Digital Transformation of Healthcare

Evaluating Predictions

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

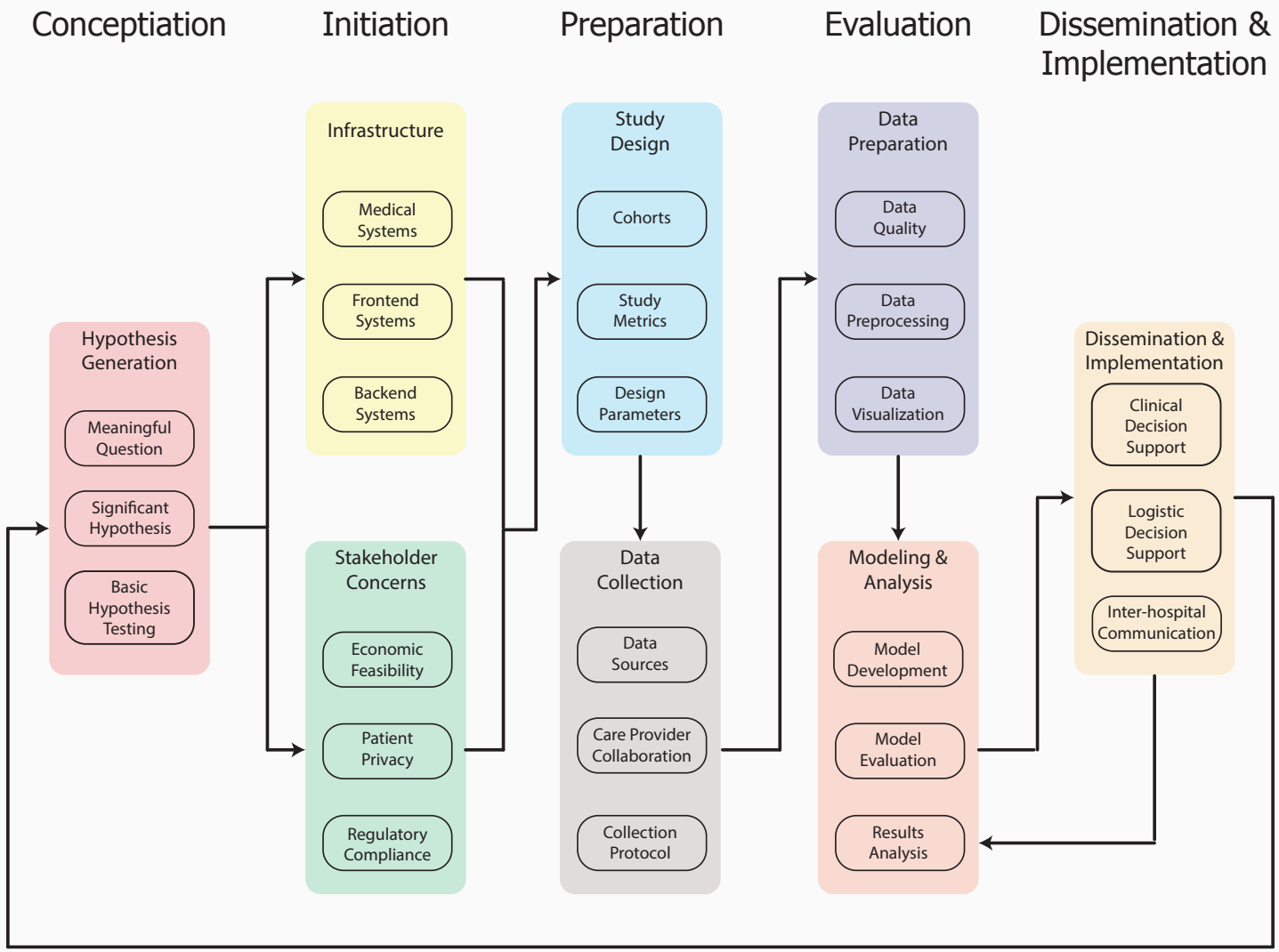
Center for Health Data Innovations

Evaluating Predictions

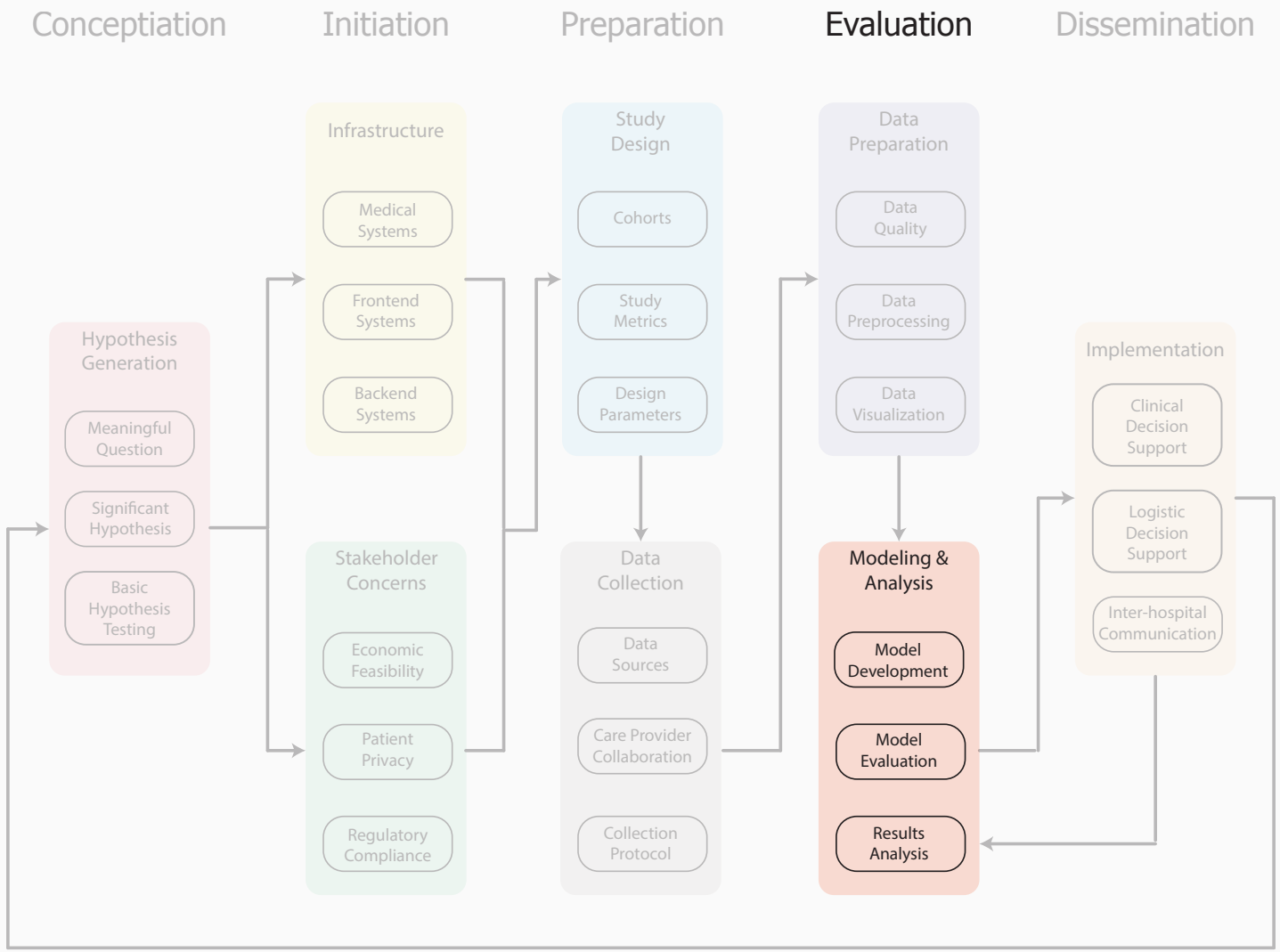
After this lecture students will be able to

- 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

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

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?

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?

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?

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?

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?

