Digital Transformation of Healthcare

Evaluating Predictions

Michoel Snow, M.D. Ph.D., Glen Ferguson, Ph.D.

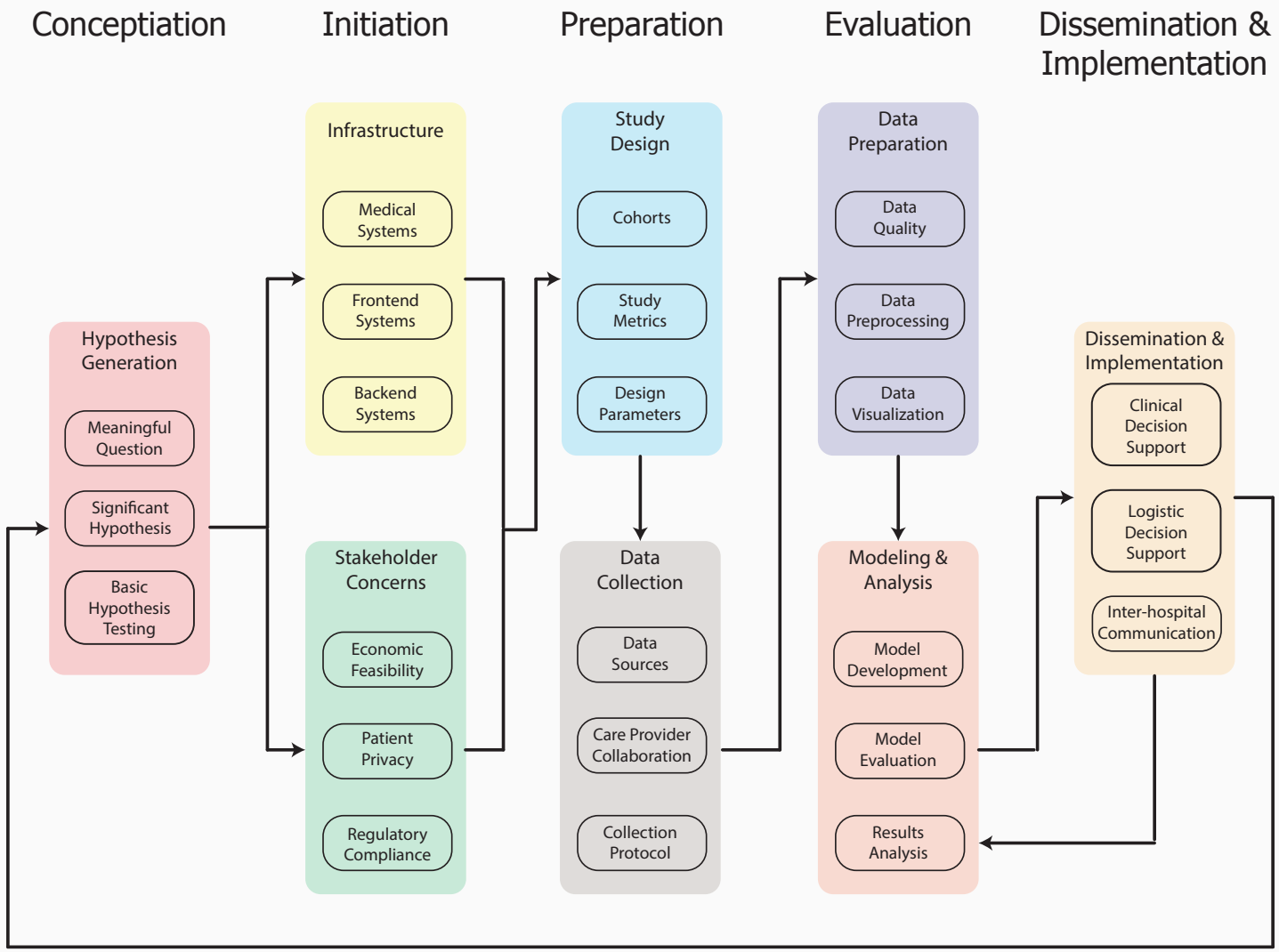
Center for Health Data Innovations

Evaluating Predictions

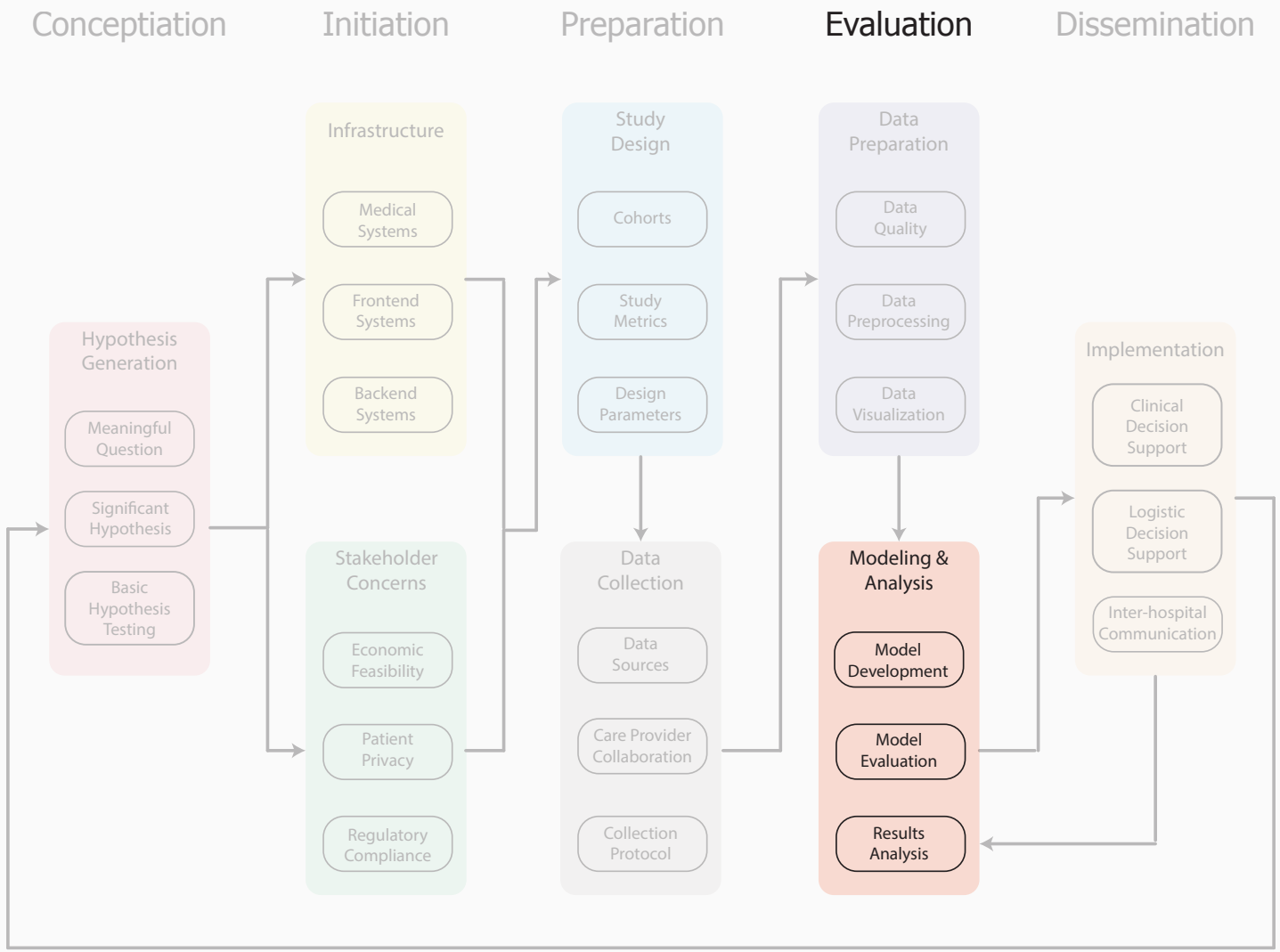
After this lecture students will be able to

