Problems with Surgical Report Cards

Frank Harrell Jr, PhD

Division of Biostatistics and Epidemiology

Dept. of Health Evaluation Sciences

University of Virginia School of Medicine

fharrell@virginia.edu

UVa Surgery Grand Rounds

25 Mar 1998

Outline

- Overview of how statistical models are used to examine variations in patient outcomes
- Which patient descriptors should be used, which should be avoided
- Methods for developing risk models
- Measuring a model’s accuracy

Outline, Continued

- Regression to the mean, or why to disbelieve some of the data
- Why risk-adjusted mortality estimates cannot be used to reliably rank surgeons
- Best measures to use in a provider report card
- Why it’s not a good idea to refuse to operate on high-risk patients

Overview of Statistical Models

- Choose a set of relevant patient measurements
- Relate these and surgeon effects to individual patient outcomes (30 day vital status)
- Subtract the effects of patient measurements on outcome
- What's left (imprecisely) measures effects of surgeons (level playing field)

Which Patient Measurements?

- Continuous measurements (age, heart pumping efficiency, weight, height) have many advantages
- Subjective assessments should be avoided if possible
- Classifications such as urgent or emergent surgery vary too much across surgeons
- Surgeons are tempted to engage in “coding creep”

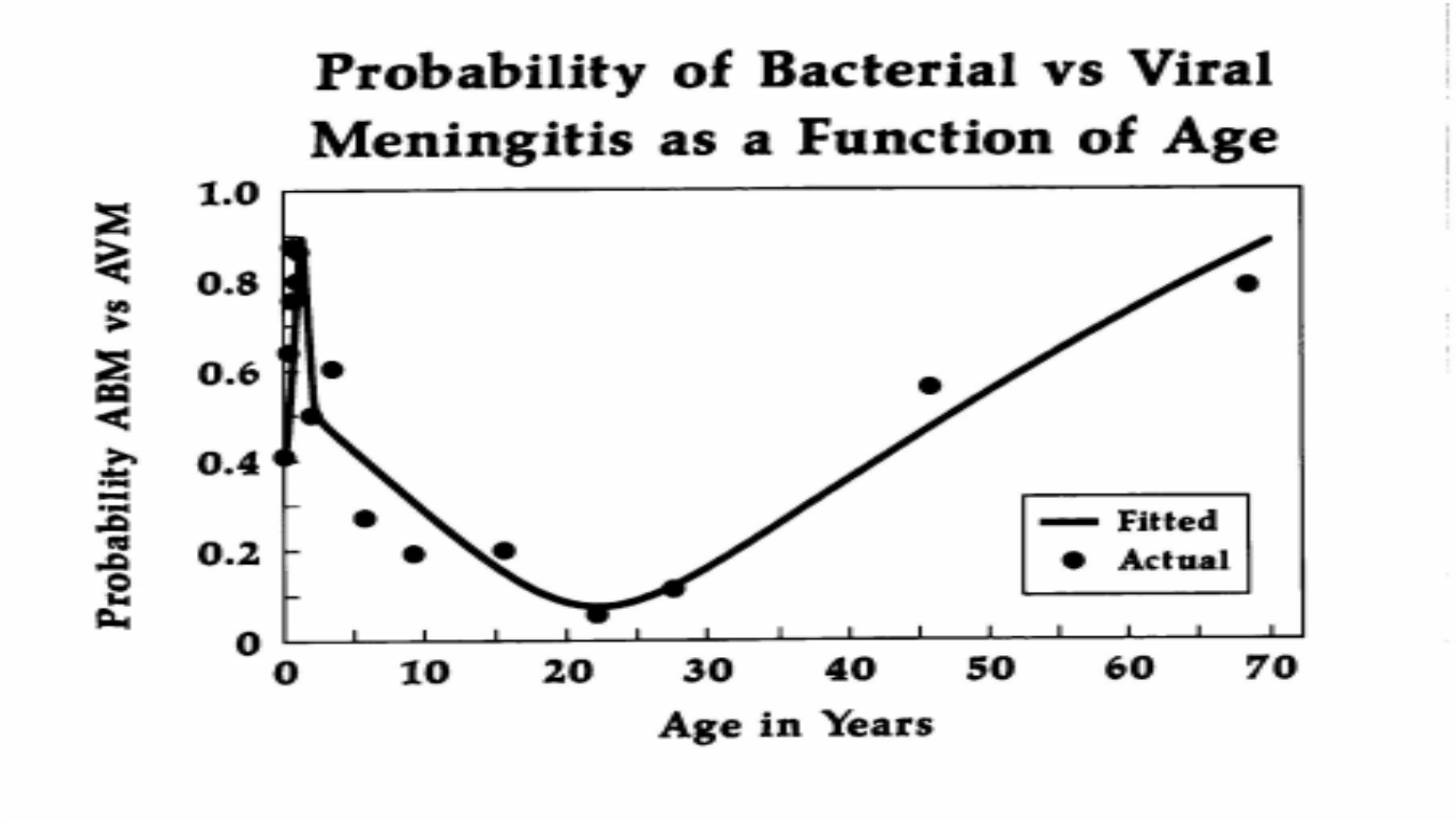
Missing Data

- Some surgical practices do not adequately characterize patients pre-operatively
- Need to determine if outcomes are worse
- Often advisable to count missing variables as if they had the most normal levels
- Makes risk-adjusted outcome worse
- Incentive to improve data collection
- Need careful statistical analysis

Model Development

- Logistic multiple regression model - patient descriptors are additive on log odds scale
- Can start with previously developed models
- Don’t assume that continuous variables are linearly related to the log odds of death
- Wrong to select variables using *P*-values
- Avoid overfitting
- See Harrell, Lee, Mark (1996) Stat in Med

