Public Reporting of Surgical Outcomes

Benjamin D. Kozower, MD, MPH
Professor of Surgery
Co-Director, Surgical Outcomes Research Center
Washington University
St. Louis, MO
Disclosures

Chair, STS General Thoracic Surgery Database
Member, STS Quality Measurement Taskforce

Disclaimers

“All models are wrong, some models are useful”
George Box
History of Outcomes Research & Public Reporting

“The trustees of hospitals should see to it that an effort is made to follow up each patient they treat, long enough to determine whether the treatment given has permanently relieved the condition or symptoms complained of.”

Dr. Ernest Amory Codman, 1918
Problems with Current Methods for Ranking Surgeons and Hospitals …..

"I have just heard that 50% of United Kingdom cardiac surgeons are below average, ….. this has got to stop!"

The Rt Hon Frank Dobson
Former Labor Secretary for Health
To Your Health

Need a new knee? Heart valve? Back surgery? This Web site could help you find the top surgeons near you.

By Lena H. Sun  July 14  Follow @bylenasun
<table>
<thead>
<tr>
<th>Select to</th>
<th>Name</th>
<th>Better Outcomes</th>
<th>Recommended Most By</th>
<th>Board Certifications</th>
<th>Surgeon's Hospitals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adams, Reid B.</td>
<td>★ ★ ★</td>
<td>---</td>
<td>---</td>
<td>Surgery</td>
</tr>
<tr>
<td></td>
<td>Faulkenberry, William L.</td>
<td>---</td>
<td>★ ★ ★</td>
<td>---</td>
<td>Surgery</td>
</tr>
<tr>
<td></td>
<td>Branum, Gene D.</td>
<td>---</td>
<td>★ ★ ★</td>
<td>---</td>
<td>Surgery</td>
</tr>
<tr>
<td></td>
<td>Garwood, Robert Allen</td>
<td>---</td>
<td>★ ★ ★</td>
<td>---</td>
<td>Surgery</td>
</tr>
<tr>
<td></td>
<td>Oates, Thomas M.</td>
<td>---</td>
<td>★ ★ ★</td>
<td>---</td>
<td>Pediatrics, Surgery</td>
</tr>
</tbody>
</table>

How confidently we say the surgeon's results are better-than-average: Very confidently = ★★★★★★ Confidently = ★★★★★ Probably = ★★★★
Surgeon Scorecard

by Sisi Wei, Olga Pierce and Marshall Allen, ProPublica, Updated July 15, 2015

Guided by experts, ProPublica calculated death and complication rates for surgeons performing one of eight elective procedures in Medicare, carefully adjusting for differences in patient health, age and hospital quality. Use this database to know more about a surgeon before your operation.

---

**SURGEONS AND HOSPITALS NEAR MY LOCATION**

Find Near Me  Find a Surgeon  Find a Hospital

---

**Eight Elective Procedures**

We focused on procedures done thousands of times a day, mostly without incident. They are scheduled in advance and generally performed on patients in stable health. We excluded patients who

**Surgeons, Not Hospitals**

Conventional wisdom tells patients to simply choose a good hospital when they need surgery. But ProPublica has found that even within “good” hospitals, performance between surgeons can
Public Reporting of Surgical Outcomes

• Routine expectation of stakeholders
  – Patients, payers, legislators, policy makers

• Ethical responsibility
  – Accountability
  – Affirm patient autonomy

• Goals
  – Education of consumers
  – Improve patient care (2 mechanisms)
  – Improve resource utilization
Public Reporting of Surgical Outcomes: Improving Quality

NY State Public Reporting for CABG

<table>
<thead>
<tr>
<th>Year</th>
<th>1989</th>
<th>1992</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-hospital mortality</td>
<td>3.5%</td>
<td>2.7%</td>
</tr>
<tr>
<td>Expected mortality (risk-adjusted)</td>
<td>2.9%</td>
<td>3.2%</td>
</tr>
</tbody>
</table>

