Methodological Considerations in Developing Hospital Composite Performance Measures

Sean M. O'Brien, PhD
Department of Biostatistics & Bioinformatics
Duke University Medical Center
sean.o'brien@dcri.duke.edu

Introduction

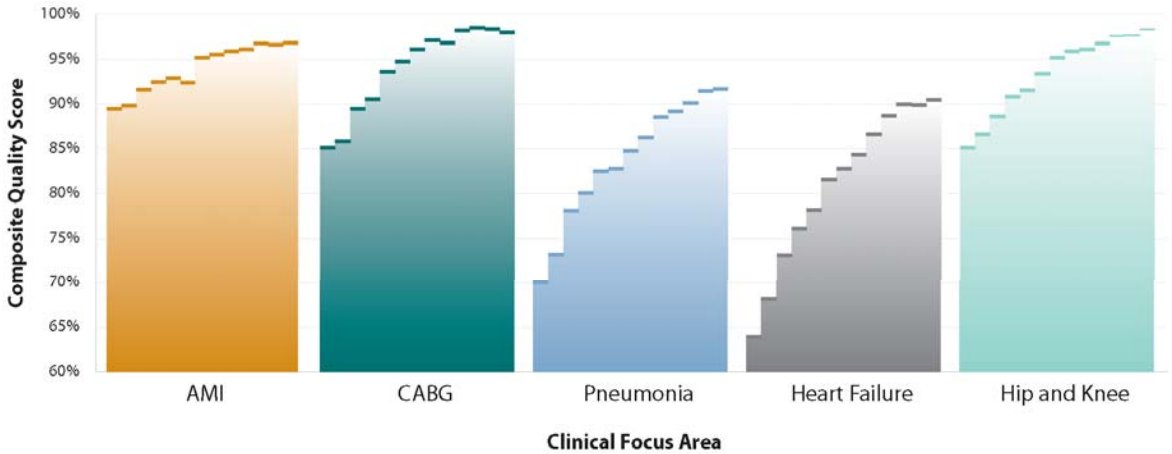
- **A “composite performance measure” is a combination of two or more related indicators**
  - e.g. process measures, outcome measures
- **Useful for summarizing a large number of indicators**
- **Reduces a large number of indicators into a single simple summary**

Example #1 of 3: CMS / Premier Hospital Quality Incentive Demonstration Project

CMS/Premier HQID Project
Sustained & Dramatic Improvement Continues

Composite Quality Score

CMS/Premier HQID Project Participants Composite Quality Score:
Trend of Quarterly Median (5th Decile) by Clinical Focus Area
October 1, 2003 - December 31, 2006 (Year 1 and Year 2 Final Data; Year 3 and Year 4 Preliminary Data)



PREMIER

source: http://www.premierinc.com/quality-safety/tools-services/p4p/hqi/images/composite-score.pdf

Example #2 of 3: US News & World Report's Hospital Rankings

2007 Rankings – Heart and Heart Surgery

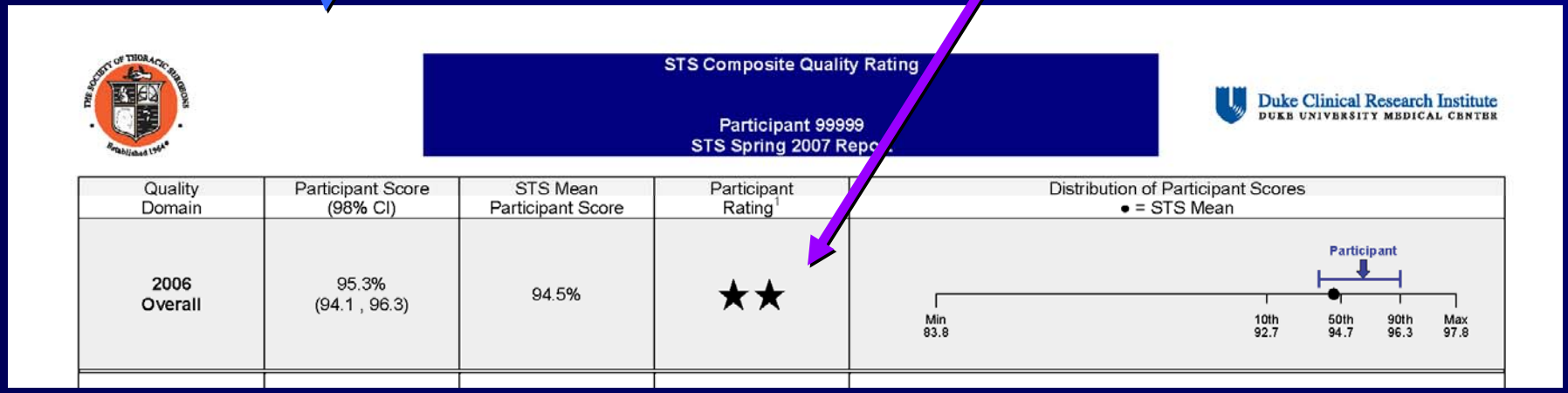
Rank	Hospital	Score
#1	Cleveland Clinic	100.0
#2	Mayo Clinic, Rochester, Minn.	79.7
#3	Brigham and Women's Hospital, Boston	50.5
#4	Johns Hopkins Hospital, Baltimore	48.6
#5	Massachusetts General Hospital, Boston	47.6
#6	New York-Presbyterian Univ. Hosp. of Columbia and Cornell	45.6
#7	Texas Heart Institute at St. Luke's Episcopal Hospital, Houston	45.0
#8	Duke University Medical Center, Durham, N.C.	42.2

source: http://www.usnews.com

Example #3 of 3: Society of Thoracic Surgeons Composite Score for CABG Quality

STS Database Participant Feedback Report

STS Composite Quality Rating



STS Composite Quality Rating

Participant 99999
STS Spring 2007 Repo

Duke Clinical Research Institute
DUKE UNIVERSITY MEDICAL CENTER

Quality Domain	Participant Score (98% CI)	STS Mean Participant Score	Participant Rating[1]	Distribution of Participant Scores • = STS Mean
2006 Overall	95.3% (94.1 , 96.3)	94.5%	★★	Participant; Min 83.8; 10th 92.7; 50th 94.7; 90th 96.3; Max 97.8

Why Composite Measures?

- Simplifies reporting
- Facilitates ranking
- More comprehensive than single measure
- More precision than single measure

Limitations of Composite Measures

- **Loss of information**
- **Requires subjective weighting**
  - No single objective methodology
- **Hospital rankings may depend on weights**
- **Hard to interpret**
  - May seem like a “black box”
  - Not always clear what is being measured

Goals

- Discuss methodological issues & approaches for constructing composite scores
- Illustrate inherent limitations of composite scores

Outline

- **Motivating Example:**
  US News & World Reports “Best Hospitals”
- **Case Study:**
  Developing a Composite Score for CABG

Motivating Example: US News & World Reports – Best Hospitals 2007

Quality Measures for Heart and Heart Surgery

Reputation Score

(Based on physician survey. Percent of physicians who list your hospital in the "top 5")

+

Mortality Index

(Risk adjusted 30-day. Ratio of observed to expected number of mortalities for for AMI, CABG etc.)

+

Structure Component
Volume
Nursing index
Nurse magnet hosp
Advanced services
Patient services
Trauma center

Motivating Example: US News & World Reports – Best Hospitals 2007

“structure, process, and outcomes each received one-third of the weight.”

- America’s Best Hospitals 2007 Methodology Report

Motivating Example: US News & World Reports – Best Hospitals 2007

Example Data – Heart and Heart Surgery

Duke University Medical Center	
Reputation	16.2%
Mortality index	0.77
Discharges	6624
Nursing index	1.6
Nurse magnet hosp	Yes
Advanced services	5 of 5
Patient services	6 of 6
Trauma center	Yes

source: usnews.com

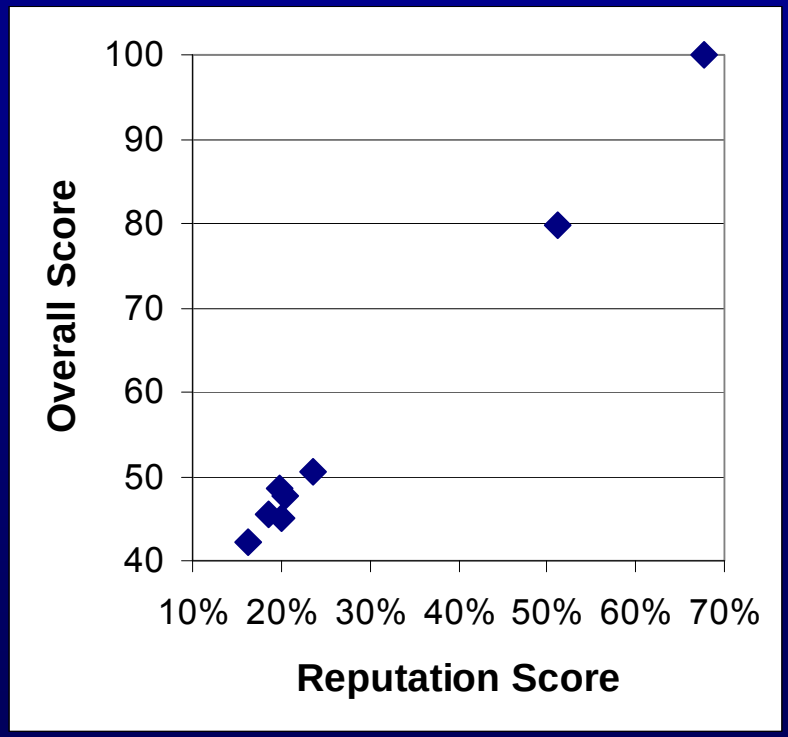
Which hospital is better?

Hospital A	
Reputation	5.7%
Mortality index	0.74
Discharges	10047
Nursing index	2.0
Nurse magnet hosp	Yes
Advanced services	5 of 5
Patient services	6 of 6
Trauma center	Yes

Hospital B	
Reputation	14.3%
Mortality index	1.10
Discharges	2922
Nursing index	2.0
Nurse magnet hosp	Yes
Advanced services	5 of 5
Patient services	6 of 6
Trauma center	Yes

Despite Equal Weighting, Results Are Largely Driven By Reputation

2007 Rank	Hospital	Overall Score	Reputation Score
#1	Cleveland Clinic	100.0	67.7%
#2	Mayo Clinic, Rochester, Minn.	79.7	51.1%
#3	Brigham and Women's Hospital, Boston	50.5	23.5%
#4	Johns Hopkins Hospital, Baltimore	48.6	19.8%
#5	Massachusetts General Hospital, Boston	47.6	20.4%
#6	New York-Presbyterian Univ. Hosp. of Columbia and Cornell	45.6	18.5%
#7	Texas Heart Institute at St. Luke's Episcopal Hospital, Houston	45.0	20.1%
#8	Duke University Medical Center, Durham, N.C.	42.2	16.2%



(source of data: http://www.usnews.com)

Lesson for Hospital Administrators (?)

- **Best way to improve your score is to boost your reputation**
  - Focus on publishing, research, etc.
- **Improving your mortality rate may have a modest impact**

Lesson for Composite Measure Developers

- No single “objective” method of choosing weights
- “Equal weighting” may not always behave like it sounds

Case Study: Composite Measurement for Coronary Artery Bypass Surgery

Background

- **Society of Thoracic Surgeons (STS) – Adult Cardiac Database**
  - Since 1990
  - Largest quality improvement registry for adult cardiac surgery
  - Primarily for internal feedback
  - Increasingly used for reporting to 3rd parties
- **STS Quality Measurement Taskforce (QMTF)**
  - Created in 2005
  - First task: Develop a composite score for CABG for use by 3rd party payers

Why Not Use the CMS HQID Composite Score?