Estimating Shape of Relationship with Outcome



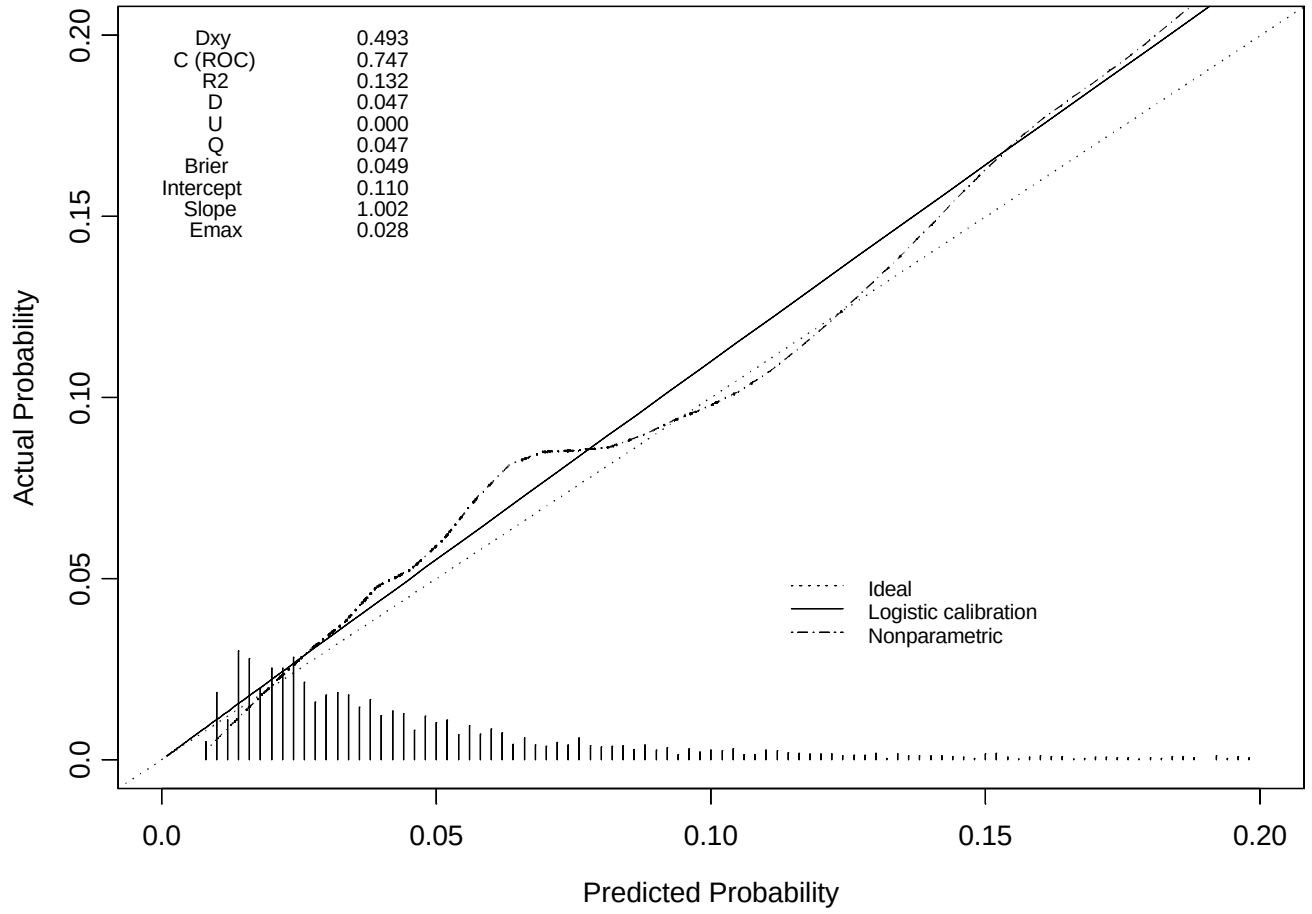
Spanos et al. (1989) JAMA

Measuring Model's Predictive Accuracy

- Goal: Forecast outcomes of new patients
- Predictive discrimination: ability of model to separate high and low-risk patients
- Calibration: agreement between predicted and observed proportions of deaths
- Validation methods: re-sampling, new patient series

