Statistical Methods for Evaluating Mammography Interpretive Performance

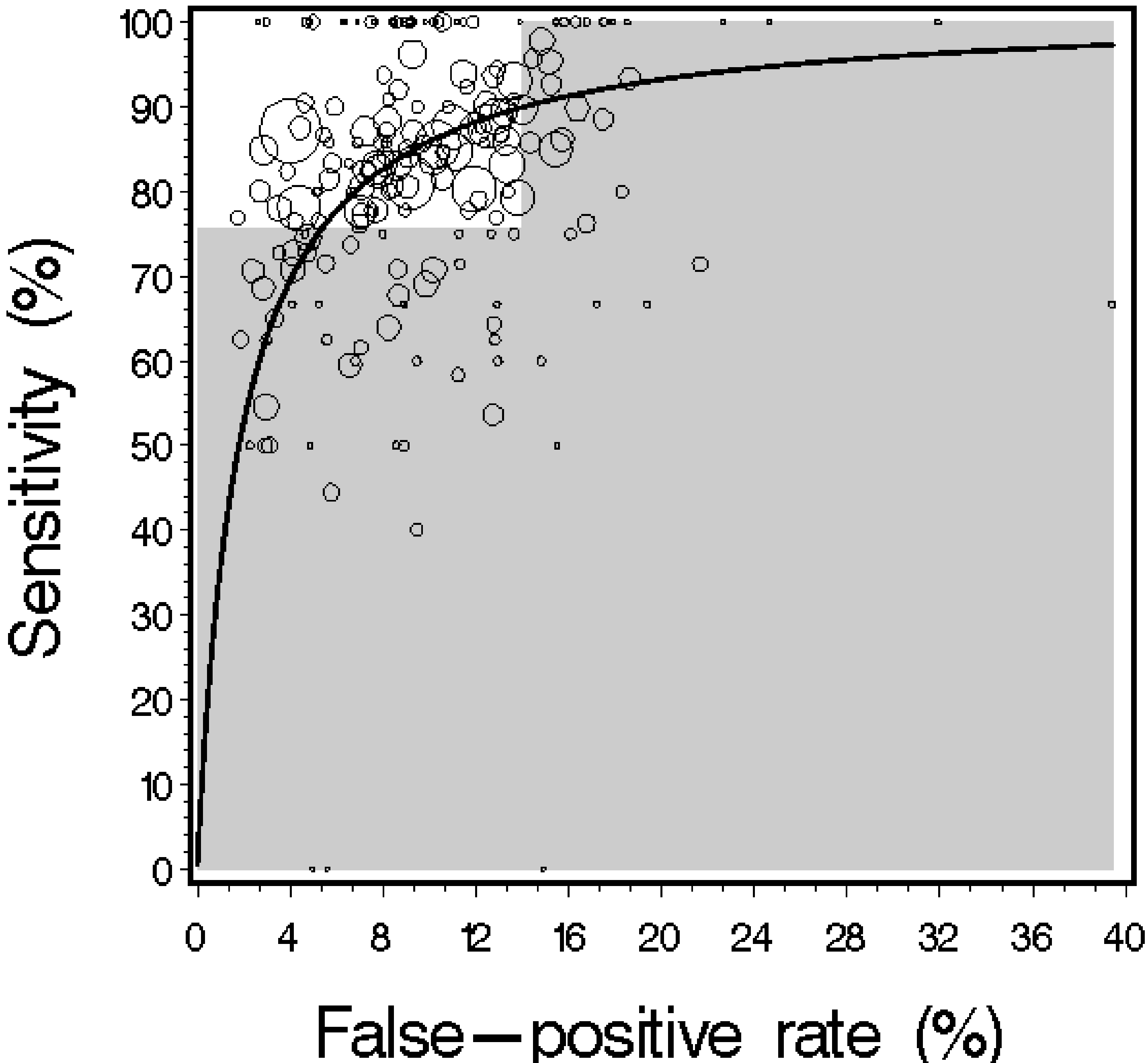
Diana L. Miglioretti, PhD
Sebastien JPA Haneuse, PhD

Group Health Center for Health Studies &
University of Washington, Seattle, WA, USA

