Modeling Rater Diagnostic Skills in Binary Classification Processes

Xiaoyan(Iris) Lin

October 13, 2016

Joint work with Don Edwards

Outline

- Motivation
- Proposed Model and Bayesian Estimation
- Cost analysis and ROC
- Simulation Study
- Mammogram Data Analysis
- Conclusion

Motivation

Subjective Judgement is Ubiquitous in the Diagnosis of Disease

For example

- Cytogenetics (chromosome testing)
- Mammograms and/or MRIs for breast cancer
- Ultrasound images
- Radiographs for fractures or tumors
- MRIs for brain/lung lesions
- Chest X-rays for TB
- Chest X-rays for COPD or emphysema
- T-SPOT Lab tests for tuberculosis
- Gleason scores (prostate cancer)
- Skin cancer or other skin condition tests

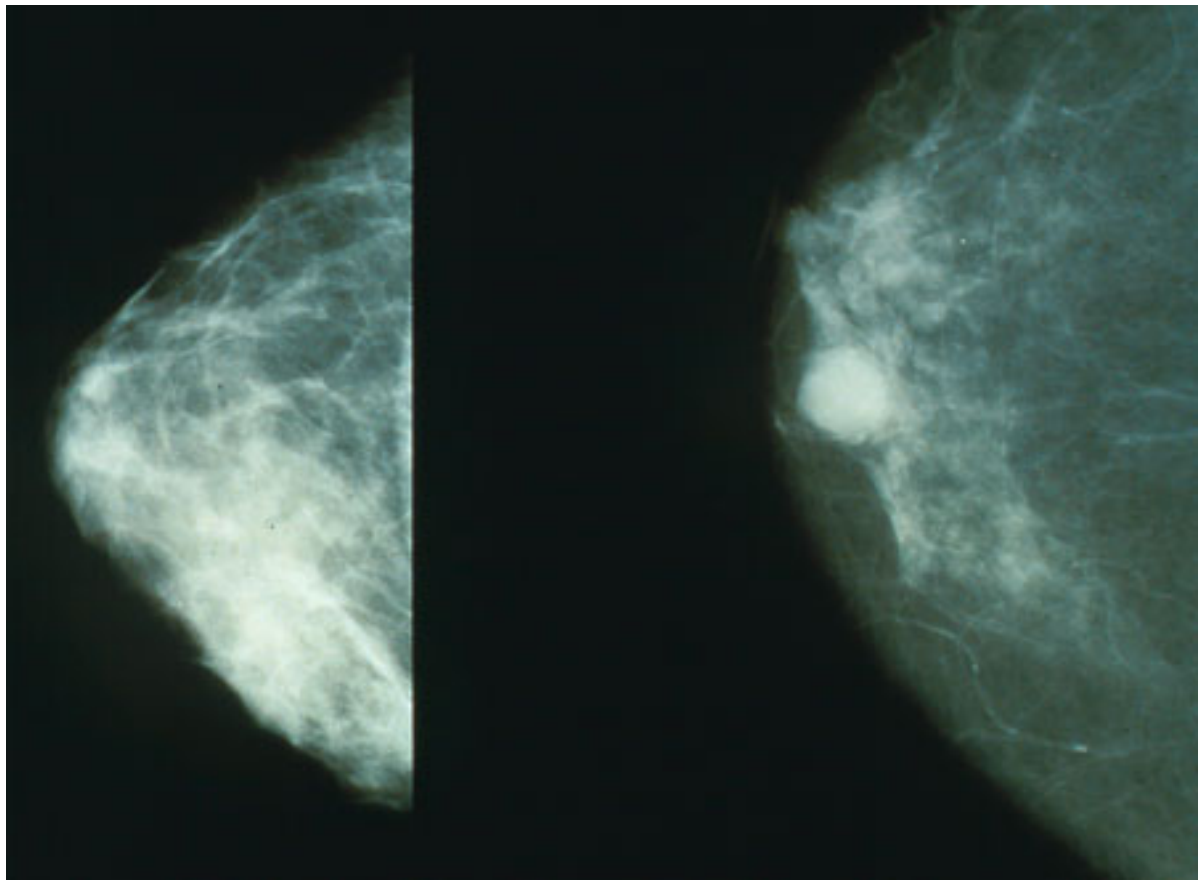


Figure: Normal (left) versus cancerous (right) mammography image

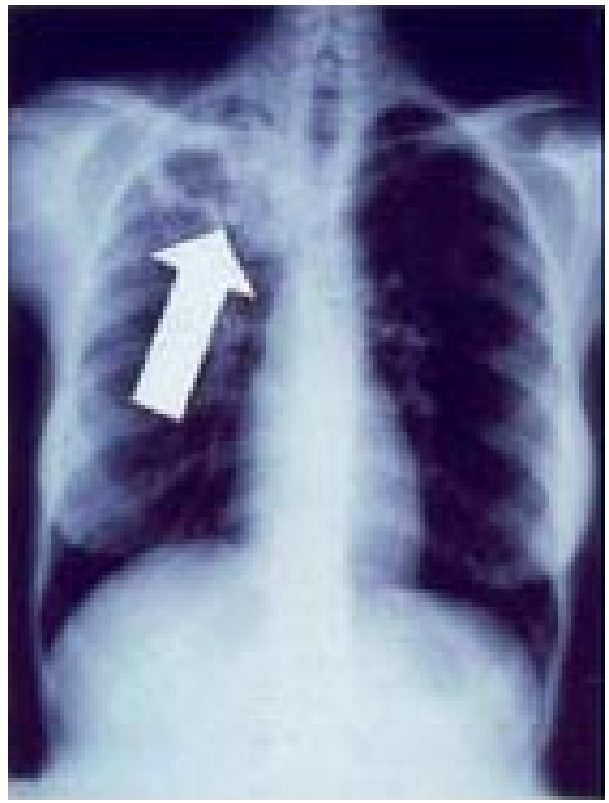


Figure: Tuberculosis creates cavities visible in x-rays like this one in the patient's right upper lobe.