Calibration Plot

NY 1994 Model Validated in 3762 Patients



Shrinkage - Disbelieving Some of the Data

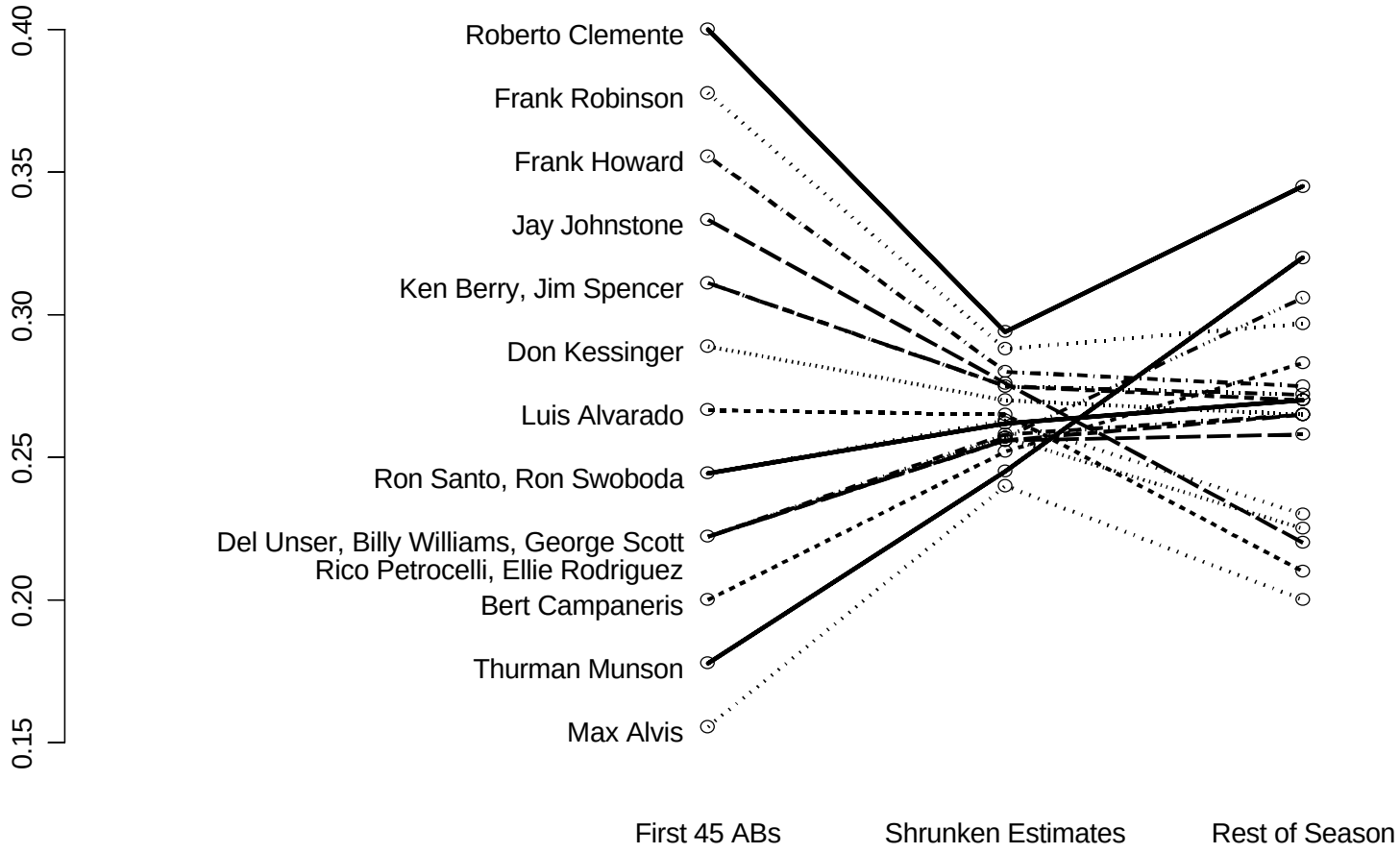
- Road intersection with ↑ fatalities
- Make any engineering change
- Fewer fatalities next year
- See John Adams (1995) Risk
- Identify surgeon by ↑ or ↓ mortality, next year her op. mort. will be closer to mean
- REGRESSION TO THE MEAN

