The Proportional Odds Model for Assessing Rater Agreement with Multiple Modalities

Elizabeth Garrett-Mayer, PhD

Assistant Professor
Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University

Learning Objectives

1) the need for standardized classification systems for cancer
2) the definition of a latent variable
3) the framework of the proportional odds model and its application to categorical ratings
4) the interpretation of rater effects and utility of the model for describing reliability of rating system

The Motivation

- Histologic classification systems help in understanding of cancer progression
- In some types of cancers, classification systems have only recently occurred
- Classification system necessary to facilitate research
  - Clinical
  - Pathologic
  - molecular

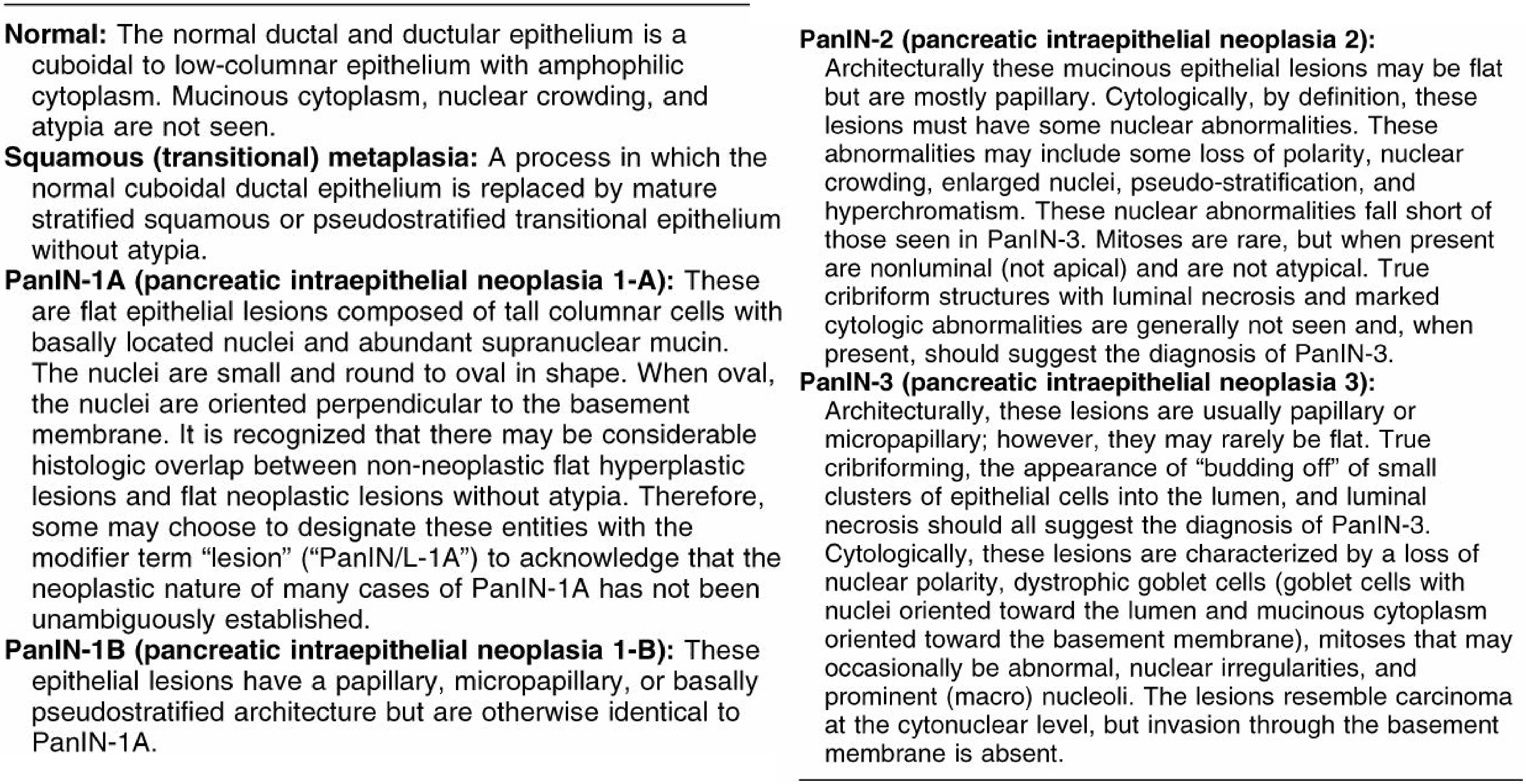
The Motivation

- Proliferative epithelial lesions in the smaller pancreatic cancer ducts *(*Hruban et al. (2001): *American Journal of Surgical Pathology)*
- Complex study design
  - First stage:
    - 8 pathologists
    - 35 microscopic slides
    - Results: Over 70 different diagnostic terms

The Motivation

- Complex study design
  - Second stage: Development of Pan-IN
    - Two classification schemes developed (next slide)
      - Illustrations
      - nomenclature
    - Same 8 pathologists (raters)
    - Same 35 slides
    - Each rater evaluated each slide twice (nomenclature and illustrations)
    - Blinded between ratings

Nomenclature



Normal: The normal ductal and ductular epithelium is a cuboidal to low-columnar epithelium with amphophilic cytoplasm. Mucinous cytoplasm, nuclear crowding, and atypia are not seen.

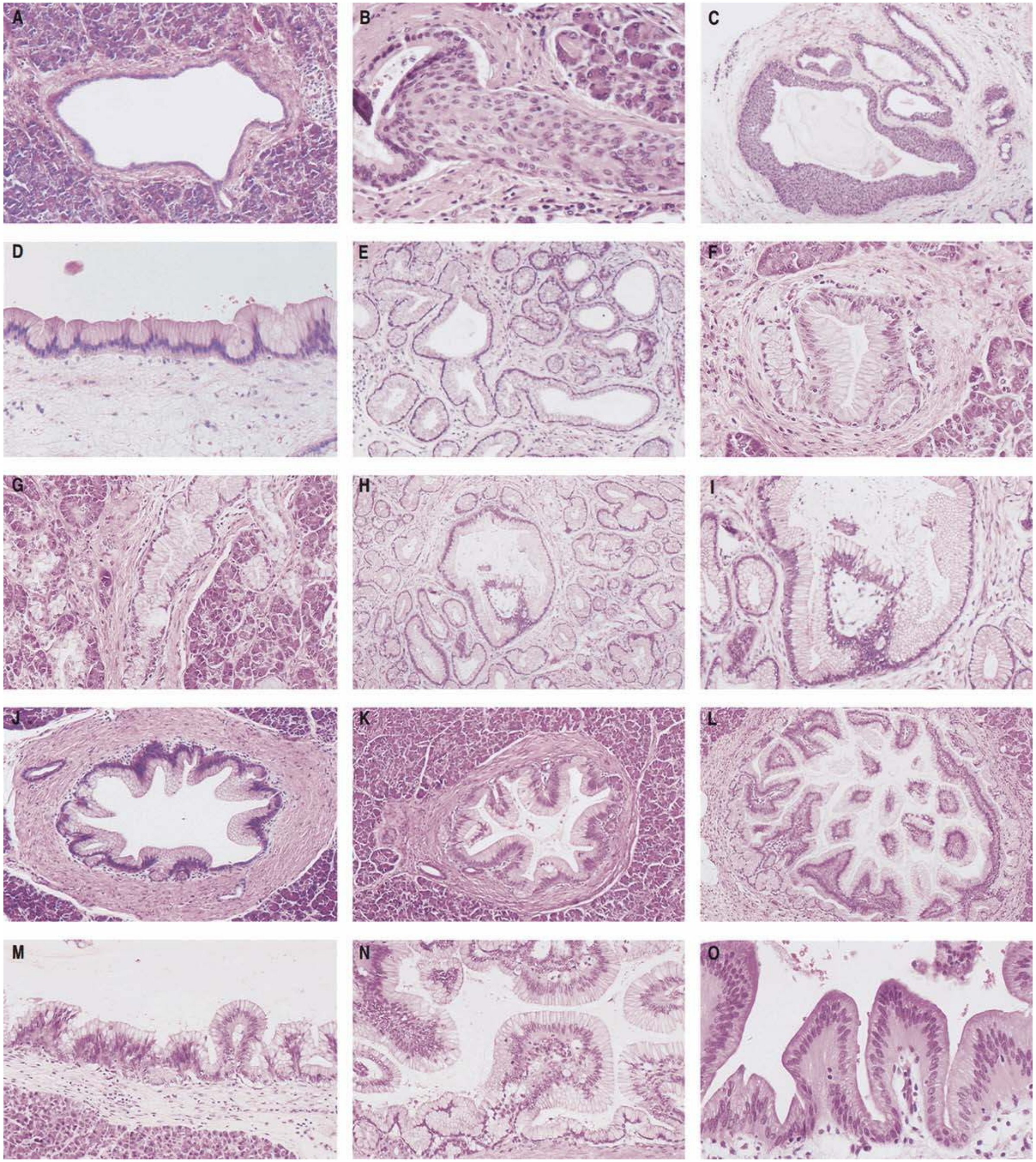
Squamous (transitional) metaplasia: A process in which the normal cuboidal ductal epithelium is replaced by mature stratified squamous or pseudostratified transitional epithelium without atypia.