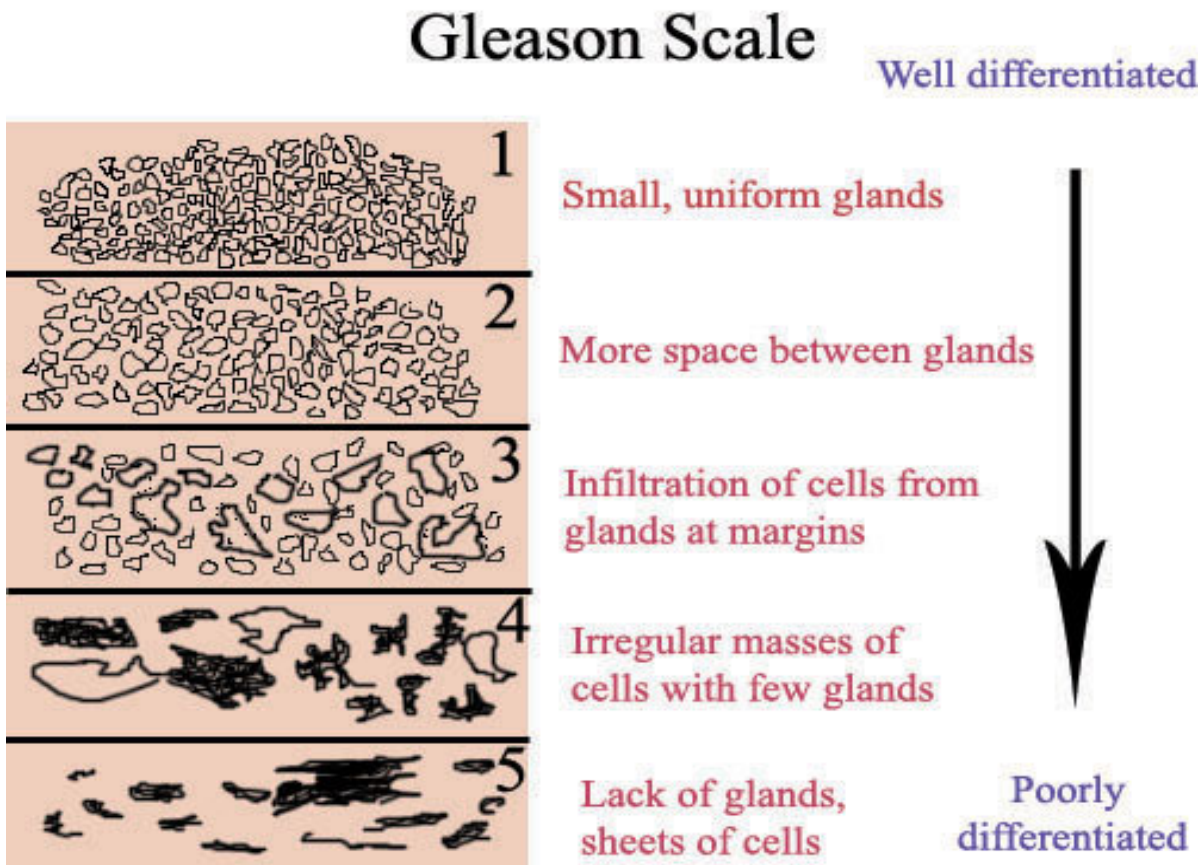


Figure: A Gleason score is given to prostate cancer based upon its microscopic appearance. Cancers with a higher Gleason score are more aggressive and have a worse prognosis.