Hannan EL et al. J Am Coll Cardiol. 2012;59
Public Reporting of Surgical Outcomes: Improving Quality

Society of Thoracic Surgeons Database: “A Decade of Change – Risk Profiles and Outcomes for CABG”

<table>
<thead>
<tr>
<th>Year</th>
<th>1990</th>
<th>1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operative mortality</td>
<td>3.9%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Expected mortality (risk-adjusted)</td>
<td>2.6%</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

Public Reporting of Surgical Outcomes

• Opportunity
  – Currently being done with poor methodology and administrative data
  – Need more relevant, patient-centered and long-term outcomes
  – Need robust methodologies to compare outcomes, “quality”, and performance

• Need for transparency
Barriers to Public Reporting: Requirements for Comparing Institutions

- We don’t have robust, accurate data
- We don’t have relevant outcomes measures
- We need sufficient statistical power to identify meaningful differences
- We don’t know how to communicate effectively to consumers

Lansky D, Health Affairs April, 2012
Gaynor JW, With Permission 2015
Impact of Case Ascertainment Methodology on Hospital Rankings

Impact of Case Ascertainment Methodology on Hospital Rankings

Predicted Risk of Mortality Models: Surgeons Need to Understand Limitations of the University HealthSystem Consortium Models

Benjamin D Kozower, MD, MPH, FACS, Gorav Ailawadi, MD, FACS, David R Jones, MD, FACS, Robert D Pates, PhD, Christine L Lau, MD, FACS, Irving L Kron, MD, FACS, George J Stukenborg, PhD

BACKGROUND: The University HealthSystem Consortium (UHC) mortality risk adjustment models are increasingly being used as benchmarks for quality assessment. But these administrative database models may include postoperative complications in their adjustments for preoperative risk. The purpose of this study was to compare the performance of the UHC with the Society of Thoracic Surgeons (STS) risk-adjusted mortality models for adult cardiac surgery and evaluate the contribution of postoperative complications on model performance.

STUDY DESIGN: We identified adult cardiac surgery patients with mortality risk estimates in both the UHC and Society of Thoracic Surgeons databases. We compared the predictive performance and calibration of estimates from both models. We then reestimated both models using only patients without any postoperative complications to determine the relative contribution of adjustments for postoperative events on model performance.

RESULTS: In the study population of 2,171 patients, the UHC model explained more variability (27% versus 13%, p < 0.001) and achieved better discrimination (C statistic = 0.88 versus 0.81, p < 0.001). But when applied in the population of patients without complications, the UHC model performance declined severely. The C statistic decreased from 0.88 to 0.49, a level of discrimination equivalent to random chance. The discrimination of the Society of Thoracic Surgeons model was unchanged (C statistic of 0.79 versus 0.81).

CONCLUSIONS: Although the UHC model demonstrated better performance in the total study population, this difference in performance reflects adjustments for conditions that are postoperative complications. The current UHC models should not be used for quality benchmarks. (J Am Coll Surg 2009;209:551–556. © 2009 by the American College of Surgeons)
Comparing Administrative & Clinical Data: Apples & Oranges

- The University HealthSystem Consortium (UHC)
  - Includes all patients in the cardiovascular service line
  - Decile rankings (no statistical reporting)

- The Society of Thoracic Surgeons (STS)
  - Includes patients operated on by cardiac surgeons
  - Risk-adjustment for common procedures
  - Tercile rankings (1-3 stars)

Kozower BD et al. JACS. 2009
Comparing Administrative & Clinical Data

- 28% difference in cardiac surgery patient populations
- STS appropriately uses preoperative risk factors
- UHC includes all diagnoses, including postoperative complications
Gold standard clinical database
Adult Cardiac, General Thoracic, Congenital
95% data accuracy, externally audited
Predictive risk models for common procedures
Barriers to Public Reporting: Requirements for Comparing Institutions

