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
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

- $D_{xy} = 0.493$
- $C_{ROC} = 0.747$
- $R^2 = 0.132$
- $D = 0.047$
- $U = 0.000$
- $Q = 0.047$
- Brier = 0.049
- Intercept = 0.110
- Slope = 1.002
- $E_{max} = 0.028$
Shrinkage - Disbelieving Some of the Data

- Road intersection with \( \uparrow \) fatalities
- Make any engineering change
- Fewer fatalities next year
- See John Adams (1995) Risk
- Identify surgeon by \( \uparrow \) or \( \downarrow \) 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

First 45 ABs
Shrunken Estimates
Rest of Season
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
# Reporting Proposal by R Galbraith

Discussion to G&S Paper

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>Cases</th>
<th>Deaths</th>
<th>Risk-adjusted mortality per 100 cases</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lewin</td>
<td>762</td>
<td>19</td>
<td>2.04–5.13</td>
<td>4–16</td>
</tr>
<tr>
<td>Lajos</td>
<td>636</td>
<td>33</td>
<td>3.56–6.99</td>
<td>11–17</td>
</tr>
<tr>
<td>Raza</td>
<td>618</td>
<td>12</td>
<td>1.15–3.56</td>
<td>1–14</td>
</tr>
<tr>
<td>Bergsland</td>
<td>613</td>
<td>5</td>
<td>0.34–2.13</td>
<td>1–10</td>
</tr>
<tr>
<td>Bhayana</td>
<td>607</td>
<td>17</td>
<td>1.87–4.89</td>
<td>3–16</td>
</tr>
<tr>
<td>Borja</td>
<td>545</td>
<td>22</td>
<td>2.85–6.38</td>
<td>7–17</td>
</tr>
<tr>
<td>Canavan</td>
<td>478</td>
<td>19</td>
<td>3.01–7.36</td>
<td>8–17</td>
</tr>
<tr>
<td>Vaughan</td>
<td>456</td>
<td>9</td>
<td>1.01–3.85</td>
<td>1–14</td>
</tr>
<tr>
<td>Britton</td>
<td>447</td>
<td>7</td>
<td>0.78–3.48</td>
<td>1–14</td>
</tr>
<tr>
<td>Cunningham</td>
<td>436</td>
<td>11</td>
<td>1.51–5.04</td>
<td>2–16</td>
</tr>
<tr>
<td>Yousuf</td>
<td>433</td>
<td>9</td>
<td>0.86–3.26</td>
<td>1–13</td>
</tr>
<tr>
<td>Tranbaugh</td>
<td>284</td>
<td>6</td>
<td>0.55–2.93</td>
<td>1–12</td>
</tr>
<tr>
<td>Ferraris</td>
<td>276</td>
<td>9</td>
<td>1.09–4.18</td>
<td>1–15</td>
</tr>
<tr>
<td>Foster</td>
<td>266</td>
<td>8</td>
<td>1.34–5.41</td>
<td>2–16</td>
</tr>
<tr>
<td>Quintos</td>
<td>259</td>
<td>6</td>
<td>0.84–4.41</td>
<td>1–15</td>
</tr>
<tr>
<td>Bennett</td>
<td>257</td>
<td>6</td>
<td>1.02–5.34</td>
<td>1–16</td>
</tr>
<tr>
<td>Older</td>
<td>222</td>
<td>13</td>
<td>2.92–8.68</td>
<td>6–17</td>
</tr>
</tbody>
</table>
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 \( \text{Prob}[\text{odds ratio} > 1.5] \)
- Concern if this probability > .9
Expected CABG Mortality and Long-Term Benefit

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