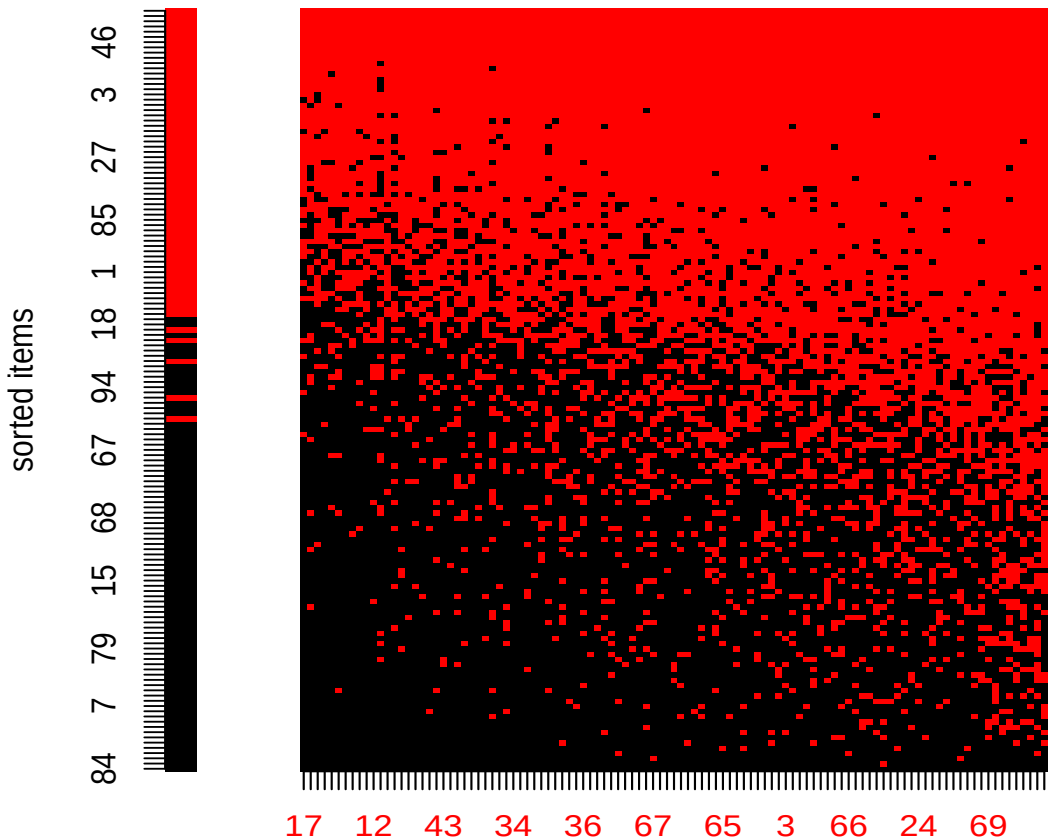
Mammogram Data

The motivating data set was collected by Beam et al. (1996), who studied variability in the interpretation of mammograms by U.S. radiologists.

- One hundred and seven radiologists for the study came from a stratified random sample of mammography facilities accredited by the U.S. Food and Drug Administration.
- Index mammograms of good technical quality were obtained from 148 women randomly sampled from a large screening program affiliated with the University of Pennsylvania. Of these, 82 were cancer-free and 64 had malignancies.

- Every radiologist interpreted all mammograms in a controlled reading environment during two 3-hour periods. The 5-point Breast Imaging Reporting and Data System (BI-RADS) scale was used for the radiologists to record the results: 1=normal, return to normal screening; 2=benign, return to normal screening; 3= probably benign, 6-month follow-up; 4=possibly malignant, biopsy recommended; 5=probably malignant, biopsy strongly recommended.

Figure below presents the dichotomized mammography data in order of sorted raters and sorted mammographies according to radiologists' rating. For the main part of the figure, color red indicates rating "BI-RADS ≥ 3" and color black indicates the otherwise. The left bar of the figure shows the true disease status with red indicating "cancer" and black indicating "cancer free".



Proposed Model and Bayesian Estimation

Proposed Model

Denote by D_i the true disease outcome of patient i, and W_{ij} the diagnostic result of patient i by rater j, both with 1 for a positive result. Our model is

$$\begin{aligned} P(D_i = 1|u_i) &= \Phi(u_i), \\ P(W_{ij} = 1|u_i) &= \Phi(a_j + b_j u_i), \end{aligned} \tag{1}$$

for $i = 1, \ldots, m;\ j = 1, \ldots, n$.

- u_i the unobserved disease severity for patient i .
- a_j the inherent tendency (bias) of clinician j to say "disease" when $u_i = 0$.
- b_j the clinician's diagnostic skill (magnifier).

Prior Specification

The rater effects a_j and b_j are independently assumed normal and truncated normal priors, respectively:

$$\begin{aligned} a_j &\overset{i.i.d}{\sim} N(\mu_a, \tau_a^{-1}), \\ b_j &\overset{i.i.d}{\sim} N(\mu_b, \tau_b^{-1}) 1_{(b_j > 0)}, \end{aligned} \tag{2}$$

for $j = 1, \ldots, n$.

To allow for the estimation of these parameters, we assign conventional diffuse hyper priors. Specifically, we assign normal priors $N(0, \tau^{-1})$ to μ_a and μ_b with τ a small value such as 0.01, and gamma priors $Ga(\theta, \theta)$ to τ_a and τ_b with θ a small value such as 0.1.

Bayesian Computation and Estimation

Computation: Jags (Just Another Gibbs Sampler)

```
model{

  for (i in 1:m) {
    D[i]~ dbern(p[i])
    probit(p[i]) <- u[i]
    u[i]~ dnorm(0,1)
    for (j in 1:n) {
   probit(q[i,j]) <- a[j]+b[j]*u[i]
    W[i,j]~ dbern(q[i,j])
    wpred[i,j]~ dbern(q[i,j]) }
   }

  for (j in 1:n) {
     b[j]~ dnorm(mub,taub)T(0,)
     a[j]~ dnorm(mua,taua)}

  mua~ dnorm(0,.001)
  mub~ dnorm(0,.001)
  taua~ dgamma(.001,.001)
  taub~ dgamma(.001,.001)
}
```

Sensitivity and Specificity

For rater j, the sensitivity and specificity are defined as

$$Se_j = P(W_{ij} = 1 | D_i = 1) = \frac{\int_u \Phi(u)\Phi(a_j + b_j u) dF(u)}{\int_u \Phi(u) dF(u)}$$

and

$$Sp_j = P(W_{ij} = 0 | D_i = 0) = \frac{\int_u \left[1 - \Phi(u)\right]\left[1 - \Phi(a_j + b_j u)\right] dF(u)}{1 - \int_u \Phi(u) dF(u)},$$

respectively, where $F(u)$ denotes the true distribution of the latent disease severity u.

Estimation

- All parameters or latent variables can be estimated by MCMC samples.
-

$$\hat{Se}_j = \frac{1}{L}\sum_{l=1}^{L}\left[\frac{\sum_{i=1}^{m}\Phi(u_i^{(l)})\Phi(a_j^{(l)}+b_j^{(l)}u_i^{(l)})}{\sum_{i=1}^{m}\Phi(u_i^{(l)})}\right] \tag{3}$$