PanIN-1A (pancreatic intraepithelial neoplasia 1-A): These are flat epithelial lesions composed of tall columnar cells with basally located nuclei and abundant supranuclear mucin. The nuclei are small and round to oval in shape. When oval, the nuclei are oriented perpendicular to the basement membrane. It is recognized that there may be considerable histologic overlap between non-neoplastic flat hyperplastic lesions and flat neoplastic lesions without atypia. Therefore, some may choose to designate these entities with the modifier term “lesion” (“PanIN/L-1A”) to acknowledge that the neoplastic nature of many cases of PanIN-1A has not been unambiguously established.

PanIN-1B (pancreatic intraepithelial neoplasia 1-B): These epithelial lesions have a papillary, micropapillary, or basally pseudostratified architecture but are otherwise identical to PanIN-1A.

PanIN-2 (pancreatic intraepithelial neoplasia 2): Architecturally these mucinous epithelial lesions may be flat but are mostly papillary. Cytologically, by definition, these lesions must have some nuclear abnormalities. These abnormalities may include some loss of polarity, nuclear crowding, enlarged nuclei, pseudo-stratification, and hyperchromatism. These nuclear abnormalities fall short of those seen in PanIN-3. Mitoses are rare, but when present are nonluminal (not apical) and are not atypical. True cribriform structures with luminal necrosis and marked cytologic abnormalities are generally not seen and, when present, should suggest the diagnosis of PanIN-3.

PanIN-3 (pancreatic intraepithelial neoplasia 3): Architecturally, these lesions are usually papillary or micropapillary; however, they may rarely be flat. True cribriforming, the appearance of “budding off” of small clusters of epithelial cells into the lumen, and luminal necrosis should all suggest the diagnosis of PanIN-3. Cytologically, these lesions are characterized by a loss of nuclear polarity, dystrophic goblet cells (goblet cells with nuclei oriented toward the lumen and mucinous cytoplasm oriented toward the basement membrane), mitoses that may occasionally be abnormal, nuclear irregularities, and prominent (macro) nucleoli. The lesions resemble carcinoma at the cytonuclear level, but invasion through the basement membrane is absent.

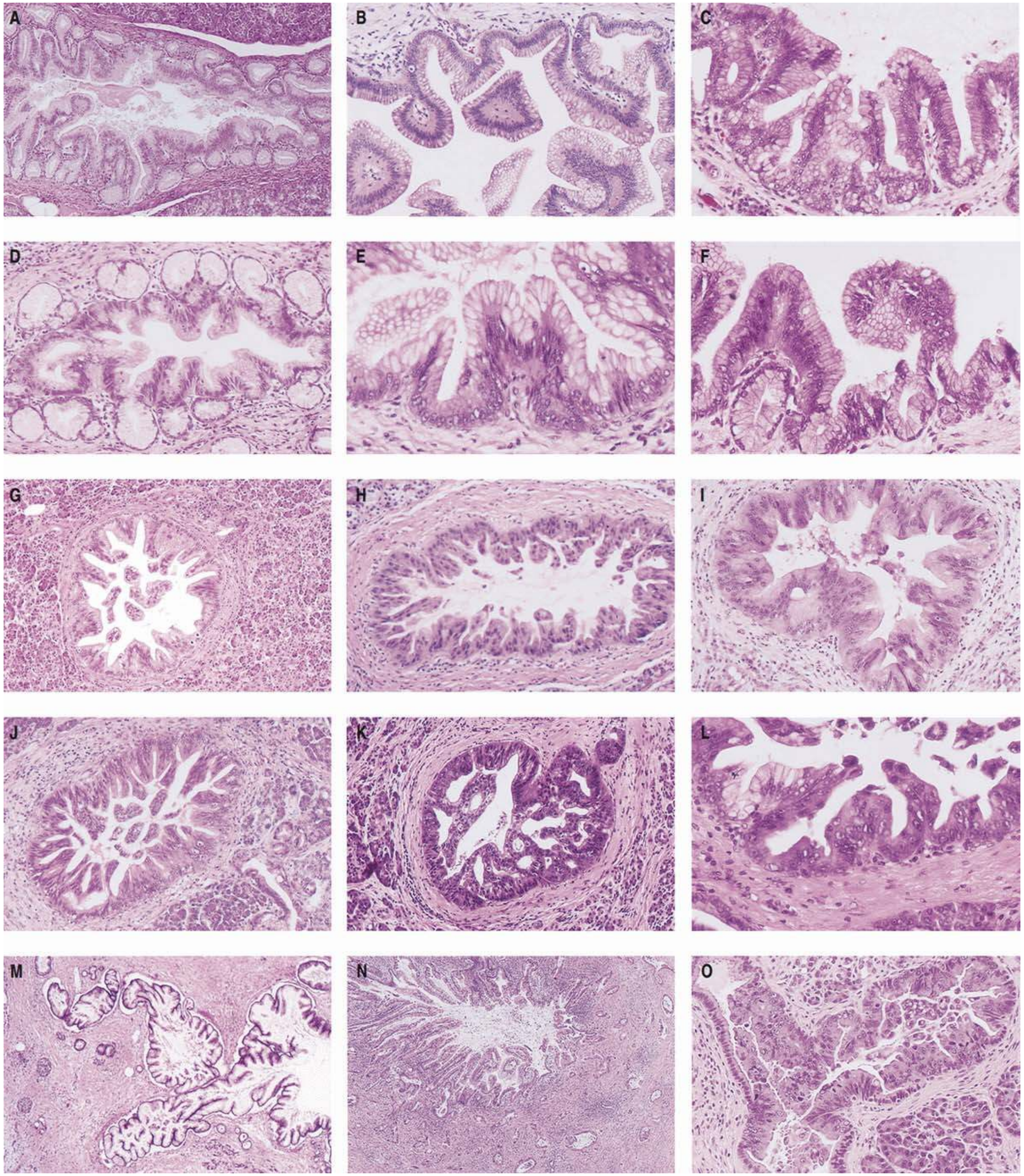


(A) Normal duct.