Breast Cancer Surveillance Consortium
BCSC
Working together to advance breast cancer research

Background and Motivation

- Extensive variability in mammography interpretation exists among radiologists in the United States.
- Interest in understanding reasons for this variability
  - Patient factors
    - Age, breast density, time since last mammogram
  - Practice and facility characteristics
    - Double reading, CAD
  - Radiologist characteristics
    - Years of experience
    - Training
    - Specialty
    - Interpretive volume (current requirement 960 mammograms over 2 years)



Sensitivity (%)
100
90
80
70
60
50
40
30
20
10
0
0
4
8
12
16
20
24
28
32
36
40
False−positive rate (%)

Background and Motivation

- Conflicting study findings on whether and how interpretive volume influences performance
- Priorities from Institute of Medicine report on *Improving Breast Imaging Quality Standards:*
  - **"Determine the effects of reader volume on interpretive accuracy, controlling for other factors that improve interpretive performance."**
  - **"More study is needed to establish the implications, advantages, and disadvantages of statistical approaches to evaluating the influence of volume on interpretive performance."**

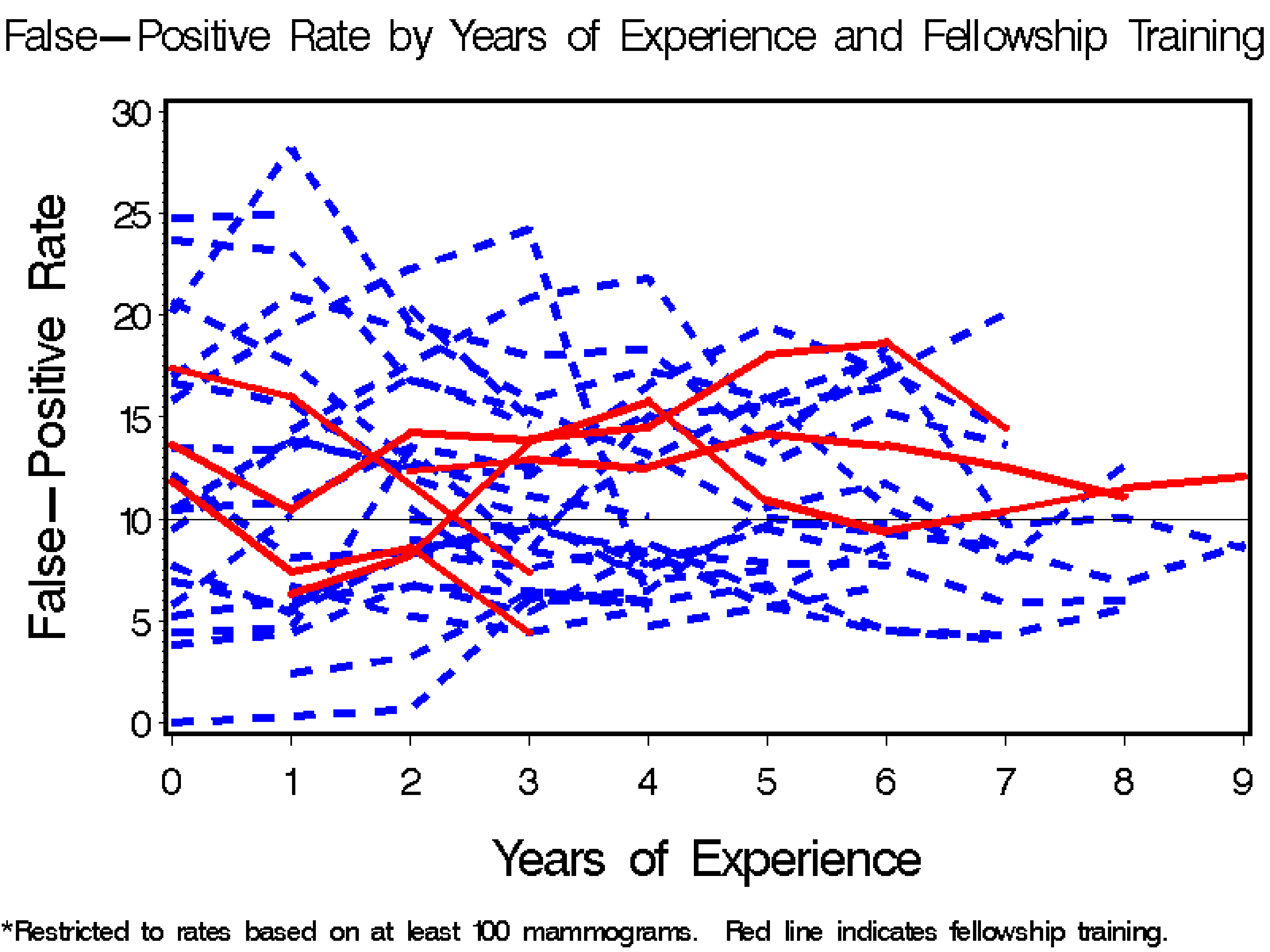
Physician characteristics associated with *clinical* screening performance

