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

1. Physician Performance Assessment
2. Hierarchical Models for Radiologist Accuracy
3. Performance Metrics
4. Related Work and Conclusions
Outline

1. Physician Performance Assessment
2. Hierarchical Models for Radiologist Accuracy
3. Performance Metrics
4. Related Work and Conclusions
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
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
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
1 Physician Performance Assessment

2 Hierarchical Models for Radiologist Accuracy

3 Performance Metrics

4 Related Work and Conclusions
Modeling Accuracy

Logistic regression:

\[
\text{logit}(S_{ij}) = X_{ij}\beta + \tau_i
\]

\[
\tau_i = W_i\gamma + \phi_i
\]

\[
\phi_i \sim N(0, \psi)
\]

- \(S_{ij}\) = sensitivity on mammogram \(i,j\)
- \(X_{ij}\) = risk factors of patient \(i,j\)
- \(W_i\) = attributes of radiologist \(i\)
- \(i = 1,\ldots,I\) radiologists
- \(j = 1,\ldots,n_i\) 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 ⇒ 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 ⇒ lower sensitivity, higher specificity)
  - academic affiliation
  - mammographic case volume
1. Physician Performance Assessment
2. Hierarchical Models for Radiologist Accuracy
3. Performance Metrics
4. Related Work and Conclusions
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

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

$$S(X_0, \beta, \tau_i) = \logit^{-1}(X_0\beta + \tau_i)$$
Performance on a Hypothetical Patient

Sensitivity on a “typical” patient:

Radiologist ID

Sensitivity (%)
Performance on a Hypothetical Patient

Specificity:

<table>
<thead>
<tr>
<th>Radiologist ID</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>43</td>
<td>43</td>
</tr>
<tr>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>97</td>
<td>97</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>101</td>
<td>101</td>
</tr>
<tr>
<td>107</td>
<td>107</td>
</tr>
<tr>
<td>109</td>
<td>109</td>
</tr>
<tr>
<td>118</td>
<td>118</td>
</tr>
<tr>
<td>119</td>
<td>119</td>
</tr>
</tbody>
</table>

Specificity (%) range from 80 to 100.
Performance on a Hypothetical Patient

Empirical sensitivity:

![Graph showing sensitivity vs radiologist ID]
Performance on a Hypothetical Patient

Performance on a hypothetical patient:

<table>
<thead>
<tr>
<th>Radiologist ID</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>44</td>
<td>43</td>
</tr>
<tr>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>94</td>
<td>97</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>101</td>
<td>101</td>
</tr>
<tr>
<td>107</td>
<td>107</td>
</tr>
<tr>
<td>109</td>
<td>109</td>
</tr>
<tr>
<td>118</td>
<td>118</td>
</tr>
<tr>
<td>119</td>
<td>119</td>
</tr>
</tbody>
</table>
Performance on a hypothetical patient:

Radiologist ID

Sensitivity (%)

Radiologist ID
Performance on a Hypothetical Patient

We have “adjusted” for differences in patient mix…

radiologist ID

Sensitivity (%) 0 20 40 60 80 100

3 10 10 3 26 23 3 19 20 23 18 20 7 11 4

Radiologist ID
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.
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:

\[ \mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} S(X_{ij}, \beta, \tau_i) \]

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

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

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

- 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:

\[
\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} S(X_{ij}, \beta, \tau_i)
\]

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

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

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

- Take \(\mu_i - \tilde{\mu}_i\)

- Performance is evaluated while adjusting for radiologist attributes.
Alternatively, take the predicted average accuracy (sensitivity or specificity) of a particular radiologist on her patients:

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

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

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

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

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:

\[ \mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} S(X_{ij}, \beta, \tau_i) \]

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

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

\[ S(X_{ij}, \beta, W_i) = E_{\tau \mid W_i} \{ S(X_{ij}, \beta, \tau) \} \]

- Take \( \mu_i - \tilde{\mu}_i \)

- Performance is evaluated while adjusting for radiologist attributes
Many radiologists had predicted specificity significantly above or below that expected; not so for sensitivity.
1 Physician Performance Assessment
2 Hierarchical Models for Radiologist Accuracy
3 Performance Metrics
4 Related Work and Conclusions
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)
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)
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


