Statistical considerations for classifying radiologists relative to targets for screening mammography interpretive performance

Rebecca Hubbard

Associate Professor
Department of Biostatistics & Epidemiology

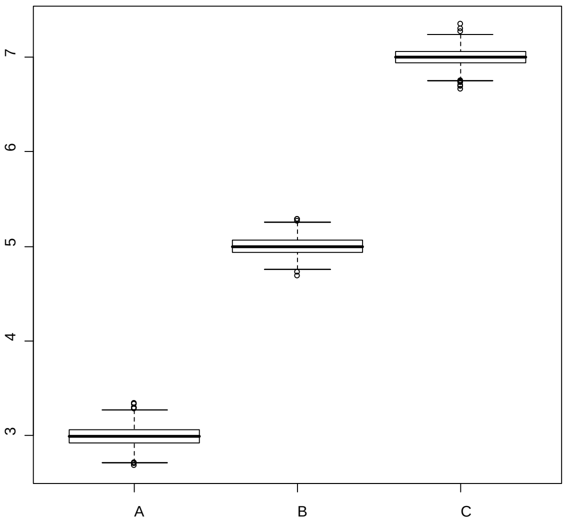
May 12, 2015

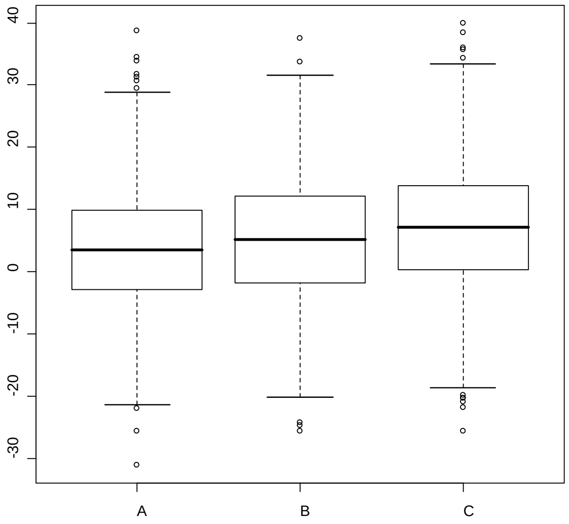
Perelman
School of Medicine
University *of* Pennsylvania

Are radiologists meeting targets for acceptable interpretive performance?

- Targets have been established for interpretive performance
- We would like to identify radiologists who are/are not meeting these targets
- Is this an achievable goal given the data that we have available?
- This is a **profiling** task with objective of classifying radiologists relative to a fixed benchmark

Challenges to estimating radiologist performance





- **Correct classification depends on our ability to distinguish between radiologists with differing performance**
  - Requires more variability between radiologists and less variability within radiologists
- **Within-provider (error) variation may be large due to small provider volume**
- **Reliability = between-provider variability/total variability**
  - Reliability >0.9 generally considered to be necessary for “high stakes” profiling (e.g., public reporting)

Objective

- Discuss relevant statistical considerations for identifying radiologists failing to meet targets for interpretive performance
- Use simulations to demonstrate performance of classification for recall and cancer detection rate
- Demonstrate how these results are modified by use of imperfect proxy for performance measures based on claims data

Conceptual framework

- **Binary event of interest observed for each patient**
  - Recall
  - Screen-detected cancer
- **Each provider has underlying, true performance**
  - Unobservable without complete data on entire patient population
  - Objective of profiling is to make inference on performance based on a finite sample

Profiling methods

- **Classification based on point estimate**
  - For each radiologist compute rate or proportion
  - Compare to guideline target and make classification
  - Can be adjusted to account for differences in patient population (case-mix) using regression methods
  - May be unstable for small patient volumes
- **Classification based on confidence intervals**
  - Compute confidence interval around point estimate
  - If confidence interval lies completely below/above target then classify as failing to meet target
  - Desired precision can be tuned by varying confidence level
  - Addresses instability in estimates for small volume providers
  - May be overly conservative