and

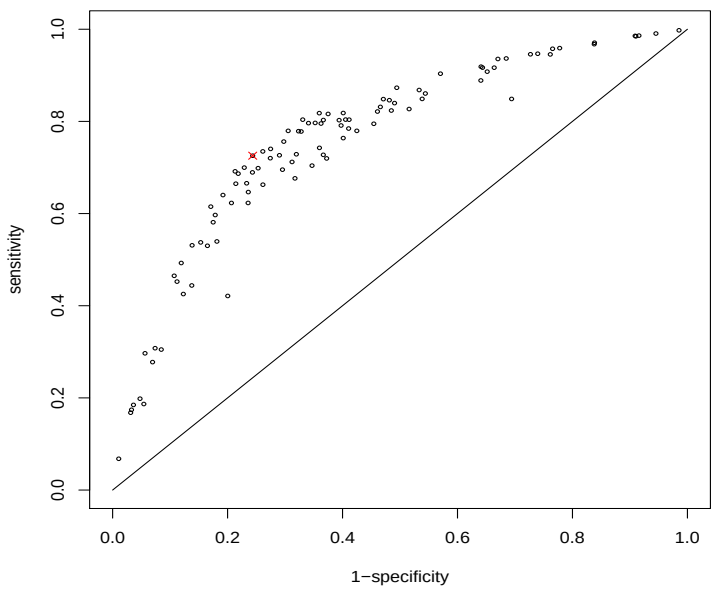
$$\hat{Sp}_j = \frac{1}{L}\sum_{l=1}^{L}\left[\frac{\sum_{i=1}^{m}\left(1-\Phi(u_i^{(l)})\right)\left(1-\Phi(a_j^{(l)}+b_j^{(l)}u_i^{(l)})\right)}{\sum_{i=1}^{m}\left(1-\Phi(u_i^{(l)})\right)}\right], \tag{4}$$

where the superscript (l) denotes the lth MCMC sample in the Gibbs sampling iterations after the burn-in period, and L is the total number of MCMC samples used for the estimation.

Cost Analysis and ROC Curve

ROC (receiver operating characteristic) Curve

- Rater-specific sensitivities and specificities can be used to identify strong raters via a simple ROC curve. The ROC curve is created by plotting true positive rate (sensitivity) over false positive rate (one minus specificity) for each rater. One simple criterion to apply is that strong raters are those with short distances to the point $(0, 1)$.



Cost

Suppose

- the cost of a false negative is a fixed constant $C_{FN}(\geq 0)$,
- the cost of a false positive is a fixed constant $C_{FP}(\geq 0)$, and
- the cost ratio of C_{FN} over C_{FP} is denoted by θ.

Then according to the model, the total expected cost TC for an individual rater is as follows (suppressing subscripts i and j):

$$\begin{aligned} TC &= C_{FN}P(W=0|D=1)P(D=1) + C_{FP}P(W=1|D=0)P(D=0) \\ &= C_{FN}P(W=0,\ D=1) + C_{FP}P(W=1,\ D=0) \\ &= C_{FN}\int_u \Phi(u)\big[1-\Phi(a+bu)\big]dF(u) + C_{FP}\int_u (1-\Phi(u))\Phi(a+bu)dF(u) \\ &\propto \theta\int_u \Phi(u)\big[1-\Phi(a+bu)\big]dF(u) + \int_u (1-\Phi(u))\Phi(a+bu)dF(u). \end{aligned}$$

A strong rater should have a small total expected cost.

Proposition 1. Given a fixed rater magnifier b, the total cost will be minimized if the rater bias a^* satisfies the following conditions:

$$\int_u \big[1-(\theta+1)\Phi(u)\big]\phi(a^*+bu)\,dF(u)=0,$$
$$\int_u \big[1-(\theta+1)\Phi(u)\big]\phi(a^*+bu)(a^*+bu)\,dF(u)<0.$$

Note: For a given rater, the optimal bias can be numerically determined via the R function *uniroot* via Proposition 2.

Proposition 2. Assume $b>0$. If the optimal rater bias a^* is used as defined in Proposition 2, then the total cost TC is a decreasing function of the rater's magnifier b.

Proposition 3. A rater with a larger magnifier b produces a uniformly better ROC curve when varying the value of a regardless of the distribution $F(u)$.

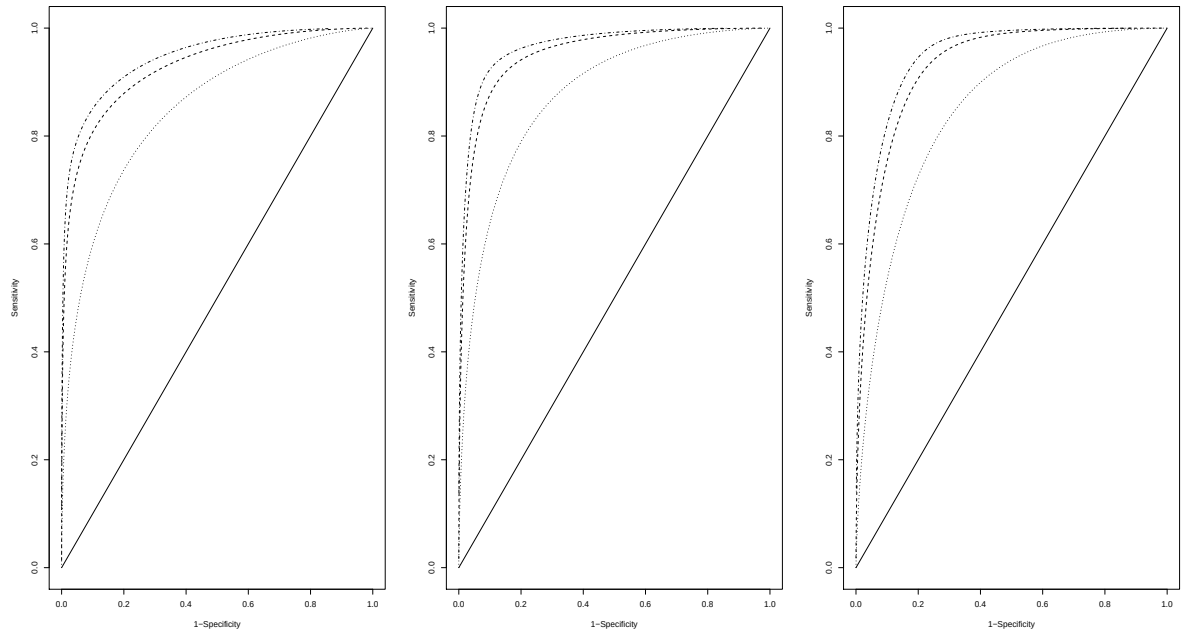


Figure: A two component normal mixture distribution $F(u) = p\Phi(u; -2, 1) + (1 - p)\Phi(u; 2, 1)$ is adopted for calculating the sensitivity and specificity, where $p = 0.8, 0.5, 0.2$ for left, middle, and right panels. For these three panels, the solid, dotted, dashed, dotdashed lines are for $b = 0, 0.5, 1, 1.5$, respectively.

