Performance Assessment for Radiologists Interpreting Screening Mammography

Dawn Woodard
School of Operations Research and Information Engineering
Cornell University

Joint work with:
Alan Gelfand
Department of Statistical Science, Duke University

William Barlow
Cancer Research and Biostatistics, Seattle, WA

Joann Elmore
University of Washington School of Medicine

Outline

Outline

Physician Performance Assessment

- When a mammogram is performed, a radiologist looks at it and decides whether to recall the patient for further testing
- There is concern about large differences in the accuracy in this recall decision between radiologists
- Database of 500,000+ mammograms performed in the U.S. from 1996-2001
  - Demographic characteristics of the patient
  - Outcome of the mammogram (false +, false -, true +, or true -)
- Radiologist surveys
  - Practice characteristics
  - Demographic characteristics
  - Level of concern about malpractice

Physician Performance Assessment

- When a mammogram is performed, a radiologist looks at it and decides whether to recall the patient for further testing
- There is concern about large differences in the accuracy in this recall decision between radiologists
- Database of 500,000+ mammograms performed in the U.S. from 1996-2001
  - Demographic characteristics of the patient
  - Outcome of the mammogram (false +, false -, true +, or true -)
- Radiologist surveys
  - Practice characteristics
  - Demographic characteristics
  - Level of concern about malpractice

Physician Performance Assessment

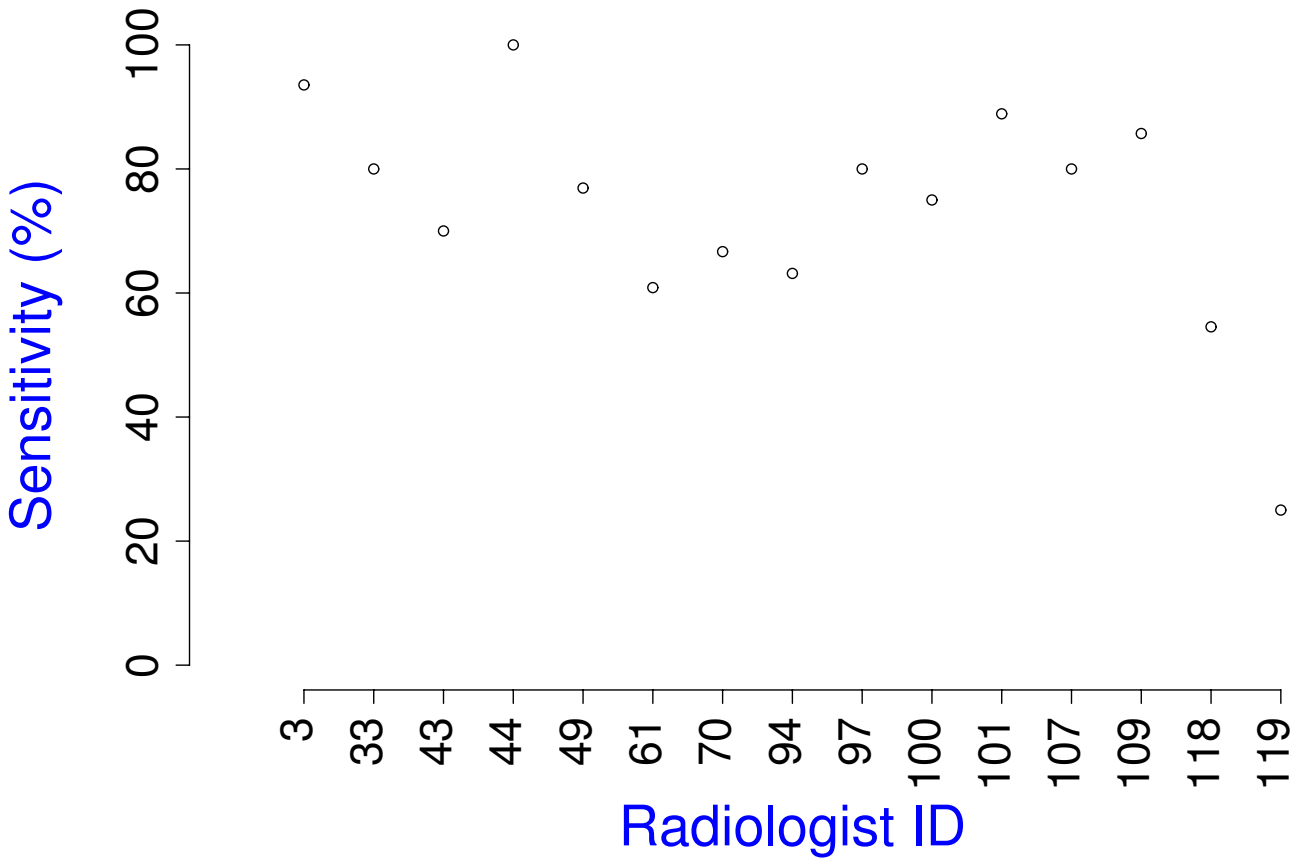
- When a mammogram is performed, a radiologist looks at it and decides whether to recall the patient for further testing
- There is concern about large differences in the accuracy in this recall decision between radiologists
- Database of 500,000+ mammograms performed in the U.S. from 1996-2001
  - Demographic characteristics of the patient
  - Outcome of the mammogram (false +, false -, true +, or true -)
- Radiologist surveys
  - Practice characteristics
  - Demographic characteristics
  - Level of concern about malpractice

Physician Performance Assessment

- When a mammogram is performed, a radiologist looks at it and decides whether to recall the patient for further testing
- There is concern about large differences in the accuracy in this recall decision between radiologists
- Database of 500,000+ mammograms performed in the U.S. from 1996-2001
  - Demographic characteristics of the patient
  - Outcome of the mammogram (false +, false -, true +, or true -)
- Radiologist surveys
  - Practice characteristics
  - Demographic characteristics
  - Level of concern about malpractice

Physician Performance Assessment

Empirical sensitivity, by radiologist:



Physician Performance Assessment

- Goal: assess physician performance while accounting for:
  - Differences in patients (case mix)
  - Differences in sample size (e.g. few cancer cases for some radiologists)

Physician Performance Assessment

- Can adjust for case mix (e.g. Salem-Schatz et al. 1994)
  - Can test whether a physician is significantly above or below average
  - Tests invalid for small sample sizes
  - Not clear how to compare one physician to another
- We build on Normand, Glickman, Gatsonis (1997): performance metrics for hospitals based on patient survival rate
  - We extend to metrics for sens. & spec. of physicians
- We use a Bayesian hierarchical modeling approach to estimate and explain accuracy differences among radiologists

Physician Performance Assessment

- Can adjust for case mix (e.g. Salem-Schatz et al. 1994)
  - Can test whether a physician is significantly above or below average
  - Tests invalid for small sample sizes
  - Not clear how to compare one physician to another
- We build on Normand, Glickman, Gatsonis (1997): performance metrics for hospitals based on patient survival rate
  - We extend to metrics for sens. & spec. of physicians
- We use a Bayesian hierarchical modeling approach to estimate and explain accuracy differences among radiologists