- We don’t have robust, accurate data
- We don’t have relevant outcomes measures
- We need sufficient statistical power to identify meaningful differences
- We don’t know how to communicate effectively to consumers

Lansky D, Health Affairs April, 2012
Gaynor JW, With Permission 2015
The Swedish Approach to Hip Replacement

- Reported health gain (EQ-5D index) after one year
- Patient satisfaction at one year
- Complications at two years
- Ten-year implant survival

- None of these measures are reported in the US
- > 90% follow up publicly reported in Sweden

Swedish Hip Arthroplasty Register 2015
http://www.shpr.se/en/
STS Database Risk Models: Predictors of Mortality and Major Morbidity for Lung Cancer Resection

Benjamin D. Kozower, MD, MPH, Shubin Sheng, PhD, Sean M. O’Brien, PhD, Michael J. Liptay, MD, Christine L. Lau, MD, David R. Jones, MD, David M. Shahian, MD, and Cameron D. Wright, MD

Departments of Surgery & Public Health Sciences, University of Virginia Health System, Charlottesville, Virginia; Duke Clinical Research Institute, Duke University, Durham, North Carolina; Department of Cardiovascular and Thoracic Surgery, Rush University, Chicago, Illinois; and Division of Thoracic Surgery and Center for Quality and Safety, Massachusetts General Hospital, Boston, Massachusetts
Hospital Performance Variation:
Mortality or Major Morbidity – STS Database
18,800 lung cancer resections from 111 hospitals

Thoracic Surgery Outcomes: Measuring Hospital Performance

• Successfully measured and compared 30 day composite outcomes

• Need long term survival for cancer care
• Need patient centered outcomes

Barriers to Public Reporting: Requirements for Comparing Institutions

- We don’t have robust, accurate data
- We don’t have relevant outcomes measures
- We need sufficient statistical power to identify meaningful differences
- We don’t know how to communicate effectively to consumers

Lansky D, Health Affairs April, 2012
Gaynor JW, With Permission 2015
Surgical Mortality as an Indicator of Hospital Quality: The Problem with Sample Size

Only CABG has the case volumes to reach the threshold of 138 cases to detect a doubling of the mortality rate.

Dimick et al JAMA 2004;292:847
• Medicare data 2009-2012
• 90-day mortality, readmission, complications
• Risk-adjustment methodology is unknown
• 5 categories, only report top three for now
Objectives

– To compare operative mortality definitions
– To determine if 90-day mortality can be used as a quality metric

Methods

– SEER-Medicare data 2006-2010
– Hierarchical GLE models to estimate 30 and 90-day mortality
Results

- 11,787 patients from 686 hospitals
- Provider volumes range from 1-383
- 33% of hospitals perform fewer than 5 resections (Medicare)
## Results

<table>
<thead>
<tr>
<th>Mortality</th>
<th>n  (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-hospital</td>
<td>355 (3.01)</td>
</tr>
<tr>
<td>30 day</td>
<td>435 (3.69)</td>
</tr>
<tr>
<td>Perioperative</td>
<td>482 (4.09)</td>
</tr>
<tr>
<td>90 day</td>
<td>812 (6.89)</td>
</tr>
</tbody>
</table>
Results: Provider Rankings

• Calculated observed/expected 90-day mortality estimates for 686 hospitals

• Hospital performance variation
  – 3% of hospitals (20/686) had an O/E 90-day mortality different than 1
    • 18 hospitals had O/E > 1
    • 2 hospitals had O/E < 1
Study Conclusions

• Currently reported measures of in-hospital and 30-day postoperative mortality do not adequately represent a patient’s true mortality risk
  – Mortality almost doubles by 90 days

• 90-day mortality is not a useful metric for evaluating hospitals following lung cancer resection
How accurate are surgeon level ratings for 90-day mortality when hospital level ratings only identify 3% of programs as different than average?