(B and C) Squamous metaplasia.

(D–I) PanIN-1A.

(J–O) PanIN-1B.



(A–F) PanIN-2.

(G–L) PanIN-3.

(M) Intraductal papillary mucinous neoplasm.

(N–O) Invasive adenocarcinoma secondarily involving a duct (cancerization of the ducts).

The Questions

- Do either of these methods work?
- Is one significantly better than the other?
- Are there discrepancies seen at one end of the scale or the other?
- What is the variation across raters?

Statistically interesting

- No gold standard to which to compare ratings
- Variability is due to both rater and method of rating
- Scale is ordinal

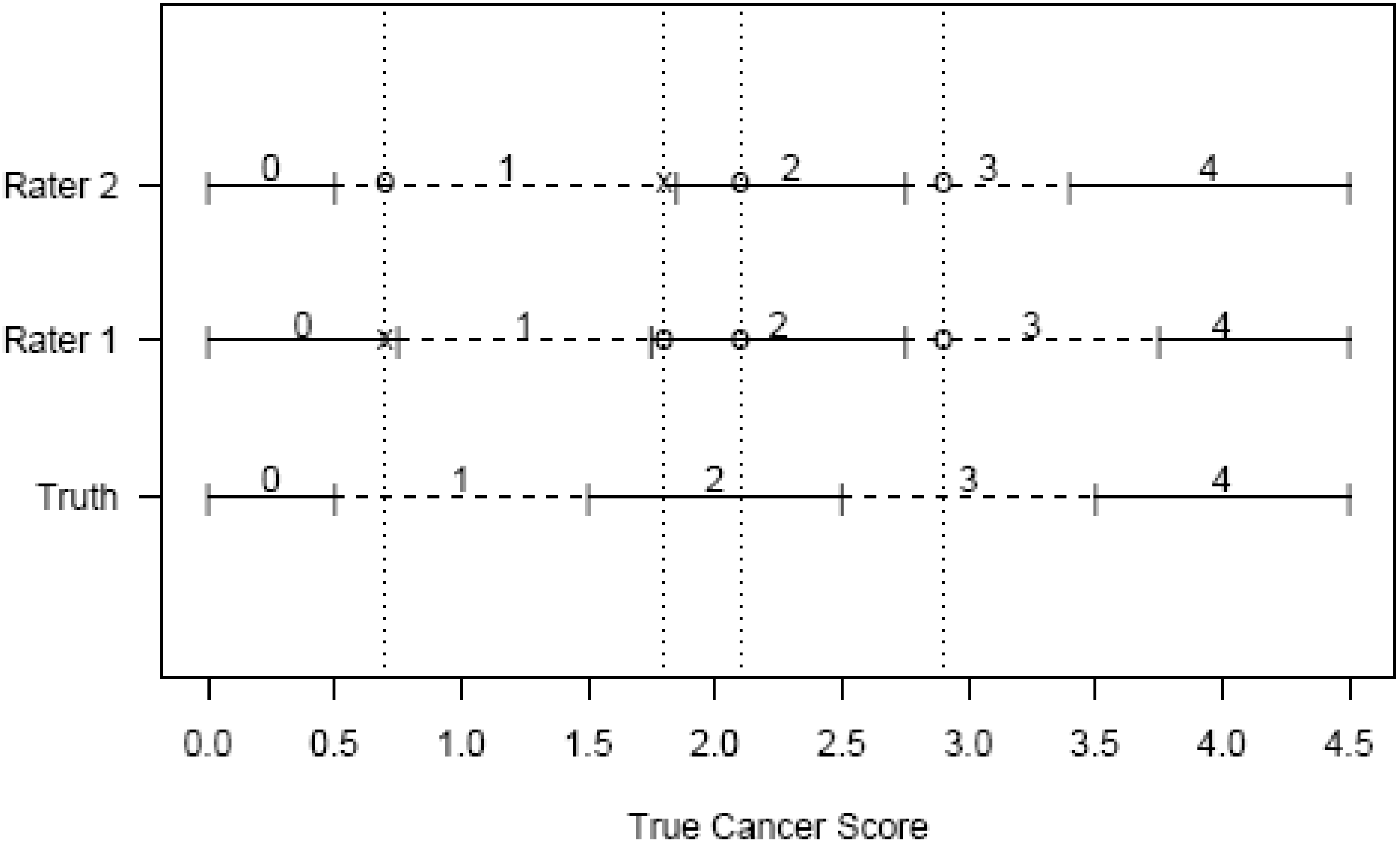
Exploit as a latent variable problem

- Is cancer progression really “categorical”?
- **More likely continuous**
- Categories represent imposition of “thresholds”
- Questions rephrased:
  - Do raters have different thresholds?
  - What is the variance of the thresholds?
  - Do thresholds and their variances vary across methods?

Hold on a second….

- What is a latent variable?
- A variable that cannot be directly measured
- Latent variable ≈ construct ≈ factor
- Examples:
  - Quality of life
  - Pain
  - Schizophrenia
  - Intelligence
  - Diet
  - Customer satisfaction
  - Knowledge
  - ***Cancer progression***

Example: 0 to 4 rating scale



Latent Variable Approach

- Treat underlying cancer progression as continuous
- Assume that ratings are ordinal and that raters “round” their ratings to nearest integer
- Estimate the thresholds of raters

Other approaches?

- Kappa?
  - Measures agreement
  - Can handle ordinal or categorical ratings
  - Can handle multiple raters
  - Or, can handle multiple modalities
  - **Cannot handle both multiple raters and modalities**
  - Can use kappa to
    - Estimate separate agreements for the two approaches
    - Estimate overall agreement, ignoring method of rating
  - Not ideal:
    - Does not allow us to compare methods directly
    - Does not allow us to assess differences due to rater effects versus variance.
    - Does not acknowledge underlying continuous variable
    - “black box”

Other approaches

- Tanner and Young (1985), Becker and Agresti (1992), Perkins and Becker (2002)
  - log-linear model
  - Partition observed data into agreement and chance components
  - Model pairwise agreements
    - Do pairs of raters have same agreement structure?
    - Do raters have same aggregate level of agreement with other raters?
  - Problems:
    - not extended to deal with multiple modalities
    - treats data as nominal
    - Does not acknowledge latent variable problem

Other approaches

- Agresti (1988)
  - Extends Tanner and Young (1985) treating data as ordinal
  - Latent **class** variable approach
  - Problem: does not acknowledge continuity
- Uebersax and Grove (1993)
  - Latent trait mixture model
  - Assumes binary classification is goal
  - Problems:
    - Imposes strong normality assumptions
    - Does not acknowledge continuity

Other approaches

- Johnson (1996)
  - Bayesian analysis
  - Underlying continuous variable
  - Assumes normality of latent trait
- Our model is similar to Johnson
  - Incorporate additional effects for modality
  - More flexible about distribution of latent trait
  - Focus on scale of latent variable

The proportional odds model

- McCullagh (1980)
- AKA “ordinal logistic regression”
- Developed to deal with thresholded data
- Goal was to estimate the association between some risk factors and an ordinal outcome of interest
- Assumes that there is a ‘proportional’ increase in risk.