Shrinkage, Continued

- Can also affect predictive accuracy of risk models: Predict patient has 0.2 chance of dying, 0.15 of similar new pts. die
- Build discounting (shrinkage) into predictions

1970 Batting Averages

Efron & Morris, Scientific American 1977

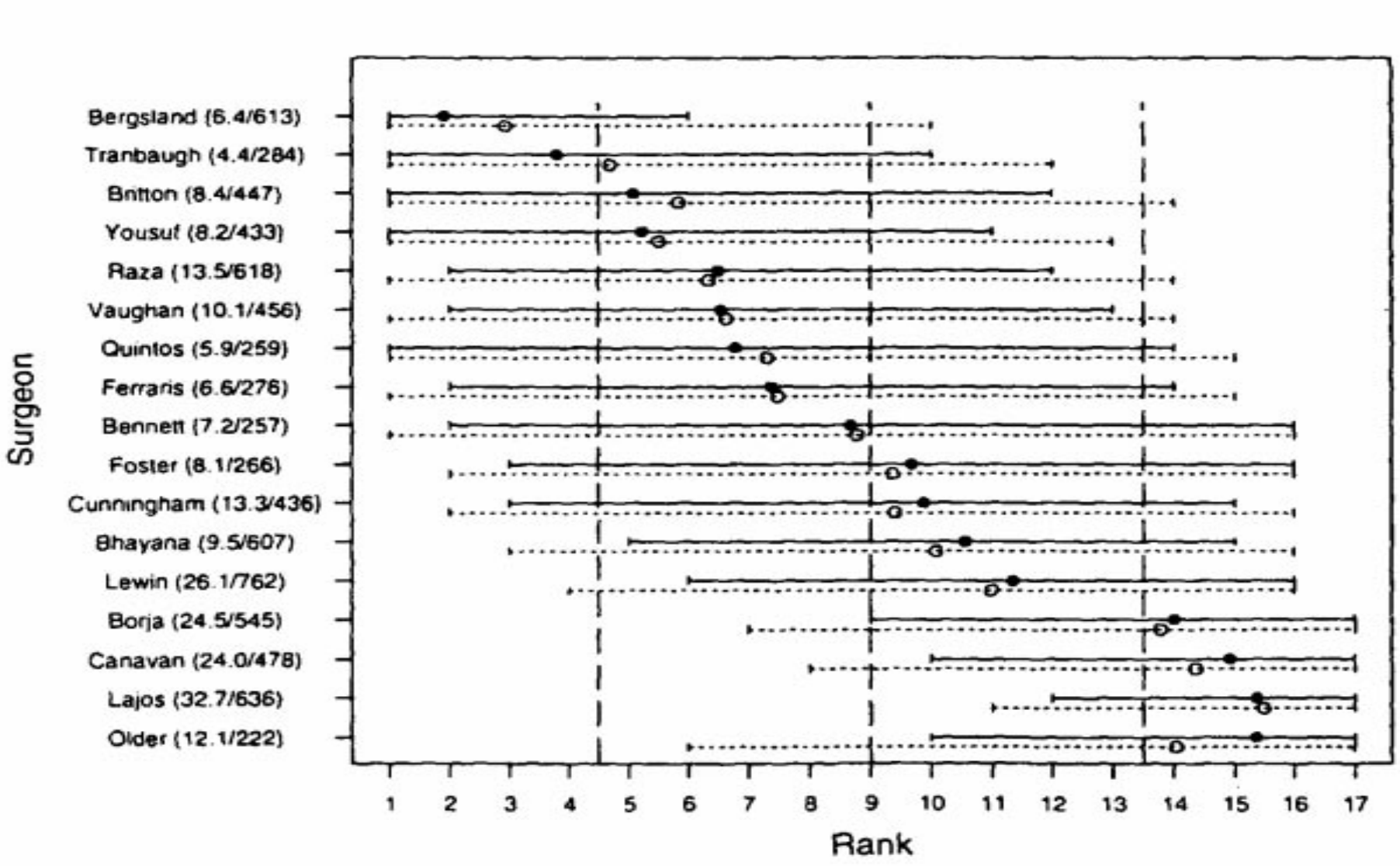


Shrinkage, Continued

- Can estimate one surgeon's outcomes more accurately by pulling to grand mean
- Pull more towards mean when
  - Surgeon patient volume is low
  - Less variation in outcomes across surgeons
- Don't base practice changes on past noise
- Base changes on outcome patterns that will replicate

Surgeons Cannot be Ranked Reliably

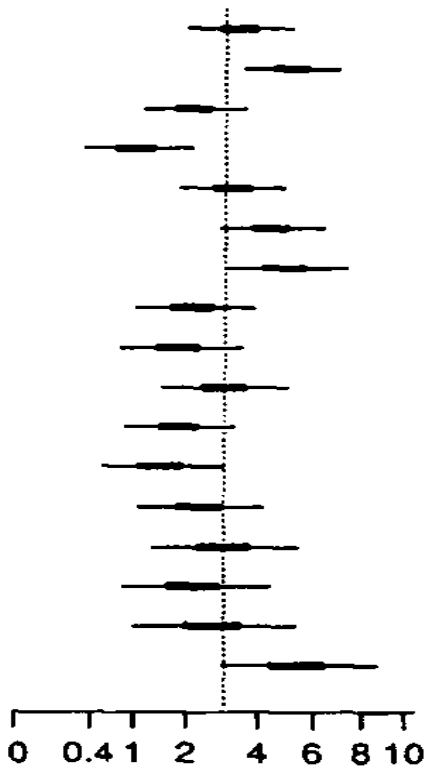
- Ranking mortality is splitting hairs
- True probability that surgeon with best (worst) results is really the best (worst) is low
- Goldstein & Spiegelhalter (1996): J Royal Stat Soc A
- NY 1990-1992 CABG, sample of 17 out of 87 surgeons, risk-adjusted


Surgeon
Bergsland (6.4/613)