Physician Performance Assessment

- Can adjust for case mix (e.g. Salem-Schatz et al. 1994)
  - Can test whether a physician is significantly above or below average
  - Tests invalid for small sample sizes
  - Not clear how to compare one physician to another
- We build on Normand, Glickman, Gatsonis (1997): performance metrics for hospitals based on patient survival rate
  - We extend to metrics for sens. & spec. of physicians
- We use a Bayesian hierarchical modeling approach to estimate and explain accuracy differences among radiologists

Outline

Modeling Accuracy

Logistic regression:

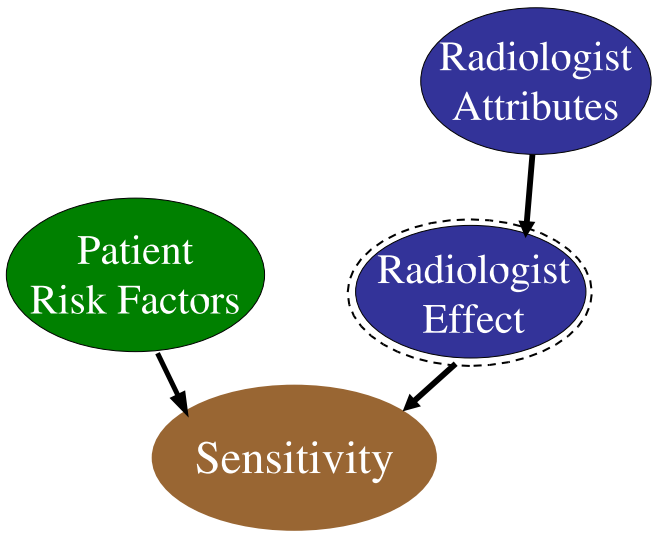
$$\mathbf{logit}(\mathbf{S_{ij}}) = \mathbf{X_{ij}}\beta + \tau_{\mathbf{i}}$$

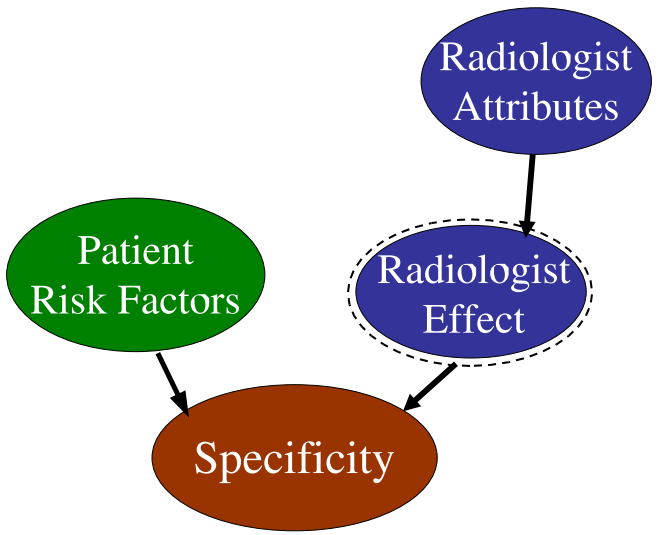
$$\tau_{\mathbf{i}} = \mathbf{W_i}\gamma + \phi_{\mathbf{i}}$$

$$\phi_{\mathbf{i}} \sim \mathbf{N}(\mathbf{0}, \psi)$$

S_{ij} = sensitivity on mammogram i,j
X_{ij} = risk factors of patient i,j
W_i = attributes of radiologist i
i = 1,...,I radiologists
j = 1,...,n_i mammograms of radiologist i with cancer present

Modeling Accuracy





Results

- Risk factors strongly related to accuracy:
  - breast density
  - mammographic history
  - menopausal status
  - history of biopsy or surgery
- Radiologist attributes strongly related to accuracy:
  - # yrs. interpreting mammograms (longer $\Rightarrow$ lower sensitivity, higher specificity)
  - academic affiliation
  - mammographic case volume

Results

- Risk factors strongly related to accuracy:
  - breast density
  - mammographic history
  - menopausal status
  - history of biopsy or surgery
- Radiologist attributes strongly related to accuracy:
  - # yrs. interpreting mammograms (longer $\Rightarrow$ lower sensitivity, higher specificity)
  - academic affiliation
  - mammographic case volume

Outline

Performance on a Hypothetical Patient

- Predict the sensitivity and specificity of each radiologist for a "typical" patient
- Or a "high-risk" or "low-risk" patient
- For a hypothetical patient with attributes X_0, the measure is

$$S(X_0, \beta, \tau_i) = \text{logit}^{-1}(X_0\beta + \tau_i)$$

Performance on a Hypothetical Patient

- Predict the sensitivity and specificity of each radiologist for a "typical" patient
- Or a "high-risk" or "low-risk" patient
- For a hypothetical patient with attributes X_0, the measure is