- **Choice of measures**
  - Some HQID measures not available in STS

    (Also, some nationally endorsed measures are not included in HQID)
- **Weighting of process vs. outcome measures**
  - HQID is heavily weighted toward process measures
  - STS QMTF surgeons wanted a score that was heavily driven by outcomes

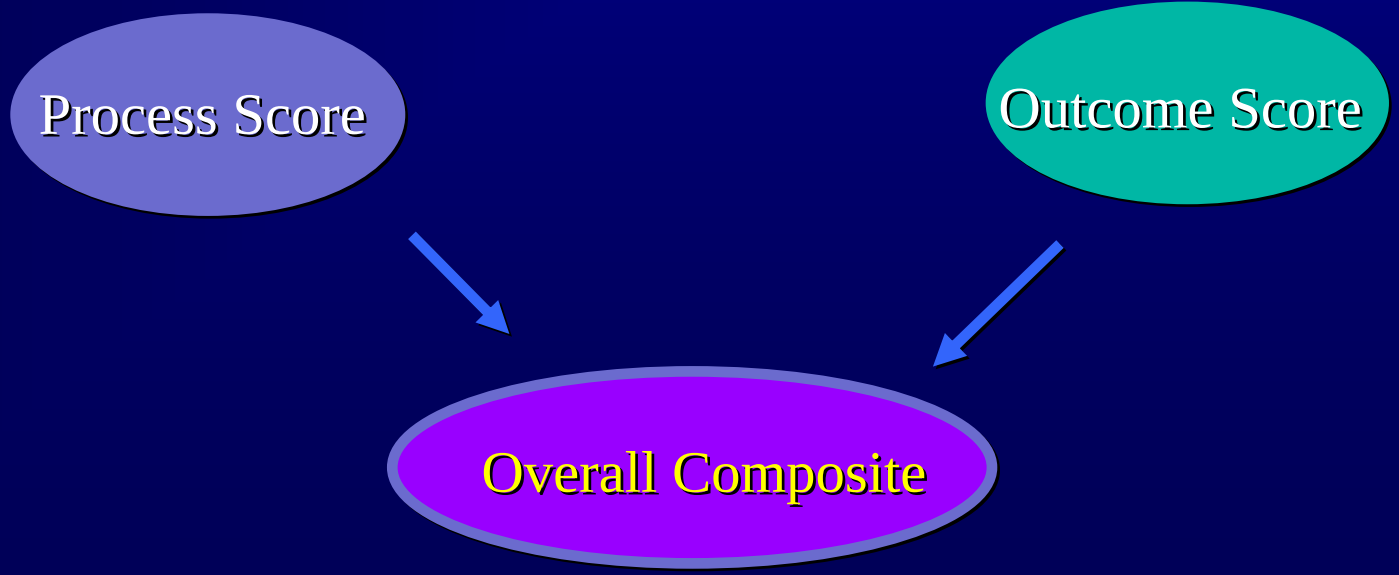
Our Process for Developing Composite Scores

- **Review specific examples of composite scores in medicine**
  - Example: CMS HQID
- **Review and apply approaches from other disciplines**
  - Psychometrics
- **Explore the behavior of alternative weighting methods in real data**
- **Assess the performance of the chosen methodology**

CABG Composite Scores in HQID (Year 1)

Process Measures (4 items)
Aspirin prescribed at discharge
Antibiotics <1 hour prior to incision
Prophylactic antibiotics selection
Antibiotics discontinued <48 hours

Outcome Measures (3 items)
Inpatient mortality rate
Postop hemorrhage/hematoma
Postop physiologic/metabolic derangement



CABG Composite Scores in HQID – Calculation of the Process Component Score

- Based on an “opportunity model”
- Each time a patient is eligible to receive a care process, there is an “opportunity” for the hospital to deliver required care
- The hospital’s score for the process component is “the percent of opportunities for which the hospital delivered the required care”

CABG Composite Scores in HQID – Calculation of the Process Component Score

Hypothetical example with N = 10 patients

Aspirin at Discharge	Antibiotics Initiated	Antibiotics Selection	Antibiotics Discontinued
9 / 9 (100%)	9 / 10 (90%)	10 / 10 (100%)	9 / 9 (100%)

$$\frac{9+9+10+9}{9+10+10+9} = 37/38 = 97.4\%$$

CABG Composite Scores in HQID – Calculation of Outcome Component

- **Risk-adjusted using 3M™ APR-DRG™ model**
- **Based on ratio of observed / expected outcomes**
- **Outcomes measures are:**
  - Survival index
  - Avoidance index for hematoma/hemmorhage
  - Avoidance index for physiologic/metabolic derangement

CABG Composite Scores in HQID – Calculation of Outcome Component – Survival Index

$$\text{survival index} = \frac{\text{observed \# of patients surviving}}{\text{expected \# of patients surviving}}$$

Interpretation:

- index <1 implies worse-than-expected survival
- index >1 implies better-than-expected survival

(Avoidance indexes have analogous definition & interpretation)

CABG Composite Scores in HQID – Combining Process and Outcomes

"Equal weight for each measure"

- 4 process measures
- 3 outcome measures
- each individual measure is weighted 1 / 7

4 / 7 x Process Score +

1 / 7 x survival index +

1 / 7 x avoidance index for hemorrhage/hematoma +

1 / 7 x avoidance index for physiologic derangment

= Overall Composite Score

Strengths & Limitations

- **Advantages:**
  - Simple
  - Transparent
  - Avoids subjective weighting
- **Disadvantages:**
  - Ignores uncertainty in performance measures
  - Not able to calculate confidence intervals
- **An Unexpected Feature:**
  - Heavily weighted toward process measures
  - As shown below…

CABG Composite Scores in HQID – Exploring the Implications of Equal Weighting

- HQID performance measures are publicly reported for the top 50% of hospitals
- Used these publicly reported data to study the weighting of process vs. outcomes

Publicly Reported HQID Data – CABG Year 1

Process Measures

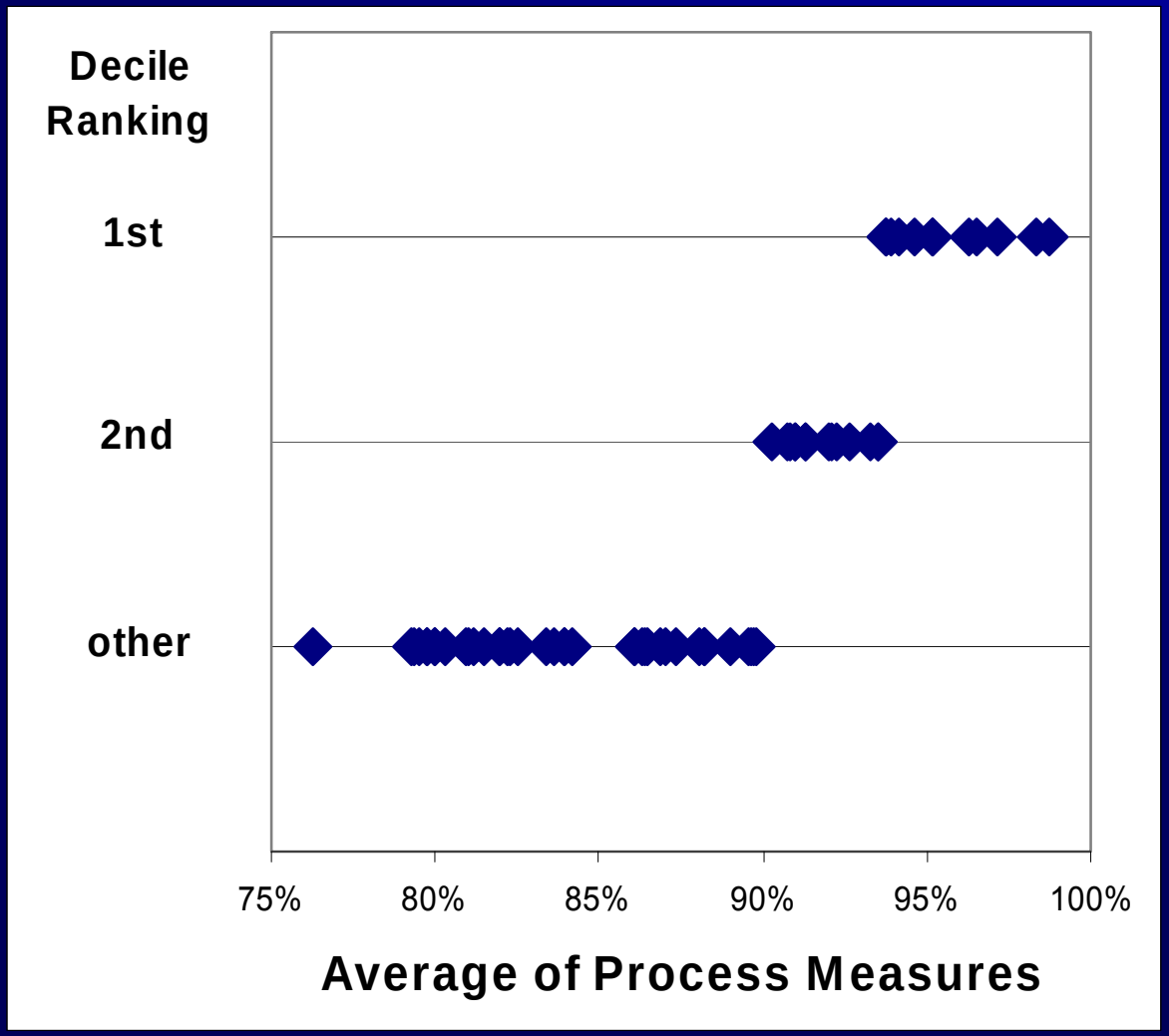
Outcome Measures

ISOLATED CORONARY ARTERY BYPASS GRAFT
CMS/Premier Hospital Quality Incentive Demonstration Project - Year 1
Top 50 % of Participants in Isolated Coronary Artery Bypass Graft (CABG)

***Hospital in top ten (10) percent of participating hospitals**
****Hospital in top twenty (20) percent of participating hospitals**
+ Estimated hospital placement, data omitted due to transmission error
Date range - Acute care inpatient discharges from October 1, 2003 - September 30, 2004
Low Sample (10 or Less) = Hospital provided service, but had ten (10) eligible patients or less during this date range
Data sorted by State (ascending order) then City (ascending order)