How is “Probably” defined and is it sufficient for public reporting?
Increasing Sample Size:
Composite Measures:
Mortality or Major Morbidity – STS Database
18,800 lung cancer resections from 111 hospitals

Barriers to Public Reporting: Requirements for Comparing Institutions

- We don’t have robust, accurate data
- We don’t have relevant outcomes measures
- We need sufficient statistical power to identify meaningful differences
- We don’t know how to communicate effectively to consumers

Lansky D, Health Affairs April, 2012
Gaynor JW, With Permission 2015
Communicating Outcomes with Consumers

• 60 million Americans use Yelp and dry cleaner ratings every month

• Healthcare is a very common search, but far from the top 10
  – 66% of internet users performed a healthcare search in the past year

Leapfrog Group, Consumer Reports Health, US News & World Report’s, HealthGrades

Examined 844 hospitals
Four Rating Systems: Leapfrog, Consumer Reports, US News and World Report’s, HealthGrades

Objectives were to compare the criteria used in ratings systems and to determine the consistency of hospital ratings
• No hospital was rated as a high performer by all 4 rating systems
• Each system has its own rating methods, a different focus to its ratings, and stresses different measures
Examined 4 different public reporting displays
  – 3 of 4 are used by publically available state reports
Displays shown to 337 adults
Respondents asked to identify which surgeon they were most likely to choose and surveyed on the importance of public reporting
This is the “Best” State Display: Which Surgeon Would You Choose?

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>Cases</th>
<th>In-Hospital</th>
<th>30-Day</th>
<th>Readmissions</th>
<th>7-Day</th>
<th>30-Day</th>
<th>Length of Stay</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>5.8</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>6.8</td>
</tr>
<tr>
<td>3</td>
<td>89</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>5.7</td>
</tr>
<tr>
<td>4</td>
<td>91</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>4.9</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>6.9</td>
</tr>
</tbody>
</table>

- ● Lower than expected (meaning that the hospital or surgeon had fewer deaths than expected after accounting for how sick the patients were in that hospital)
- ○ Same as expected (meaning that the hospital or surgeon had as many deaths as expected after accounting for how sick the patients were in that hospital)
- ○ Higher than expected (meaning that the hospital or surgeon had more deaths than expected after accounting for how sick the patients were in that hospital)

Donelan K et al. Ann Thorac Surg. 2011;91
Fig 3. Personal importance of public ratings: “If you have to select a surgeon, how important is it to you to have data like these available to you?”

Donelan K et al. Ann Thorac Surg. 2011;91
Surgeon Preference by Display Format

The public does not understand current displays of reported outcomes

The “best surgeon” was correctly identified between 13% and 66% of the time

---

**Table 4. Surgery Results: Surgeon Preference for Display Type**

<table>
<thead>
<tr>
<th>Display</th>
<th>Surgeon 1 (%)</th>
<th>Surgeon 2 (%)</th>
<th>Surgeon 3 (%)</th>
<th>Surgeon 4 (%)</th>
<th>Surgeon 5 (%)</th>
<th>Need More Information (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>14</td>
<td>44</td>
<td>16*a</td>
<td>N/A</td>
<td>N/A</td>
<td>26</td>
</tr>
<tr>
<td>B</td>
<td>53*a</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>66*a</td>
<td>17</td>
</tr>
<tr>
<td>D</td>
<td>32</td>
<td>10</td>
<td>2</td>
<td>22*a</td>
<td>13</td>
<td>21</td>
</tr>
</tbody>
</table>

Surgeon preference for each display is shown ("which surgeon would you choose given the information shown"). *Indicates the surgeon with the lowest risk-adjusted mortality in the display.

N/A = not available.
Public Reporting of Surgical Outcomes

STS Public Reporting Online

Heart Surgery Outcomes - Public Access

STS believes the public has a right to know the quality of surgical outcomes and considers public reporting an ethical responsibility of the specialty. STS Public Reporting Online enables STS National Database participants to voluntarily report to each other and the public their heart surgery composite star ratings and the component ratings from which those are derived.