BCSC
Working together to advance breast cancer research

Characteristic	Association	Reference
Years of Experience	↓ FP, no Δ TP ↓ FP, ↓ TP ↓ FP ↓ FP	Smith-Bindman, 2005 Barlow, 2004 Elmore, 2002 Tan, 2006
Volume	↓ FP (middle vol), no Δ TP ↑ FP, ↑ TP ↑ PPV >4,000 ↓ FP, no Δ CDR ↓ FP, ↑ or no Δ TP no Δ CDR or Recall, ↑ PPV ↑ CDR	Smith-Bindman (US), 2005 Barlow (US), 2004 Miglioretti (US), 2007 Théberge (Quebec), 2005 Kan (BC), 2000 Coldman (Canada), 2006 Rickard (South Wales), 2006
Screening Focus	↑ FP, ↑ TP no Δ FP or TP	Smith-Bindman, 2005 Barlow, 2004
Specialists	↓ Recall, ↑ CDR no Δ Recall or CDR	Sickles, 2002 (*N*=10) Leung, 2007 (*N*=9)

Statistical issues that could account for conflicting study findings

- Model assumptions
  - *E.g.,* variability among radiologists does not depend on volume
  - Expect more experienced radiologists to perform more similarly than less experienced radiologists
- Differences in regression frameworks used
  - Conditional/cluster-specific
  - Marginal/population-averaged


False−Positive Rate by Years of Experience and Fellowship Training
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
False−Positive Rate
Years of Experience
*Restricted to rates based on at least 100 mammograms. Red line indicates fellowship training.

Importance of Accounting for Clustering within Radiologists

- Mammography performance data are *clustered*
  - Radiologists have different skill levels and thresholds
  - Interpretations made by the same radiologist are correlated
- For valid inference, it is necessary to adjust for correlation among interpretations made by the same radiologist.
  - Naïve methods (chi-square, logistic regression) provide biased standard errors
- Example:
  - 50,000 mammograms interpreted by 10 radiologists (5 experienced, 5 non-experienced)
  - Tempting to think of as 50,000 independent observations
  - Reality is that sample size is closer to “10” independent observations

Common Regression Methods for Clustered Binary Data

- Conditional (cluster-specific) Models
  - logit(P(recall | $\boldsymbol{x}_{ij}$, z_i)) = $\boldsymbol{x}_{ij}\,\beta^c + z_i$
  - z_i = radiologist-specific effect to account for correlation
  - Random effects model: $z_i \sim \text{Normal}(0, \sigma^2)$
  - Conditional logistic regression: z_i fixed effect
- Marginal (population-averaged) Models
  - logit(P(recall | $\boldsymbol{x}_{ij}$)) = $\boldsymbol{x}_{ij}\,\beta^M$
  - Generalized Estimating Equations (GEE)
    - Robust standard errors take into account correlation
  - Likelihood-based approaches
    - Fully parameterized model for association
- β^C = average effect for an individual radiologist <u>or</u> average effect controlling for z
- β^M = population-averaged effect