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

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

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

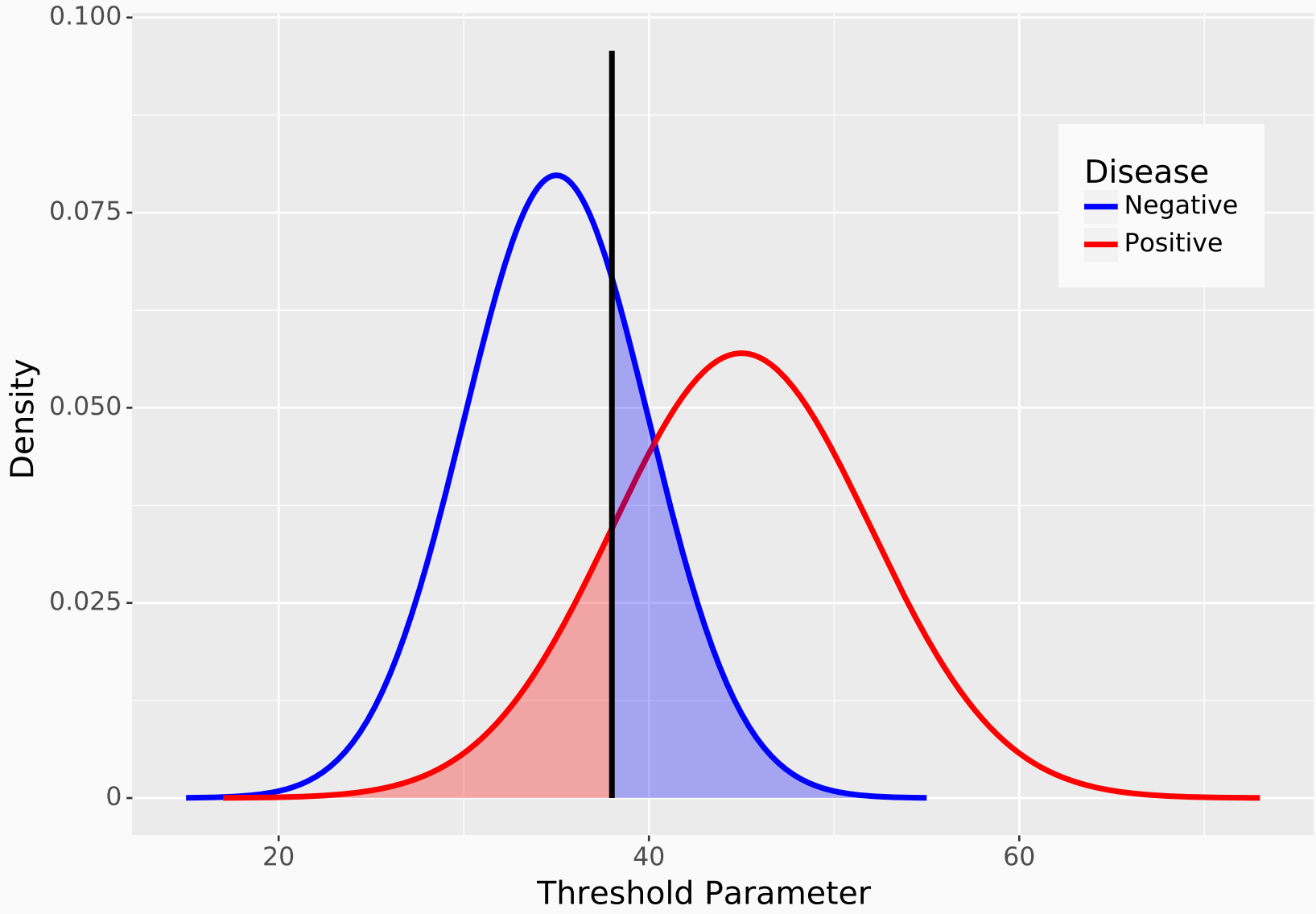