Isolated CABG:
Search or browse star ratings for surgery groups
Search or browse star ratings for hospitals

Isolated AVR:
Search or browse star ratings for surgery groups
Search or browse star ratings for hospitals
The Society of Thoracic Surgeons Composite Score for Lobectomy for Lung Cancer


STS Quality Measurement Task Force
STS General Thoracic Surgery Database Task Force
Duke Clinical Research Institute

Public Reporting of Surgical Outcomes

- Develop a platform for thoracic surgery
  - Start with lobectomy for lung cancer
  - Most common procedure
  - Use 3 years of rolling data

- Two domain composite
  - Operative mortality
  - Major complications
Methods: Estimation of Composite Scores & Star Ratings

• Composite score = (1-risk adjusted mortality rate) + (1-risk adjusted complication rate)

• Outcomes weighted inversely by their respective standard deviations

• Star ratings (3 categories)
  – High (3-star) and low (1-star) performers
    • 95% credible intervals outside average scores
  – Easy for the public to comprehend

Results: Patient Population

22,362 lobectomies at 234 programs

- Non-elective status (1,211)
- Missing data (n=281)
- Benign disease (n=122)

20,657 at 231 programs (92.4%)
## Composite Outcomes

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operative mortality</td>
<td>1.5%</td>
</tr>
<tr>
<td>Major complications</td>
<td>9.6%</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>4.3%</td>
</tr>
<tr>
<td>Unexpected return to OR</td>
<td>3.9%</td>
</tr>
<tr>
<td>Reintubation</td>
<td>3.4%</td>
</tr>
<tr>
<td>Pulmonary embolus</td>
<td>0.5%</td>
</tr>
<tr>
<td>Initial Vent Support &gt;48 Hours</td>
<td>0.5%</td>
</tr>
<tr>
<td>Bronchopleural fistula</td>
<td>0.4%</td>
</tr>
<tr>
<td>Tracheostomy</td>
<td>1.0%</td>
</tr>
<tr>
<td>ARDS</td>
<td>0.7%</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

**Length of stay (median):** 4 days, IQR (3,7)
Mortality is weighted approximately 4 times that of a major complication in the quality measure.
Reliability of Star Ratings (Signal / Noise Ratio)

- Reliability is a key measure for profiling
- Describes the confidence in distinguishing performance between providers
- Signal is the proportion of variability in measured performance explained by real differences in performance
- Noise is the random statistical variation

Reliability of Star Ratings Based on Program Volume

<table>
<thead>
<tr>
<th>Number of participants</th>
<th>No Minimum</th>
<th>$\geq30$ cases</th>
<th>$\geq50$ cases</th>
<th>$\geq100$ cases</th>
<th>$\geq150$ cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability (95% Crl)</td>
<td>42% [33%-50%]</td>
<td>51% [41%-60%]</td>
<td>55% [45%-64%]</td>
<td>60% [48%-70%]</td>
<td>65% [50%-78%]</td>
</tr>
</tbody>
</table>
Table: Number of STS Participants Identified as Different from STS Average

<table>
<thead>
<tr>
<th>Criterion for categorizing participants</th>
<th>Below STS-Average (1-star)</th>
<th>STS-Average (2-star)</th>
<th>Above STS-Average (3-star)</th>
</tr>
</thead>
<tbody>
<tr>
<td>98% Credible interval</td>
<td>5 (2.9%)</td>
<td>160 (93.0%)</td>
<td>7 (4.1%)</td>
</tr>
<tr>
<td>95% Credible interval</td>
<td>8 (4.7%)</td>
<td>152 (88.4%)</td>
<td>12 (7.0%)</td>
</tr>
<tr>
<td>90% Credible interval</td>
<td>12 (7.0%)</td>
<td>142 (82.6%)</td>
<td>18 (10.5%)</td>
</tr>
<tr>
<td>80% Credible interval</td>
<td>16 (9.3%)</td>
<td>128 (74.4%)</td>
<td>28 (16.3%)</td>
</tr>
</tbody>
</table>
Distribution of Composite Scores
## Construct Validity:
Complication Rates Vary Across Star Ratings