- Calculate common classification and regression metrics
- Describe the role of simple classification metrics
- Evaluate the implementation of metrics for a study
- Articulate the information underlying common compound classification metrics
- Classify regression metrics
- Connect regression metric outcomes to facets of the associated models
- Identify transition points which can affect data quality
- Discuss methods for measuring and evaluating data quality

Bioinformatics Pipeline



Evaluating Predictions



Metrics for Evaluation of Classification Models

Terms and Questions

Terms

- Accuracy
- Specificity
- Sensitivity
- Positive Predictive Value
- Negative Predictive Value
- Likelihood Ratio
- ROC & AUC
- F1 Score

Questions

- Is accuracy a useful metric?
- What information is conveyed by sensitivity vs specificity ?
- What information do the PPV and NPV add?
- Intuitively, how do sensitivity, specificity, likelihood ratios and ROC connect?
- Is the F1 score a more robust metric than the ROC and AUC?

Clinical Cases

- Low dose CT for detecting lung cancer (LDCT)[1]
- Ultrasound detection of abdominal aortic aneurysms (AAA)[2]
- Blood pressure monitoring in adolescents using home machines (HTN)[3]
- Detecting suicidality among adolescent outpatients by clinicians versus trained raters using the Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS-PL) [4]

---

[1] National Lung Screening Trial Research Team. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine, 365(5), 395-409.

[2] Thompson, S. G., Ashton, H. A., Gao, L., Buxton, M. J., Scott, R. A. P., & Multicentre Aneurysm Screening Study (MASS) Group. (2012). Final followup of the Multicentre Aneurysm Screening Study (MASS) randomized trial of abdominal aortic aneurysm screening. British Journal of Surgery, 99(12), 1649-1656.

[3] Stergiou, G. S., Nasothimiou, E., Giovas, P., Kapoyiannis, A., & Vazeou, A. (2008). Diagnosis of hypertension in children and adolescents based on home versus ambulatory blood pressure monitoring. Journal of hypertension, 26(8), 1556-1562.

[4] Holi, M. M., Pelkonen, M., Karlsson, L., Tuisku, V., Kiviruusu, O., Ruuttu, T., & Marttunen, M. (2008). Detecting suicidality among adolescent outpatients: evaluation of trained clinicians' suicidality assessment against a structured diagnostic assessment made by trained raters. BMC psychiatry, 8(1), 97.

Confusion Matrix

Lung Cancer

Low-Dose CT	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound	p	n
p′	600	734
n′	61	25,480

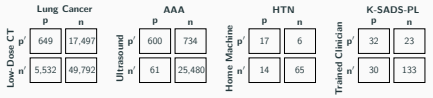
HTN

Home Machine	p	n
p′	17	6
n′	14	65

K-SADS-PL

Trained Clinician	p	n
p′	32	23
n′	30	133

- What is the accuracy of these tests?
- Are these good screening tests and/or diagnostic tests?
- How does the PPV and NPV affect your opinion of their utility?
- When are sensitivity, specificity, PPV and NPV appropriate tests?
- How are sensitivity, specificity, PPV and NPV affected by prevalence?



- What is the accuracy of these tests?
- Are these good screening tests and/or diagnostic tests?
- How does the PPV and NPV affect your opinion of their utility?
- When are sensitivity, specificity, PPV and NPV appropriate tests?
- How are sensitivity, specificity, PPV and NPV affected by prevalence?

Parameter	Interpretation	Appropriate for
Accuracy	Overall proximity of test to reality	Balanced sample sizes
Sensitivity	Chance of a false negative	Cheap further testing/Severe disease
Specificity	Chance of a false positive	Expensive further testing/Mild disease
PPV	Sensitivity diagnostic utility	Balanced prevalence
NPV	Specificity diagnostic utility	Balanced prevalence

Case	**TP**	**FP**	**TN**	**FN**	**Sens**	**Spec**	**PPV**	**NPV**	**Acc**	**F1**
LDCT	649	17,497	49,792	5,532	10	74	4	90	69	6
AAA	600	734	25,480	61	91	97	45	100	97	60
HTN	17	6	65	14	55	92	74	82	80	63
KSADS	32	23	133	30	52	85	58	82	76	55

Combined Statistics

Lung Cancer

Low-Dose CT	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound	p	n
p′	600	734
n′	61	25,480

HTN

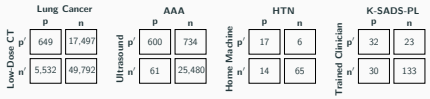
Home Machine	p	n
p′	17	6
n′	14	65

K-SADS-PL

Trained Clinician	p	n
p′	32	23
n′	30	133

- What are 4 'sensible' pairings of the base stats
- What are the different ways to combine the base stats into summary statistics (hint: what are the basic ways to combine any numbers)?
  - Work through each of the four clinical cases
- What determines the split of positive cases into TP vs FN and negative cases into TN vs FP?