Simulation Study

Data Generation

- We generated 200 data sets for two sample size scenarios: $m = 50$ patients and $n = 50$ raters, and $m = 150$ patients and $n = 100$ raters.
- The raters had the same true diagnostic biases and skills across all of the data sets. True diagnostic biases a_j, $j = 1, \ldots, n$, were independently sampled from a standard normal distribution while true diagnostic magnifiers b_j, $j = 1, \ldots, n$, were independently sampled from a truncated normal distribution $N(1.2, .3^2)1_{(b_j > 0)}$.
- For each data set, patients' unobserved disease severity u_i, $i = 1, \ldots, m$, were independently sampled from a bimodal normal distribution $0.5N(-\sqrt{0.8}, 0.2) + 0.5N(\sqrt{0.8}, 0.2)$.
- Then based on the true simulated a_j, b_j, and u_i, D_i and W_{ij} were generated for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

Simulation Results

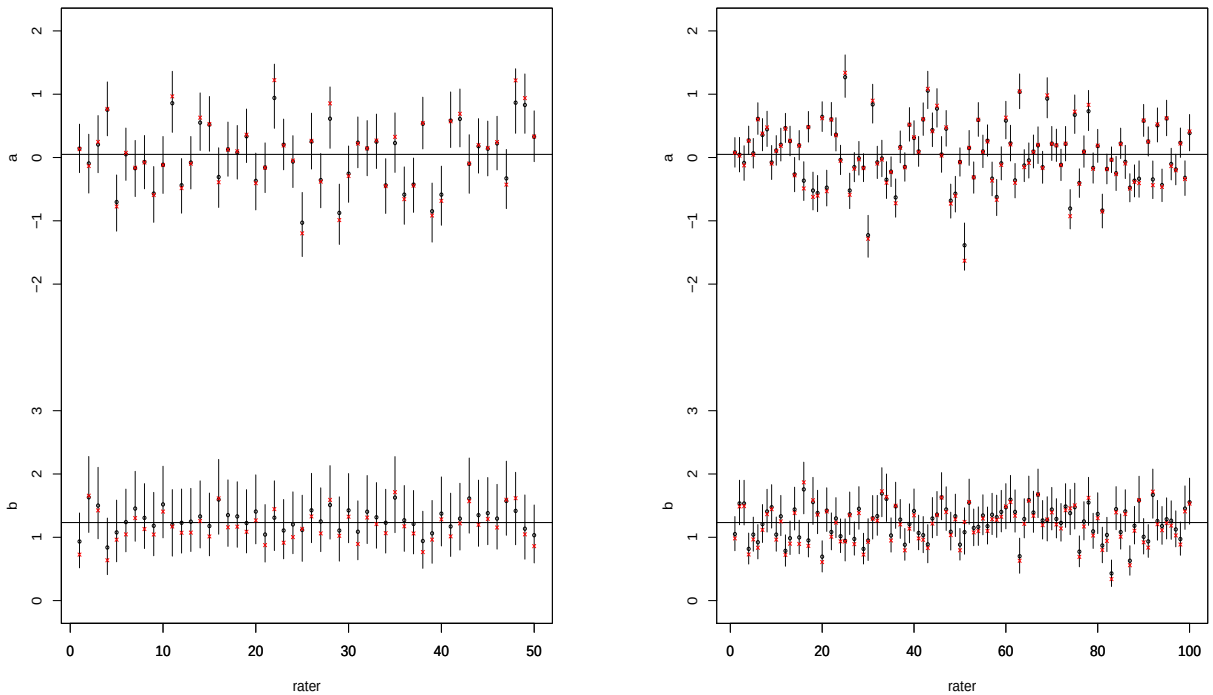


Figure: Plot of point estimates and 95% credible intervals of a_j and b_j for the raters. Cross shows the true values of a_j and b_j.

.