$$S(X_0, \beta, \tau_i) = \text{logit}^{-1}(X_0\beta + \tau_i)$$

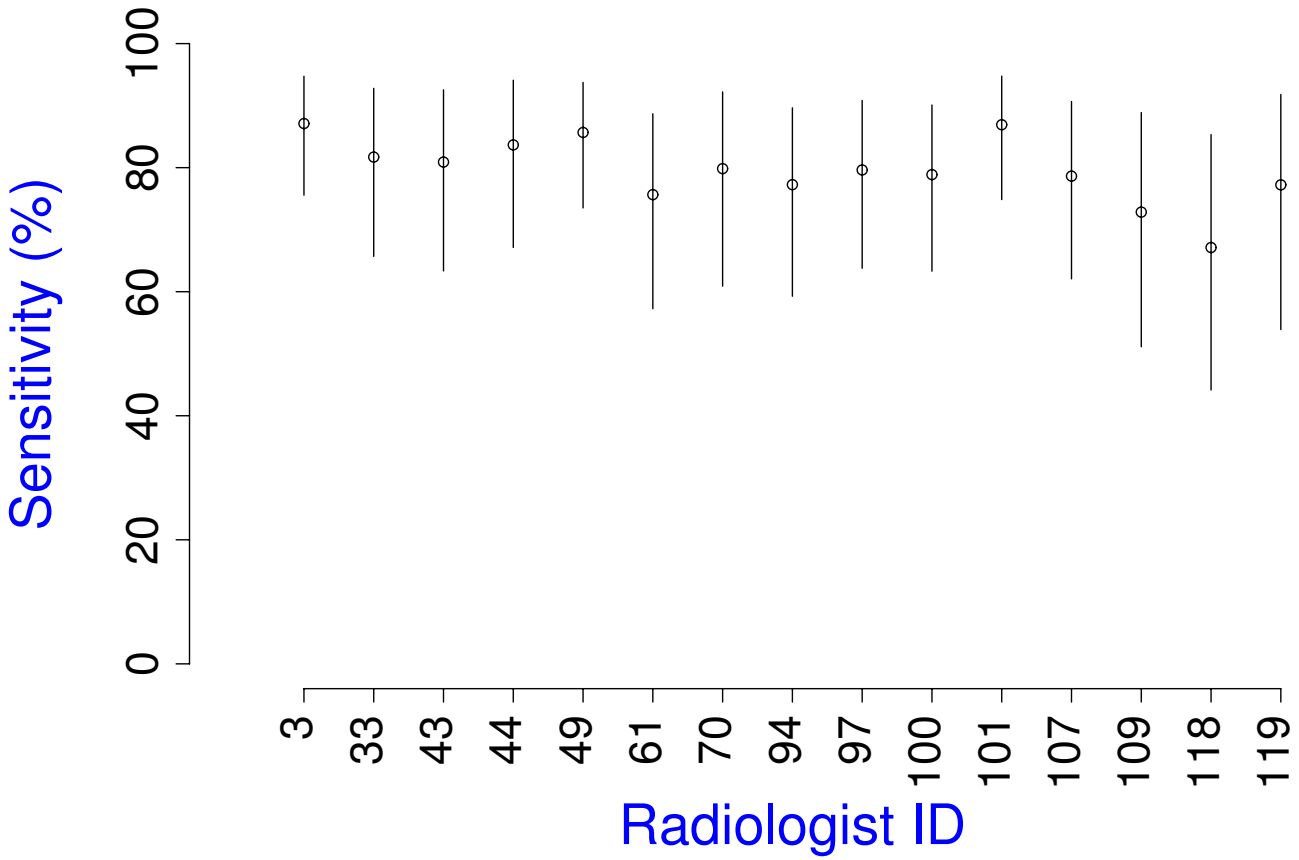
Performance on a Hypothetical Patient

- Predict the sensitivity and specificity of each radiologist for a "typical" patient
- Or a "high-risk" or "low-risk" patient
- For a hypothetical patient with attributes X_0, the measure is

$$S(X_0, \beta, \tau_i) = \text{logit}^{-1}(X_0\beta + \tau_i)$$

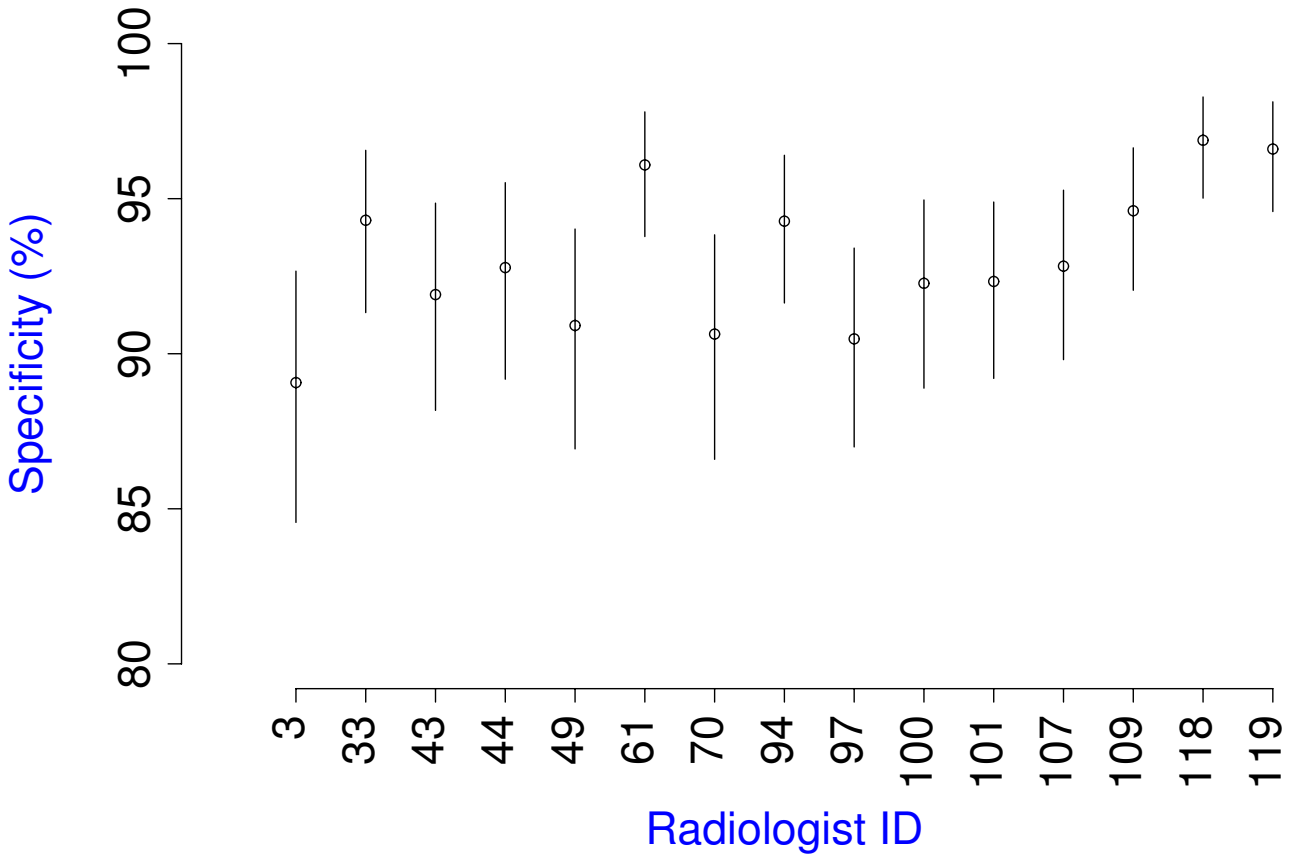
Performance on a Hypothetical Patient

Sensitivity on a “typical” patient:



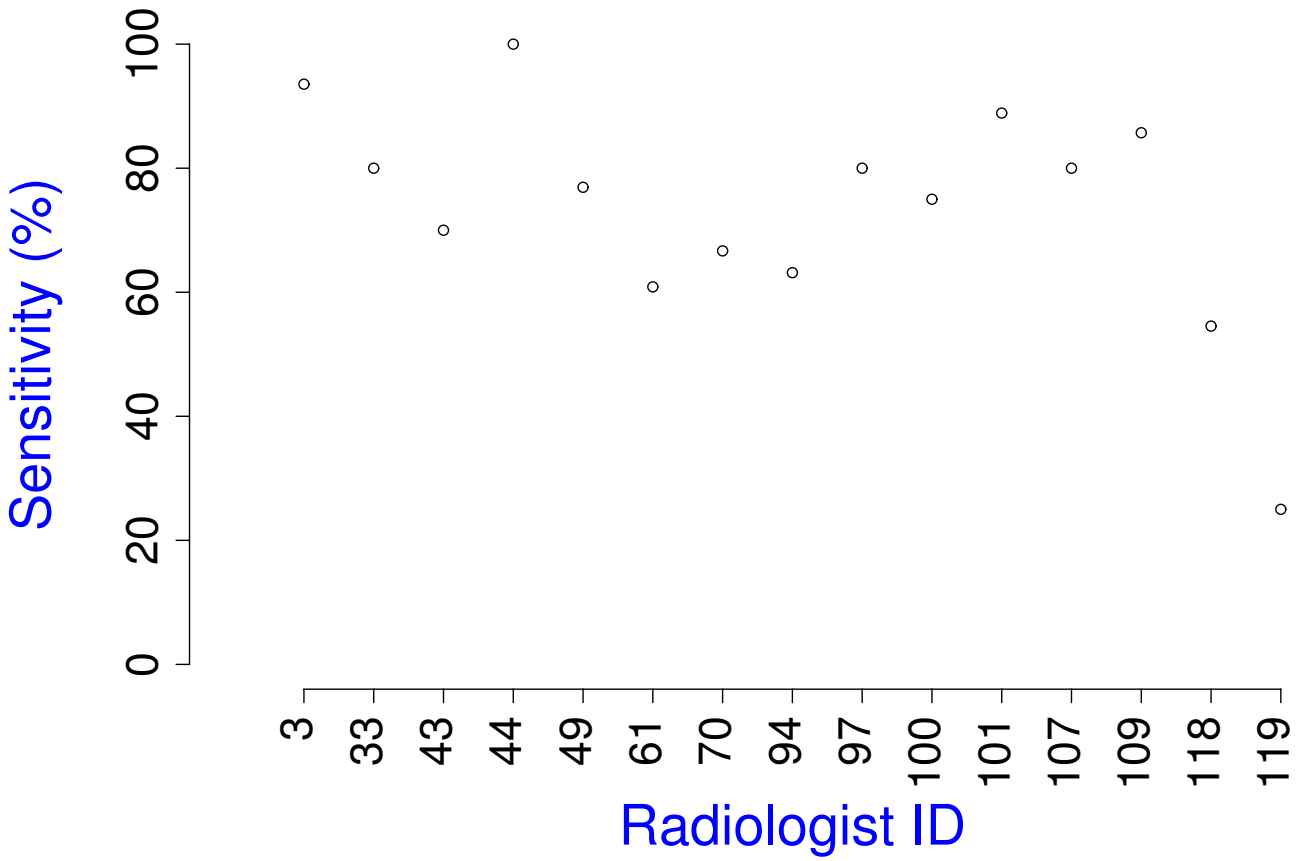
Performance on a Hypothetical Patient

Specificity:



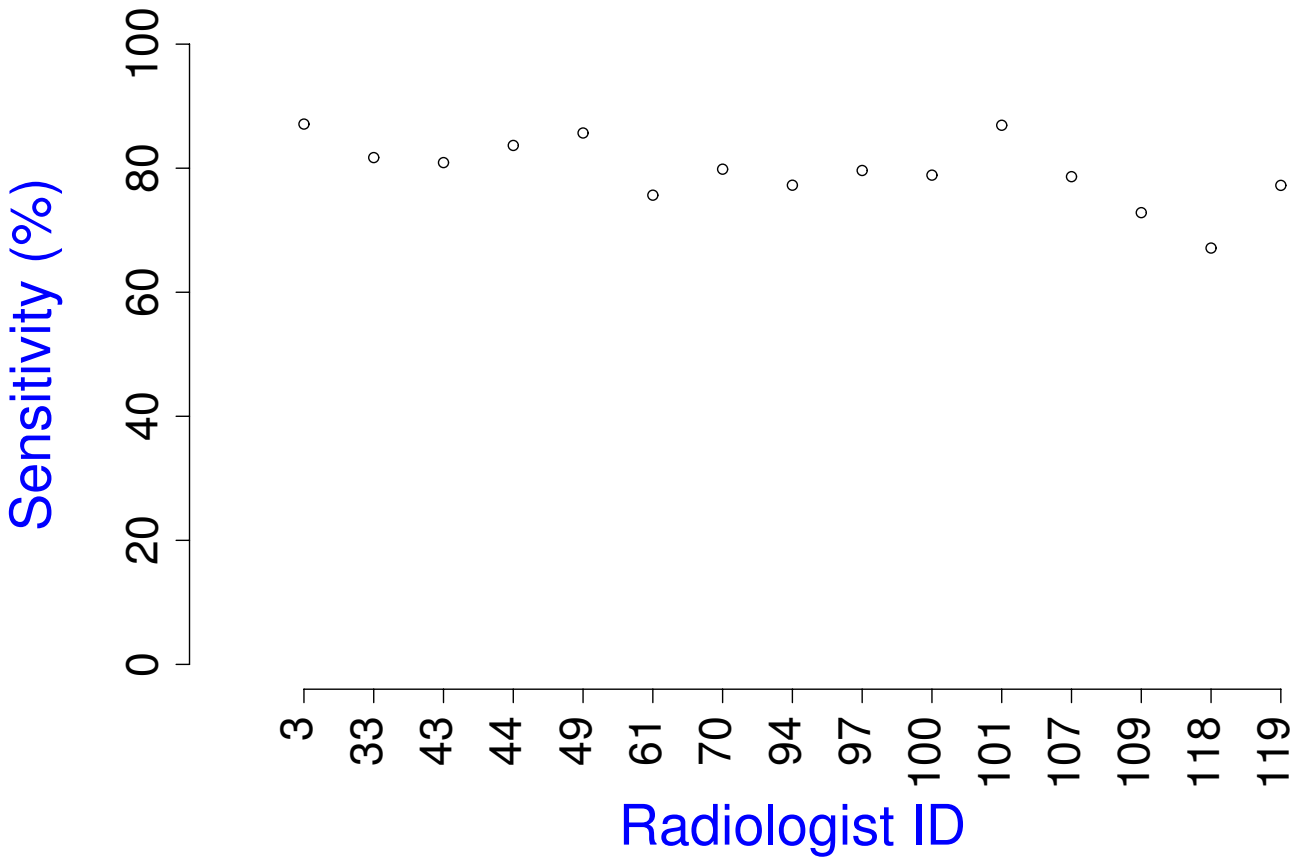
Performance on a Hypothetical Patient

Empirical sensitivity:



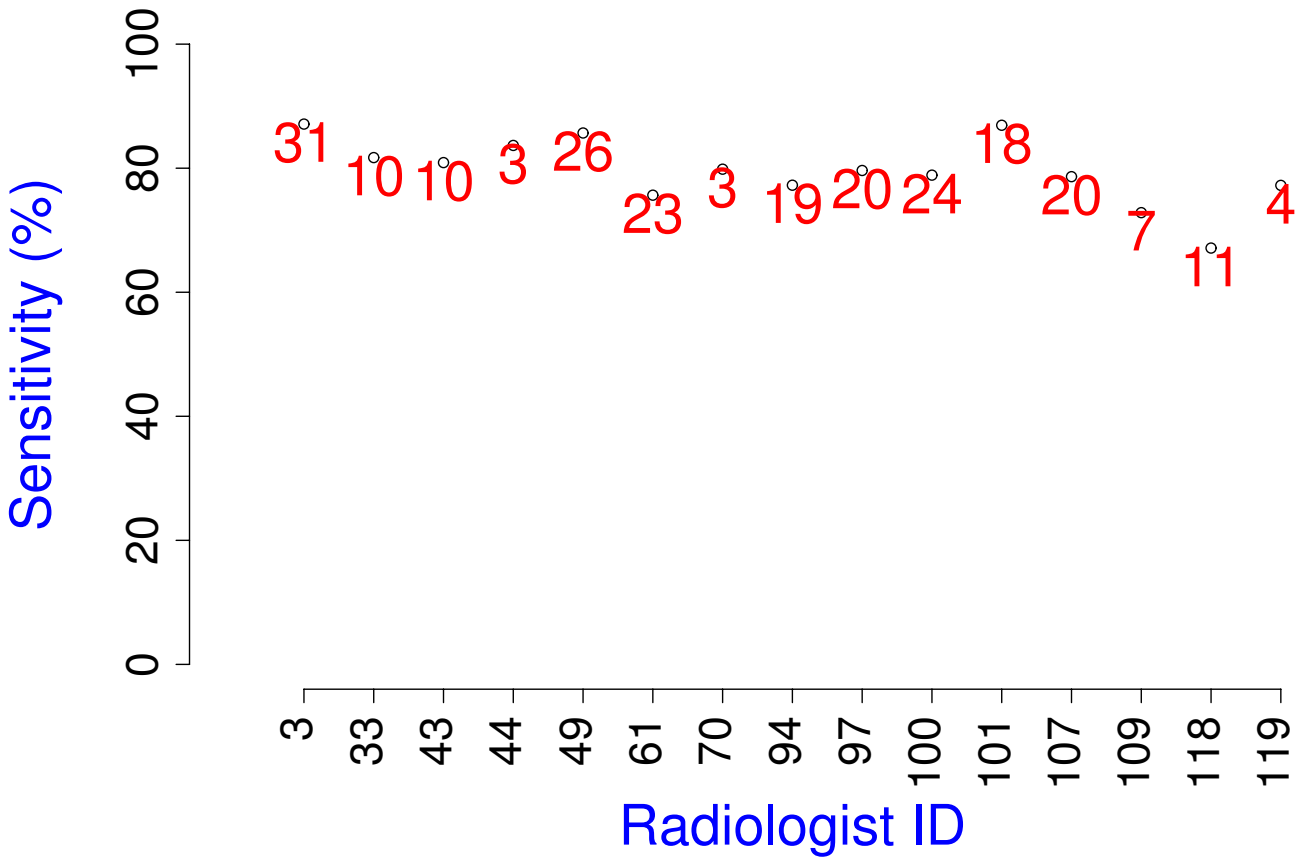
Performance on a Hypothetical Patient

Performance on a hypothetical patient:



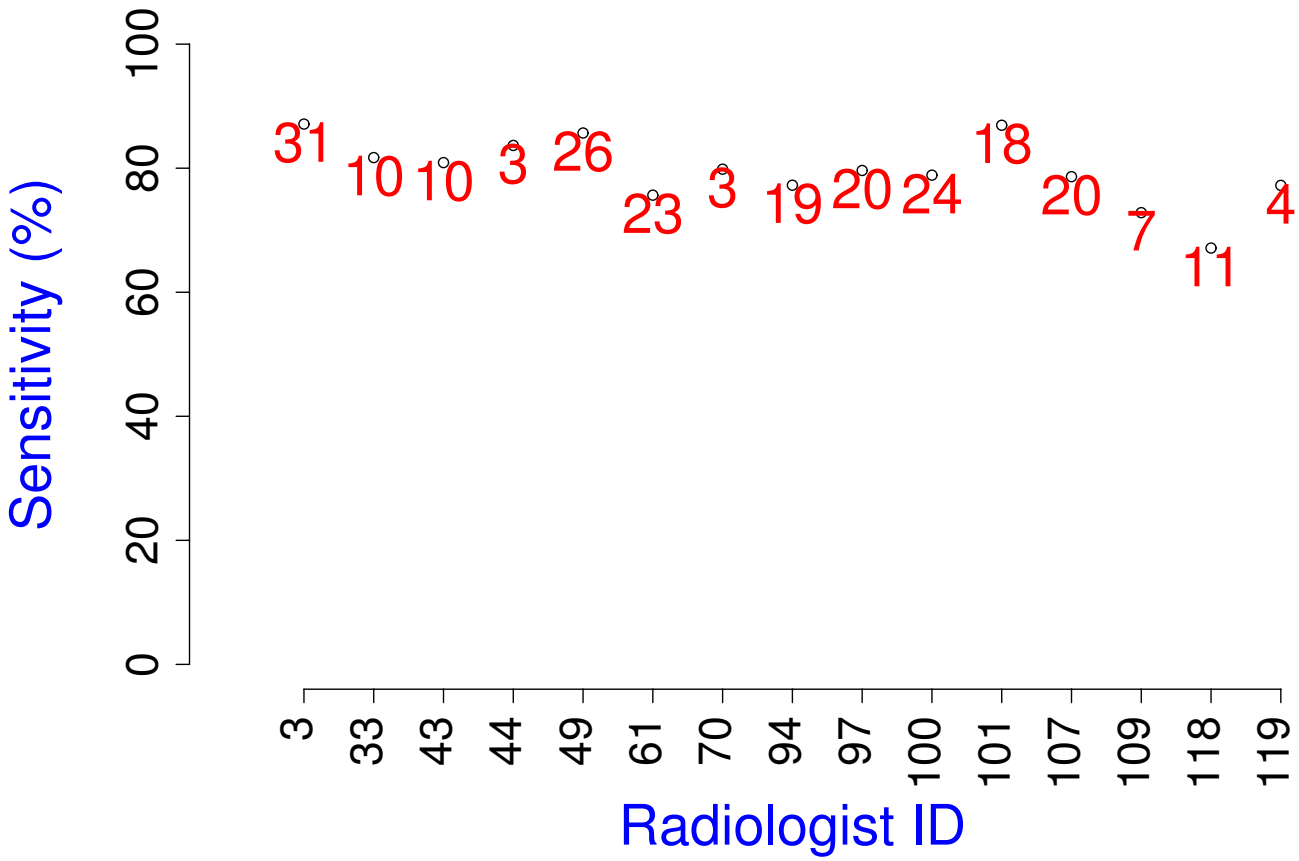
Performance on a Hypothetical Patient

Performance on a hypothetical patient:



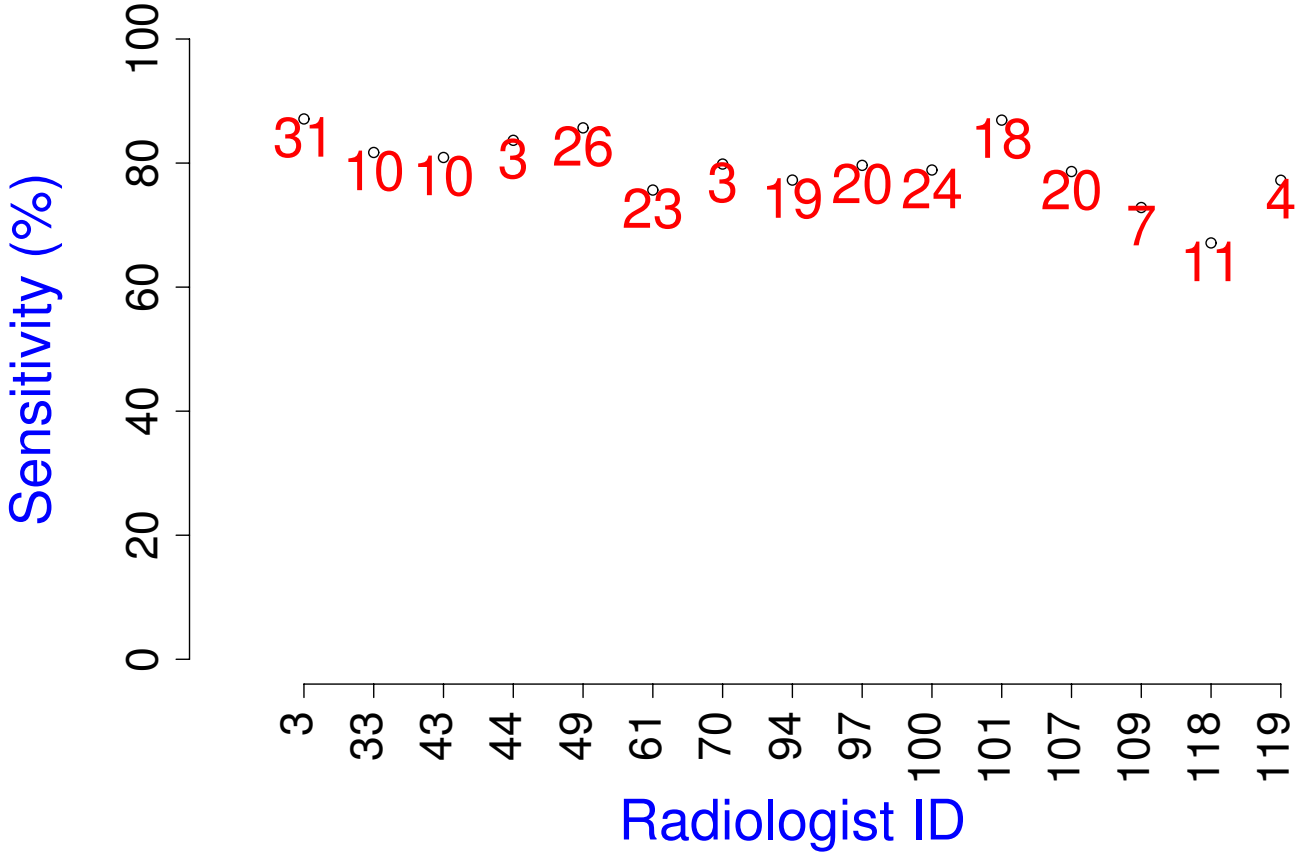
Performance on a Hypothetical Patient

We have “adjusted” for differences in patient mix...



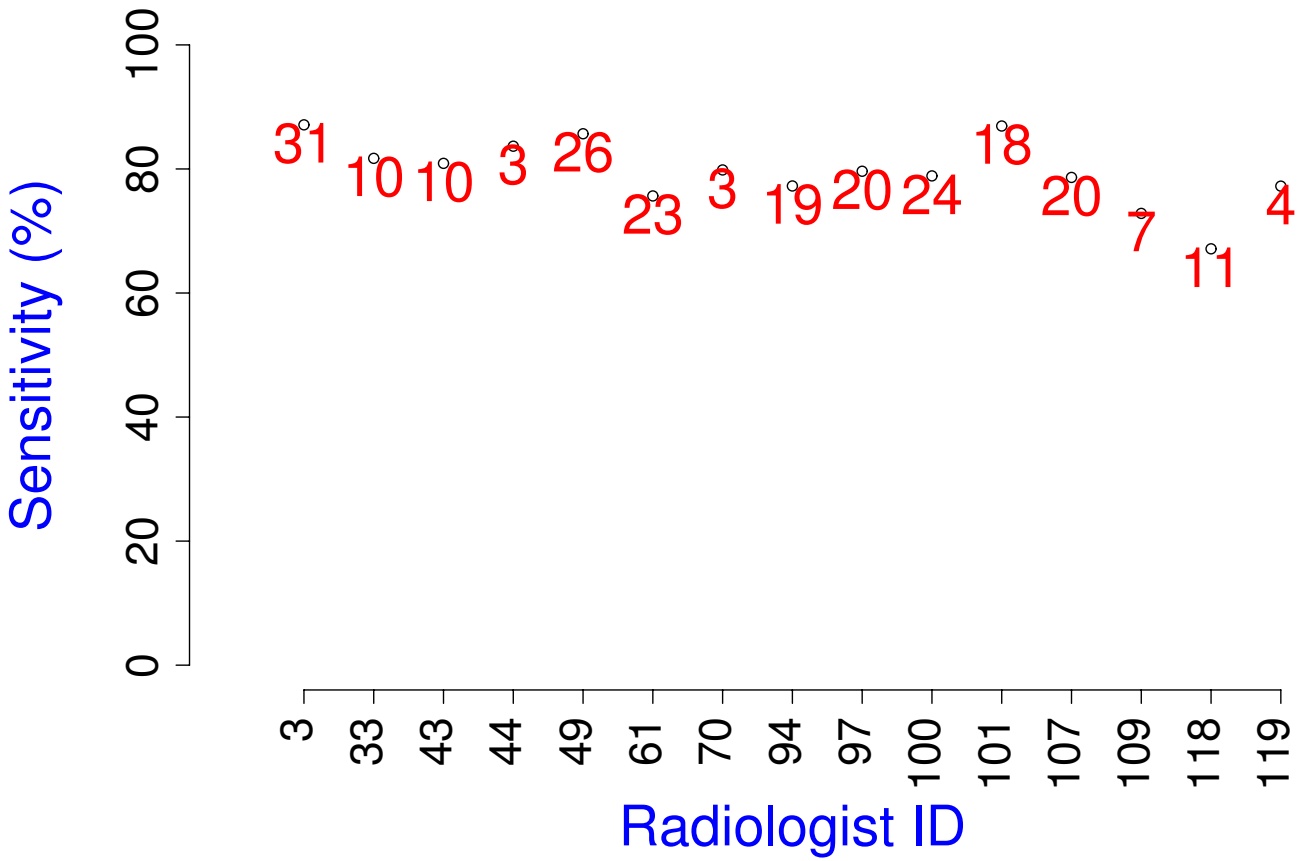
Performance on a Hypothetical Patient

...and shrunk the sensitivities towards the common mean...