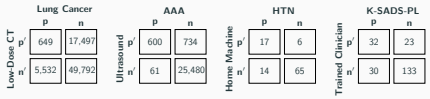
Combined Statistics



	Lung Cancer		AAA		HTN		K-SADS-PL	
	p	n	p	n	p	n	p	n
Low-Dose CT p′	649	17,497						
Low-Dose CT n′	5,532	49,792						
Ultrasound p′			600	734				
Ultrasound n′			61	25,480				
Home Machine p′					17	6		
Home Machine n′					14	65		
Trained Clinician p′							32	23
Trained Clinician n′							30	133

- What are 4 'sensible' pairings of the base stats
- What are the different ways to combine the base stats into summary statistics (hint: what are the basic ways to combine any numbers)?
  - Work through each of the four clinical cases
- What determines the split of positive cases into TP vs FN and negative cases into TN vs FP?

- sens/spec, PPV/NPV, sens/PPV, spec/NPV
- the 4 basic operations are add, subtract, multiply and divide **add Ex**
- **Add** adding just gives you the numbers themselves without an idea of how they each contribute. averaging sens/spec or ppv/npv is a good summary of how they perform. Averaging sens/ppv tells you how well you can predict TP taking into account the reliability of the test and the prevalence of the disease, and specifically ignoring the effects of TN.
- **Subtract** Not really a helpful metric as the difference between values doesn't tell you much about the values themselves. F1 score combines the typical mean and the difference between the values
- **multiply** Similar to averaging and F1 but punishes if both lower values
- **divide** dividing two probabilities gives you an odds ratio, i.e., how much more likely the numerator is to happen than the denominator.



Lung Cancer	p	n
p′	649	17,497
n′	5,532	49,792

AAA	p	n
p′	600	734
n′	61	25,480

HTN	p	n
p′	17	6
n′	14	65

K-SADS-PL	p	n
p′	32	23
n′	30	133

- What are 4 'sensible' pairings of the base stats
- What are the different ways to combine the base stats into summary statistics (hint: what are the basic ways to combine any numbers)?
  - Work through each of the four clinical cases
- What determines the split of positive cases into TP vs FN and negative cases into TN vs FP?

$$LR+ = \frac{sensitivity}{1 - specificity} = \frac{P(T+ \mid D+)}{P(T+ \mid D-)}$$

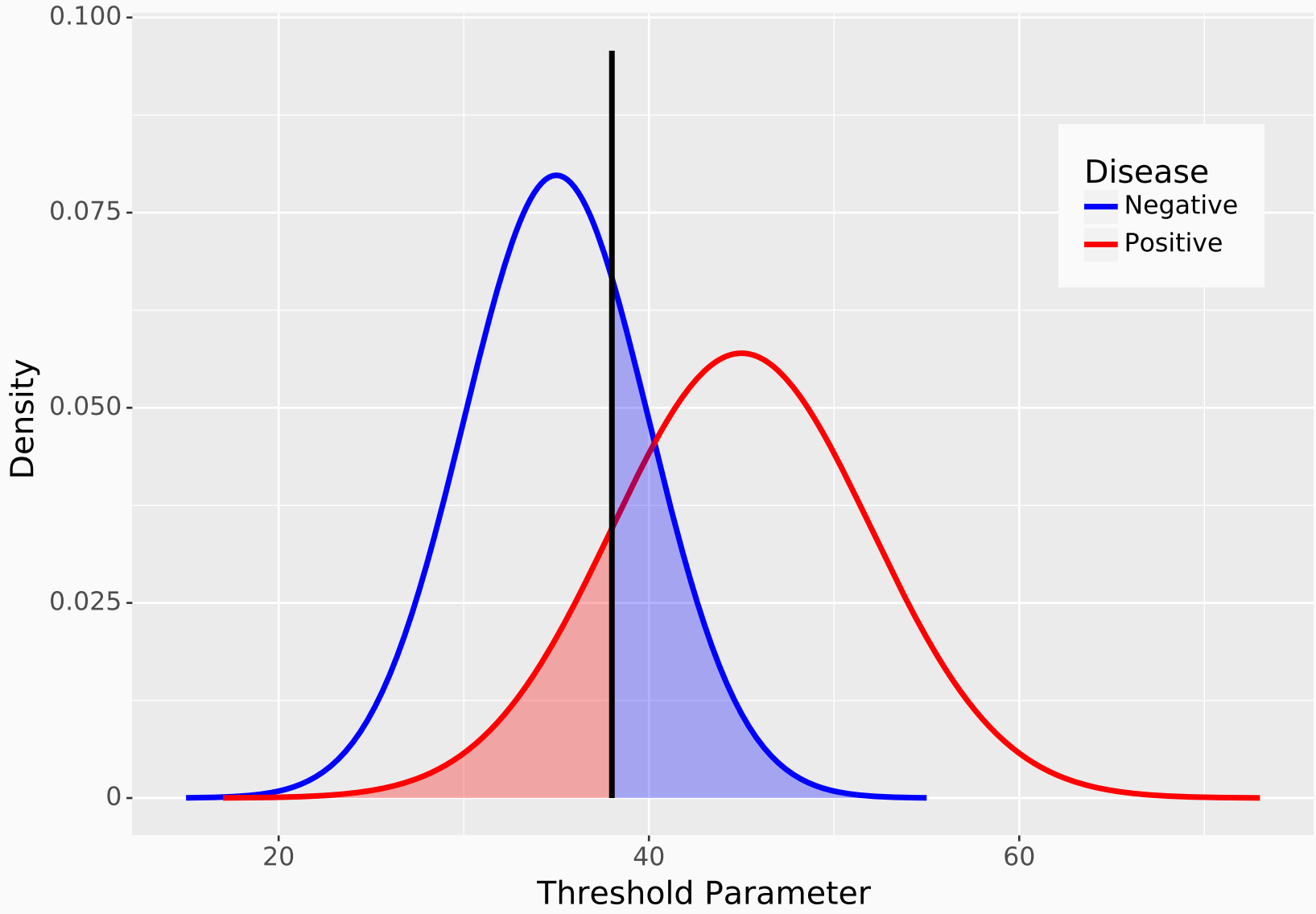
$$LR- = \frac{1 - sensitivity}{specificity} = \frac{P(T- \mid D+)}{P(T- \mid D-)}$$