Tranbaugh (4.4/284)
Britton (8.4/447)
Yousuf (8.2/433)
Raza (13.5/618)
Vaughan (10.1/456)
Quintos (5.9/259)
Ferraris (6.6/276)
Bennett (7.2/257)
Foster (8.1/266)
Cunningham (13.3/436)
Bhayana (9.5/607)
Lewin (26.1/762)
Borja (24.5/545)
Canavan (24.0/478)
Lajos (32.7/636)
Older (12.1/222)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Rank

Reporting Proposal by R Galbraith

Discussion to G&S Paper

Surgeon	*Cases*	*Deaths*	*Risk-adjusted mortality per 100 cases*	*Rank*
Lewin	762	19	2.04–5.13	4–16
Lajos	636	33	3.56–6.99	11–17
Raza	618	12	1.15–3.56	1–14
Bergsland	613	5	0.34–2.13	1–10
Bhayana	607	17	1.87–4.89	3–16
Borja	545	22	2.85–6.38	7–17
Canavan	478	19	3.01–7.36	8–17
Vaughan	456	9	1.01–3.85	1–14
Britton	447	7	0.78–3.48	1–14
Cunningham	436	11	1.51–5.04	2–16
Yousuf	433	9	0.86–3.26	1–13
Tranbaugh	284	6	0.55–2.93	1–12
Ferraris	276	9	1.09–4.18	1–15
Foster	266	8	1.34–5.41	2–16
Quintos	259	6	0.84–4.41	1–15
Bennett	257	6	1.02–5.34	1–16
Older	222	13	2.92–8.68	6–17