The proportional odds model (POM)

$$\log\left(\frac{P(Y_i > k)}{P(Y_i \leq k)}\right) = \beta x_i + \alpha_k \quad ; \quad k = 1, \ldots, K - 1$$

- Example:
  - Outcome: "how is your health?" (5=excellent, 4=very good, 3=good, 2=fair, 1=poor)
  - Predictor: diabetes (1=yes, 0=no)
- β: log odds ratio of higher rating for diabetics versus non-diabetics (2,3,4,5 vs. 1; 3,4,5 vs. 1,2 ; 4,5 vs. 1,2,3; 5 vs. 1,2,3,4)
- α_k: nuisance parameter. calibration factor.

POM for rater agreement in Pan-IN

- β_i: latent variable – represents true cancer progression for patient i
- **$\boldsymbol{\alpha}_{\mathbf{k}}$: threshold parameters**
- Three categories:
  - 1A and 1B lumped together
  - none were "normal"
  - also had an "other" category that is treated as missing (not ordinal)
- Two thresholds: between 1 and 2, and between 2 and 3.
- Model must include
  - rater effects (8 raters)
  - method effect (illustration versus nomenclature)
- Additional complication: latent variable?!

Latent variable POM

$$\log\left(\frac{P(Y_{ijm} > k)}{P(Y_{ijm} \le k)}\right) = \beta_i + \alpha_{jk} + m\delta_{jk} \quad ; \quad k = 1,2$$

- Y_{ijm} = rating of slide *i* by rater *j* by method *m*
- *i=1,...,35* slides
- *j=1,...8* raters
- *m=1* if illustrations, *0* if nomenclature
- **$\boldsymbol{\beta_i}$ = true cancer score of patient *i***

Latent variable POM

$$\log\left(\frac{P(Y_{ijm} > k)}{P(Y_{ijm} \le k)}\right) = \beta_i + \alpha_{jk} + m\delta_{jk} \quad ; \quad k = 1,2$$

α_{j1}: lower nomenclature effect for rater j

α_{j2}: upper nomenclature effect for rater j

$\alpha_{j1} + \delta_{j1}$: lower illustrations effect for rater j

$\alpha_{j2} + \delta_{j2}$: upper illustrations effect for rater j

δ_{j1}: difference between lower effects for rater j

δ_{j2}: difference between upper effects for rater j

Model Assumptions and Estimation

- Modeling not standard due to
  - latent variable
  - hierarchical assumptions
    - Do not estimate fixed rater effects
    - Assume rater effects come from common distribution
    - “random” effects
- MCMC estimation procedure
- WinBugs software
- Regression parameter assumptions standard
- Latent variable modeling
  - normal or uniform
  - post hoc rescaling

Model Assumptions

$$\alpha_{j1} \sim N(0, \sigma^2_{\alpha 1}) I(-\infty, \alpha_{j2})$$

$$\alpha_{j2} \sim N(\alpha_2, \sigma^2_{\alpha 2}) I(\alpha_{j1}, \infty)$$

$$\delta_{jk} \sim N(\delta_k, \sigma^2_{\delta k})$$

$$\beta_i \sim N(\mu, \tau^2) \text{ or } \beta_i \sim U[a, b]$$

$$\alpha_2 \sim N(0,100)$$

$$1 / \sigma^2_{\alpha k} \sim Gamma(0.01,0.01)$$

$$1 / \sigma^2_{\delta k} \sim Gamma(0.01,0.01)$$

Interpretation of β_i

$$\log\left(\frac{P(Y_{ij0} > 1)}{P(Y_{ij0} \le 1)}\right) = \beta_i + \alpha_{j1} \quad \text{and} \quad E(\alpha_{j1}) = 0$$

$$\Rightarrow\ E\left(\log\left(\frac{P(Y_{ij0} > 1)}{P(Y_{ij0} \le 1)}\right)\right) = \beta_i$$

$$\Rightarrow\ \hat{P}(Y_{i\cdot 0} > 1) = \frac{e^{\hat{\beta}_i}}{1 + e^{\hat{\beta}_i}}$$

***β_i* represents (on the logit scale) the probability that slide *i* is rated a 2 or a 3 by nomenclature.**

Example: $\beta_i = 0$ means that the estimated probability that slide *i* is a 2 or 3 is 0.50

Estimating Thresholds

- Not so much interested in overall thresholds.
- How do raters vary?
- For each j, solve the following two equations *separately* for *β* for each rater (*j*) and method (*m*):

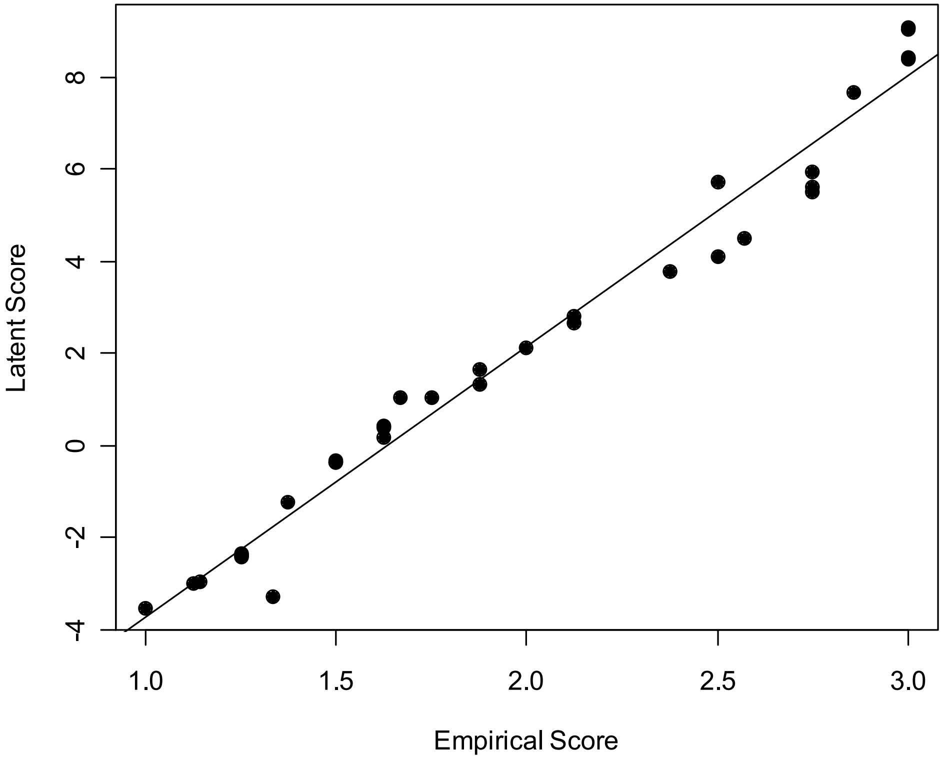
$$2P(Y_{ijm} > 1) = 1 + P(Y_{ijm} > 2)$$ ⇦ 1 v 2 threshold

$$2P(Y_{ijm} > 2) = P(Y_{ijm} > 1)$$ ⇦ 2 v 3 threshold

- Solution to first equation: for what β the rater would be equally likely to rate a 1 or a 2.
- Solution to second equation: for what β the rater would be equally likely to rate a 2 or a 3.
- No closed form solution