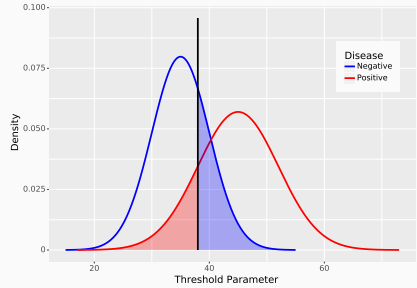
Likelihood Ratio	Approximate Change in Probability(%)
0.1	-45
0.2	-30
0.5	-15
1	0
2	+15
5	+30
10	+45

Change in post test probability $\approx 0.2 \times \ln LR$

McGee, Steven. "Simplifying likelihood ratios." Journal of general internal medicine 17.8 (2002): 647-650. APA

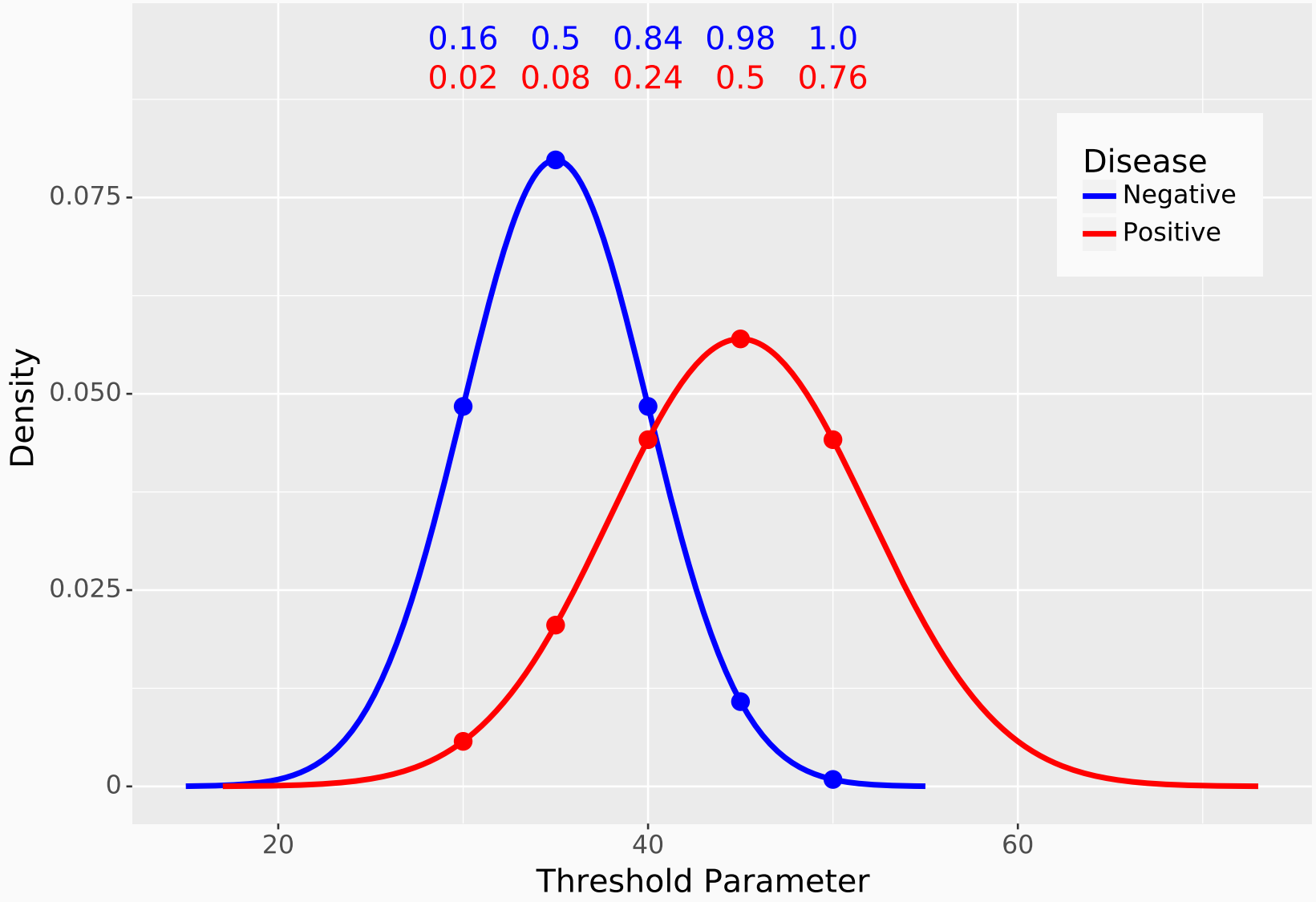
Hypothesis Testing

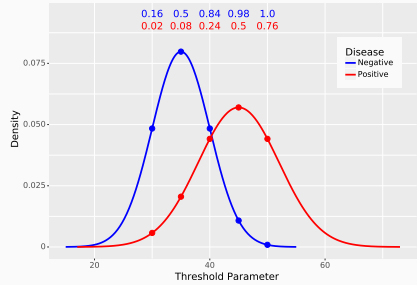




- these two curves represent the positive and negative cases
- As the threshold parameter from right to left, your sensitivity increases but you specificity decreases
- In order to calculate specific values we need to know what the areas under the curve are, for different thresholds