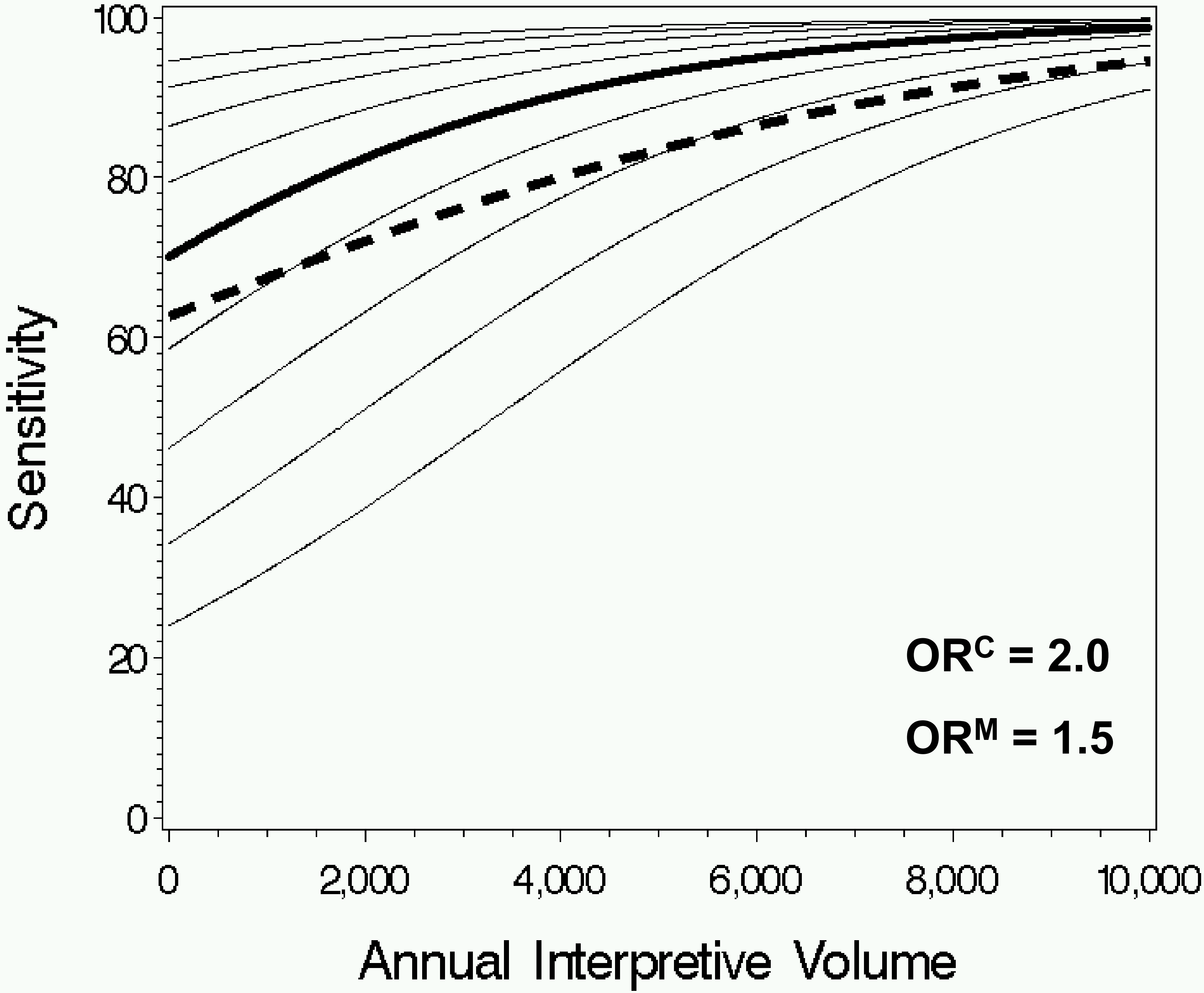
Radiologist-Specific vs. Population-Averaged Effects

- Example: Model for effect of high vs. low interpretive volume on sensitivity
- Radiologist-specific odds ratio
  - Change in odds of a *true positive* assessment if a radiologist was high-volume compared to low-volume
- Population-averaged odds ratio
  - Sensitivity of mammography interpreted by the population of high-volume compared to low-volume radiologists
- Answer different scientific questions but both have meaning (and both may be of interest!)
  - Volume: increase volume vs. stop practicing