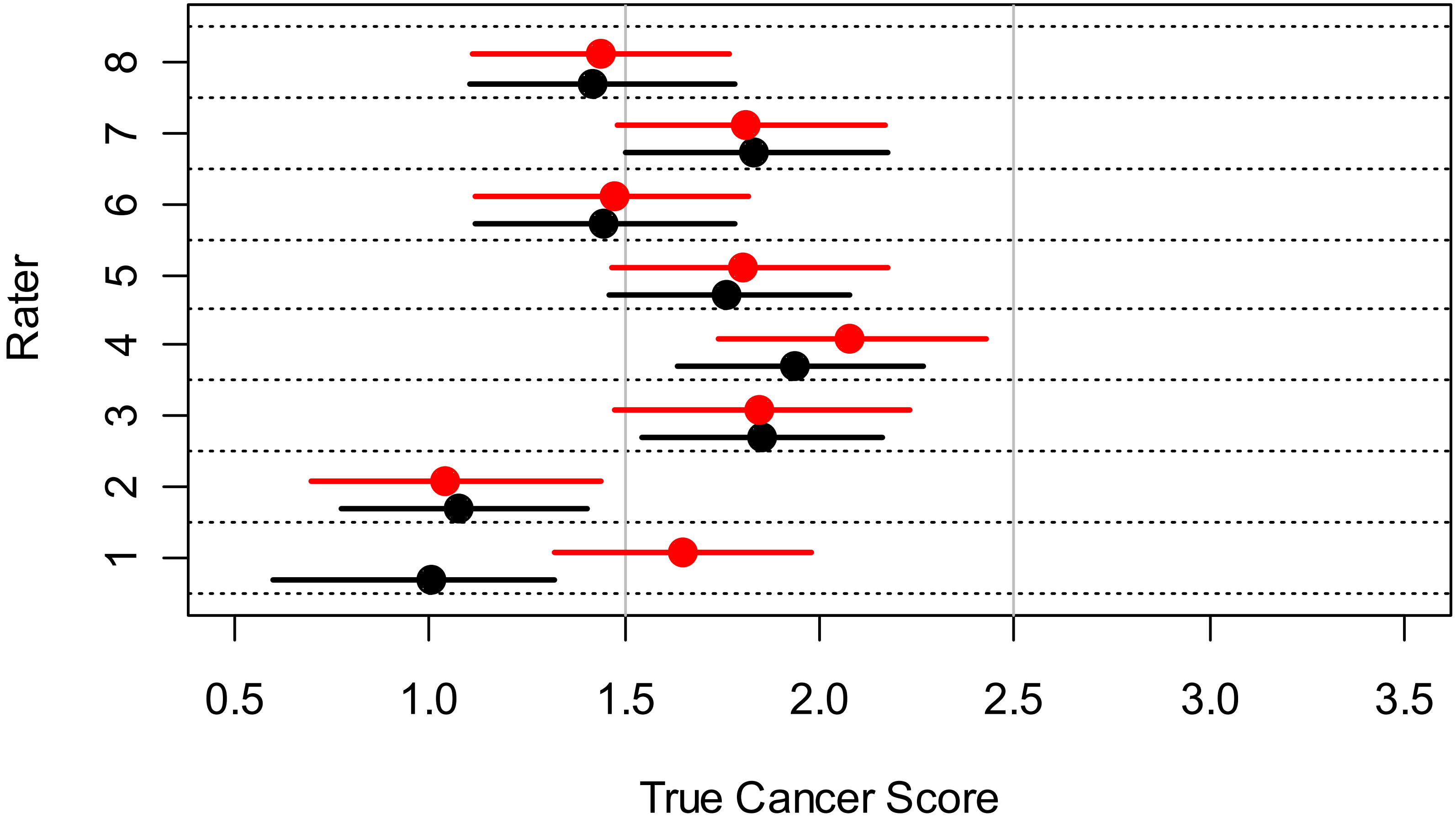
Rescaling

- Latent variable (β) is not in scale of grading system
- Most interpretable if thresholds are in terms of original units
- Can we recalibrate?
- Interpolation
  - Assume linear relationship between β and empirical means
  - Interpolate any other quantities on β-scale of interest
- Other approaches