Hypothesis Testing





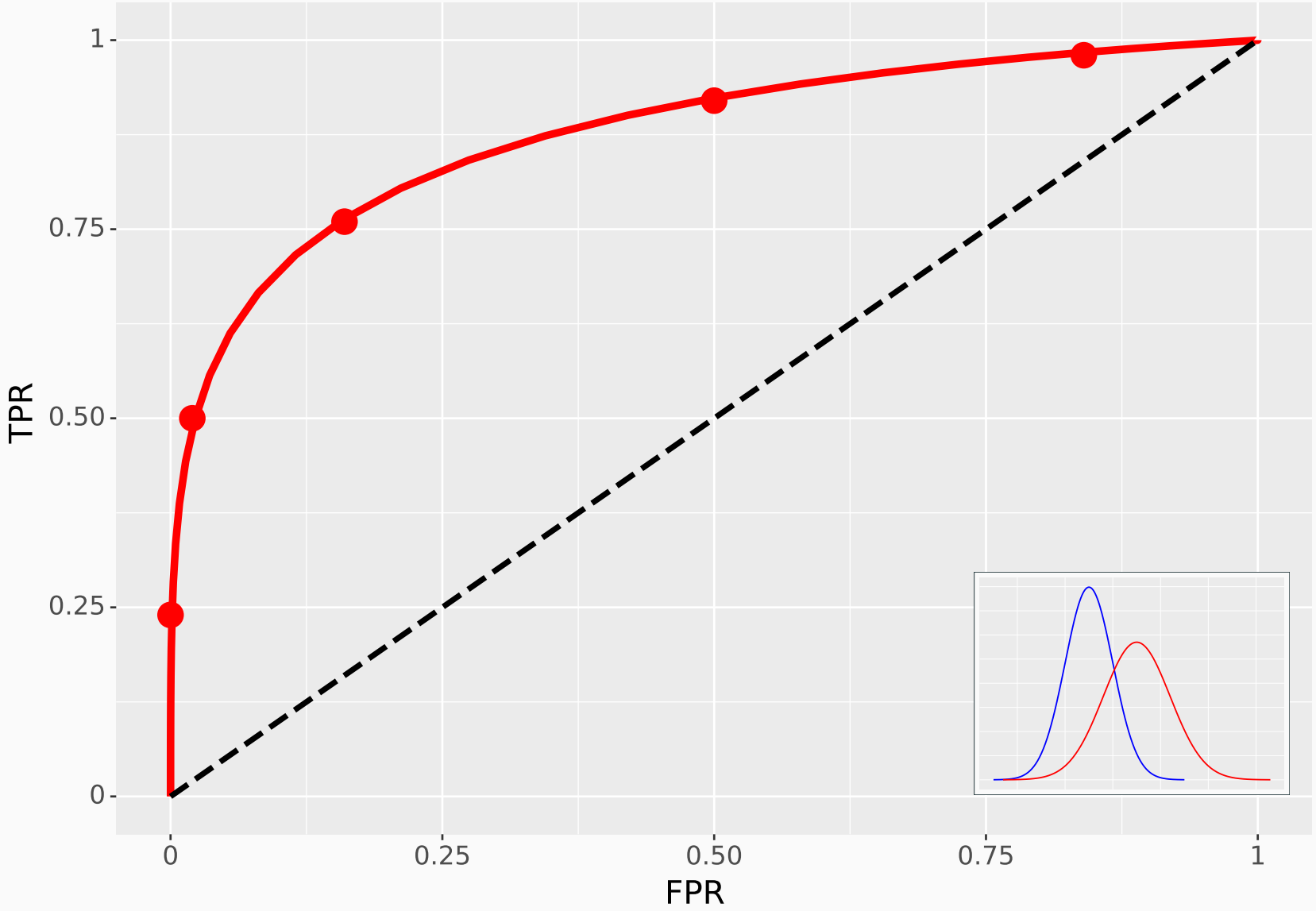
- Let's calculate the odds ratio for various points along the curve

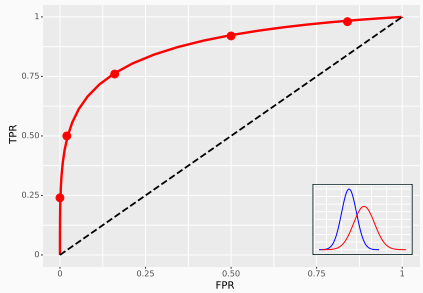
N/P	Sens	Spec	Odds	PPV	NPV	F1
0.16/002	0.98	0.16	1.17	0.54	0.89	0.70
0.5/0.08	0.92	0.50	1.84	0.65	0.86	0.76
0.84/0.24	0.76	0.84	4.78	0.83	0.78	0.79
0.98/0.5	0.50	0.98	26.32	0.96	0.66	0.66
1.0/0.76	0.24	1.00	∞	1	0.57	0.39

N/P	PPV-25%P	NPV-25%P	PPV-75%P	NPV-75%P	f1-25%P	f1-75%P
0.16/002	0.28	0.96	0.78	0.73	0.44	0.87
0.5/0.08	0.38	0.95	0.85	0.68	0.54	0.88
0.84/0.24	0.61	0.91	0.93	0.54	0.68	0.84
0.98/0.5	0.89	0.86	0.99	0.40	0.64	0.66
1.0/0.76	1	0.80	1.00	1	0.31	0.39

- both of these curves have an area of 1, so we need to multiply each metric by the percentage of patients to see how prevalence affects the outcome