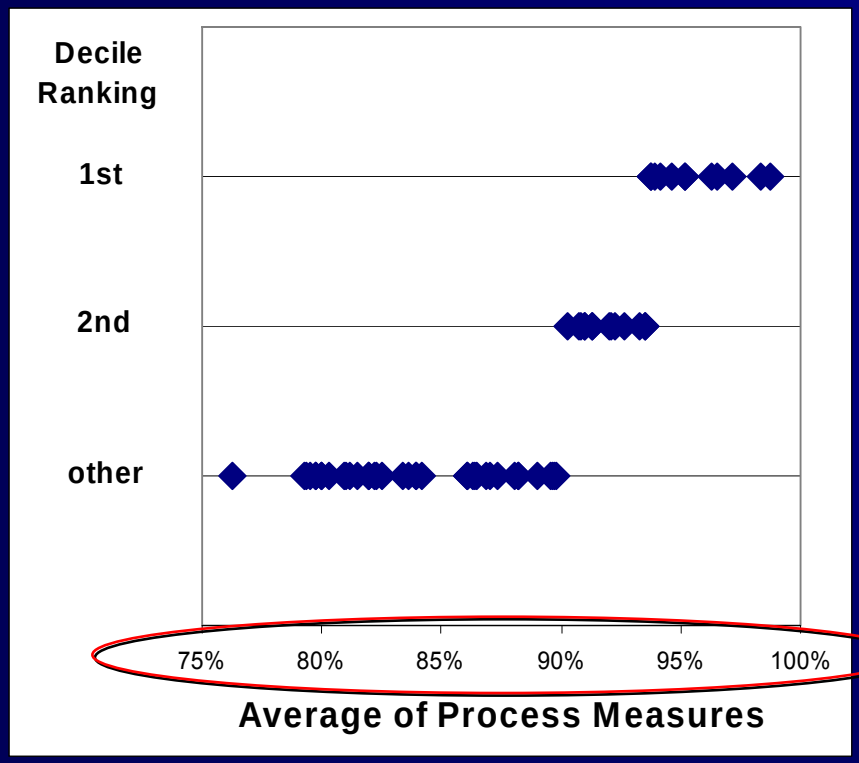
State	City	Hospital	Medicare Provider #	Aspirin prescribed at discharge	Prophylactic antibiotic received within 1 hour prior to surgical incision	Prophylactic antibiotic selection for surgical patients	Prophylactic antibiotic discontinued within 24 hours after surgery end time	Post-op physiologic /metabolic derangement avoidance index	Post-op hemorrhage /hematoma avoidance index	Survival Index	Total Case Count
				% Patients Received	% Patients Received	% Patients Received	% Patients Received	Occurrence rate expressed as Avoidance Index, can exceed 100%	Occurrence rate expressed as Avoidance Index. can exceed 100%	Mortality rate expressed as Survival Index. can exceed 100%	
AL	Dothan	SOUTHEAST ALABAMA MEDICAL CENTER	010001	97.80%	87.62%	98.73%	61.86%	100.00%	99.93%	100.10%	360
AL	Opelika	EAST ALABAMA MEDICAL CENTER AND SNF*	010029	100.00%	96.79%	99.20%	99.06%	99.92%	99.93%	98.63%	271
CA	Fullerton	ST JUDE MEDICAL CENTER	050168	93.38%	72.96%	98.74%	59.49%	99.93%	99.94%	100.74%	185
CA	Glendale	GLENDALE ADVENTIST MEDICAL CENTER**	050239	96.61%	78.76%	100.00%	88.54%	99.74%	99.94%	101.61%	128
CA	Lynwood	ST FRANCIS MEDICAL CENTER**	050104	92.94%	87.65%	100.00%	92.50%	100.00%	99.94%	98.48%	90
CA	Mission Viejo	MISSION HOSPITAL REGIONAL MEDICAL CENTER*	050567	99.44%	95.65%	98.91%	94.57%	100.00%	100.00%	100.02%	196
CA	Orange	ST JOSEPH HOSPITAL	050069	99.31%	59.73%	80.54%	90.34%	100.00%	100.00%	101.66%	155
CA	Rancho Mirage	EISENHOWER MEDICAL CENTER	050573	96.55%	86.41%	100.00%	22.05%	99.95%	100.00%	100.00%	206
CO	Grand Junction	ST MARYS HOSPITAL AND MEDICAL CENTER*	060023	95.12%	92.59%	98.77%	88.46%	100.00%	100.00%	101.54%	86
FL	Miami	SOUTH MIAMI HOSPITAL	100154	92.50%	90.00%	100.00%	44.58%	w Sample (10 or Less)	100.00%	106.21%	98
FL	Tampa	ST JOSEPH'S HOSPITAL	100075	87.67%	67.91%	99.29%	94.44%	100.00%	100.00%	102.15%	289
FL	Venice	VENICE REGIONAL MEDICAL CENTER**	100070	90.82%	85.44%	97.67%	100.00%	100.00%	100.00%	101.23%	109
HI	Honolulu	KUAKINI MEDICAL CENTER	120007	100.00%	92.86%	100.00%	35.14%	99.83%	99.94%	96.27%	121
IA	Mason City	MERCY MEDICAL CENTER-NORTH IOWA**	160064	100.00%	87.88%	98.18%	77.02%	100.00%	100.00%	101.28%	185
KY	Lexington	CENTRAL BAPTIST HOSPITAL	180103	95.00%	91.60%	98.82%	60.58%				912

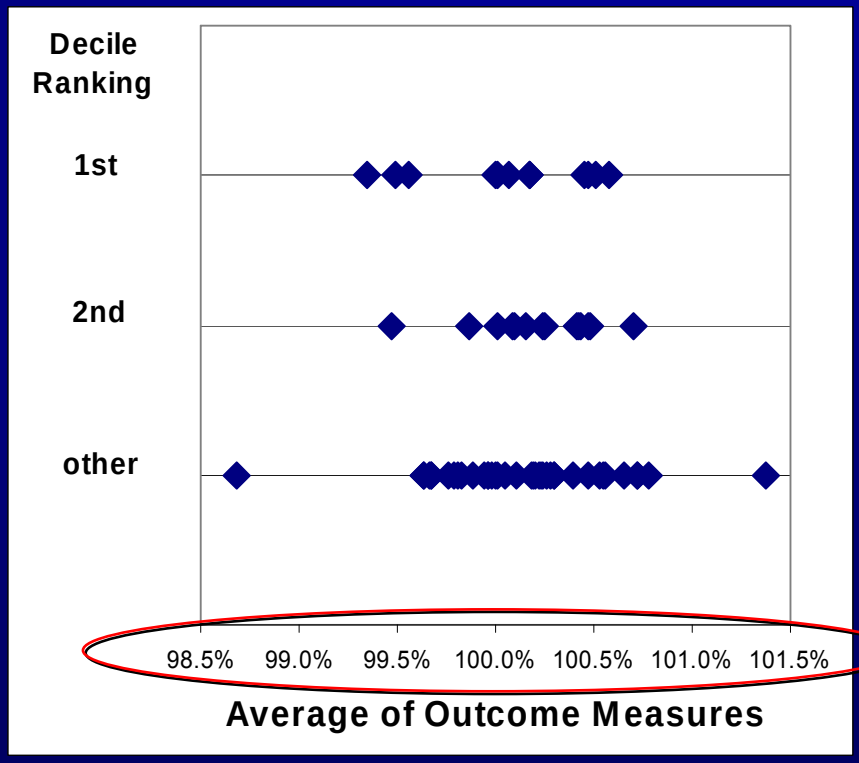
Process Performance vs. Overall Composite Decile Ranking



Outcome Performance vs. Overall Composite Decile Ranking

Explanation:
Process Measures Have Wider Range of Values





- The amount that outcomes can increase or decrease the composite score is small relative to process measures

Process vs. Outcomes: Conclusions

- **Outcomes will only have an impact if a hospital is on the threshold between a better and worse classification**
- **This weighting may have advantages**
  - Outcomes can be unreliable
    - *Chance variation*
    - *Imperfect risk-adjustment*
  - Process measures are actionable
- **Not transparent**

Lessons from HQID

- **Equal weighting may not behave like it sounds**
- **If you prefer to emphasize outcomes, must account for unequal measurement scales, e.g.**
  - standardize the measures to a common scale
  - or weight process and outcomes unequally

Goals for STS Composite Measure

- **Heavily weight outcomes**
  - Use statistical methods to account for small sample sizes & rare outcomes
- **Make the implications of the weights as transparent as possible**
- **Assess whether inferences about hospital performance are sensitive to the choice of statistical / weighting methods**

Outline

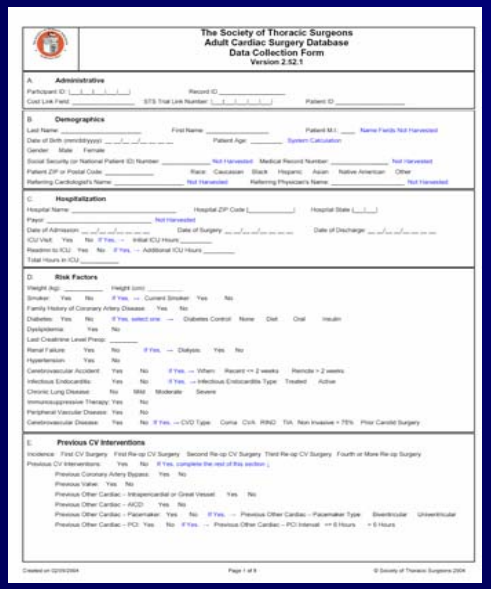
- Measure selection
- Data
- Latent variable approach to composite measures
- STS approach to composite measures

The STS Composite Measure for CABG – Criteria for Measure Selection

- **Use Donabedian model of quality**
  - Structure, process, outcomes
- **Address three temporal domains**
  - Preoperative, intraoperative, postoperative
- **Choose measures that meet various criteria for validity**
  - Adequately risk-adjusted
  - Adequate data quality

The STS Composite Measure for CABG – Criteria for Measure Selection

Captured In STS



The Society of Thoracic Surgeons
Adult Cardiac Surgery Database
Data Collection Form
Version 2.52.1

A Administrative

B Demographics

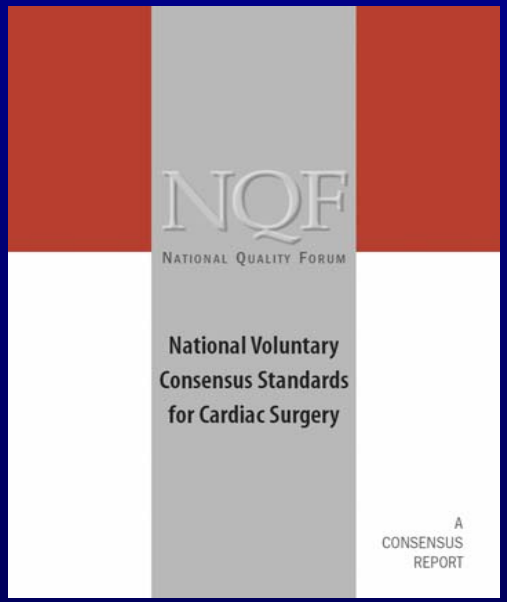
C Hospitalization

D Risk Factors

E Previous CV Interventions

+

Endorsed by NQF



Process Measures

- Internal mammary artery (IMA)
- Preoperative betablockers
- Discharge antiplatelets
- Discharge betablockers
- Discharge antilipids

Risk-Adjusted Outcome Measures

- Operative mortality
- Prolonged ventilation
- Deep sternal infection
- Permanent stroke
- Renal failure
- Reoperation

NQF Measures Not Included In Composite

- **Inpatient Mortality**
  - Redundant with operative mortality
- **Participation in a Quality Improvement Registry**
- **Annual CABG Volume**

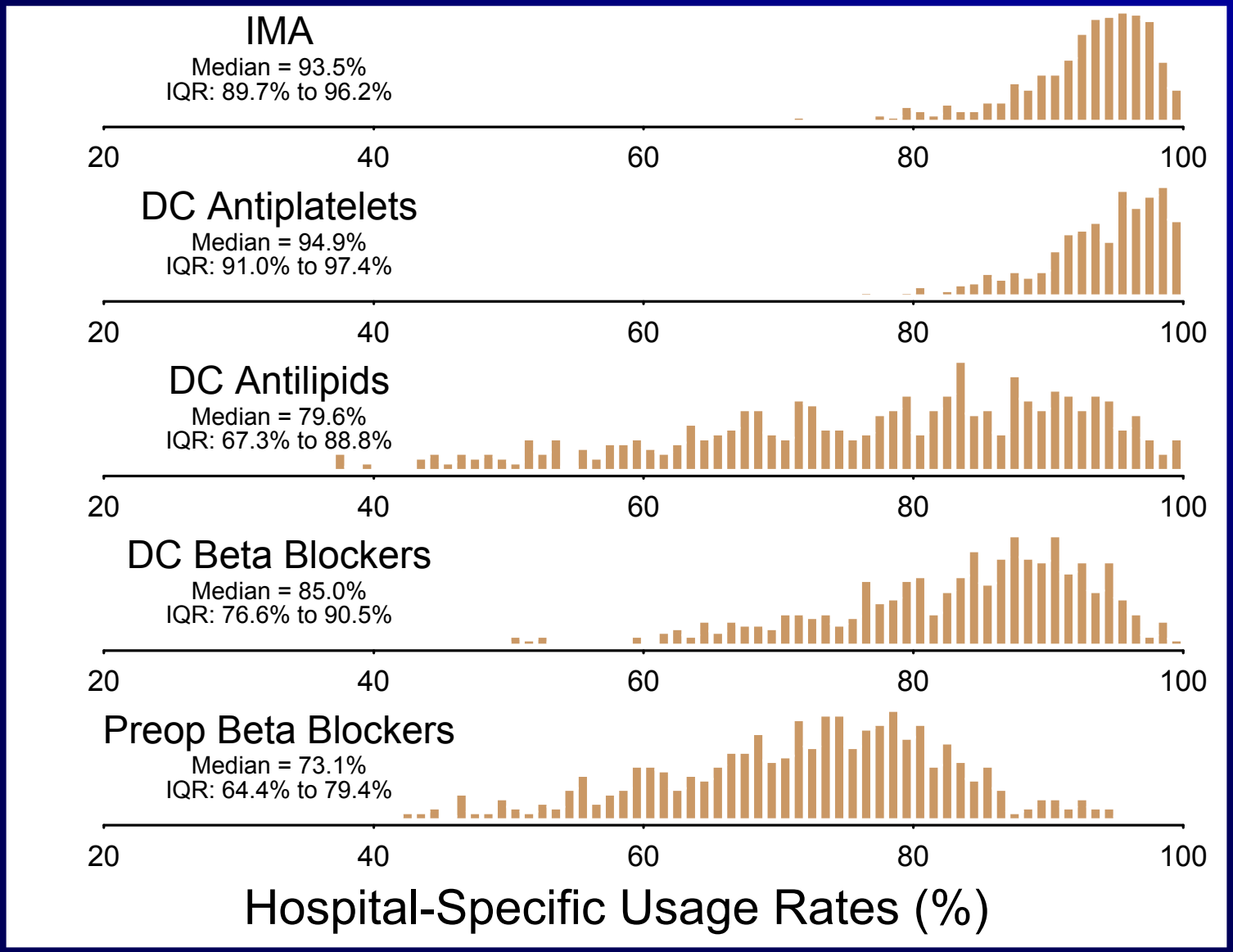
Other Measures Not Included in Composite