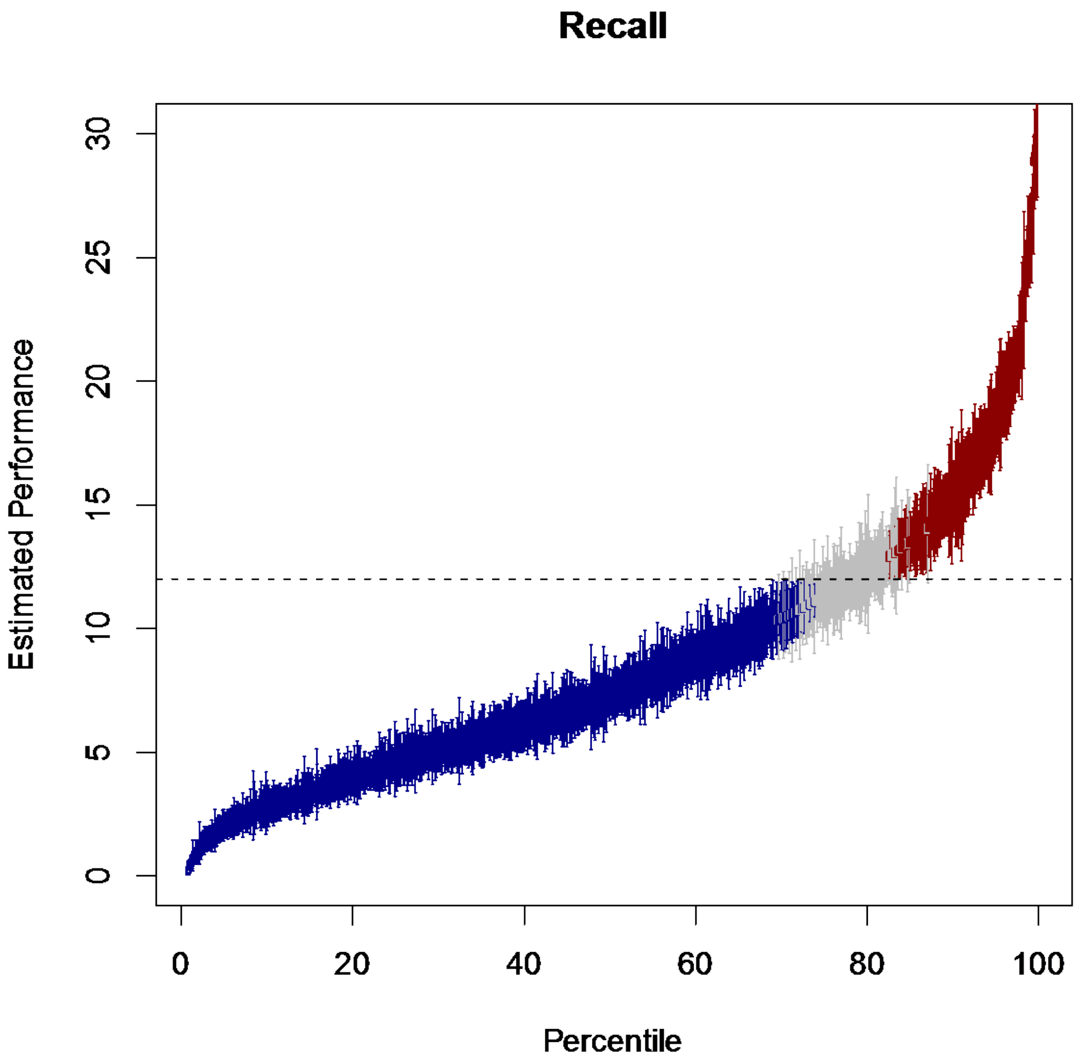
Simulation study design

- We conducted a statistical simulation study to demonstrate performance of classification relative to guideline targets for recall and cancer detection rate using point estimates and CIs
- Simulation study parameters chosen to generate data following real-world distribution of radiologist screening volume, CDR and recall
- Performance of classification evaluated in terms of sensitivity
  - **Sensitivity** = Proportion of radiologists failing to meet target successfully identified as failing to meet target
  - **Specificity** = Proportion of radiologists meeting target successfully identified as meeting target
- Evaluate sensitivity and specificity for
  - CDR relative to threshold of 2.5 per 1000
  - Recall relative to threshold of 12%