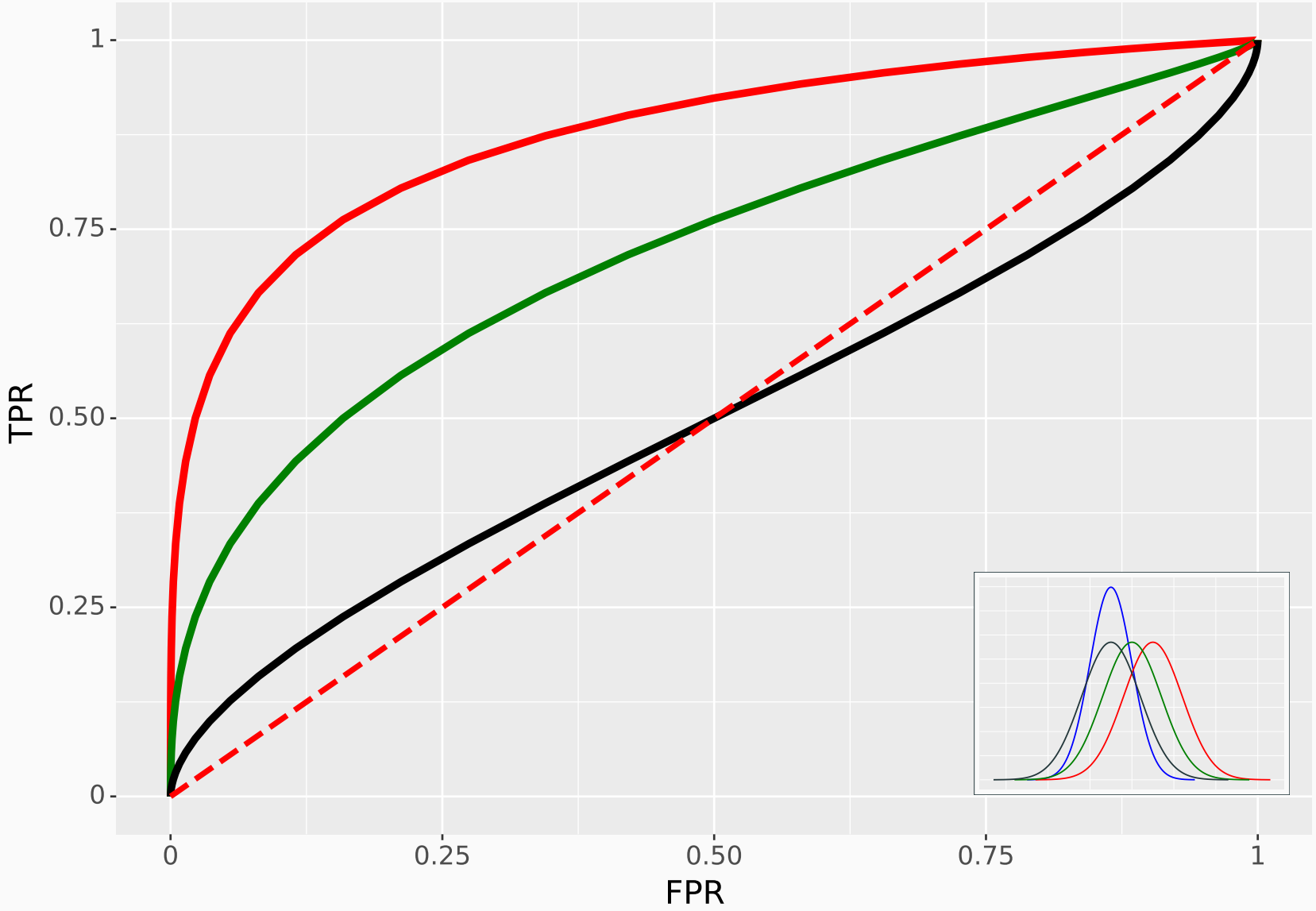
Receiver Operating Characteristic



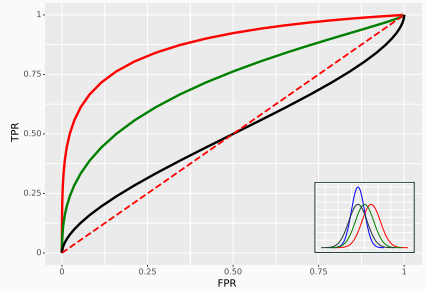


1. What does the Area Under the Curve (AUC) correspond to? - Given a positive test result what are the chances that the subject is truly positive irrespective of prevalence?
2. What does the diagonal correspond to? - equal probabilities
3. PPV is threshold dependent, while AUC is threshold independent but variable dependent