- **HQID measures, not captured in STS**
  - Antibiotics Selection & Timing
  - Post-op hematoma/hemmorhage
  - Post-op physiologic/metabolic derangment
- **Structural measures**
- **Patient satisfaction**
- **Appropriateness**
- **Access**
- **Efficiency**

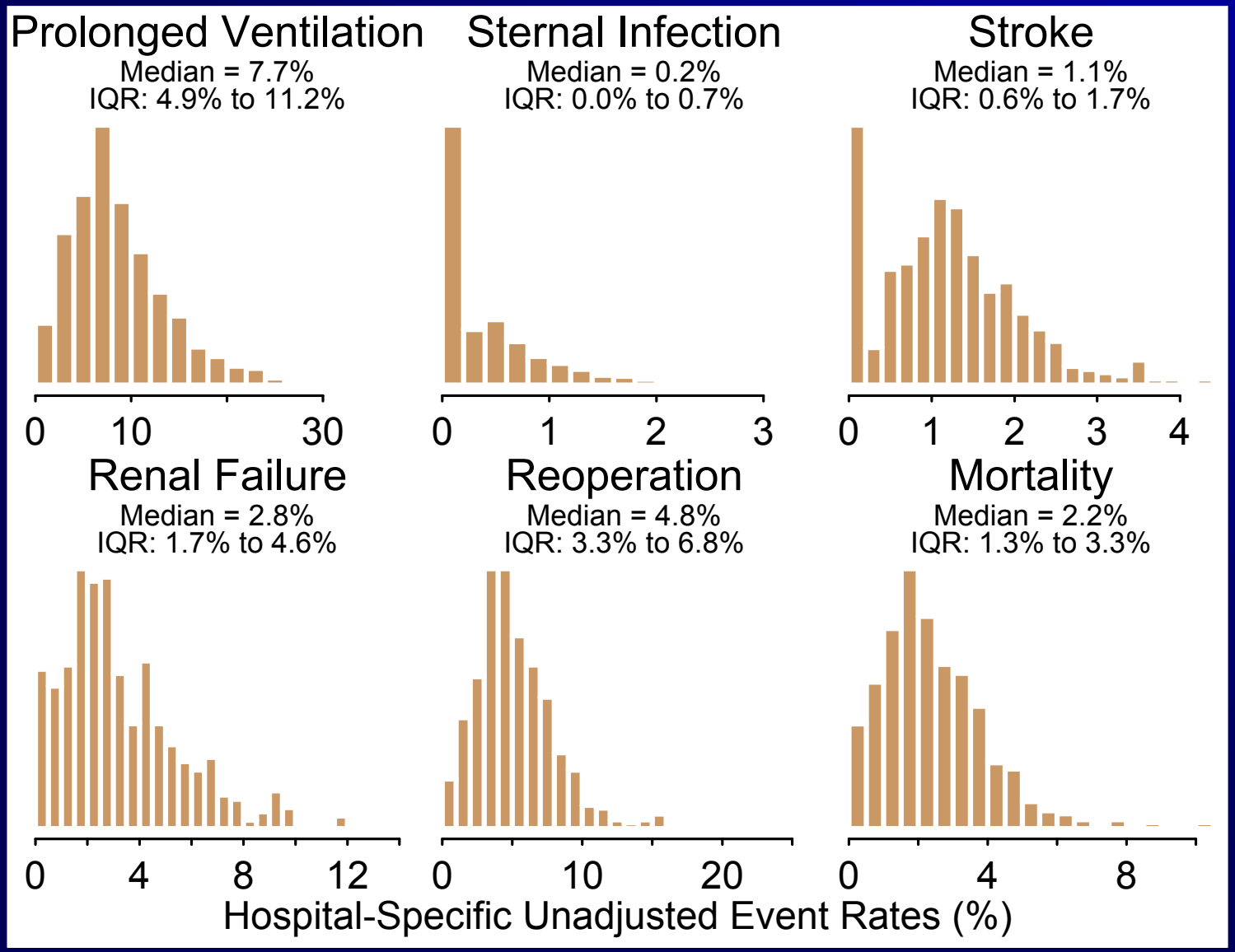
Data

- **STS database**
- **133,149 isolated CABG operations during 2004**
- **530 providers**
- **Inclusion/exclusion:**
  - Exclude sites with >5% missing data on any process measures
  - For discharge meds– exclude in-hospital mortalities
  - For IMA usage – exclude redo CABG
- **Impute missing data to negative (e.g. did not receive process measure)**

Distribution of Process Measures in STS

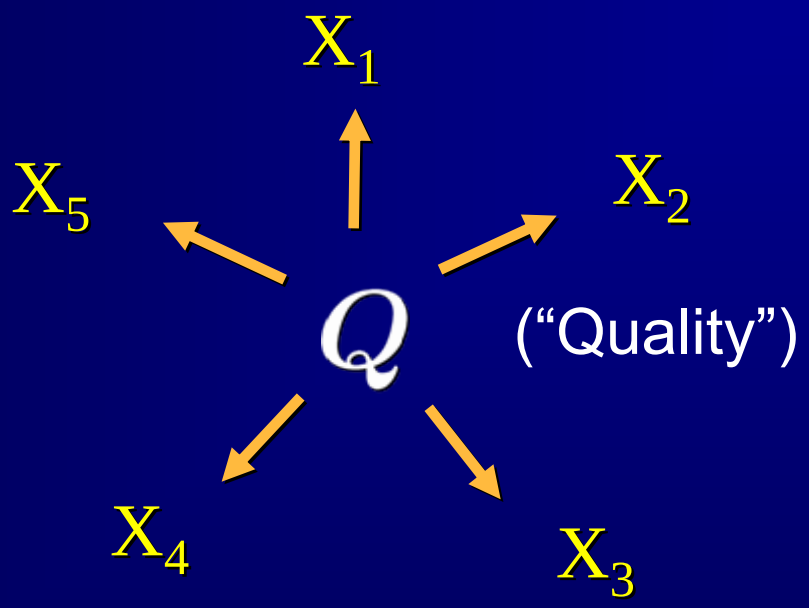


Distribution of Outcomes Measures in STS



Latent Variable Approach to Composite Measures

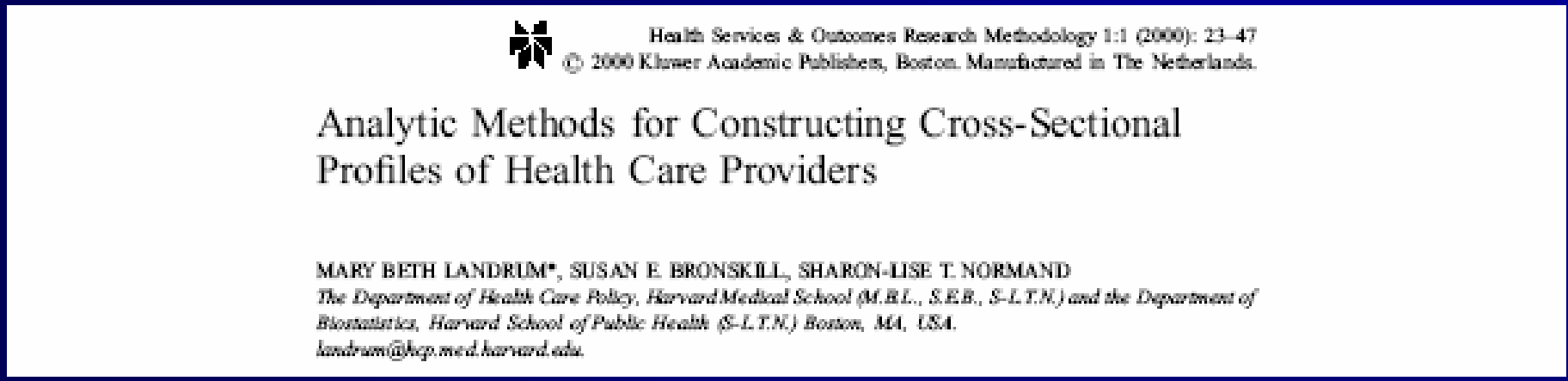
- **Psychometric approach**
  - Quality is a “latent variable”
    - *Not directly measurable*
    - *Not precisely defined*
  - Quality indicators are the observable manifestations of this latent variable
  - Goal is to use the observed indicators to make inferences about the underlying latent trait


X_1
X_2
X_3
X_4
X_5
Q
("Quality")

Common Modeling Assumptions

- **Case #1: A single latent trait**
  - All variables measure the same thing (unidimensionality)
  - Variables are highly correlated (internal consistency)
  - Imperfect correlation is due to random measurement error
  - Can compensate for random measurement error by collecting lots of variables and averaging them
- **Case #2: More than a single latent trait**
  - Can identify clusters of variables that describe a single latent trait (and meet the assumptions of Case #1)
  - NOTE: Measurement theory does not indicate how to reduce multiple distinct latent traits into a single dimension
    - *Beyond the scope of measurement theory*
    - *Inherently normative, not descriptive*

Models for A Single Latent Trait



Health Services & Outcomes Research Methodology 1:1 (2000): 23–47

Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers

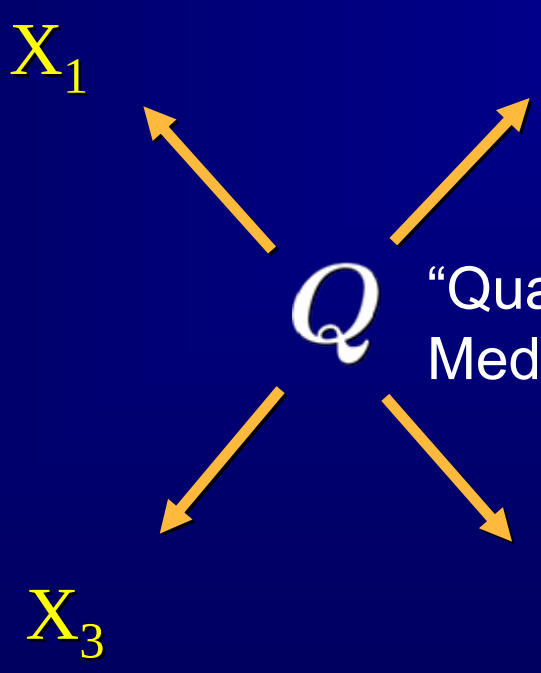
MARY BETH LANDRUM*, SUSAN E. BRONSKILL, SHARON-LISE T. NORMAND
The Department of Health Care Policy, Harvard Medical School (M.B.L., S.E.B., S-L.T.N.) and the Department of Biostatistics, Harvard School of Public Health (S-L.T.N.) Boston, MA, USA.
landrum@hcp.med.harvard.edu.