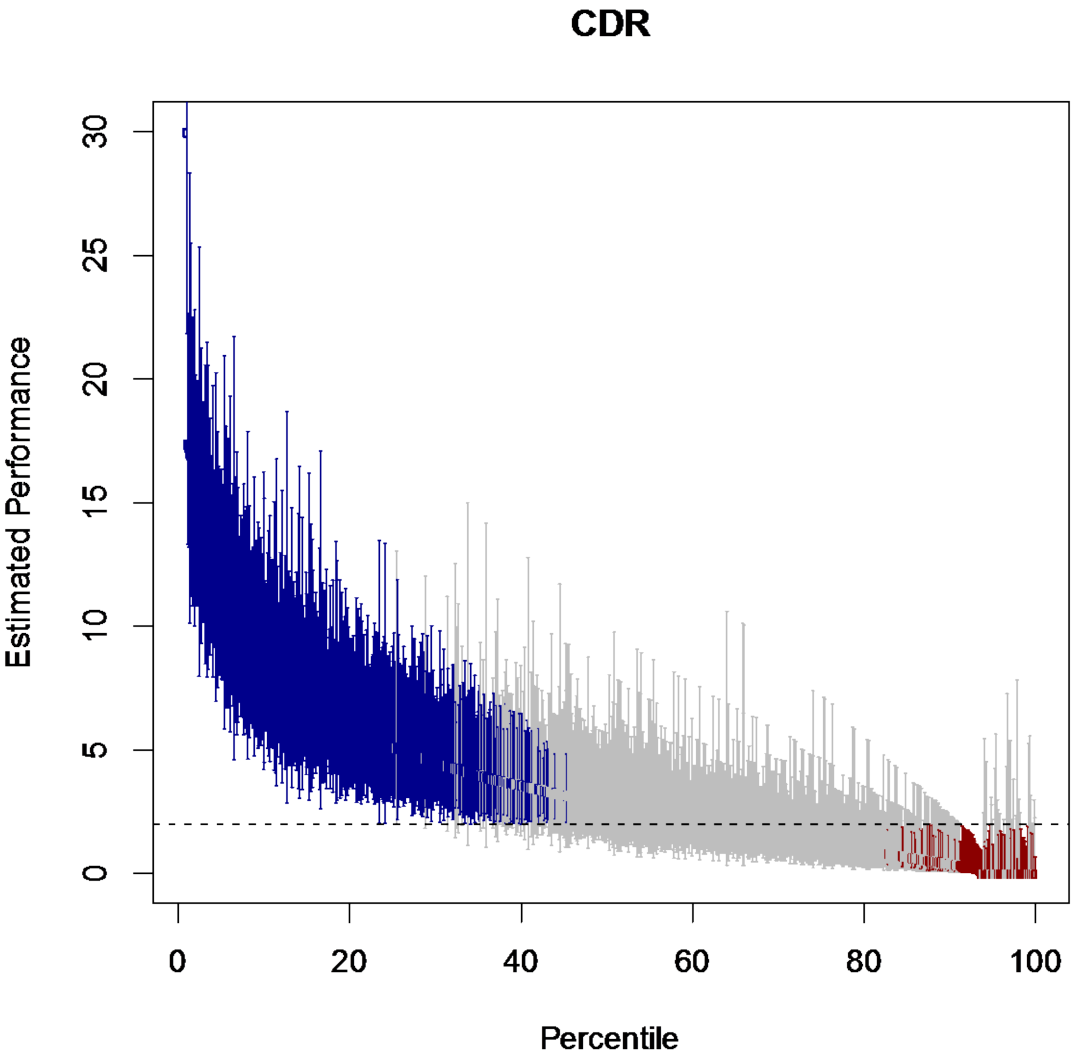
Simulation study design

- Average of N = 1000 radiologists per simulation
- Patient volumes ~ Gamma(3.4,458.4), truncated at 480
    - Mean = 1557, SD = 845
- True radiologist performance measures
    - Recall ~ Beta(2.5, 26.4)
        - Mean = 8.5%, SD = 5.1%
        - Reliability = 0.981
    - CDR ~ Beta(1.36, 372.69)
        - Mean = 3.6/1000, SD = 3.1/1000
        - Reliability = 0.807
- Repeat simulations 1000 times

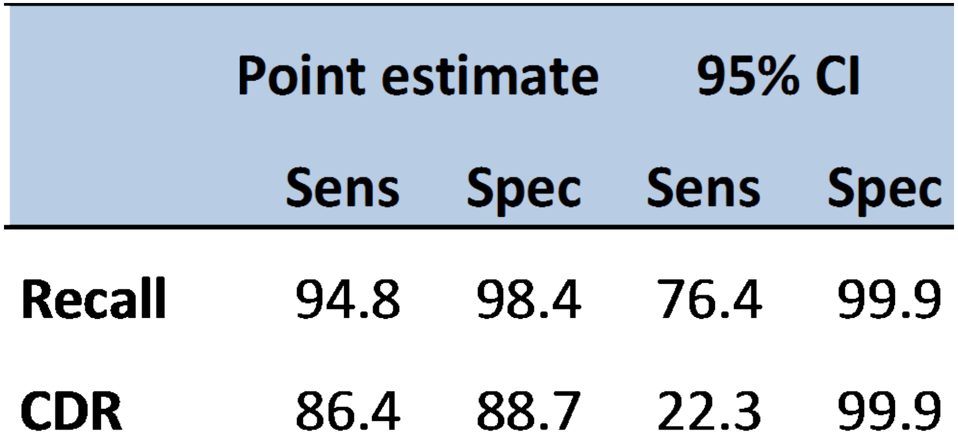
Recall classification for simulated population