<table>
<thead>
<tr>
<th></th>
<th>1- Star</th>
<th>2 - Star</th>
<th>3 - Star</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operative Mortality</td>
<td>3.2% [1.6% - 5.9%]</td>
<td>1.6% [0.6% - 3.5%]</td>
<td>0.9% [0.4% - 1.6%]</td>
</tr>
<tr>
<td>(95% CrI)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major Complications</td>
<td>17.1% [11.3% - 24.2%]</td>
<td>10.1% [5.1% - 16.9%]</td>
<td>6.5% [3.7% - 9.6%]</td>
</tr>
<tr>
<td>(95% CrI)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Risk Adjusted Outcomes: Star Ratings

STS General Thoracic Surgery Database
Composite Quality Rating
Participant 99999
STS Period Ending 12/31/2013

Quality Domain | Participant Score (95% CI) | STS Mean Participant Score | Participant Rating
--- | --- | --- | ---
Jan 2011 - Dec 2013 Overall | 93.8% (92.5, 95.0) | 94.5% | ★★

Jan 2011 - Dec 2013 Absence of Mortality

Jan 2011 - Dec 2013 Absence of Major Complication

Jan 2011 - Dec 2013 Absence of Mortality

Jan 2011 - Dec 2013 Absence of Major Complication

Eligible Procedures | Detail | Count | Percent
--- | --- | --- | ---
400 | Mortality | 13 |

Jan 2011 - Dec 2013 Absence of Major Complication

Jan 2011 - Dec 2013 Absence of Major Complication

Percentages represent the proportion that the specific complication contributed to the total number of patients with major complication.

This information is intended to facilitate and focus process and quality improvement initiatives by providers.
Limitations

• Generalizability
  – STS outcomes are better than national benchmarks
  – Captures < 50% of lobectomies in the US

• Short term outcome measure
  – Vital status out to 5-years added to STS Database in 2015
  – Lack patient reported outcomes
STS Study Conclusions

- Developed a two-domain quality measure for lobectomy for lung cancer
- Identified 12% of participating programs as outliers
- Will present risk-adjusted outcomes and star ratings in the public reporting effort
- Encourage participation in voluntary public reporting
Unintended Consequences of Public Reporting

- Misclassification of surgeons or hospitals
- Risk aversion for high risk patients with reduced access to care
- Confusion for the public and payers
- Potential for serious financial consequences
Distribution of Composite Scores:
Potential for misclassification
Comparison to the mean, not to nearest neighbor
Most surgeons (62%) refused to operate on at least one high-risk CABG patient over the prior year, primarily because of public reporting.

“We must address the underlying incentives for “case-selection creep” by improving risk adjustment methods for the highest-risk patients and by highlighting centers and physicians who undertake high-risk procedures in appropriate patients.”

Resnic FS. JACC. 2009;53(10)
Public Reporting of Surgical Outcomes: Conclusions

• The primary goals are:
  – Inform patients and stakeholders
  – Improve the quality of care

• Current public reporting initiatives have serious limitations
  – Opportunity for surgeons to lead

• Need accurate data

• Important and relevant outcomes
Public Reporting of Surgical Outcomes: Conclusions

• Need collaboration between key stakeholders
• Need to minimize potential consequences
• Reliable and transparent methodology
  – Iterative process
The Search for Quality
Acknowledgements

• University of Virginia Surgical Outcomes Research Center
  – George Stukenborg, PhD
  – Tim McMurry, PhD

• Society of Thoracic Surgeons
  – David Shahian, MD
  – J William Gaynor, MD