“Latent Trait Logistic Model”
Landrum et al. 2000

Example of latent trait logistic model applied to 4 medication measures

(Preop Betablocker)

(Discharge Betablocker)

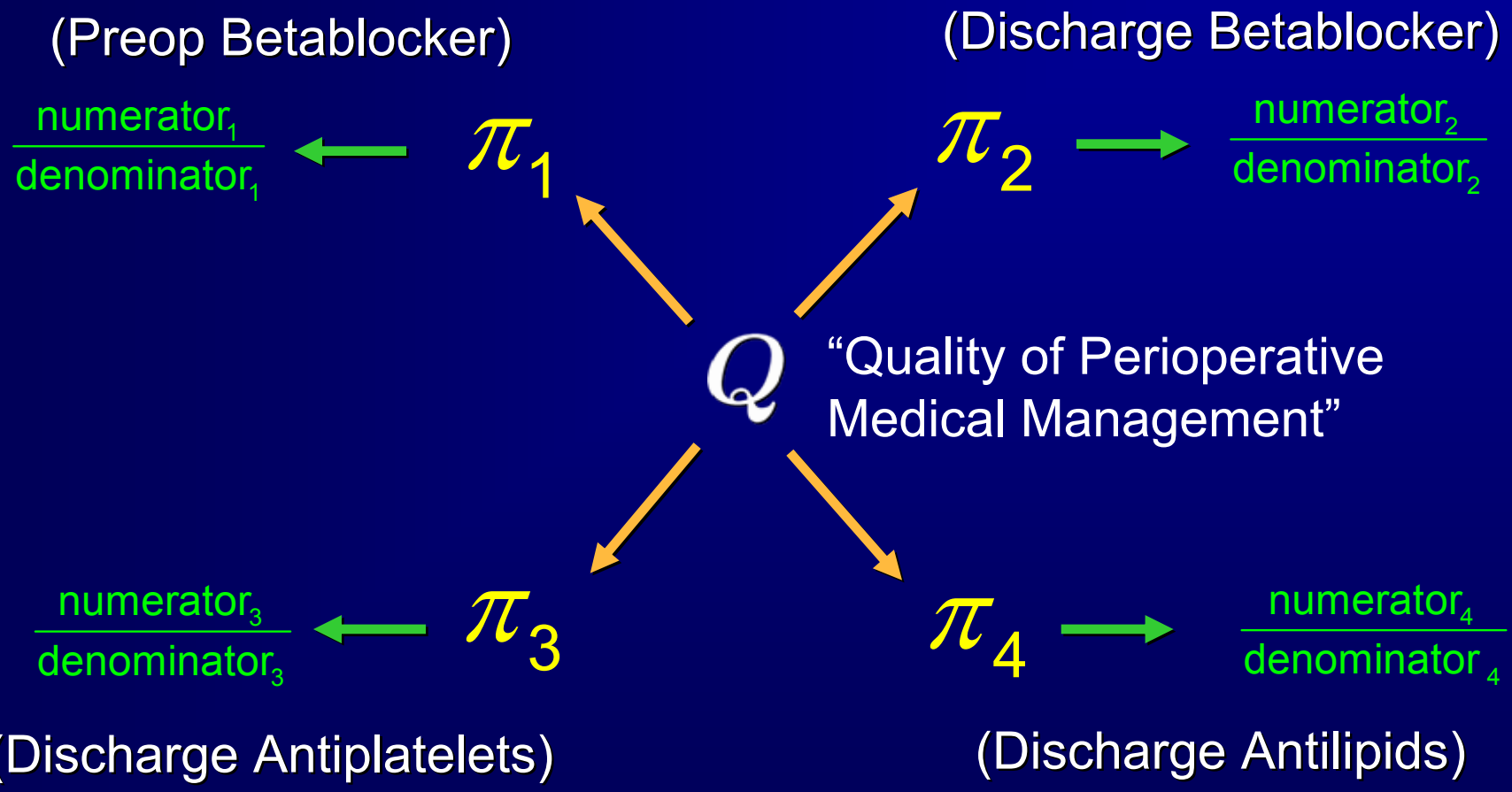


"Quality of Perioperative Medical Management"

(Discharge Antiplatelets)

(Discharge Antilipids)

Example of latent trait logistic model applied to 4 medication measures



π denotes underlying true probability

Technical Details of Latent Trait Analysis

$$\text{(preop betablockers)} \quad \log[\pi_1/(1-\pi_1)] = \alpha_1 + \beta_1 Q$$

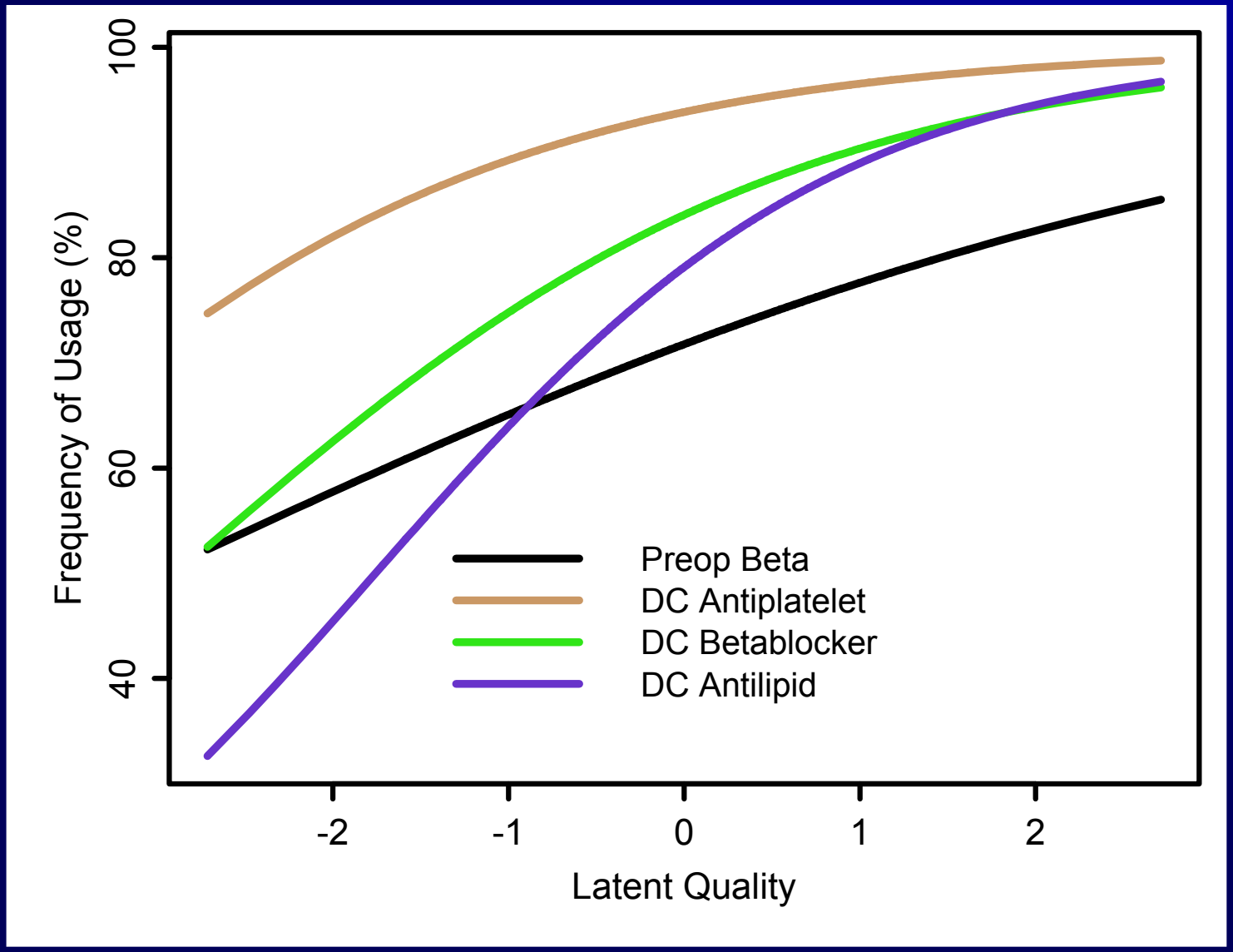
$$\text{(discharge betablockers)} \quad \log[\pi_2/(1-\pi_2)] = \alpha_2 + \beta_2 Q$$

$$\text{(discharge antiplatelets)} \quad \log[\pi_3/(1-\pi_3)] = \alpha_3 + \beta_3 Q$$

$$\text{(discharge antilipids)} \quad \log[\pi_4/(1-\pi_4)] = \alpha_4 + \beta_4 Q$$

- Q is an unobserved latent variable
- Goal is to estimate Q for each participant
- Use observed numerators and denominators

Latent trait logistic model



Latent Trait Analysis

Advantages:

- **Quality can be estimated efficiently**
  - Concentrates information from multiple variables into a single parameter
- **Avoids having to determine weights**

Latent Trait Analysis

Disadvantages:

- **Hard for sites to know where to focus improvement efforts because weights are not stated explicitly**
- **Strong modeling assumptions**
  - A single latent trait (unidimensionality)
  - Latent trait is normally distributed
  - One major assumption is not stated explicitly but can be derived by examining the model
    - *100% correlation between the individual items*
    - *A very unrealistic assumption!!*

Model did not fit the data

Table 1. Correlation between hospital log-odds parameters under IRT model

	DISCHARGE ANTILIPIDS	DISCHARGE BETABLOCKER	PREOPERATIVE BETABLOCKER
DISCHARGE ANTIPLATELETS	1.00	1.00	1.00
DISCHARGE ANTILIPIDS		1.00	1.00
DISCHARGE BETABLOCKER			1.00

Table 2. Estimated correlation between hospital log-odds parameters

	DISCHARGE ANTILIPIDS	DISCHARGE BETABLOCKER	PREOPERATIVE BETABLOCKER
DISCHARGE ANTIPLATELETS	0.38	0.30	0.15
DISCHARGE ANTILIPIDS		0.34	0.19
DISCHARGE BETABLOCKER			0.50

Model Also Did Not Fit When Applied to Outcomes

	INFEC	STROKE	RENAL	REOP	MORT
VENT	0.46	0.15	0.49	0.49	0.50
INFECT		0.16	0.16	0.54	0.65
STROKE			0.40	0.43	0.43
RENAL				0.44	0.54
REOP					0.61

Latent Trait Analysis – Conclusions

- **Model did not fit the data!**
- **Each measure captures something different**
  - # latent variables = # of measures?
- **Cannot use latent variable models to avoid choosing weights**