CDR classification for simulated population



Sensitivity and specificity for radiologist classification



	Point estimate		95% CI	
	Sens	Spec	Sens	Spec
Recall	94.8	98.4	76.4	99.9
CDR	86.4	88.7	22.3	99.9

Considerations under outcome misclassification

- We also explored classification based on Medicare claims
- This introduces an additional challenge since classification of an exam as resulting in recall or CDR is imperfect
- Algorithm operating characteristics for proxy recall and CDR are known
  - Sensitivity: probability of event based on claims given truly was an event
  - Specificity: probability of no event based on claims given truly was no event
- How does using an imperfect proxy for outcomes affect radiologist classification relative to targets?

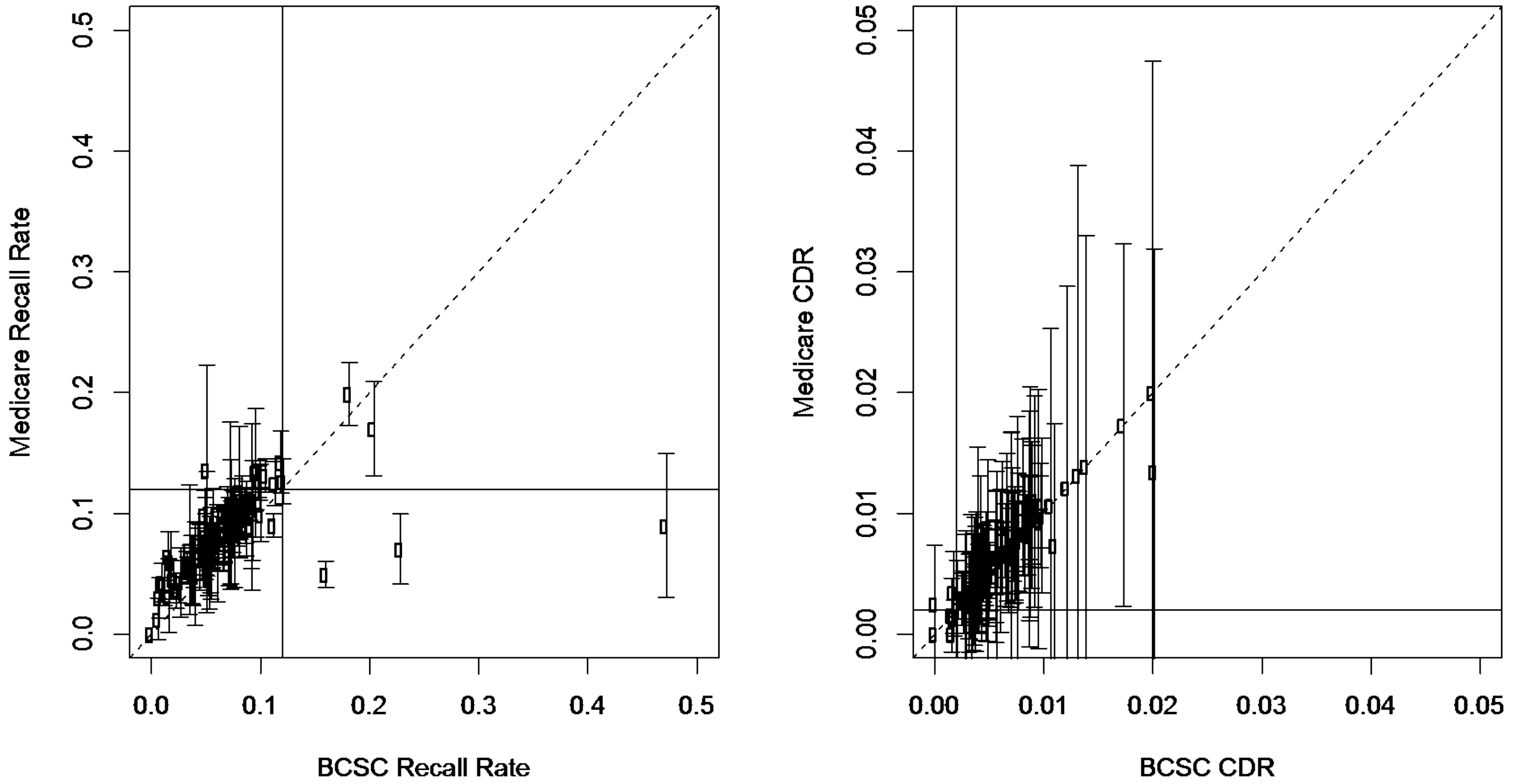
Claims-based algorithms

- Claims-based algorithms for recall and screen-detected cancer
  - Outcomes based on ICD-9 and HCPCS codes for breast imaging and breast cancer diagnosis around time of screening mammogram
- Algorithm performance:
  - Recall: Sensitivity = 82.6%, Specificity = 96.7%
  - CDR: Sensitivity = 94.0%, Specificity = 99.9%

Medicare-linked BCSC data

- Clinical data on mammography interpretation and cancer outcomes available from the Breast Cancer Surveillance Consortium
- Linked to Medicare claims
- Data on 134,330 screening mammograms from 2003 – 2005 performed at 106 mammography facilities
- Volume ranged from 52 to 5,925 mammograms per facility

Claims-based performance estimates

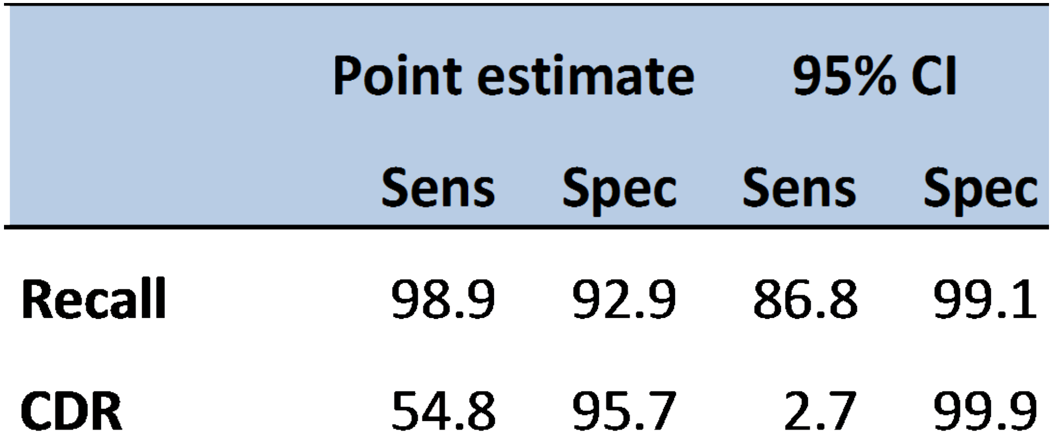


ML point estimates and 95% confidence intervals

Comparison of Medicare and BCSC estimates

- Imperfect specificity results in slight inflation of recall and CDR estimates