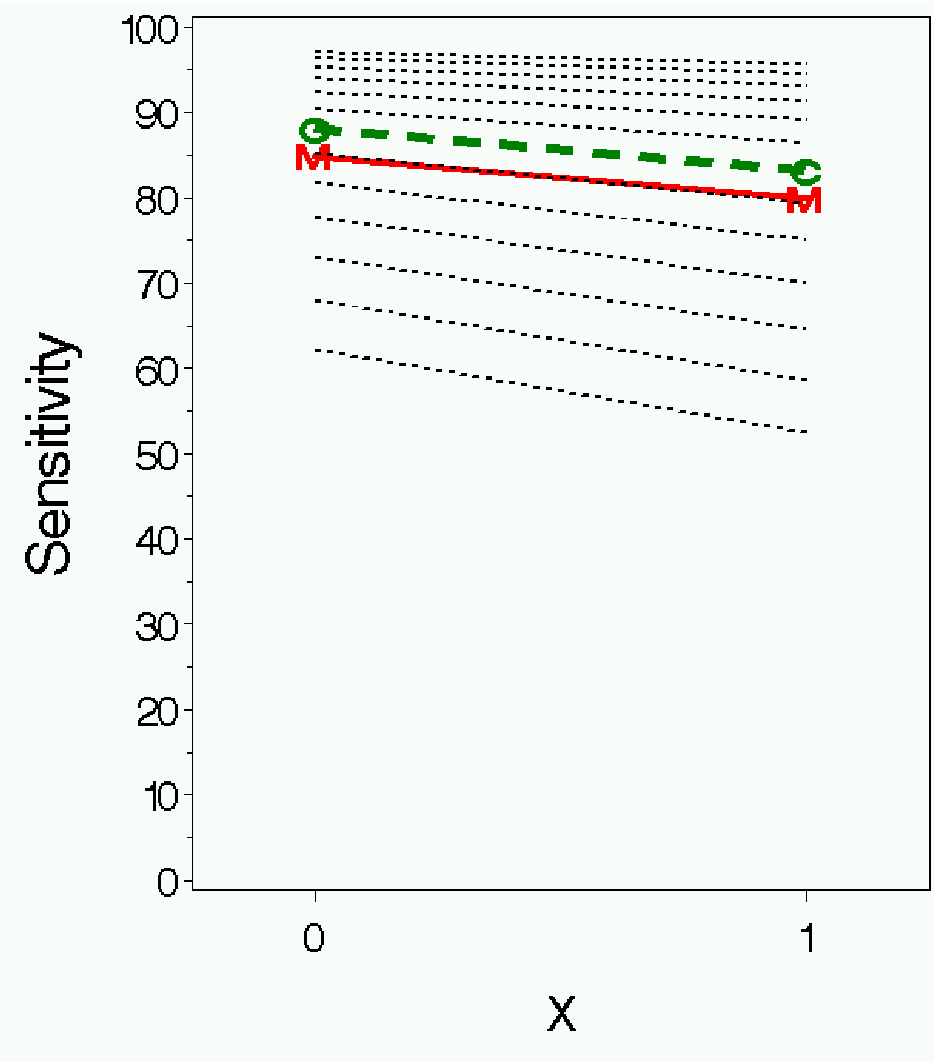
Relationship between Conditional and Marginal Models

