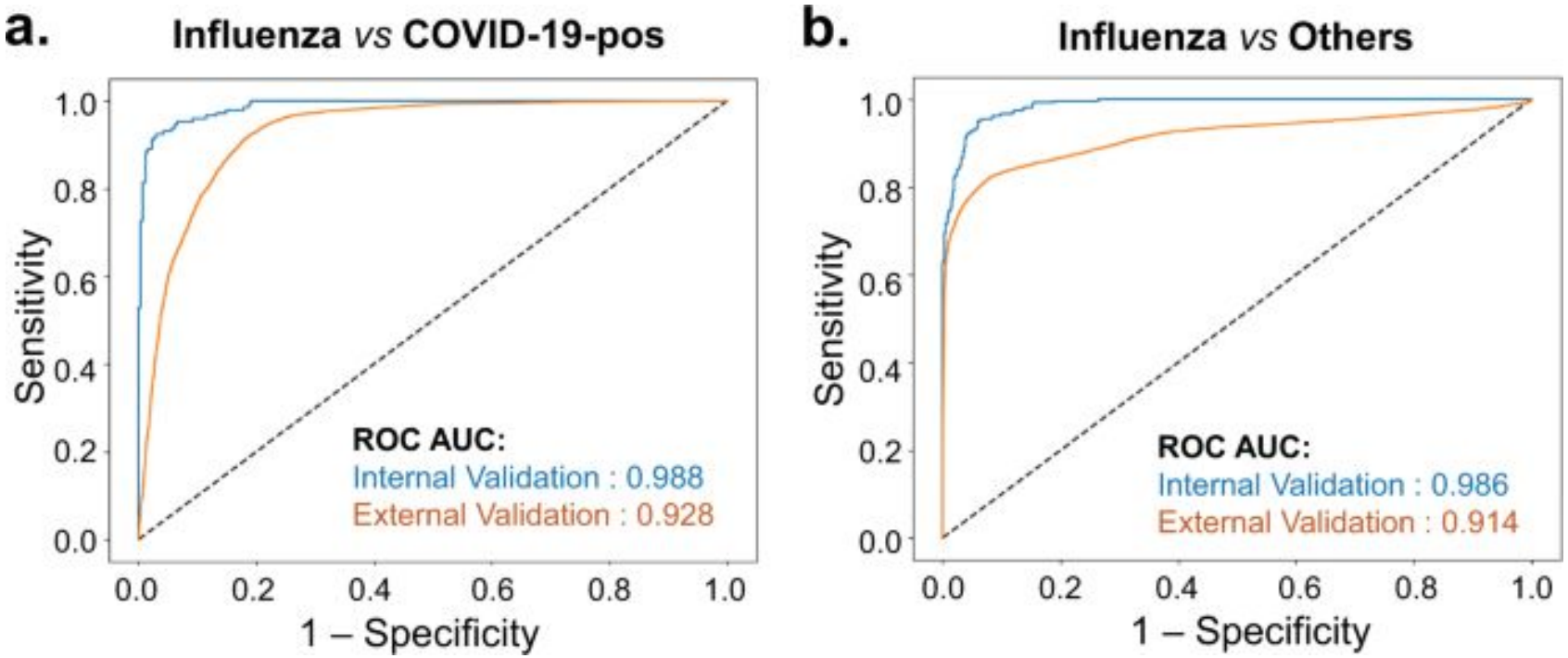
"A vital sign-based prediction algorithm for differentiating COVID-19 versus seasonal influenza in hospitalized patients"