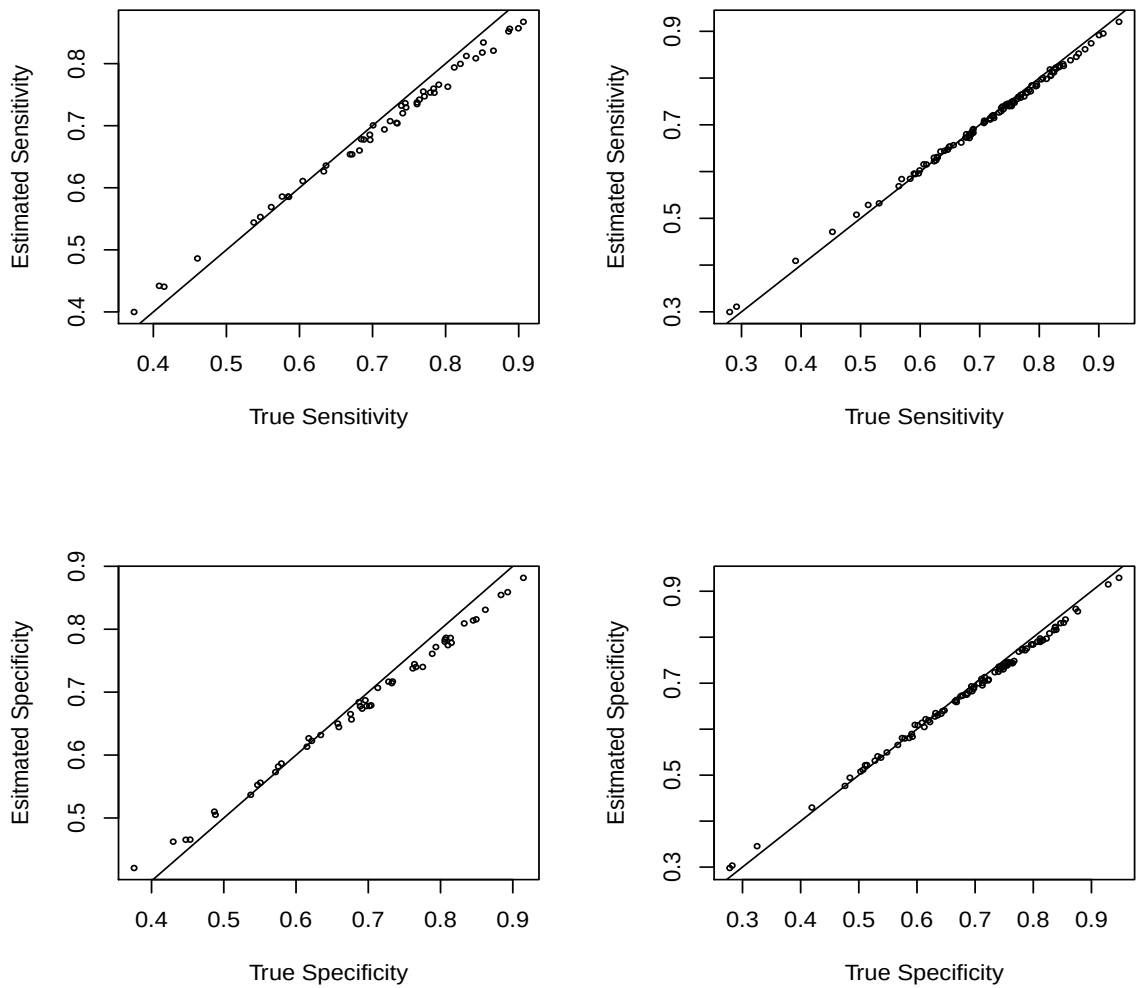


Figure: Plot of point estimates of sensitivity and specificity vs. true values of sensitivity and specificity.

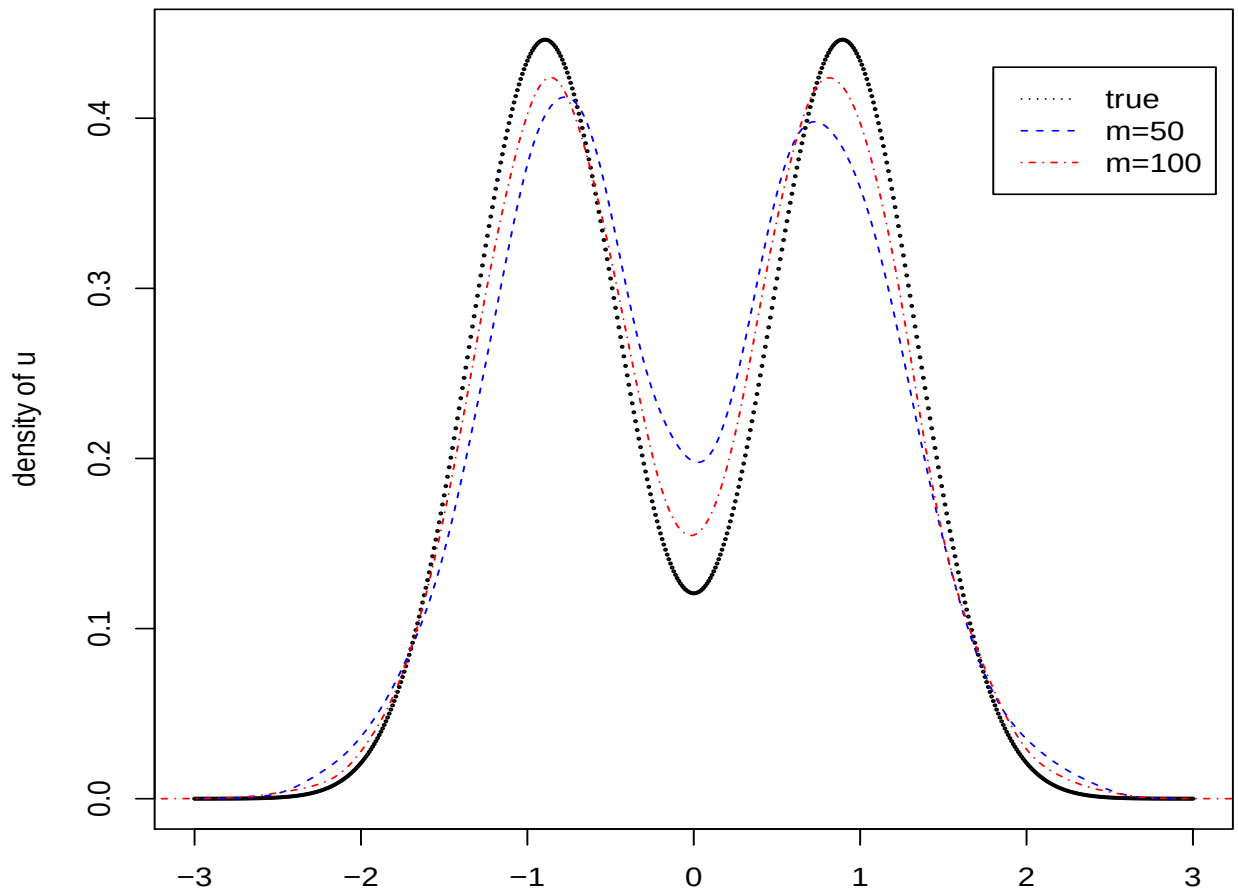


Figure: Estimation of the distribution of patient disease severity u.