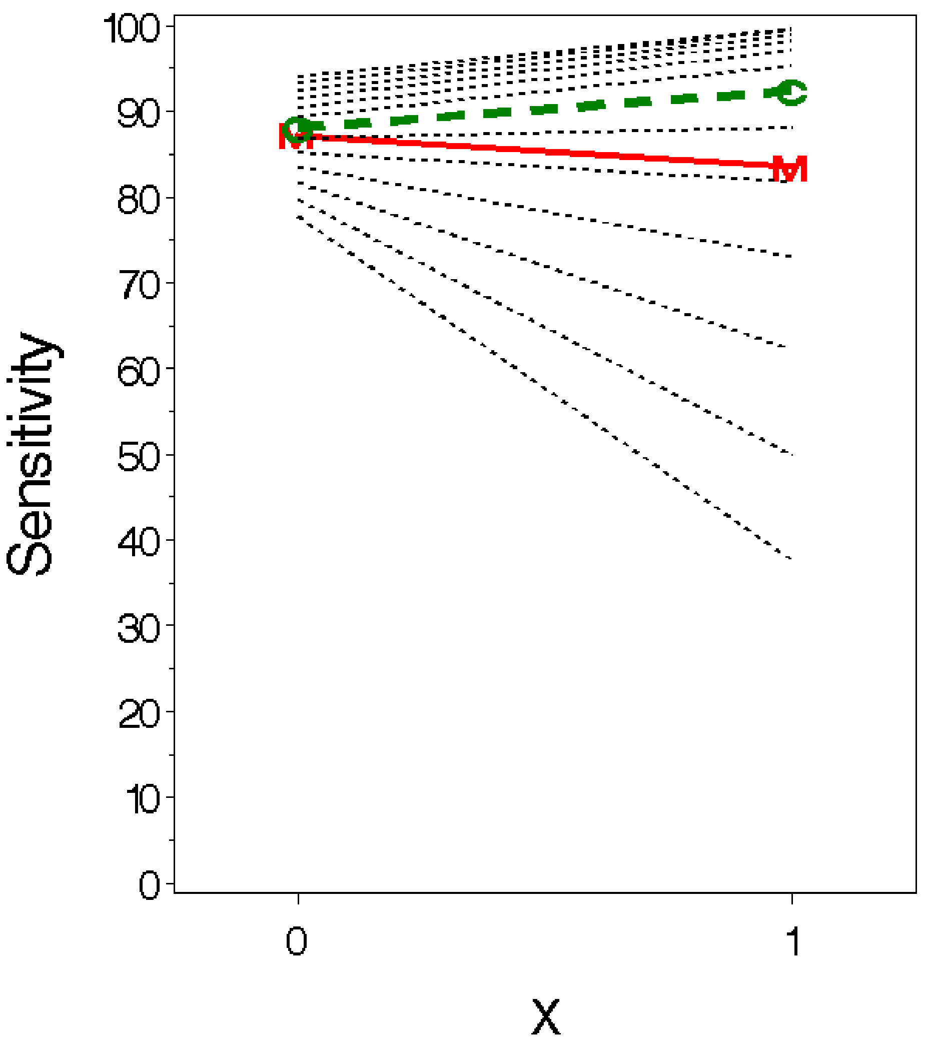
- Constant random effect variance:
  - Marginal OR is attenuated towards 1.0 relative to conditional OR
  - If conditional model is correctly specified, marginal model will have correct type I error rate

- If random effect variance depends on *X*:
  - Relative to conditional OR, marginal OR may be attenuated, amplified, or even in opposite direction!