The STS Composite Method

The STS Composite Method

Step 1. Quality Measures are Grouped Into 4 Domains

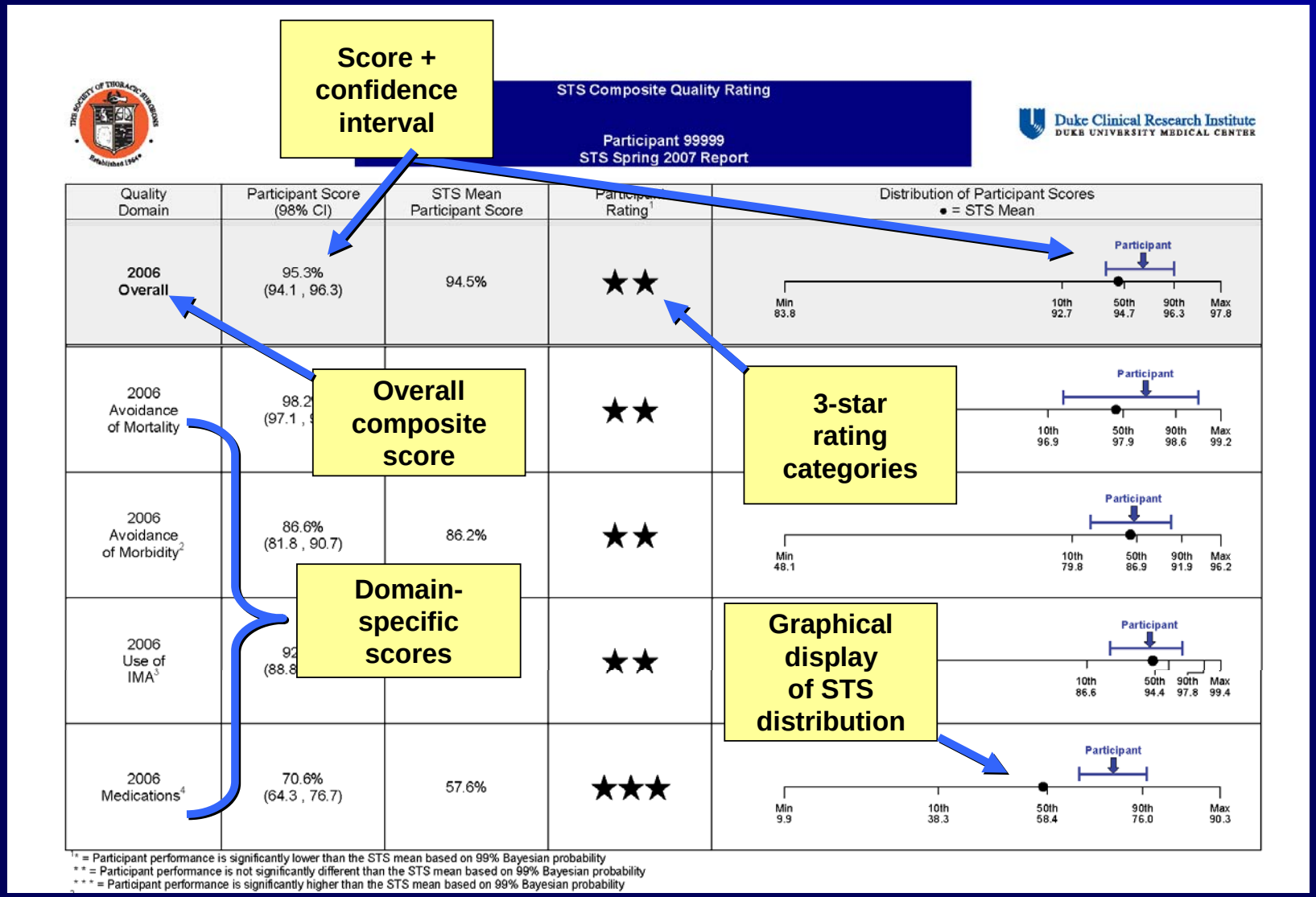
Step 2. A Summary Score is Defined for Each Domain

Step 3. Hierarchical Models Are Used to Separate True Quality Differences From Random Noise and Case Mix Bias

Step 4. The Domain Scores are Standardized to a Common Scale

Step 5. The Standardized Domain Scores are Combined Into an Overall Composite Score by Adding Them

Preview: The STS Hospital Feedback Report



Step 1. Quality Measures Are Grouped Into Four Domains

Perioperative Medical Care Bundle
Preop B-blocker
Discharge B-blocker
Discharge Antilipids
Discharge ASA

Operative Technique
IMA Usage

Risk-Adjusted Mortality Measure
Operative Mortality

Risk-Adjusted Morbidity Bundle
Stroke
Renal Failure
Reoperation
Sternal Infection
Prolonged Ventilation

Of Course Other Ways of Grouping Items Are Possible…

*Taxonomy of Animals in a Certain Chinese Encyclopedia**

a) **Those that belong to the Emperor**
b) **Embalmed ones**
c) **Tame ones**
d) **Suckling pigs**
e) **Sirens**
f) **Fabulous ones**
g) **Stray dogs**
h) **Those included in the present classification**
i) **Frenzied ones**
j) **Innumerable ones**
k) **Those drawn with a very fine camelhair brush**
l) **Others**
m) **Those that have just broken a water pitcher**
n) **Those that from a long way off look like flies**

*According to Michel Foucault, *The Order of Things*, 1966

Step 2. A Summary Measure Is Defined for Each Domain

Perioperative Medical Care Bundle

Operative Technique

Risk-Adjusted Mortality Measure

Risk-Adjusted Morbidity Bundle

- **Medications**
  - "all-or-none" composite endpoint

    Proportion of patients who received ALL four medications (except where contraindicated)
- **Morbidities**
  - "any-or-none" composite endpoint

    Proportion of patients who experienced AT LEAST ONE of the five morbidity endpoints

All-Or-None / Any-Or-None

Advantages:

- **No need to determine weights**
- **Reflects important values**
  - Emphasizes systems of care
  - Emphasizes high benchmark
- **Simple to analyze statistically**
  - Using methods for binary (yes/no) endpoints

Disadvantages:

- **Choice to treat all items equally may be criticized**

Step 2. A Summary Measure Is Defined for Each Domain

Perioperative Medical Care Bundle	Operative Technique	Risk-Adjusted Mortality Measure	Risk-Adjusted Morbidity Bundle
Proportion of patients who received all 4 medications	Proportion of patients who received an IMA	Proportion of patients who experienced operative mortality	Proportion of patients who experienced at least one major morbidity

Step 3. Use Hierarchical Models to Separate True Quality Differences from Random Noise

- **proportion of successful outcomes**

  = numerator / denominator

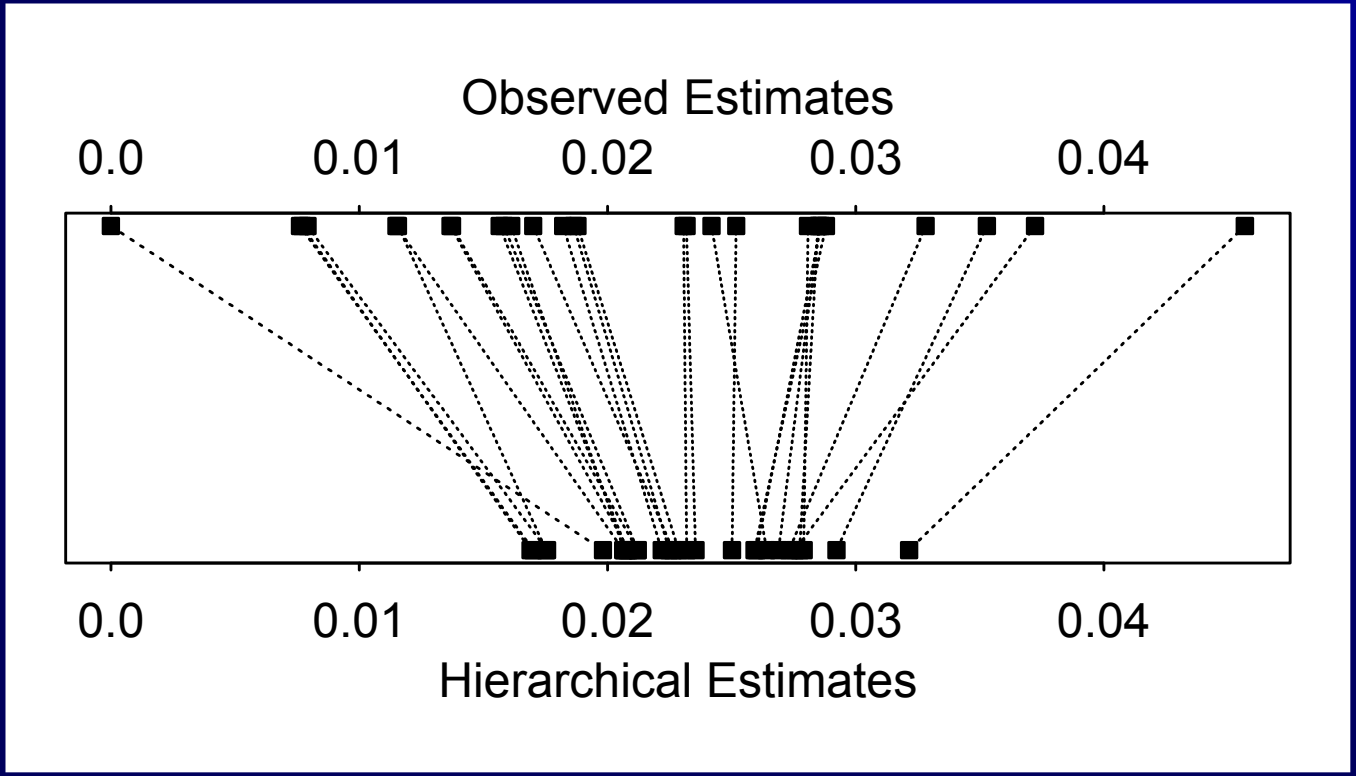
  = “true probability” + random error

- **Hierarchical models estimate the true probabilities**

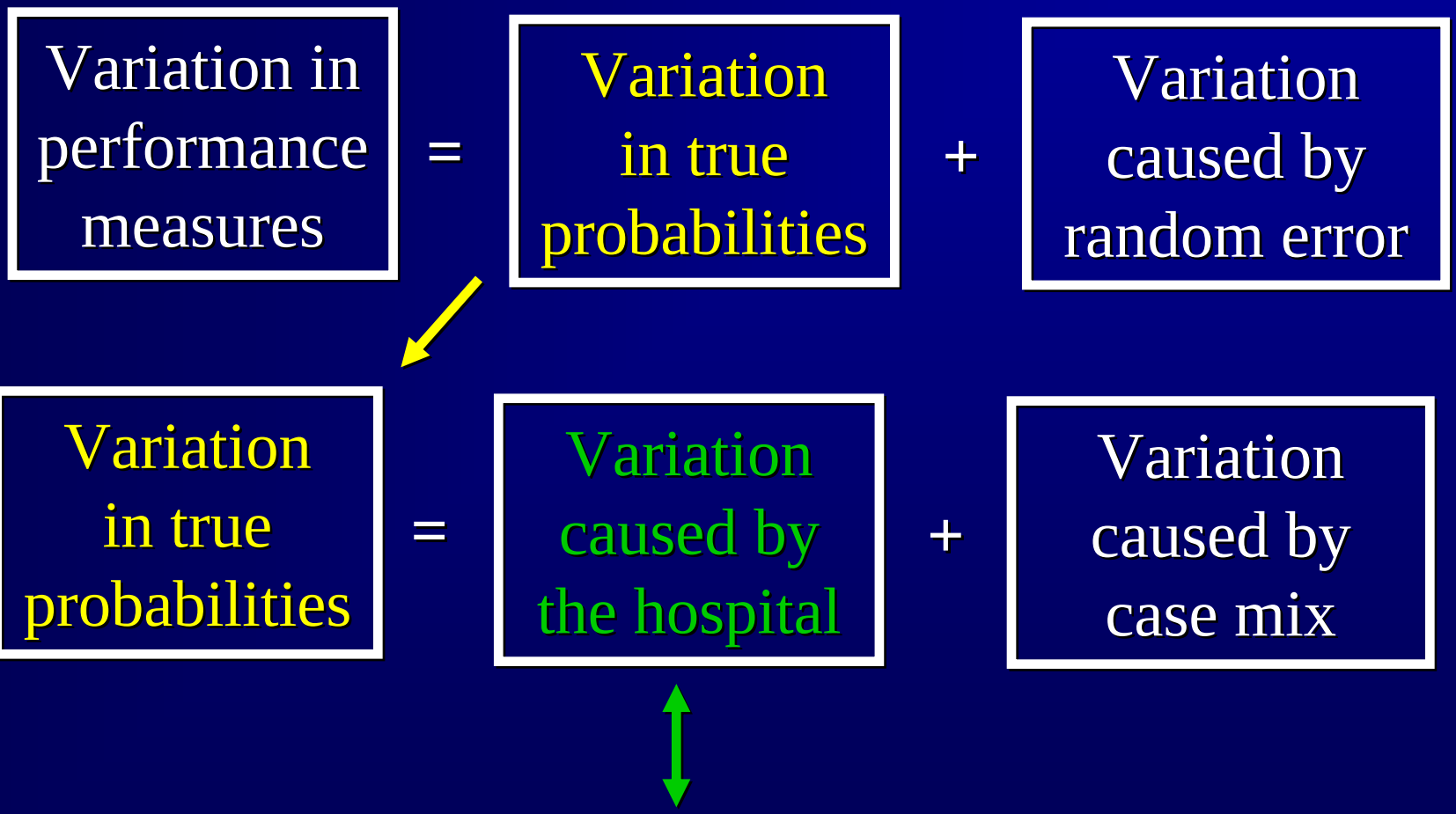
Variation in performance measures	=	Variation in true probabilities	+	Variation caused by random error