Analysis for the Mammogram Data

Results of Analysis

- The posterior mean of μ_a is 0.241, showing that there is a slight positive diagnostic bias of U.S. radiologists.
- The posterior mean of μ_b, the overall mean of diagnostic magnifier of U.S. radiologists, is 1.727.
- The posterior means of $\tau_a^{-1/2}$ and $\tau_b^{-1/2}$ are 0.494 and 0.395, reflecting the variation of diagnostic bias and diagnostic skill of the U.S. radiologists.

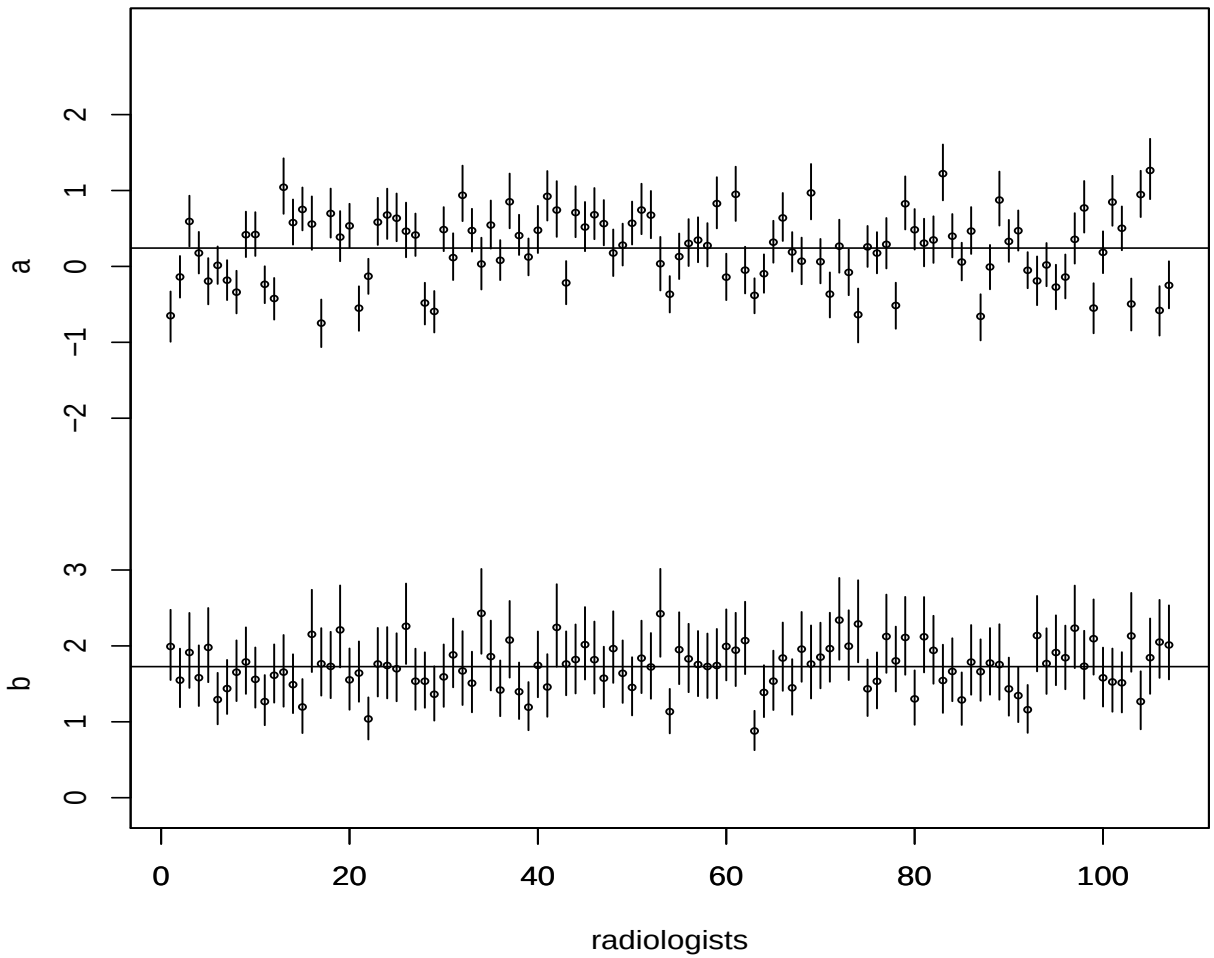


Figure: Estimation of diagnostic bias a_j and diagnostic magnifier b_j for the 107 radiologists in the mammogram data

Table: Three identified strongest raters for the mammogram data based upon different values of cost ratio θ. For each identified rater, a^* is the solved optimal diagnostic bias.