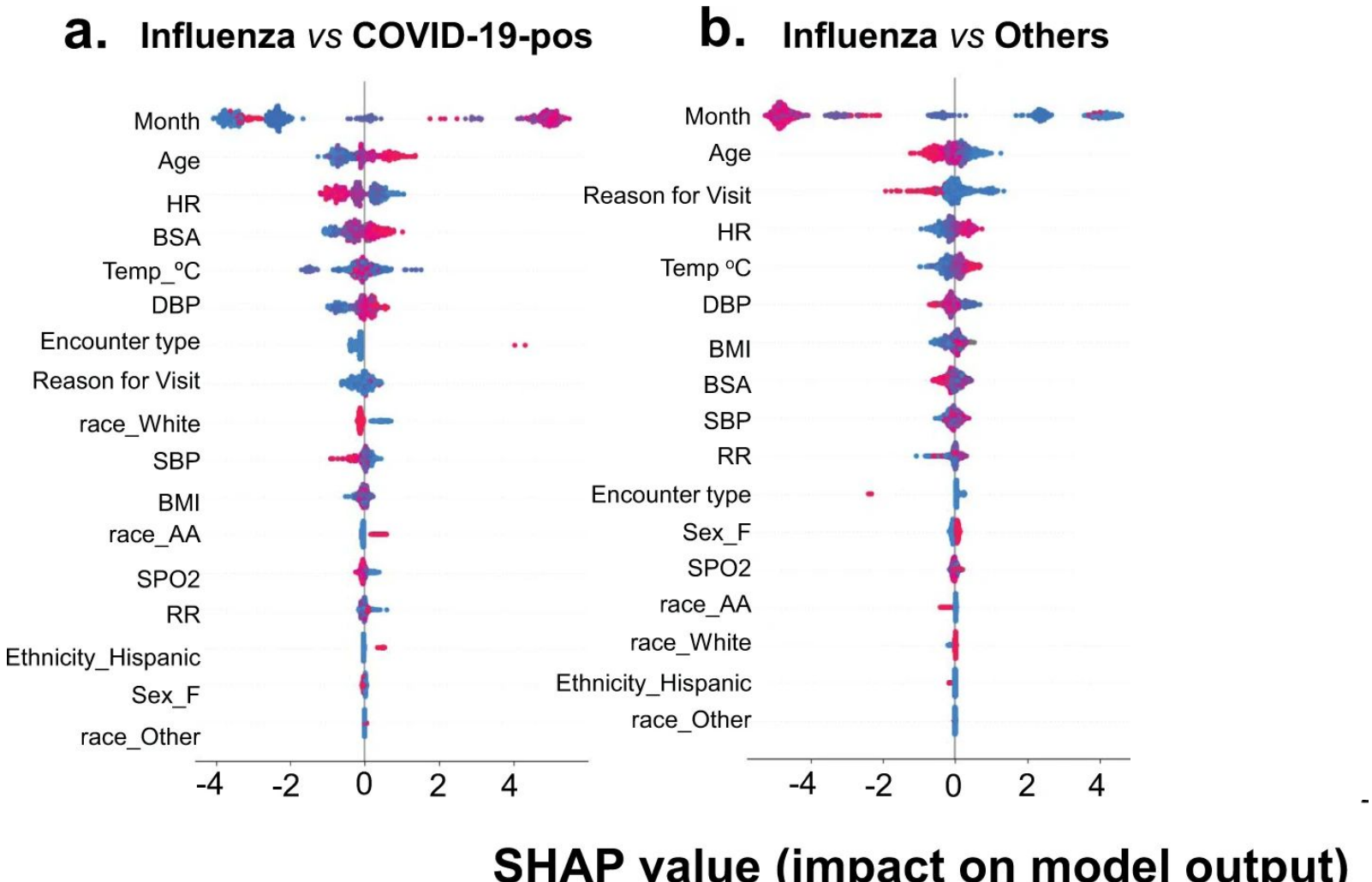
https://www.nature.com/articles/s41746-021-00467-8

- Built a model from 3883 patients
  - 19% tested positive for COVID
  - 49% tested negative for COVID
  - 31% tested positive for influenza
- Internal validation:
  - https://www.nature.com/articles/s41746-021-00467-8/tables/3
- External validation:
  - https://www.nature.com/articles/s41746-021-00467-8/tables/4
  -

"A vital sign-based prediction algorithm for differentiating COVID-19 versus seasonal influenza in hospitalized patients"

https://www.nature.com/articles/s41746-021-00467-8




a. Influenza vs COVID-19-pos
Month
Age
HR
BSA
Temp_°C
DBP
Encounter type
Reason for Visit
race_White
SBP
BMI
race_AA
SPO2
RR
Ethnicity_Hispanic
Sex_F
race_Other
-4
-2
0
2
4
b. Influenza vs Others
Month
Age
Reason for Visit
HR
Temp °C
DBP
BMI
BSA
SBP
RR
Encounter type
Sex_F
SPO2
race_AA
race_White
Ethnicity_Hispanic
race_Other
-4
-2
0
2
4
SHAP value (impact on model output)

Finally, let's head to to R!