Home Machine	p	n
p′	17	6
n′	14	65

K-SADS-PL

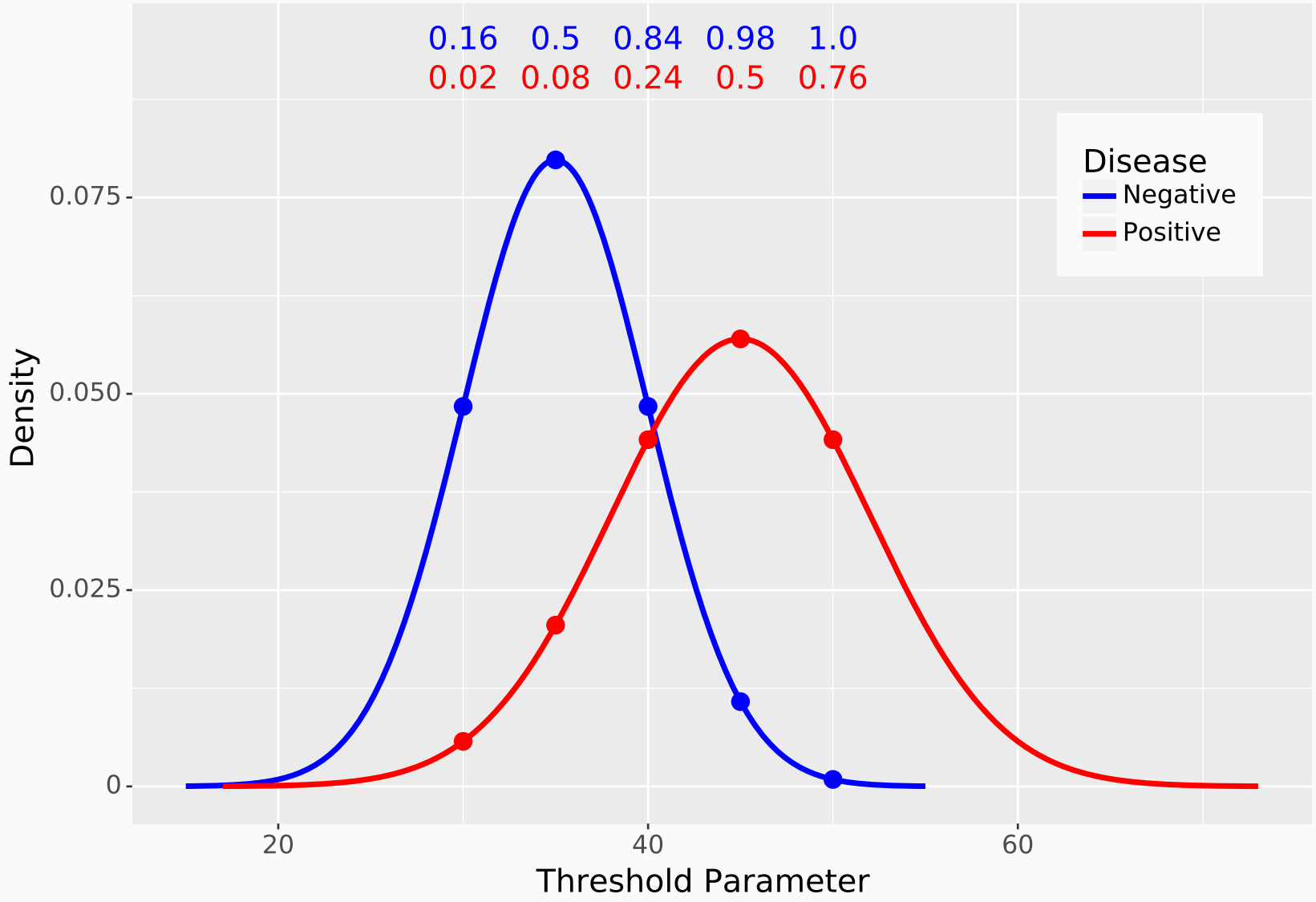
Trained Clinician	p	n
p′	32	23
n′	30	133

- 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?