Example of Hierarchical Models

Figure. Mortality Rates in a Sample of STS Hospitals



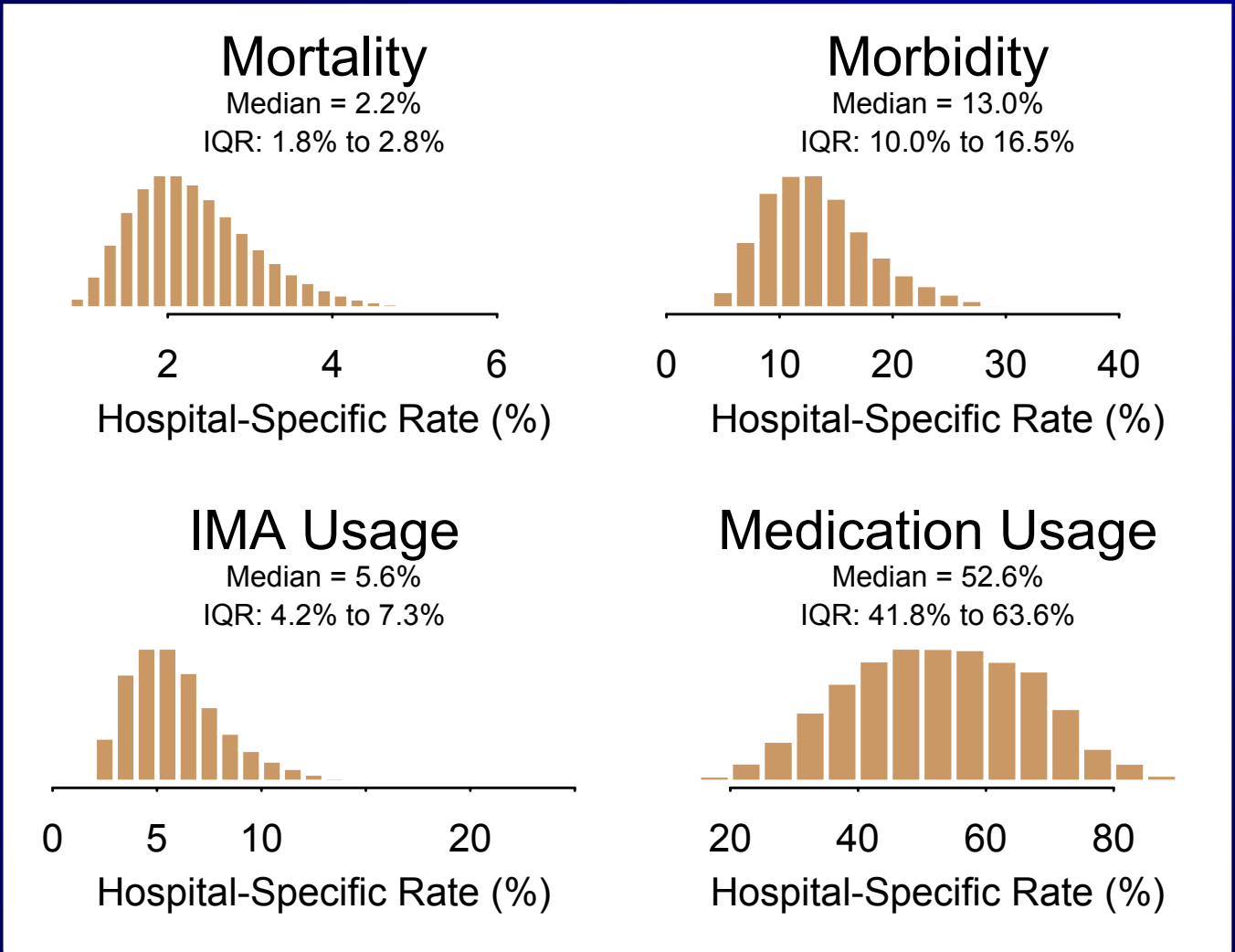
Step 3. Use Hierarchical Models to Separate True Quality Differences from Case Mix



Advantages of Hierarchical Model Estimates

- **Less variable than a simple proportion**
  - Shrinkage
- **Borrows information across hospitals**
  - Our version also borrows information across measures
- **Adjusts for case mix differences**

Estimated Distribution of True Probabilities (Hierarchical Estimates)



Step 4. The Domain Scores Are Standardized to a Common Scale

Step 4a. Consistent Directionality

Directionality…

Needs to be consistent in order to sum the measures

Solution…

Measure success instead of failure

Worse **Better**

IMA usage rate
All-or-none medication adherence

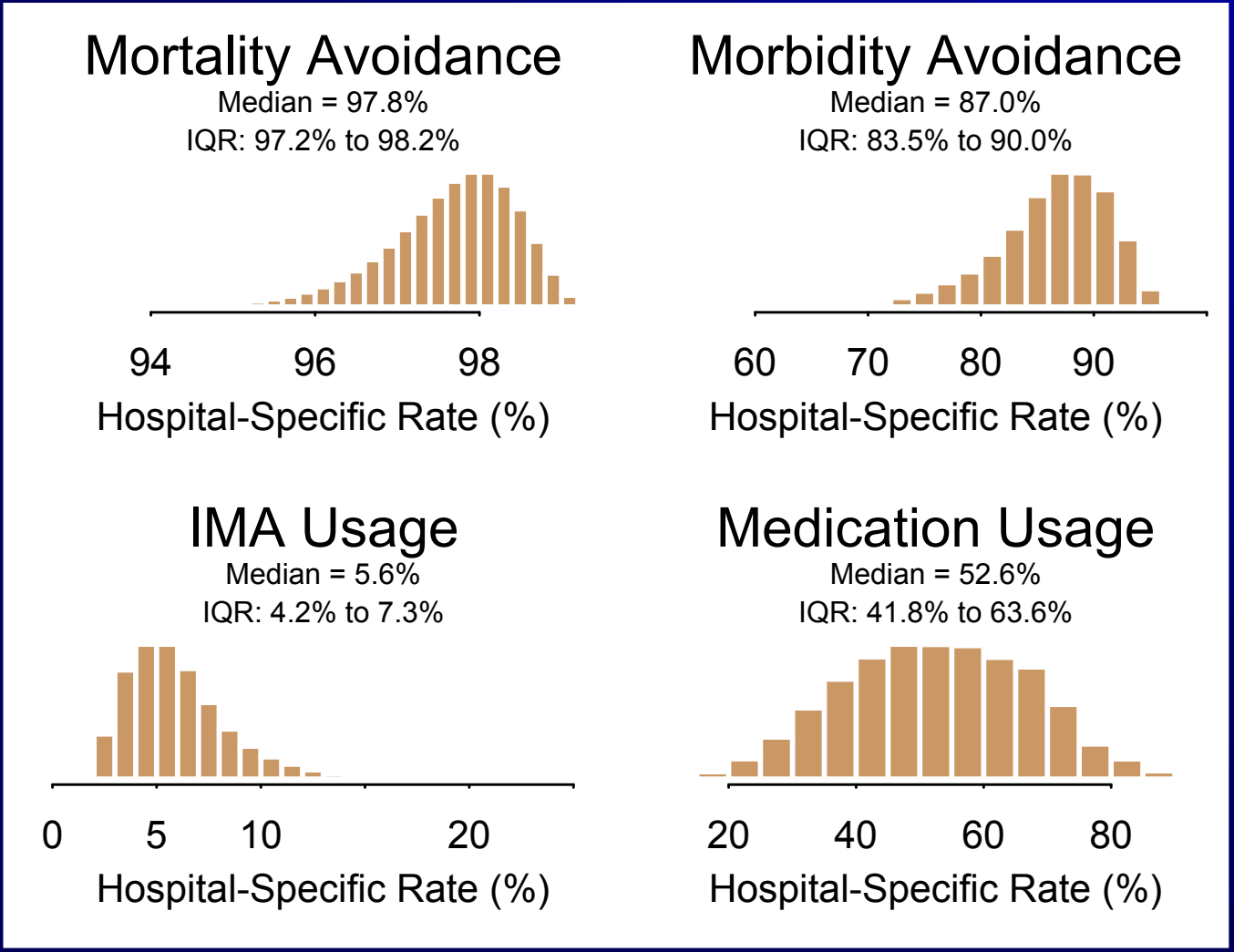
Better **Worse**

Risk-Adjusted Mortality Rate
Risk-Adjusted Any-Morbidity Rate

$$\text{Probability of NO mortality} = 1 - \text{Probability of mortality}$$

$$\text{Probability of NO morbidity} = 1 - \text{Probability of morbidity}$$

Step 4a. Consistent Directionality



Step 4b. Standardization

Each measure is re-scaled by dividing by its standard deviation (sd)

Notation

π_{meds} = Probability of receiving all medications

π_{IMA} = Probability of receiving an IMA

π_{mort} = Probability of NO operative mortality

π_{morb} = Probability of NO major morbidity

Step 4b. Standardization

Each measure is re-scaled by dividing by its standard deviation (sd)

$$\text{standardized meds measure} = \pi_{\text{meds}} / \text{sd}_{\text{meds}}$$

$$\text{standardized IMA measure} = \pi_{\text{IMA}} / \text{sd}_{\text{IMA}}$$

$$\text{standardized mort measure} = \pi_{\text{mort}} / \text{sd}_{\text{mort}}$$

$$\text{standardized morb measure} = \pi_{\text{morb}} / \text{sd}_{\text{morb}}$$

Step 5. The Standardized Domain Scores Are Combined By Adding Them

$$\text{Composite} = \left(\frac{\hat{\pi}_{\text{mort}}}{\text{sd}_{\text{mort}}}\right) + \left(\frac{\hat{\pi}_{\text{morb}}}{\text{sd}_{\text{morb}}}\right) + \left(\frac{\hat{\pi}_{\text{IMA}}}{\text{sd}_{\text{IMA}}}\right) + \left(\frac{\hat{\pi}_{\text{meds}}}{\text{sd}_{\text{meds}}}\right)$$