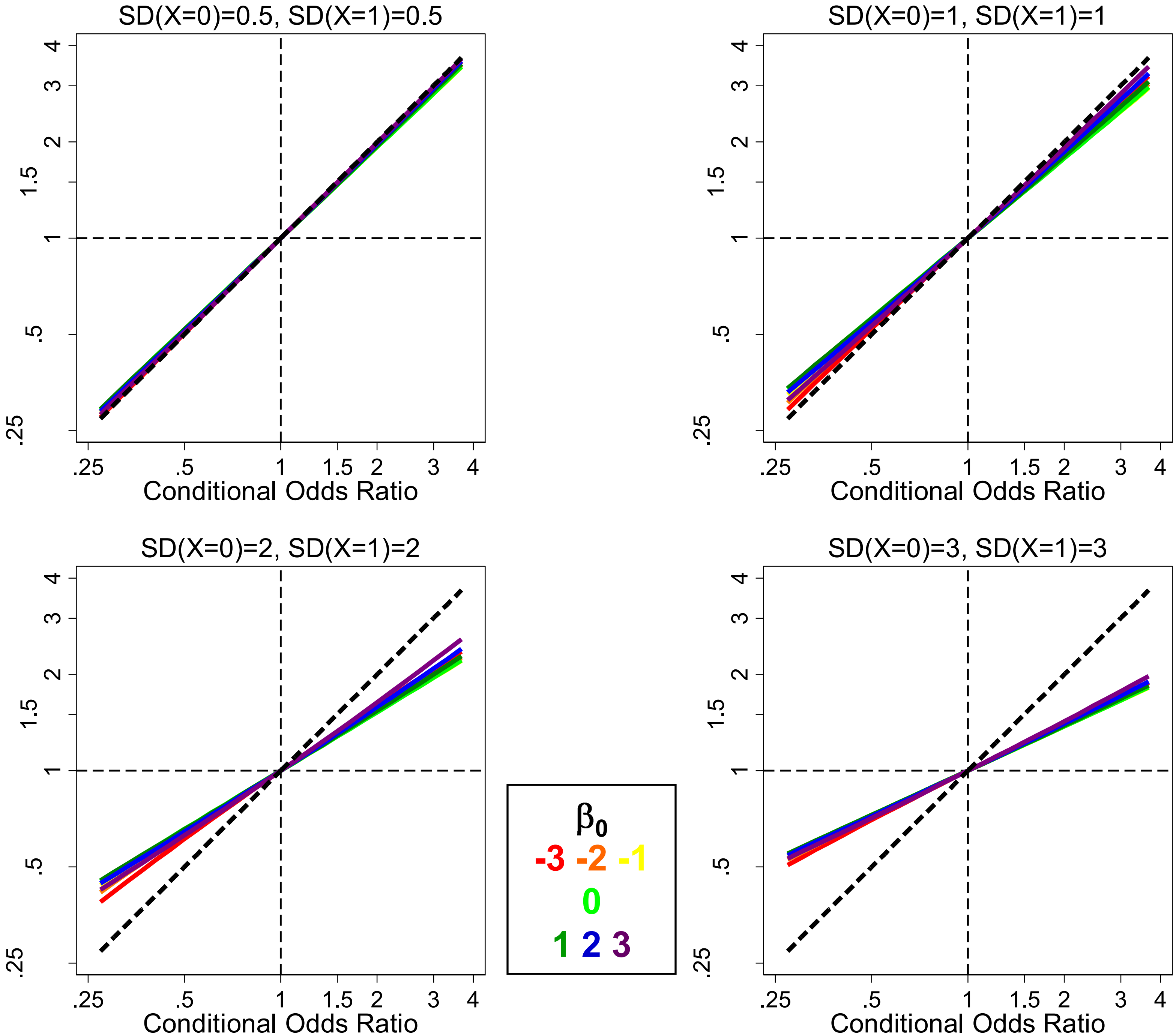
100
80
60
40
20
0
Sensitivity
0
2,000
4,000
6,000
8,000
10,000
Annual Interpretive Volume
OR^C = 2.0
OR^M = 1.5