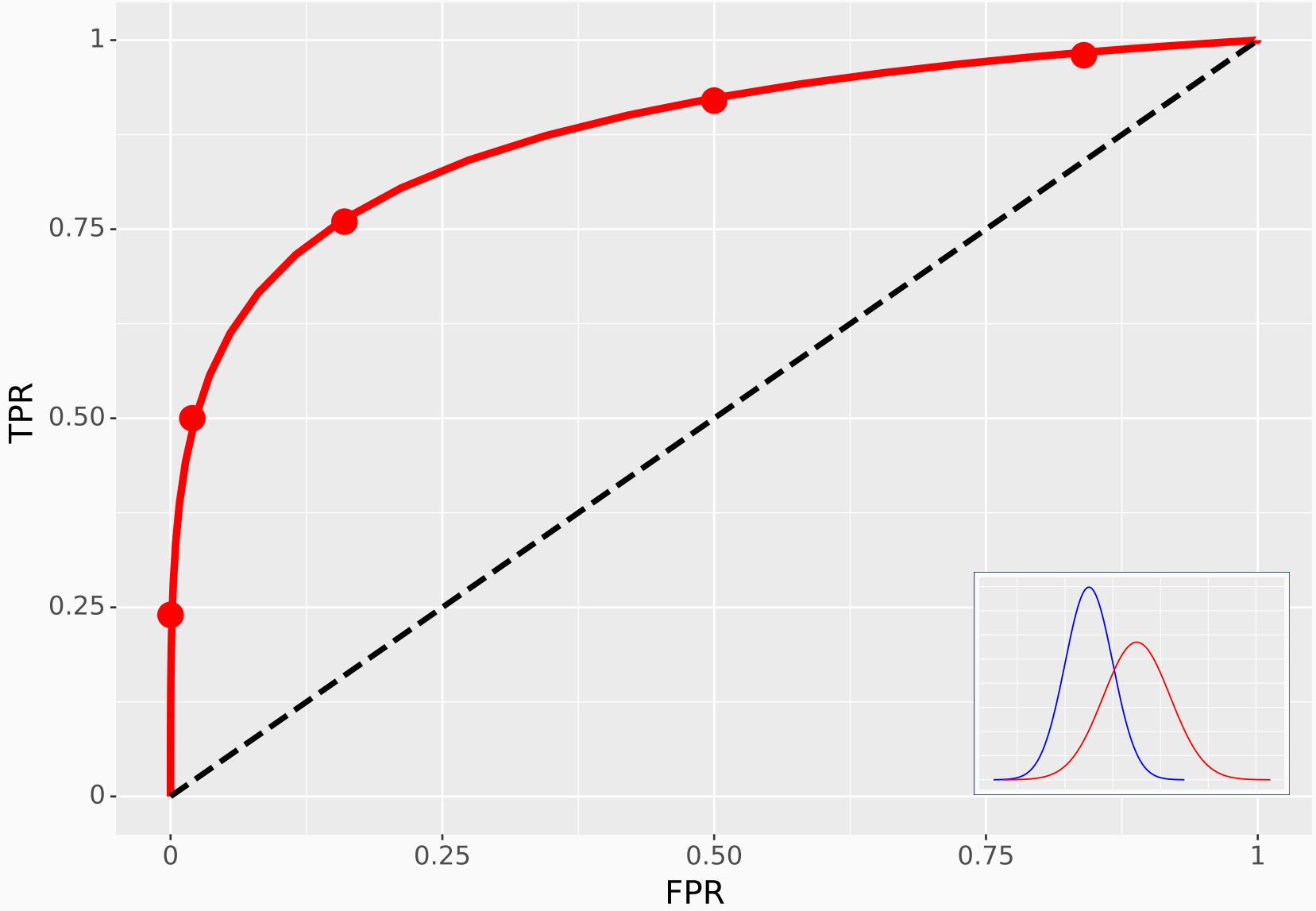
Hypothesis Testing



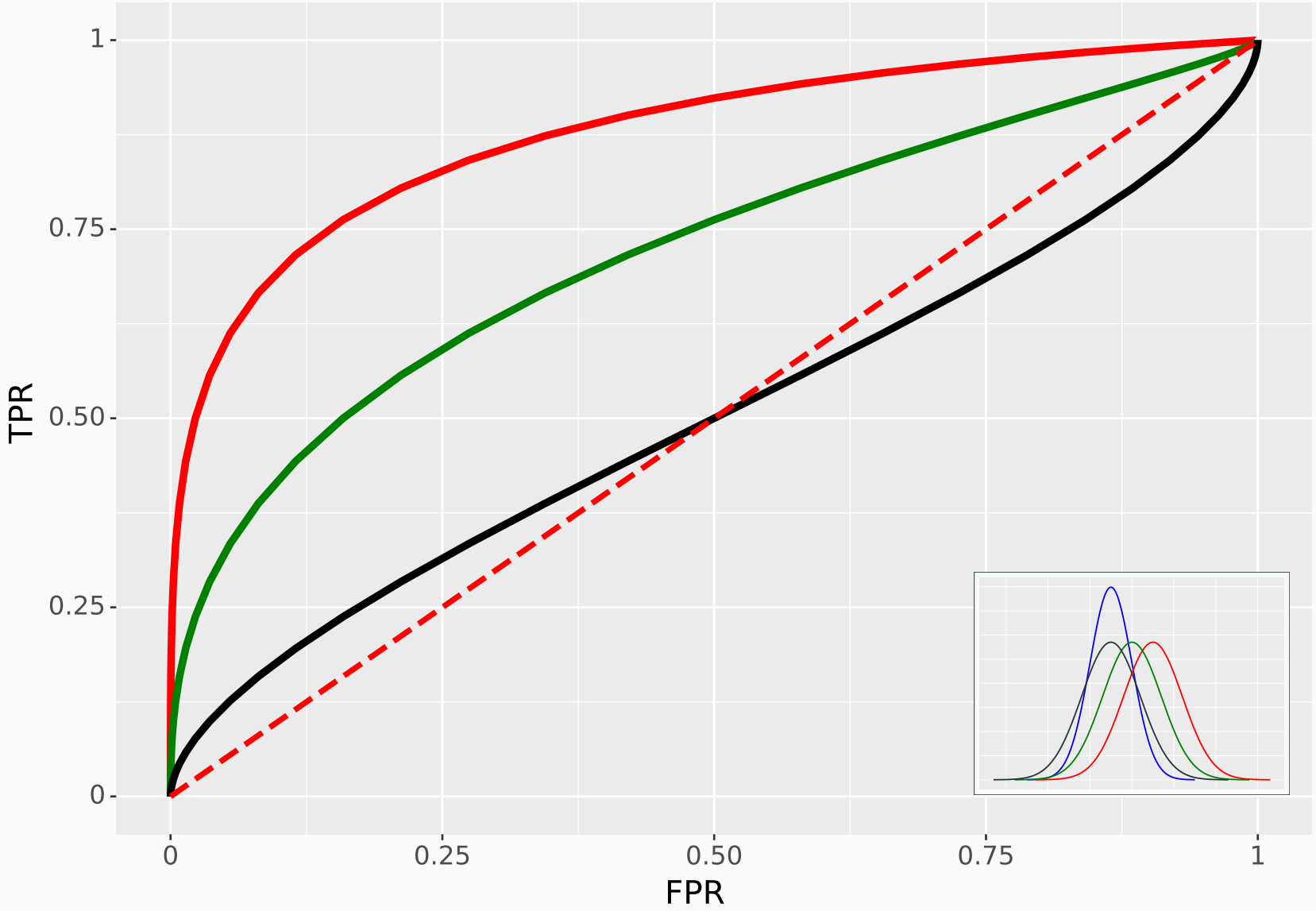