Receiver Operating Characteristic



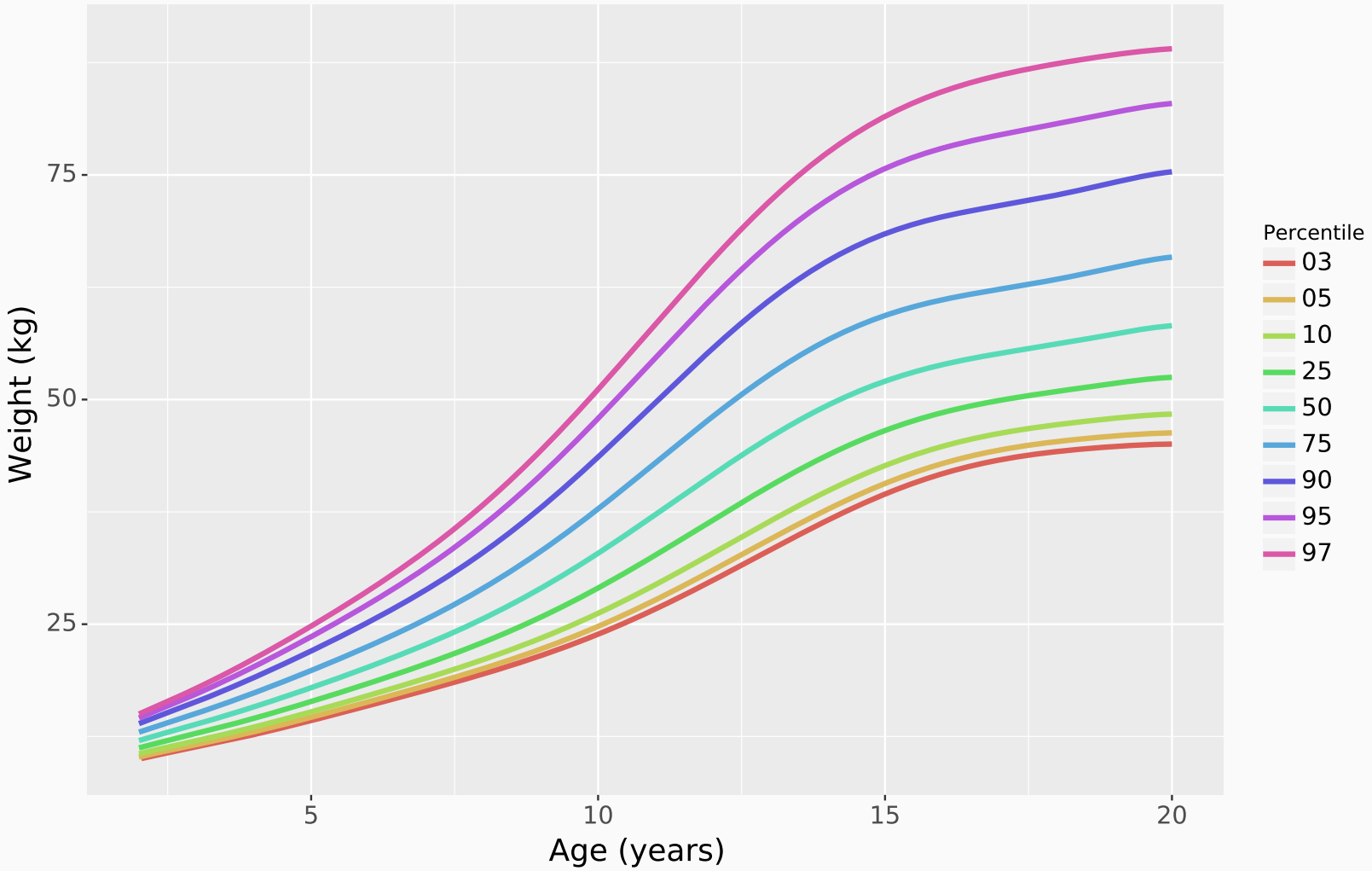
Receiver Operating Characteristic



Test	Information	Situation	Prev dep
Sensitivity	Chance of a false negative	Cheap further testing, severe disease	No
Specificity	Chance of false positive	Expensive further testing, mild disease	No
PPV, NPV	how good is test at pulling out true cases	balanced prevalence	Yes
Likelihood	odds of true result over false result	medium to low suspicion	No
ROC-AUC	how likely is a random test going to be from a true case		No
F1	How good are you at spotting TP	maximizing TP matter the most	Yes

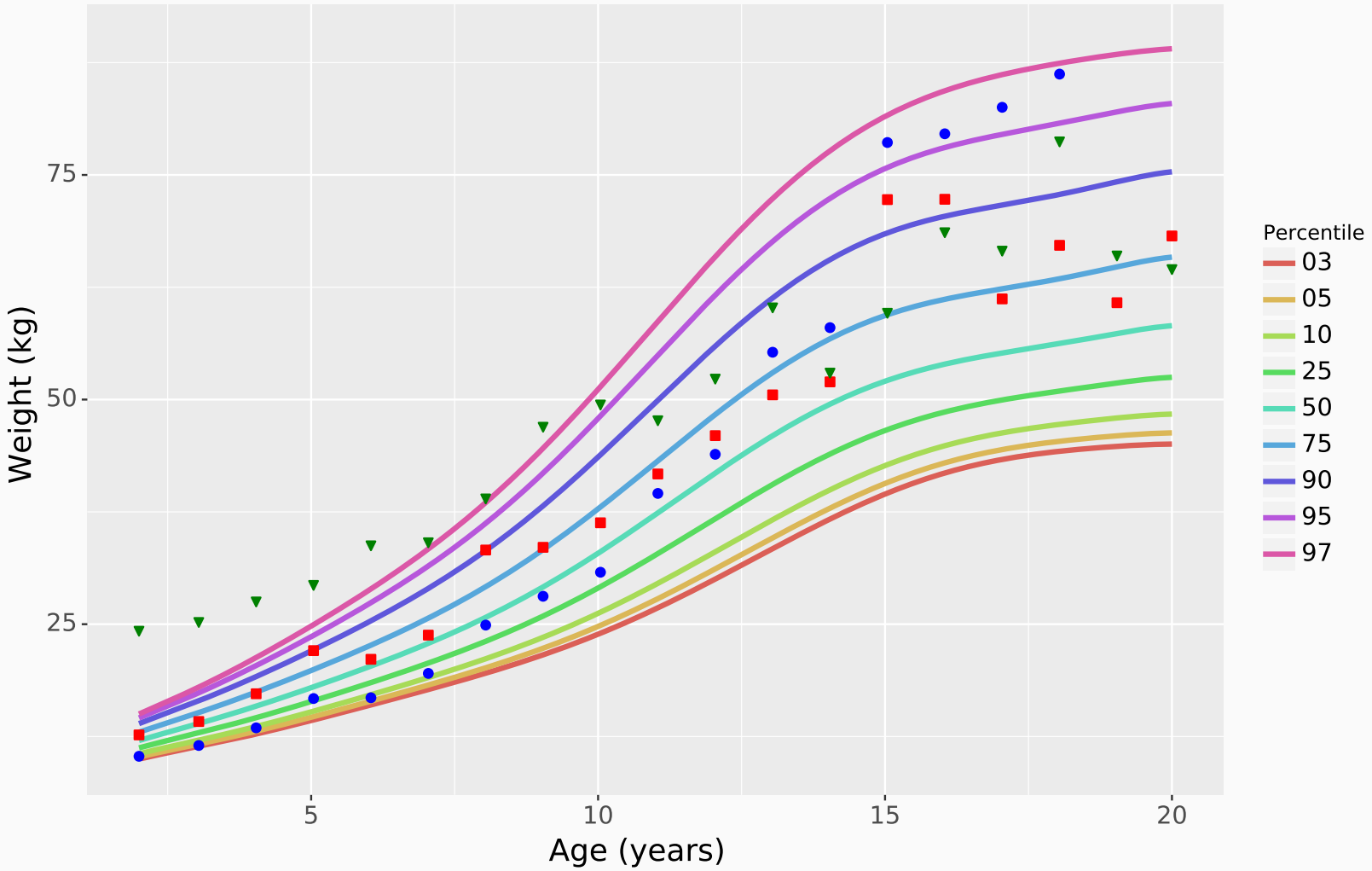
Metrics for Evaluation of Regression Models