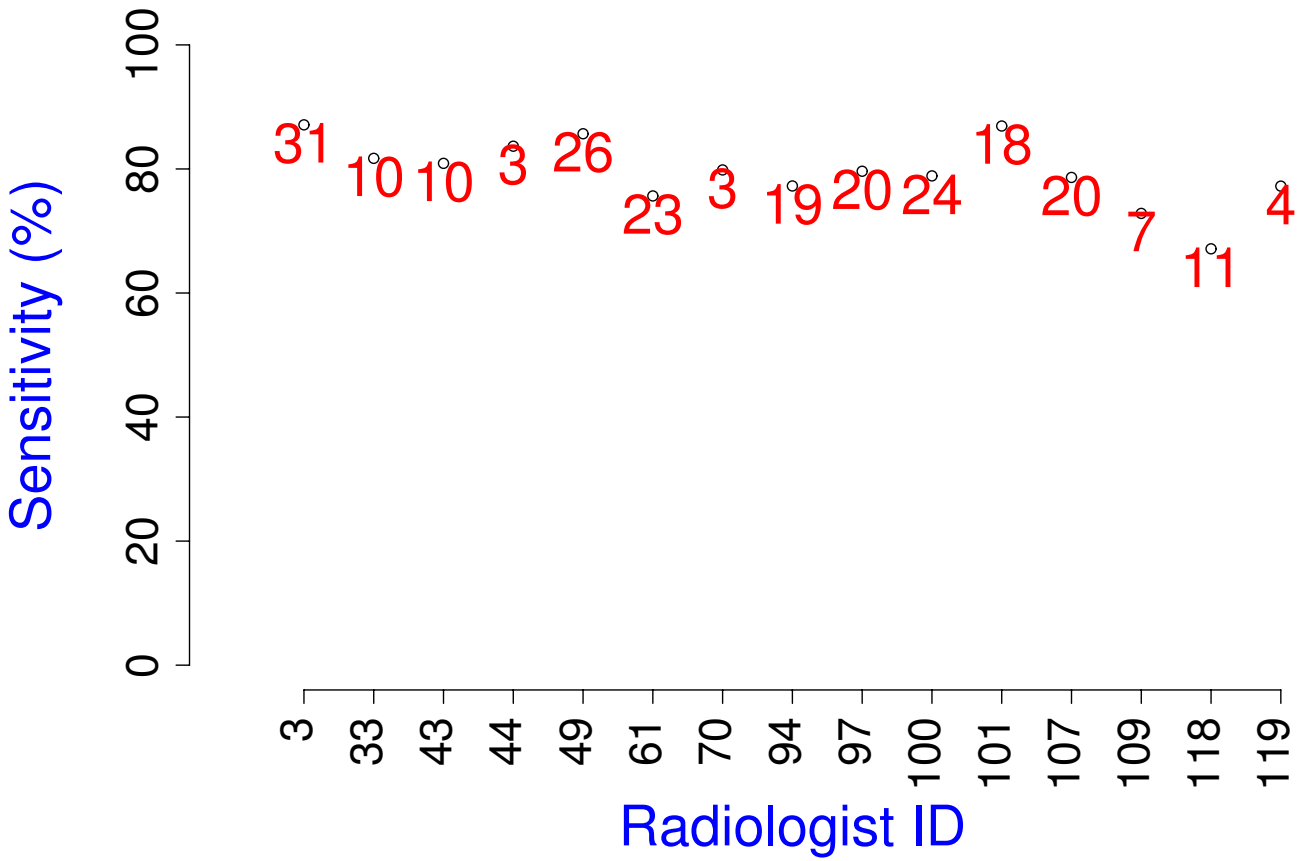
Performance on a Hypothetical Patient

...where the amount of shrinkage depends on the sample size.



Performance on a Hypothetical Patient

We have not adjusted for differences in radiologist attributes.



Performance Relative to a Standard

- Alternatively, take the predicted average accuracy (sensitivity or specificity) of a particular radiologist on her patients:

$$\mu_i = \frac{1}{n_i} \Sigma_{j=1}^{n_i} S(X_{ij}, \beta, \tau_i)$$

- Compare to that expected for a radiologist with the same attributes and patient mix:

$$\tilde{\mu}_i = \frac{1}{n_i} \Sigma_{j=1}^{n_i} S(X_{ij}, \beta, W_i)$$

$$S(X_{ij}, \beta, W_i) = E_{\tau|W_i}\{S(X_{ij}, \beta, \tau)\}$$

- Take $\mu_i - \tilde{\mu}_i$
- Performance is evaluated while adjusting for radiologist attributes

Performance Relative to a Standard

- Alternatively, take the predicted average accuracy (sensitivity or specificity) of a particular radiologist on her patients:

$$\mu_i = \frac{1}{n_i} \Sigma_{j=1}^{n_i} S(X_{ij}, \beta, \tau_i)$$

- Compare to that expected for a radiologist with the same attributes and patient mix:

$$\tilde{\mu}_i = \frac{1}{n_i} \Sigma_{j=1}^{n_i} S(X_{ij}, \beta, W_i)$$

$$S(X_{ij}, \beta, W_i) = E_{\tau|W_i}\{S(X_{ij}, \beta, \tau)\}$$

- Take $\mu_i - \tilde{\mu}_i$
- Performance is evaluated while adjusting for radiologist attributes

Performance Relative to a Standard

- Alternatively, take the predicted average accuracy (sensitivity or specificity) of a particular radiologist on her patients:

$$\mu_i = \frac{1}{n_i} \Sigma_{j=1}^{n_i} S(X_{ij}, \beta, \tau_i)$$

- Compare to that expected for a radiologist with the same attributes and patient mix:

$$\tilde{\mu}_i = \frac{1}{n_i} \Sigma_{j=1}^{n_i} S(X_{ij}, \beta, W_i)$$

$$S(X_{ij}, \beta, W_i) = E_{\tau|W_i}\{S(X_{ij}, \beta, \tau)\}$$

- Take $\mu_i - \tilde{\mu}_i$
- Performance is evaluated while adjusting for radiologist attributes