Hypothesis Testing



Receiver Operating Characteristic

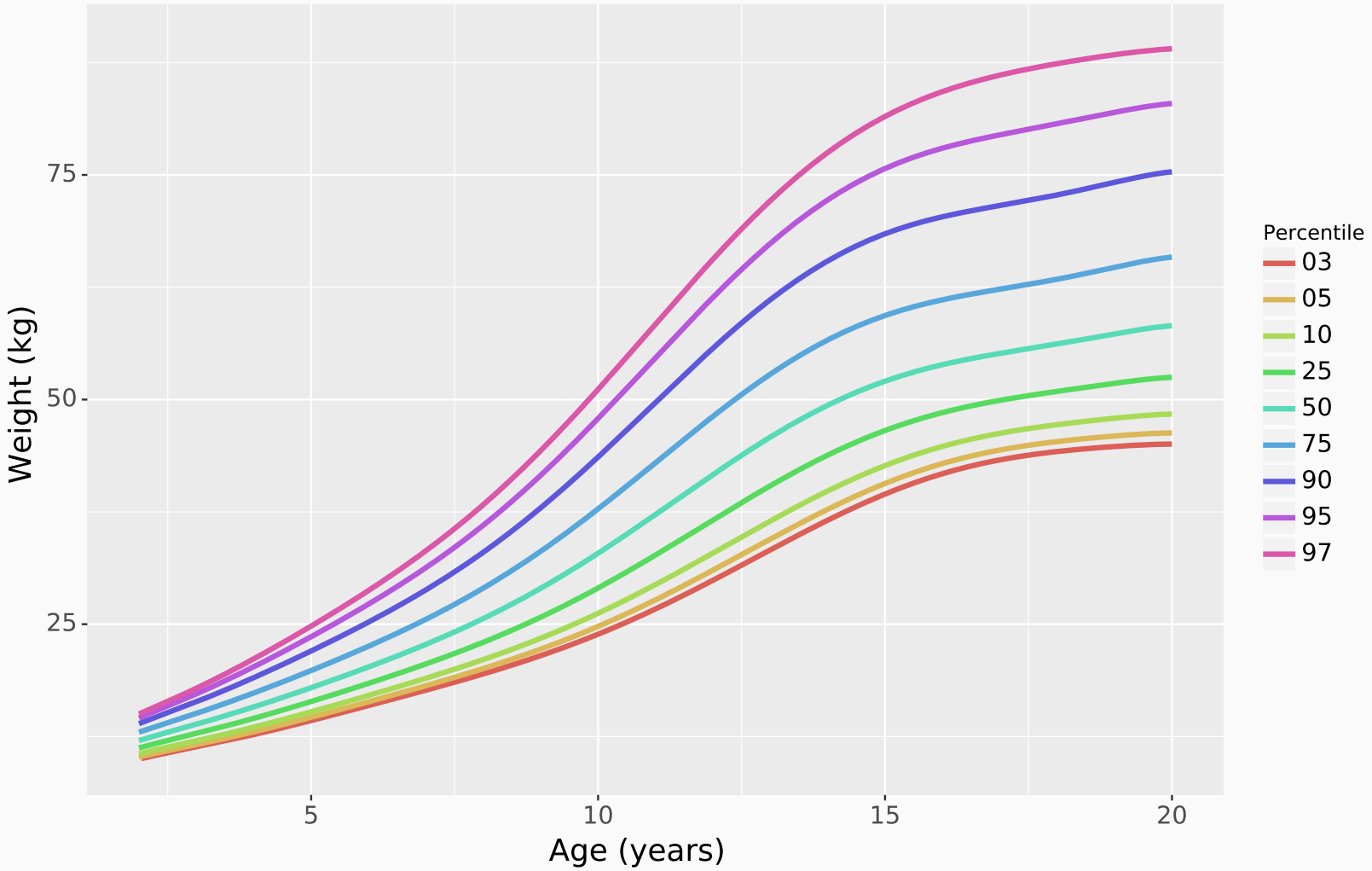