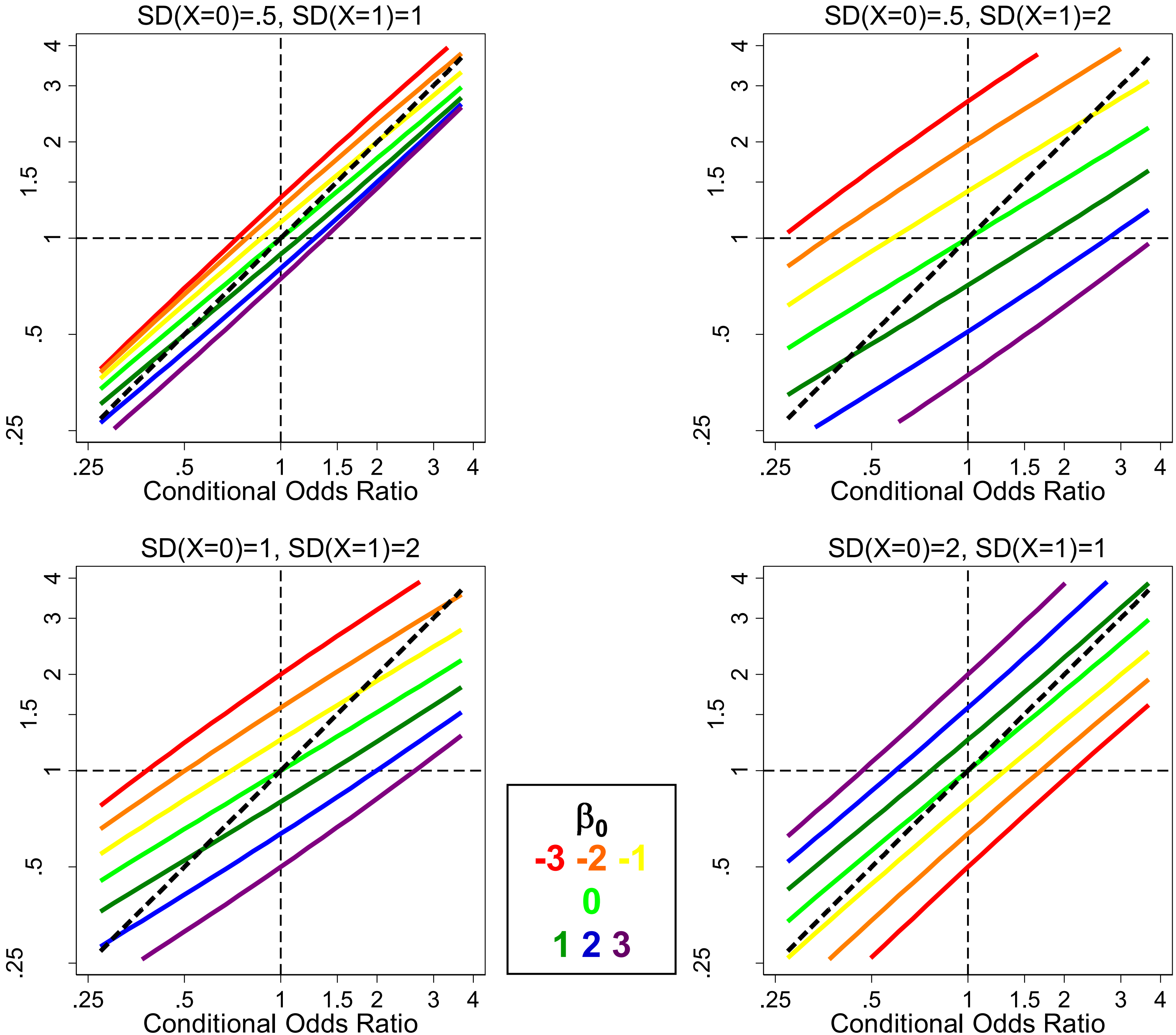




OR^M = 0.71, OR^C = 0.67,
σ_0 = 1, σ_1 = 1
z_i = -1.5σ to 1.5 σ by .25

OR^M = 0.71, OR^C = 1.7
σ_0 = 0.5, σ_1 = 2
Z_i = -1.5σ to 1.5 σ by .25


SD(X=0)=0.5, SD(X=1)=0.5
SD(X=0)=1, SD(X=1)=1
SD(X=0)=2, SD(X=1)=2
SD(X=0)=3, SD(X=1)=3
.25 .5 1 1.5 2 3 4
Conditional Odds Ratio
β_0
-3 -2 -1
0
1 2 3


SD(X=0)=.5, SD(X=1)=1
SD(X=0)=.5, SD(X=1)=2
SD(X=0)=1, SD(X=1)=2
SD(X=0)=2, SD(X=1)=1
.25 .5 1 1.5 2 3 4
Conditional Odds Ratio
β0
-3 -2 -1
0
1 2 3

Summary and Conclusions

- Marginal and conditional models may give different results, because they are modeling different probabilities
  - Marginal effects attenuated if random effect variance constant
  - Marginal effects may be amplified, attenuated, or even in the opposite direction if the random effect variation depends on the covariate of interest
- If interest is in conditional inference
  - Important to take into account differences in RE variation
    - Assuming constant variance can lead to bias
  - Easy to do using standard software
- If interest is in marginal effects
  - May be important to understand mechanism for generating those effects
- Often important to understand reasons for differences in marginal and conditional results