Performance Relative to a Standard

- Alternatively, take the predicted average accuracy (sensitivity or specificity) of a particular radiologist on her patients:

$$\mu_i = \frac{1}{n_i} \Sigma_{j=1}^{n_i} S(X_{ij}, \beta, \tau_i)$$

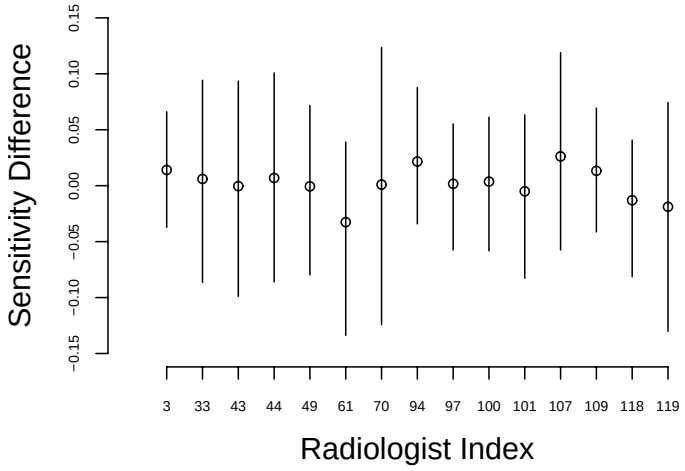
- Compare to that expected for a radiologist with the same attributes and patient mix:

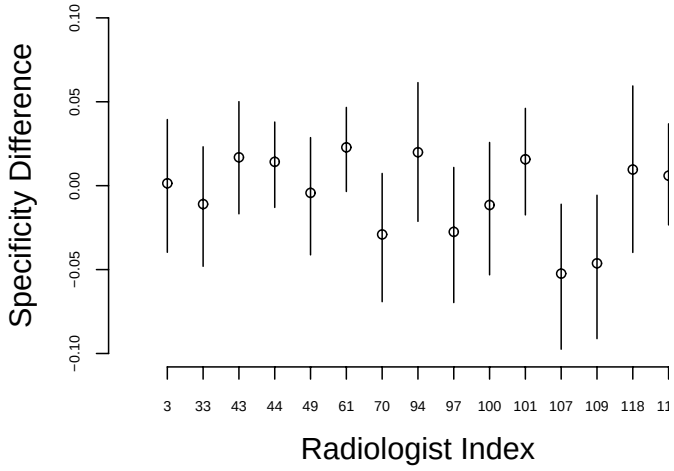
$$\tilde{\mu}_i = \frac{1}{n_i} \Sigma_{j=1}^{n_i} S(X_{ij}, \beta, W_i)$$

$$S(X_{ij}, \beta, W_i) = E_{\tau|W_i}\{S(X_{ij}, \beta, \tau)\}$$

- Take $\mu_i - \tilde{\mu}_i$
- Performance is evaluated while adjusting for radiologist attributes

Performance Relative to a Standard





Many radiologists had predicted specificity significantly above or below that expected; not so for sensitivity

Outline

Modeling Dependence of Sensitivity and Specificity

- Sensitivity and specificity are known to be dependent
- Paliwal et al. (2006) fit independent logistic regression models to the prob. of cancer given recall status, & prob. of recall
  - This induces dependence between sens. & spec.
  - Is their assumption reasonable?
- Puggioni, Gelfand, and Elmore (2008) fit independent logistic regression models to a different set of 3 conditional probabilities
  - Also induces dependence between sens. & spec.
  - Is their assumption reasonable?
- More desirable: a multinomial logit model for the 4 outcomes
  - Numerical difficulties for obtaining MLEs or Bayesian estimates (Puggioni et al., 2008)

Modeling Dependence of Sensitivity and Specificity

- Sensitivity and specificity are known to be dependent
- Paliwal et al. (2006) fit independent logistic regression models to the prob. of cancer given recall status, & prob. of recall
  - This induces dependence between sens. & spec.
  - Is their assumption reasonable?
- Puggioni, Gelfand, and Elmore (2008) fit independent logistic regression models to a different set of 3 conditional probabilities
  - Also induces dependence between sens. & spec.
  - Is their assumption reasonable?
- More desirable: a multinomial logit model for the 4 outcomes
  - Numerical difficulties for obtaining MLEs or Bayesian estimates (Puggioni et al., 2008)

Modeling Dependence of Sensitivity and Specificity

- Sensitivity and specificity are known to be dependent
- Paliwal et al. (2006) fit independent logistic regression models to the prob. of cancer given recall status, & prob. of recall
  - This induces dependence between sens. & spec.
  - Is their assumption reasonable?
- Puggioni, Gelfand, and Elmore (2008) fit independent logistic regression models to a different set of 3 conditional probabilities
  - Also induces dependence between sens. & spec.
  - Is their assumption reasonable?
- More desirable: a multinomial logit model for the 4 outcomes
  - Numerical difficulties for obtaining MLEs or Bayesian estimates (Puggioni et al., 2008)

Modeling Dependence of Sensitivity and Specificity

- Sensitivity and specificity are known to be dependent
- Paliwal et al. (2006) fit independent logistic regression models to the prob. of cancer given recall status, & prob. of recall
  - This induces dependence between sens. & spec.
  - Is their assumption reasonable?
- Puggioni, Gelfand, and Elmore (2008) fit independent logistic regression models to a different set of 3 conditional probabilities
  - Also induces dependence between sens. & spec.
  - Is their assumption reasonable?
- More desirable: a multinomial logit model for the 4 outcomes
  - Numerical difficulties for obtaining MLEs or Bayesian estimates (Puggioni et al., 2008)

Conclusions

- Physician performance metrics must include accurate estimates of uncertainty
- Bayesian modeling of patient-level sensitivity and specificity provides estimates of performance measures while fully accounting for uncertainty
- Can be used in other screening settings
  - cardiac (ECHO) examinations
  - clinical breast examinations

Conclusions

- Physician performance metrics must include accurate estimates of uncertainty
- Bayesian modeling of patient-level sensitivity and specificity provides estimates of performance measures while fully accounting for uncertainty
- Can be used in other screening settings
  - cardiac (ECHO) examinations
  - clinical breast examinations

Conclusions

- Physician performance metrics must include accurate estimates of uncertainty
- Bayesian modeling of patient-level sensitivity and specificity provides estimates of performance measures while fully accounting for uncertainty
- Can be used in other screening settings
  - cardiac (ECHO) examinations
  - clinical breast examinations

Bibliography

- Normand, S. T., Glickman, M. E., and Gatsonis, C. A. (1997). Statistical methods for profiling providers of medical care: issues and applications. *J. of the American Statistical Association* 92: 803-814.

- Paliwal, P., Gelfand, A. E., Abraham, L., Barlow, W. E., Elmore, J. G. (2006). Examining accuracy of screening mammography using an event order model. *Statistics in Medicine* 25: 267-283.

- Puggioni, G., Gelfand, A. E., and Elmore, J. G. (2008). Joint modeling of sensitivity and specificity. *Statistics in Medicine* 27:1745-1761.

- Salem-Schatz, S., Morre, G., Rucker, M., Pearson, S. D. (1994). The case for case-mix adjustment in practice profiling: when good apples look bad. *J. of the American Medical Association* 272: 871-874.

- Woodard, D. B., Gelfand, A. E., Barlow, W. E., Elmore, J. G. (2007). Performance assessment for radiologists interpreting screening mammography. *Statistics in Medicine* 26: 1532-1551.