- Provider-level estimates based on claims agree well with gold-standard
- However, agreement between the two sources does not ensure correct classification of providers
- Evaluation of claims-based measures often includes only operating characteristics, but this does not address error in profiling due to sampling variability
- We repeated the simulation study incorporating error due to imperfect classification

Sensitivity and specificity under misclassification



	Point estimate		95% CI	
	Sens	Spec	Sens	Spec
Recall	98.9	92.9	86.8	99.1
CDR	54.8	95.7	2.7	99.9

Conclusions

- Reliability provides a good first indication of the likely success of profiling
- With or without misclassification of outcomes, performance is reasonable for recall because it is relatively common
- Classification of radiologists on CDR is challenging because outcome is rare
  - Profiling based on point estimates works reasonably well when there is no misclassification of events
  - Misclassification of events at the level of our Medicare claims-based algorithm resulted in low sensitivity
- Incorporating uncertainty through CIs results in decrease in sensitivity, increase in specificity
- The purpose of profiling should be considered when choosing an approach/determining acceptable levels of radiologist misclassification

Acknowledgments

NCI-funded grant R21CA158510 and Breast Cancer Surveillance Consortium (HHSN261201100031C)

Rhondee Benjamin-Johnson
Rebecca Smith-Bindman
Weiwei Zhu
Tracy Onega
Joshua Fenton

References

Hubbard RA, Benjamin-Johnson R, Onega, T, Smith-Bindman R, Zhu W, Fenton JJ. 2015. Classification accuracy of Medicare claims-based methods for identifying providers failing to meet performance targets. *Statistics in Medicine*. 34(1):93-105.

Thank you