Receiver Operating Characteristic



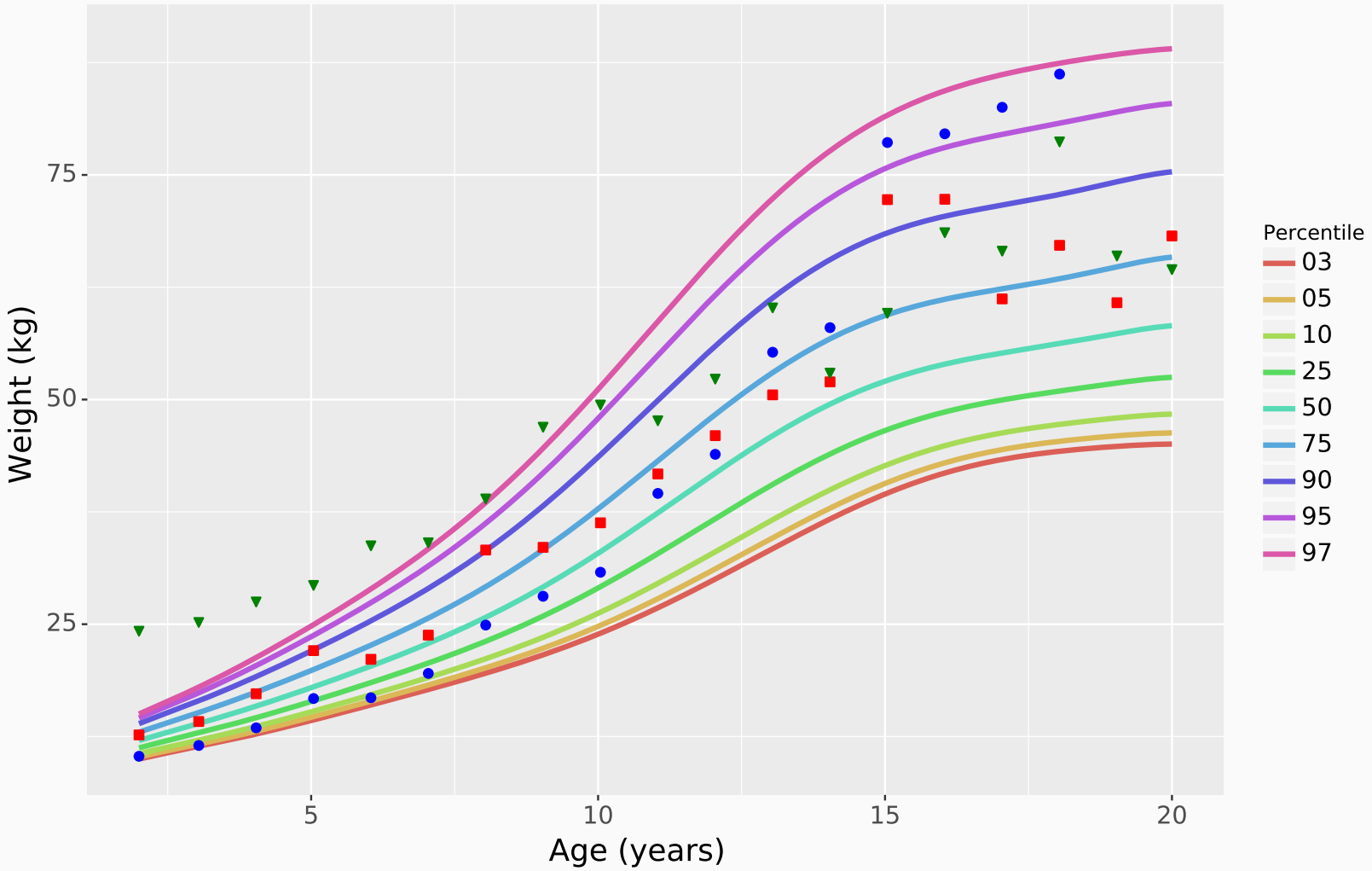
Metrics for Evaluation of Regression Models