Growth Curves



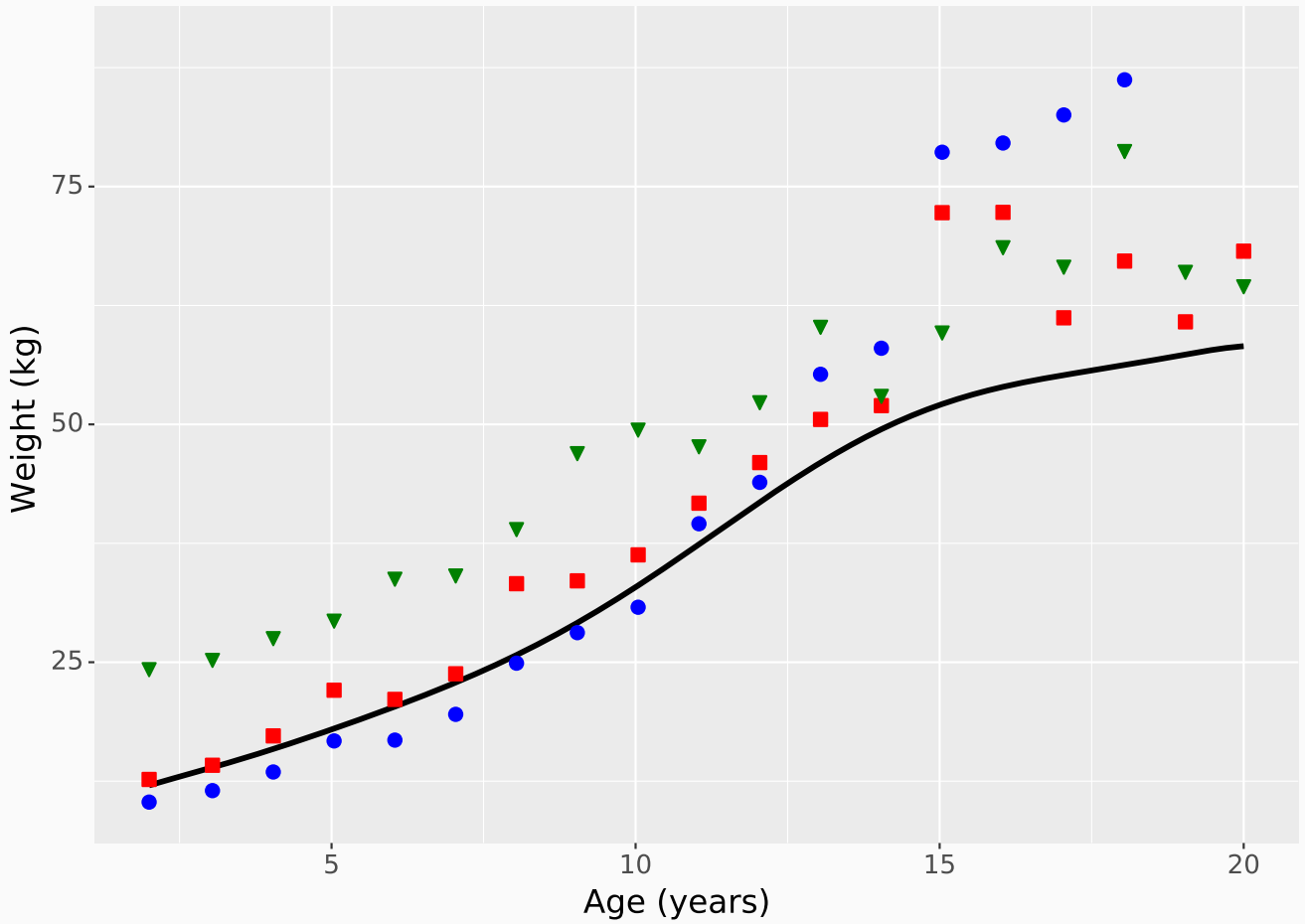
Centers for Disease Control and Prevention, National Center for Health Statistics. CDC growth charts: United States.

Growth Curves



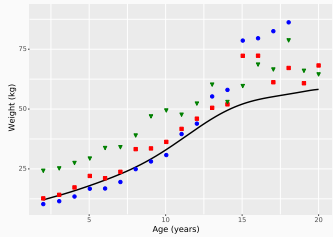
Centers for Disease Control and Prevention, National Center for Health Statistics. CDC growth charts: United States.

Regression Metrics



- What aspects of a model's predictions should I care about?
- What aspects of the model's predictions can I evaluate?

Regression Metrics

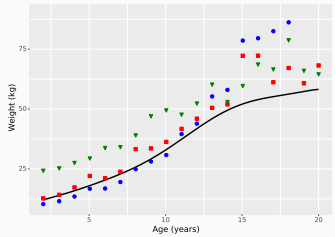


- What aspects of a model's predictions should I care about?
- What aspects of the model's predictions can I evaluate?

- Accuracy (bias), precision (variance)
- Average distance of errors
- Worst case error
- Do large errors matter more than small errors
- Maximal distance of errors
- Difference between my model and some standard model

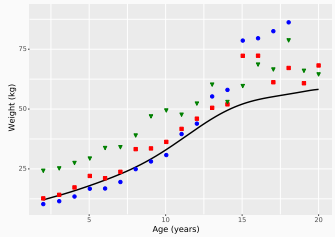
difference, squared difference, min/max, variance of predictions, relative difference (percentage error)

Regression Metrics



- What aspects of a model's predictions should I care about?
- What aspects of the model's predictions can I evaluate?

Equal weighting of errors	MAE	$\frac{1}{n}\sum_{i=0}^{n-1}\lvert y_i - \hat{y}_i \rvert$
	MAPE	$\frac{100}{n}\sum_{i=0}^{n-1}\frac{\lvert y_i - \hat{y}_i \rvert}{y_i}$
Unequal weighting of errors	MSE	$\frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \hat{y}_i\right)^2$
	RMSE	$\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \hat{y}_i\right)^2}$
	MSLE	$\frac{1}{n}\sum_{i=0}^{n-1}\left(\ln\left(1 = y_i\right) - \ln\left(1 + \hat{y}_i\right)\right)^2$
Data Variance	R^2	$1 - \frac{\sum_{i=0}^{n-1}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=0}^{n-1}\left(y_i - \bar{y}_i\right)^2}$
	Explained Var	$1 - \frac{Var\,(y - \hat{y})}{Var\,(y)}$



- What aspects of a model's predictions should I care about?
- What aspects of the model's predictions can I evaluate?

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Age	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
True	12	13	15	18	20	22	25	29	33	37	41	45	49	52	53	55	56	57	58
blue	10	11	13	16	16	19	24	28	30	39	43	55	58	78	79	82	86	90	94
red	12	14	17	22	21	23	33	33	36	41	45	50	51	72	72	61	67	60	68
green	24	25	27	29	33	34	38	46	49	47	52	60	52	59	68	66	78	65	64

	MAE	MAPE	MSE	RMSE	MSLE	R2	EV
blue	11.528	24.643	293.544	17.133	0.067	-0.12	0.227
red	5.653	13.793	61.967	7.872	0.022	0.764	0.886
green	11.962	42.712	160.046	12.651	0.136	0.39	0.935