θ	rater	a^*	a	b	sensitivity	specificity
0.5	74	-1.208	-0.639	2.287	0.614	0.861
			(-1.007,-0.291)	(1.749, 2.924)	(0.563, 0.661)	(0.824, 0.892)
	1	-1.144	-0.643	2.0312	0.600	0.862
			(-0.9966,0.3097)	(1.573,2.575)	(0.546,0.651)	(0.821,0.895)
	99	-1.162	-0.557	2.109	0.618	0.853
			(-0.892,-0.230)	(1.635, 2.666)	(0.565, 0.665)	(0.812, 0.886)
1	34	-0.223	0.032	2.453	0.614	0.861
			(-0.319, 0.388)	(1.882, 3.131)	(0.563, 0.661)	(0.824, 0.892)
	53	-0.226	0.034	2.416	0.703	0.789
			(-0.312, 0.383)	(1.886, 3.039)	(0.660, 0.742)	(0.739, 0.831)
	93	-0.250	-0.181	2.161	0.672	0.809
			(-0.500, 0.145)	(1.668, 2.745)	(0.623, 0.717)	(0.762, 0.849)
2	42	0.777	0.745	2.246	0.788	0.658
			(0.397, 1.137)	(1.707, 2.848)	(0.751, 0.822)	(0.593, 0.718)
	26	0.787	0.478	2.281	0.756	0.713
			(0.135, 0.832)	(1.773, 2.861)	(0.717, 0.792)	(0.655, 0.766)
	79	0.743	0.830	2.120	0.802	0.625
			(0.486, 1.175)	(1.648, 2.620)	(0.765, 0.837)	(0.554, 0.690)
5	105	3.062	1.273	1.860	0.866	0.468
			(0.879,1.709)	(1.359,2.413)	(0.831,0.896)	(0.390,0.543)
	83	2.957	1.217	1.532	0.877	0.416
			(0.859,1.615)	(1.076, 2.026)	(0.841,0.907)	(0.336,0.494)
	13	2.988	1.059	1.666	0.850	0.490
			(0.707,1.423)	(1.203,2.1645)	(0.813,0.882)	(0.411,0.566)

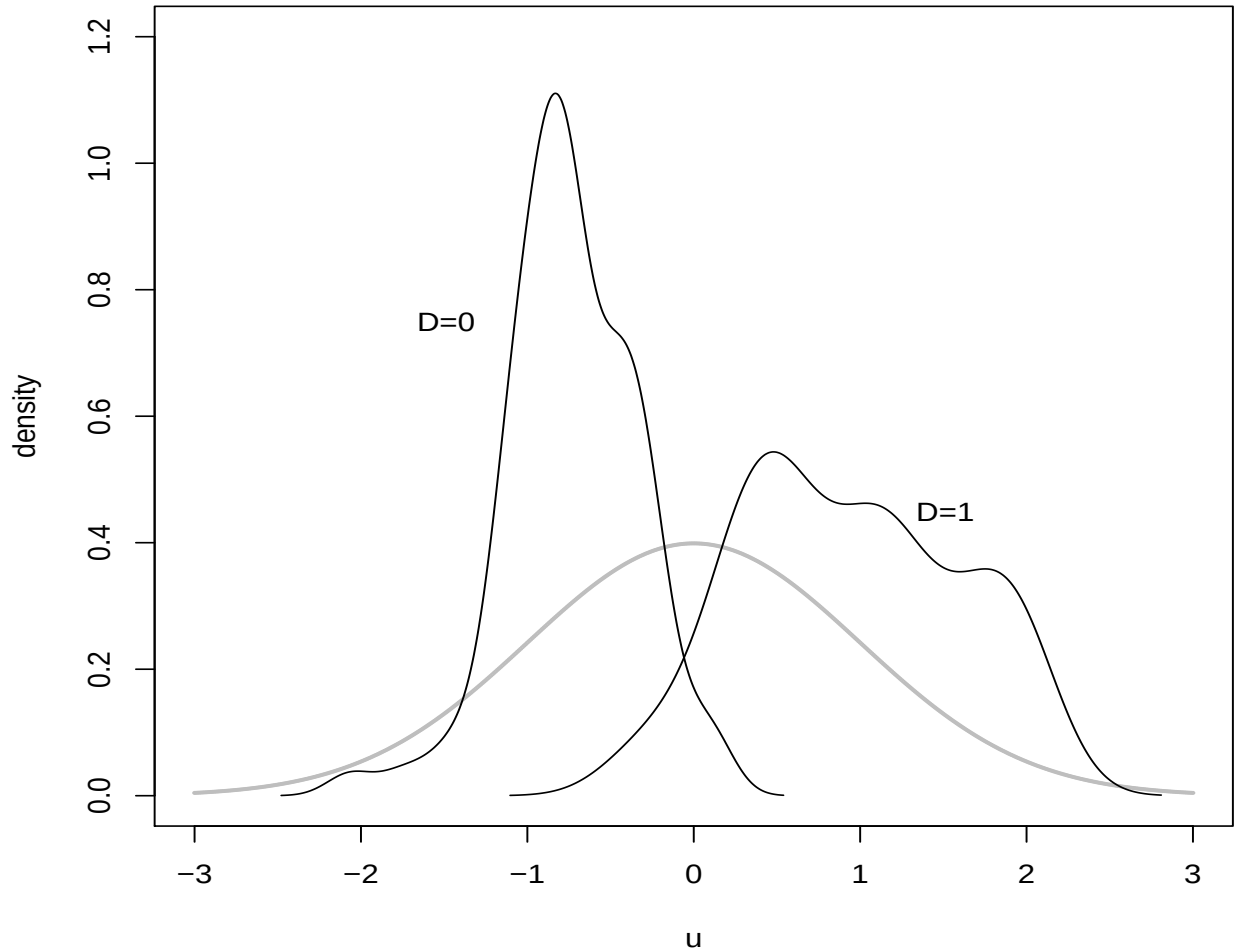


Figure: Estimation of the distribution of patient latent disease severity u for the mammogram data.

Conclusion

- A novel statistical method is provided to objectively assess individual and group diagnostic skills with dichotomous rating data
- A statistical guide is provided to optimize raters' performance through the adjustment of diagnostic bias and skill based on the cost theory.
- It can also assess the distribution of patient latent disease severity, and facilitate the epidemiology study of disease.

Future Research

- Investigate some other semi- or non-parametric link.
- Propose model-based agreement measures for measuring the inter-rater agreement.
- Augment the model by adding regression analysis on diagnostic bias and skill.
- Extend the binary-outcome model to ordinal-outcome model.

Thank you!