Growth Curves



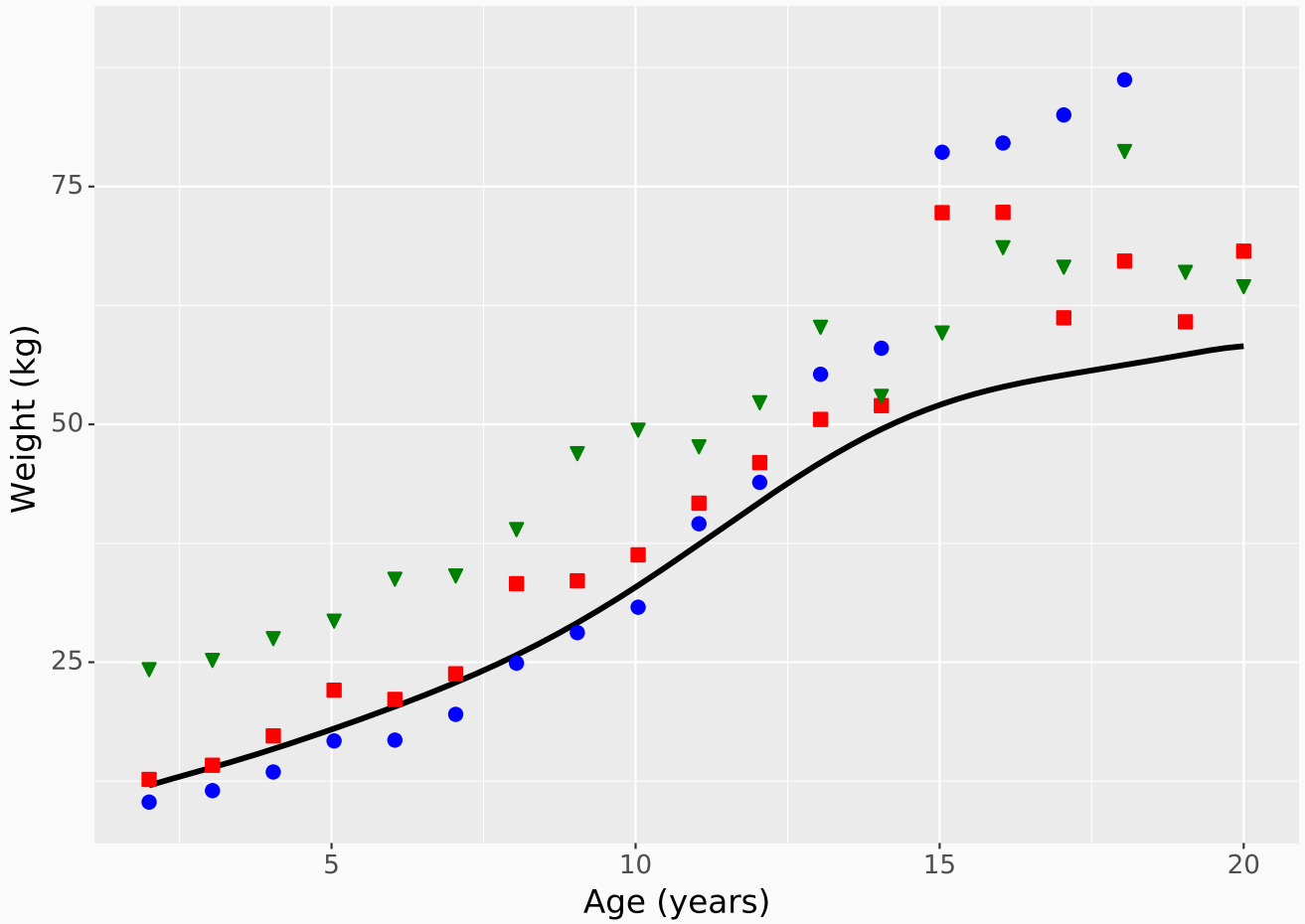
Centers for Disease Control and Prevention, National Center for Health Statistics. CDC growth charts: United States.

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?