Problematic Statistical Measures

- Z-scores and P-values
- Multiple comparison problems (false positives)
- Not tied to relevant mortality differences
- Confidence intervals are better but still have confusing interpretation, difficult to use average or median mortality as a reference point

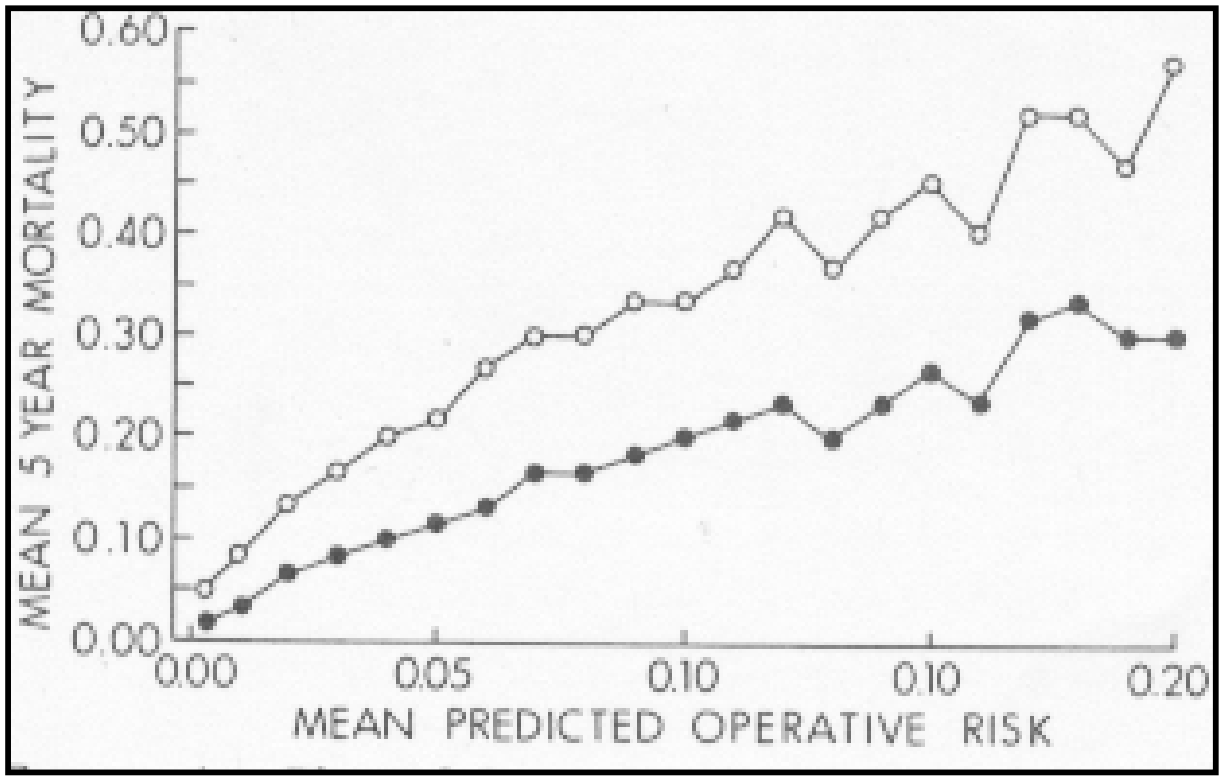
Better Report Card Measures

- Modification of Normand, Glickman, Gatsonis (1997) J Amer Statist Assoc
- Use Bayesian mixed effects logistic model (uses shrinkage)
- Examine each surgeon's effect (coefficient = log odds ratio)

Better Report Cards

- Estimate surgeon's odds of death relative to odds of death for "median" surgeon
- Compute Prob[odds ratio > 1.5]
- Concern if this probability > .9

Expected CABG Mortality and Long-Term Benefit



Med

CABG

Califf, Harrell, Lee *et al*. *Circ* 78 Supp I 1988

Summary

- Choose patient descriptors carefully to increase data quality, minimize gaming, maximize discrimination
- Risk models must be derived carefully; avoid fitting noise
- Regression to the mean is a dominant effect in operative mortality

Summary, Continued

- Shrinkage of risk-adjusted mortality estimates is necessary
- Almost futile to rank fine differences
- Identify problems by the probability of a large relative odds of death
- Not operating on high-risk patients may not benefit a surgeon’s risk-adjusted mortality or the patient