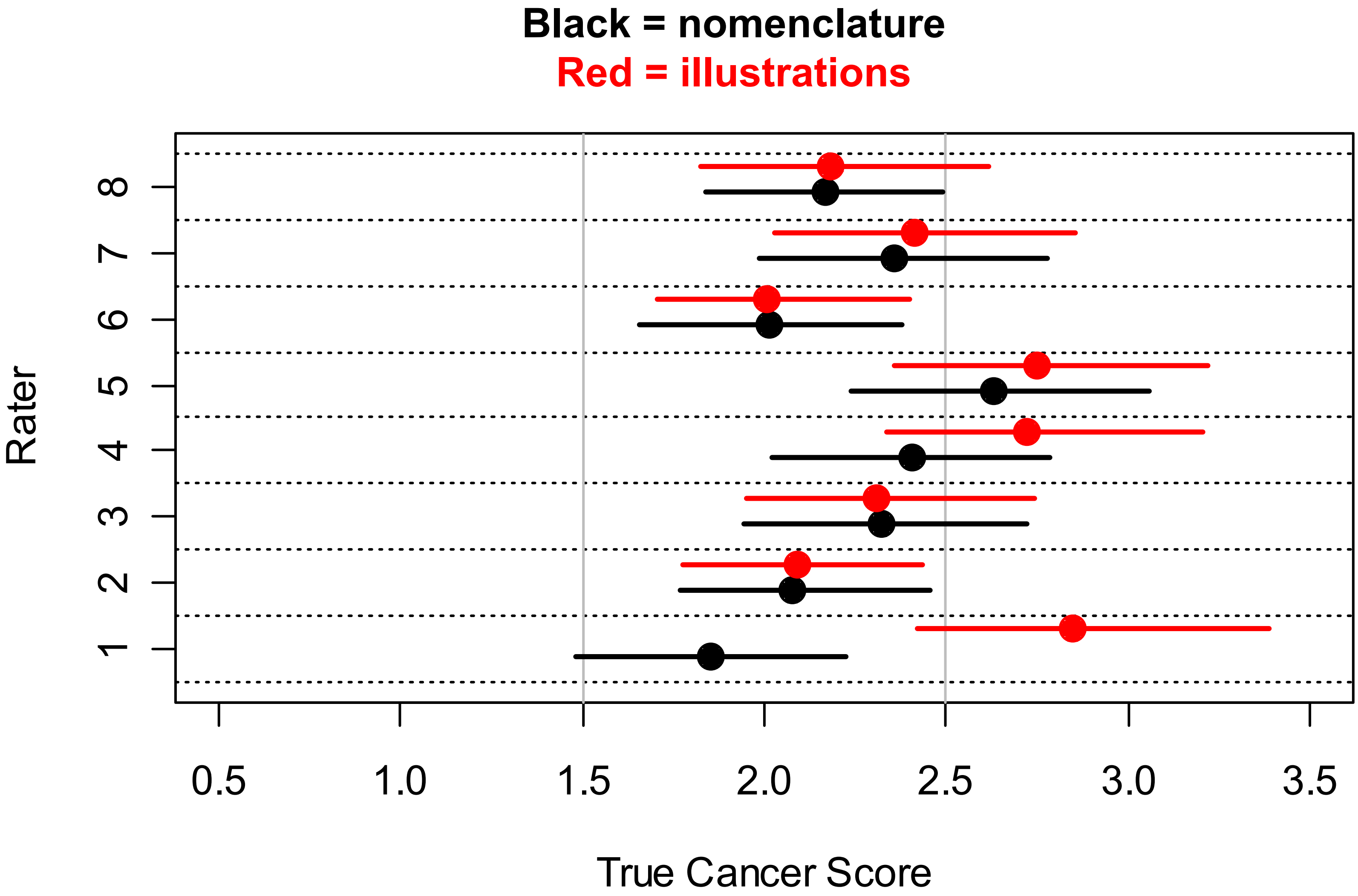


Results: 1 v 2 thresholds

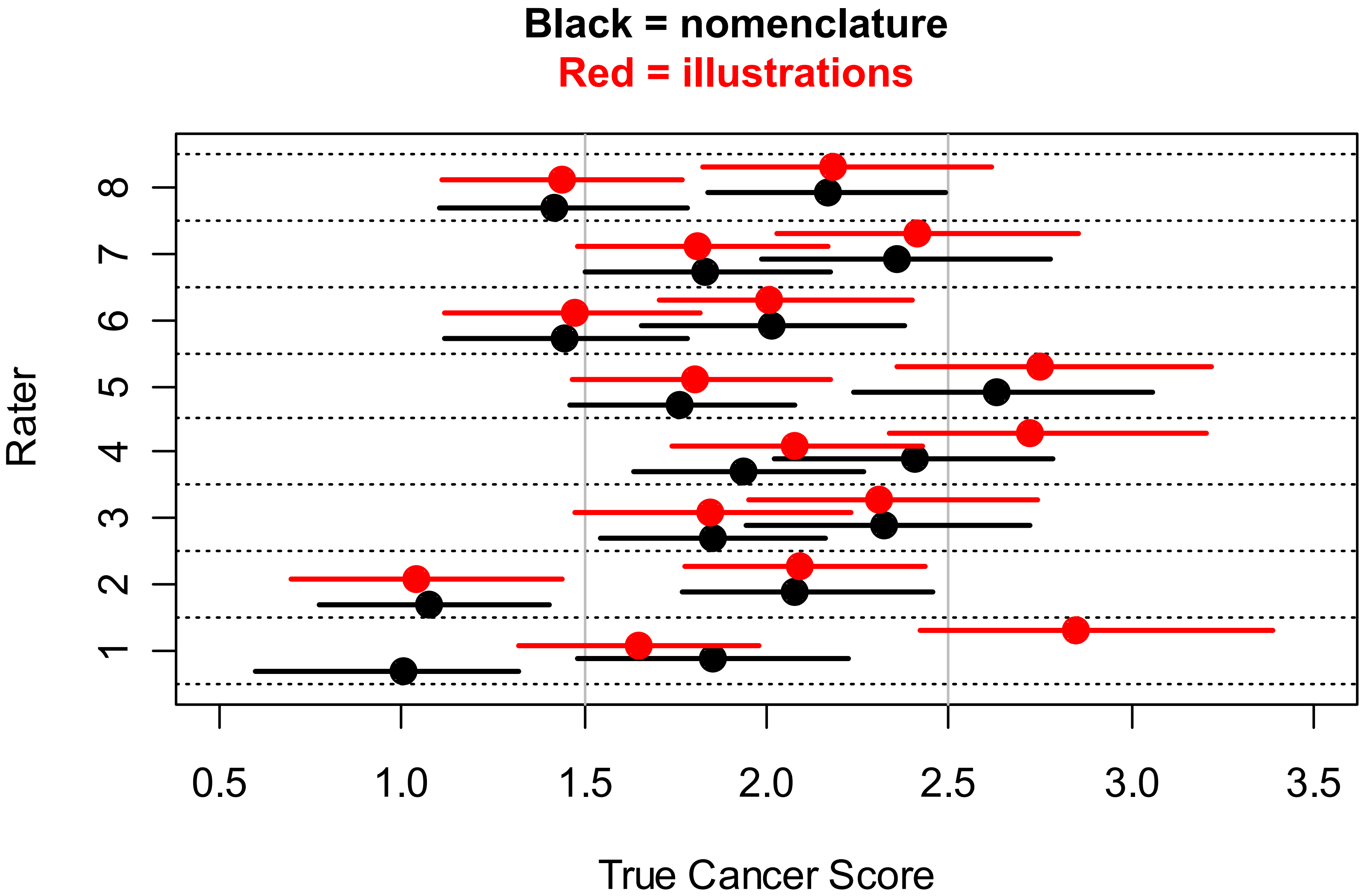
Black = nomenclature
Red = illustrations



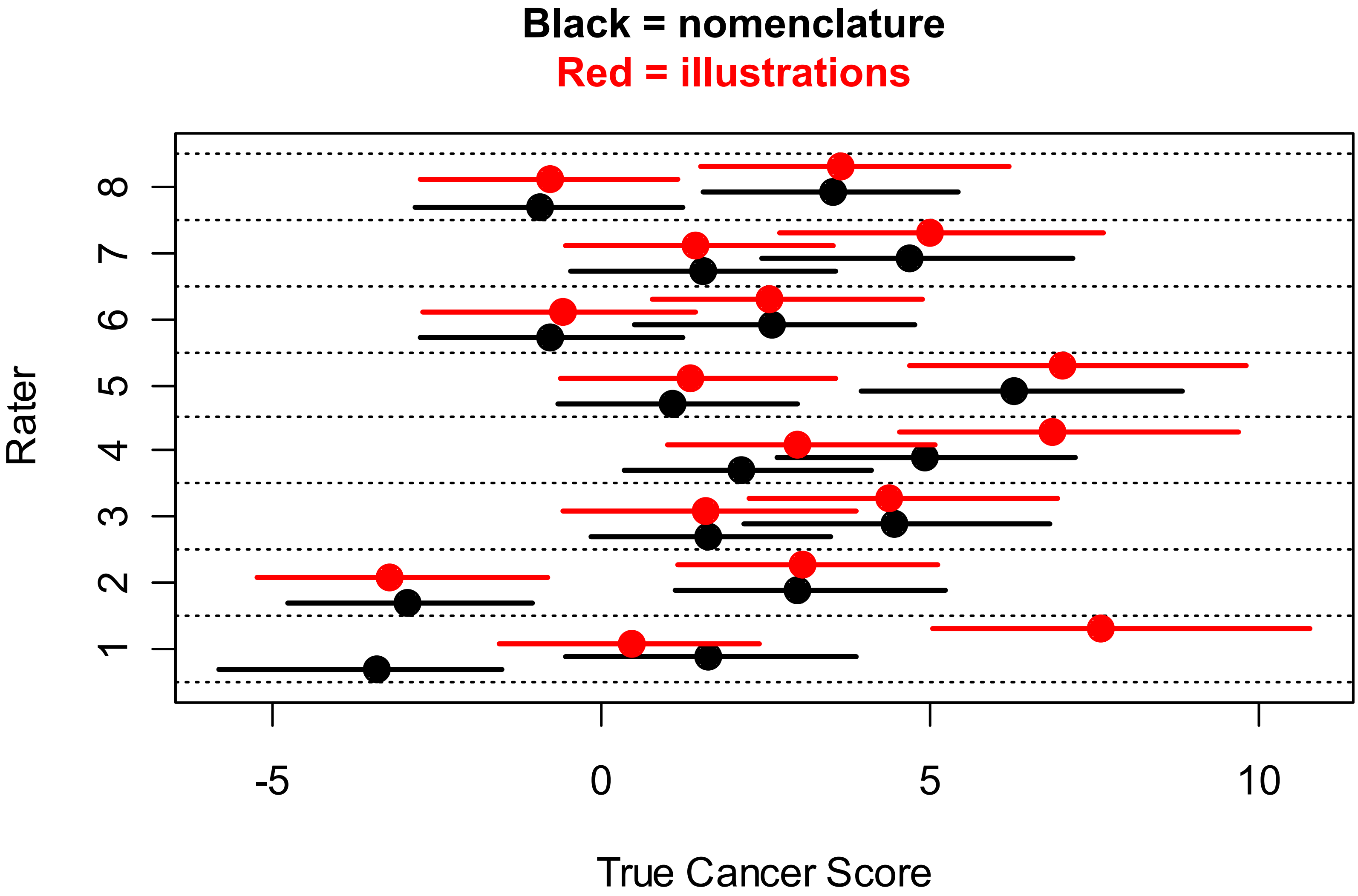
Results: 2 v 3 thresholds



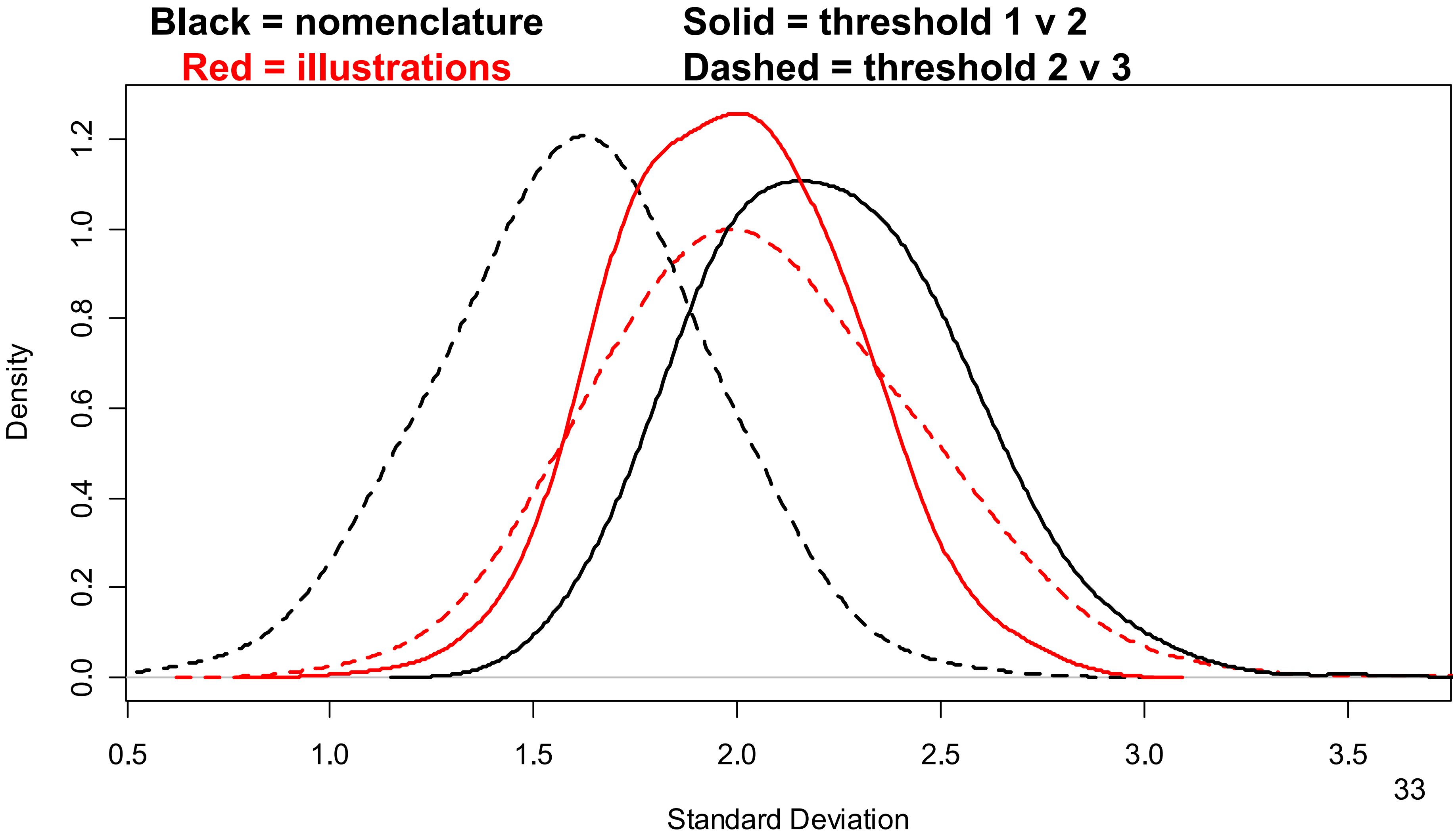
Results: All thresholds



Results: All thresholds



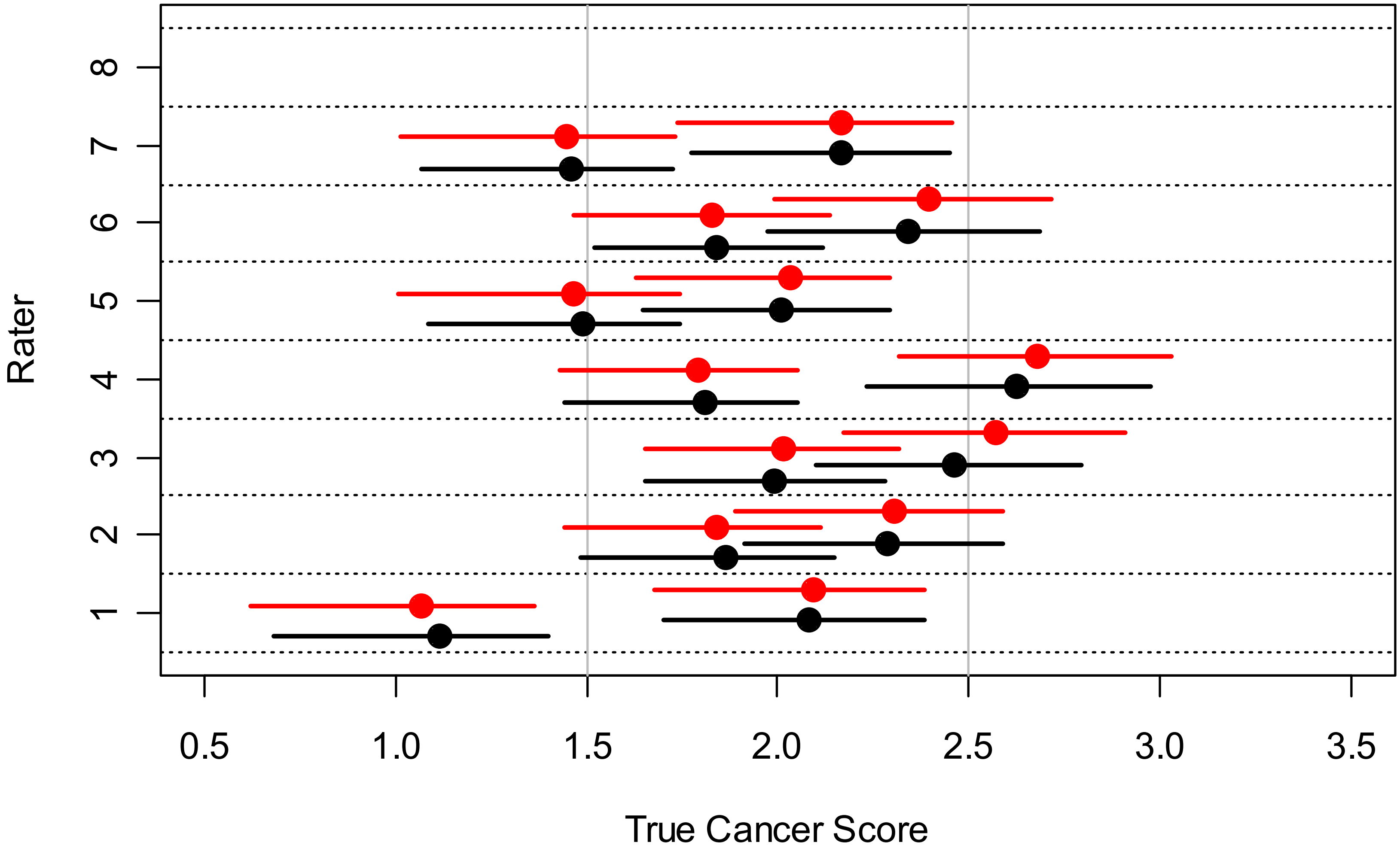
Standard Deviation of Thresholds



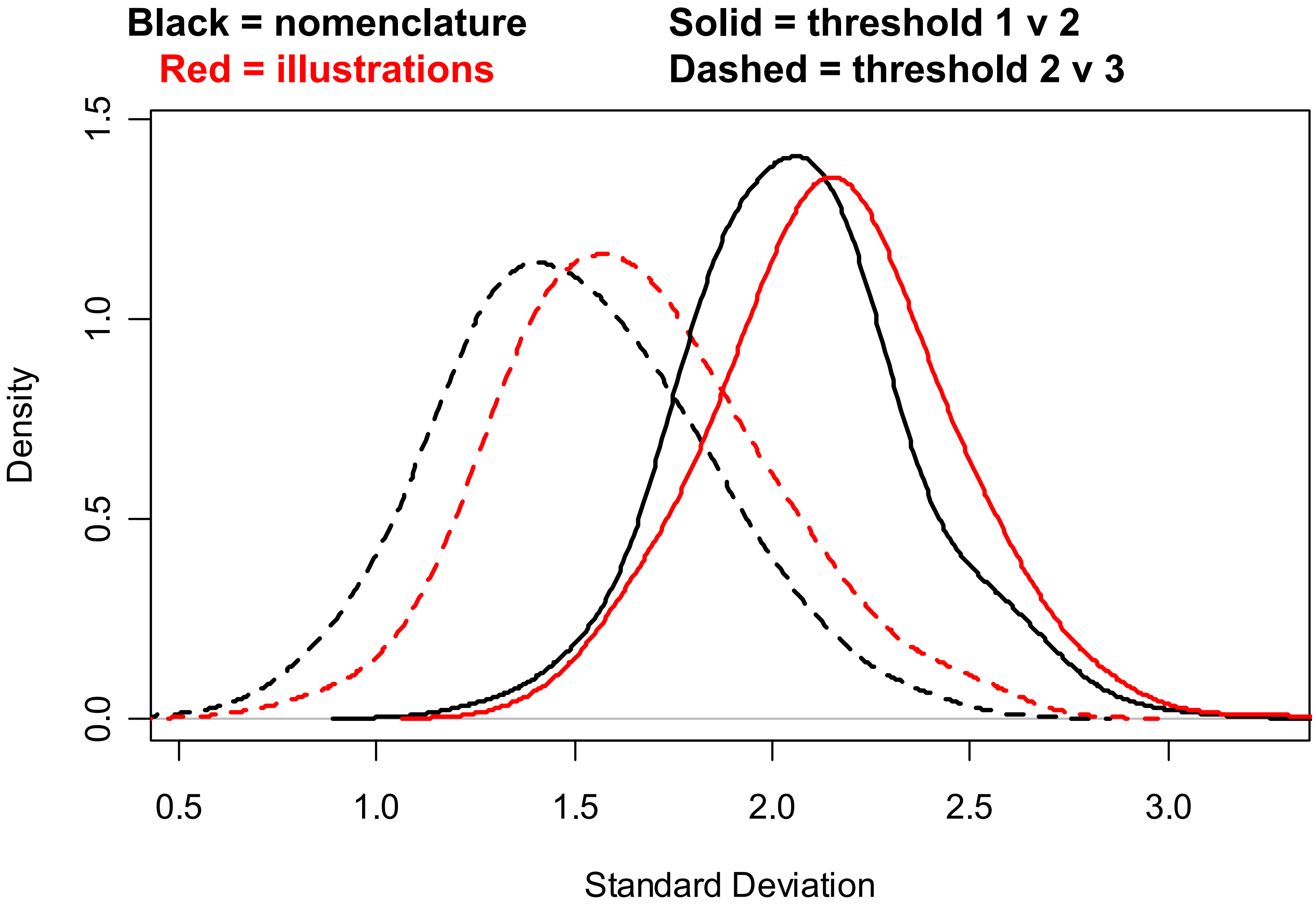
Anything look odd?

- Rater 1 had misunderstood the rating system for nomenclature
- Instead of coding PanIN 1A and 1B as “1”, he coded them as “1” and “2”, and PanIN 2 as “3”, etc.
- Effect: his nomenclature ratings tend to be biased upwards

Revised Results



Revised Results



Revised Results

- Standard Deviations across two methods look about the same as each other now
    - illustrations slightly higher
    - not ‘interestingly’ different
- Otherwise, things are similar
    - Rater thresholds were not sensitive to exclusion
    - True cancer scores were not sensitive

Sensitivity Analysis

- Assumed distribution of β
- Uniform versus Normal?
- Inferences appear to be the same

The Questions

- Do either of these methods work?
  - sort of: there is considerable variability across raters
- Is one significantly better than the other?
  - no, but nomenclature looks a little bit better
- Are there discrepancies seen at one end of the scale or the other?
  - yes, raters have better agreement at high end of scale versus low end
- What is the variation in rater thresholds?
  - considerable, with a range of thresholds of approximately 1

Interpretation

- Both methods appear to perform approximately equally well
- Bias?
  - Both methods taught simultaneously.
  - Better design:
    - separate raters for each method
    - Teach rater only one method
- After removing rater 1, nomenclature looks slightly better
- Large variability in rater thresholds
- Greater variability in 1 vs. 2 than in 2 vs. 3
- Suggests room for improvement

Utility of Model

- Not only when multiple raters, multiple modalities
- If only nomenclature OR illustrations still would have been useful
- Some key ideas:
  - Comparisons of rater thresholds
  - Assessment of variability of thresholds
  - Rescaling to original metric for interpretability
- Useful for any ordinal rating system where underlying variable can be considered continuous