where $\hat{\pi}$ denotes the hierarchical estimate of π

Step 5. The Standardized Domain Scores Are Combined By Adding Them

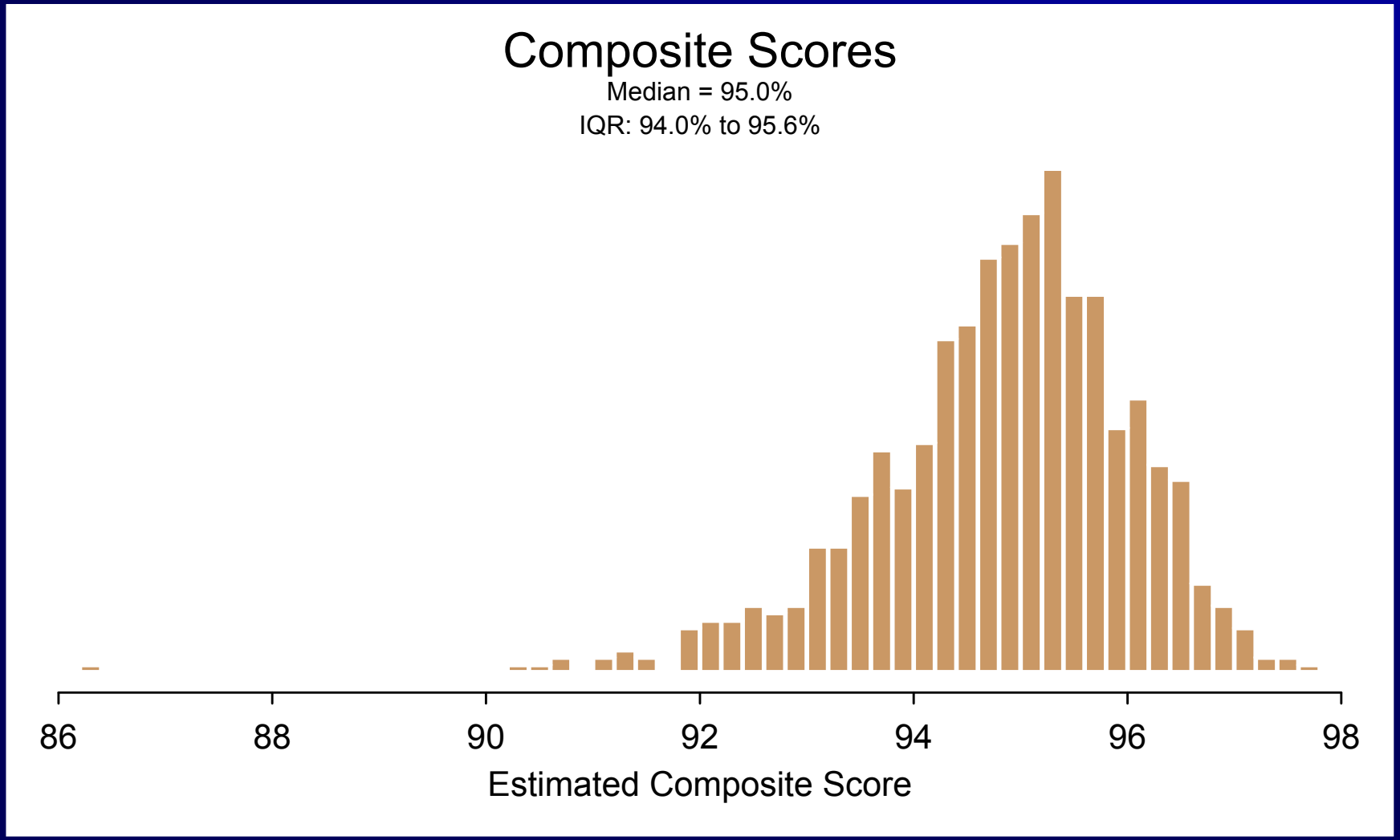
…then rescaled again (for presentation purposes)

$$\text{Composite} = \frac{1}{c} \times \left[\left(\frac{\hat{\pi}_{\text{mort}}}{\text{sd}_{\text{mort}}} \right) + \left(\frac{\hat{\pi}_{\text{morb}}}{\text{sd}_{\text{morb}}} \right) + \left(\frac{\hat{\pi}_{\text{IMA}}}{\text{sd}_{\text{IMA}}} \right) + \left(\frac{\hat{\pi}_{\text{meds}}}{\text{sd}_{\text{meds}}} \right) \right]$$

$$\text{where } c = \left(\frac{1}{\text{sd}_{\text{mort}}} + \frac{1}{\text{sd}_{\text{morb}}} + \frac{1}{\text{sd}_{\text{IMA}}} + \frac{1}{\text{sd}_{\text{meds}}} \right)$$

(This guarantees that final score will be between 0 and 100.)

Distribution of Composite Scores



(Fall 2007 harvest data. Rescaled to lie between 0 and 100.)

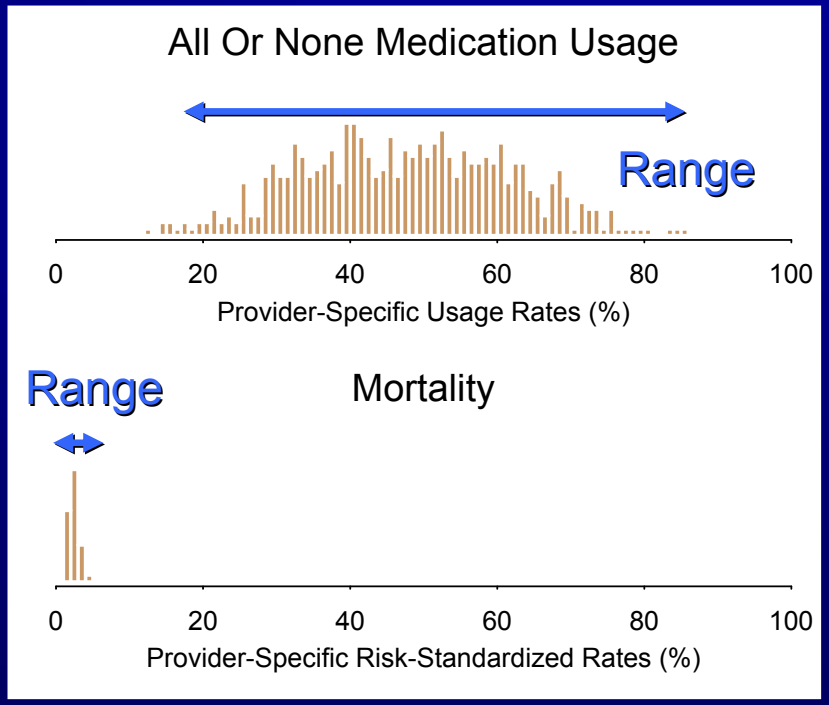
Goals for STS Composite Measure

- **Heavily weight outcomes**
  - Use statistical methods to account for small sample sizes & rare outcomes
- **Make the implications of the weights as transparent as possible**
- **Assess whether inferences about hospital performance are sensitive to the choice of statistical / weighting methods**

Exploring the Implications of Standardization

If items were NOT standardized

- **Items with a large scale would disproportionately influence the score**
  - example: medications would dominate mortality
- **A 1% improvement in mortality would have the same impact as 1% improvement in any other domain**



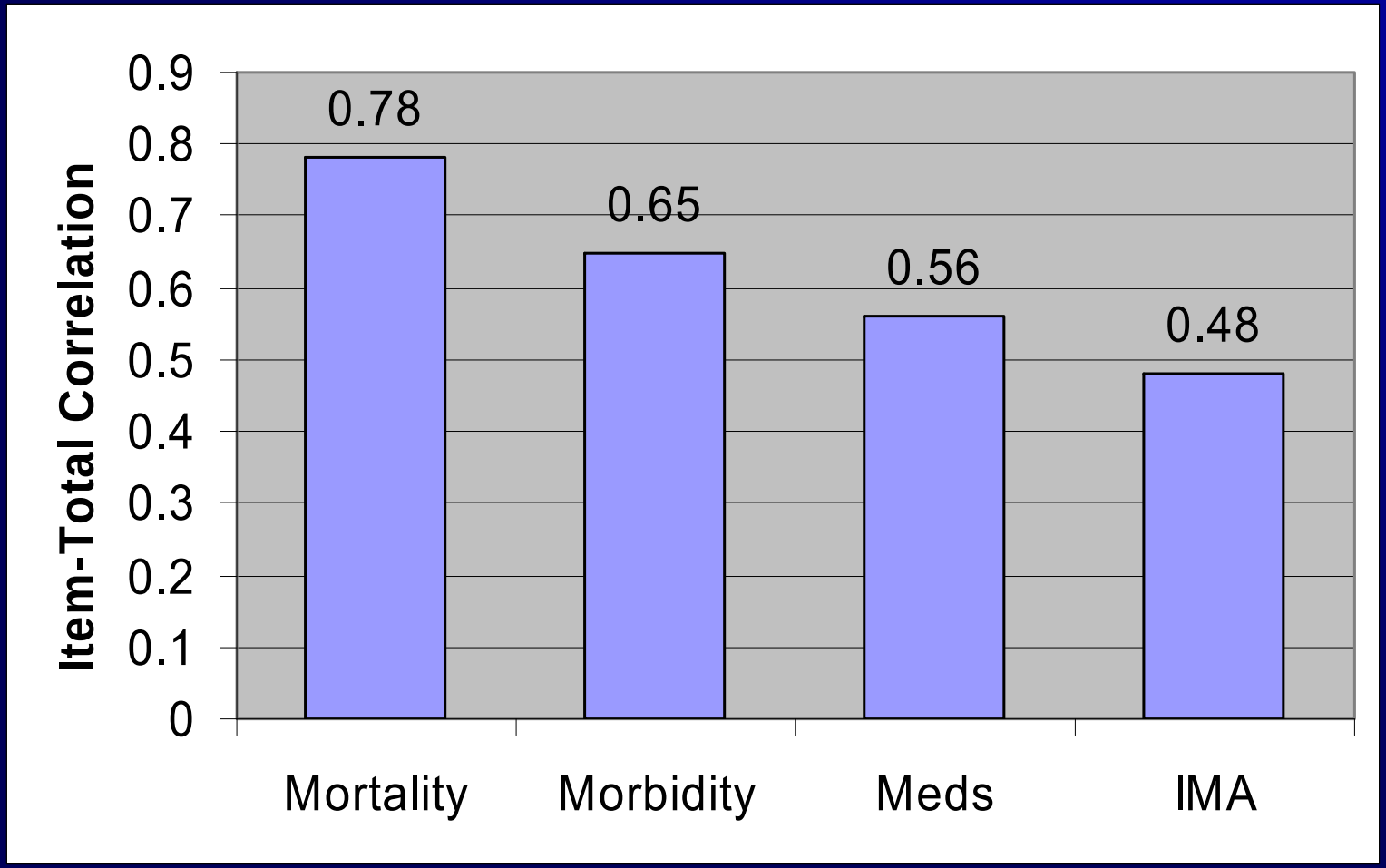
Exploring the Implications of Standardization

$$\text{Composite} = \left(\frac{\hat{\pi}_{\text{mort}}}{0.5}\right) + \left(\frac{\hat{\pi}_{\text{morb}}}{4.2}\right) + \left(\frac{\hat{\pi}_{\text{IMA}}}{5.8}\right) + \left(\frac{\hat{\pi}_{\text{meds}}}{14.3}\right)$$

After standardizing

- **A 1-point difference in mortality has same impact as:**
  - 8% improvement in morbidity rate
  - 11% improvement in use of IMA
  - 28% improvement in use of all medications

Composite is weighted toward outcomes…



Sensitivity Analyses

Key Question

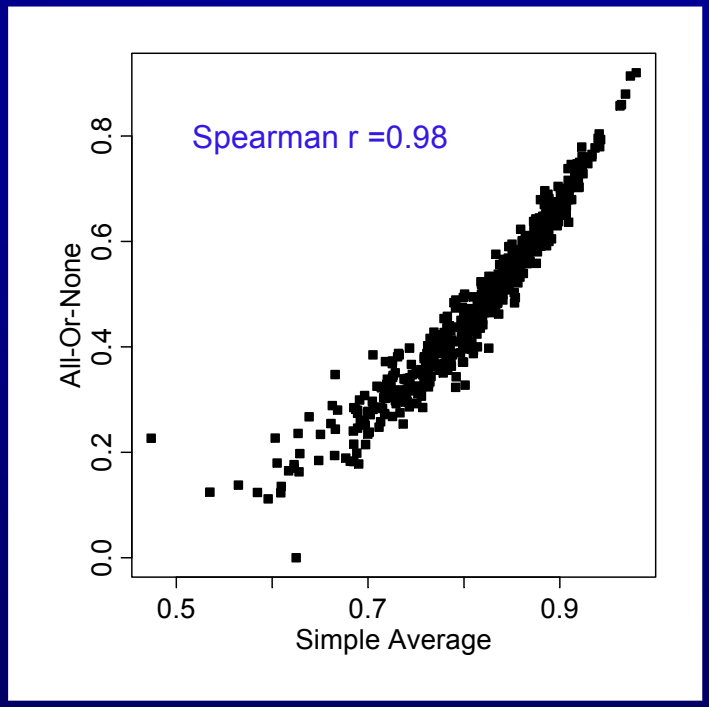
- **Are inferences about hospital quality sensitive to the choice of methods?**
  - If not, then stakes are not so high…

Analysis

- **Calculate composite scores using a variety of different methods and compare results**

Sensitivity Analysis: Within-Domain Aggregation
Opportunity Model vs. All-Or-None Composite

- **Agreement between methods**
  - Spearman rank correlation = 0.98
  - Agree w/in 20 %-tile pts = 99%
  - Agree on top quartile = 93%
  - Pairwise concordance = 94%
- **1 hospital's rank changed by 23 percentile points places**
- **No hospital was ranked in the top quartile by one method and bottom half by the other**

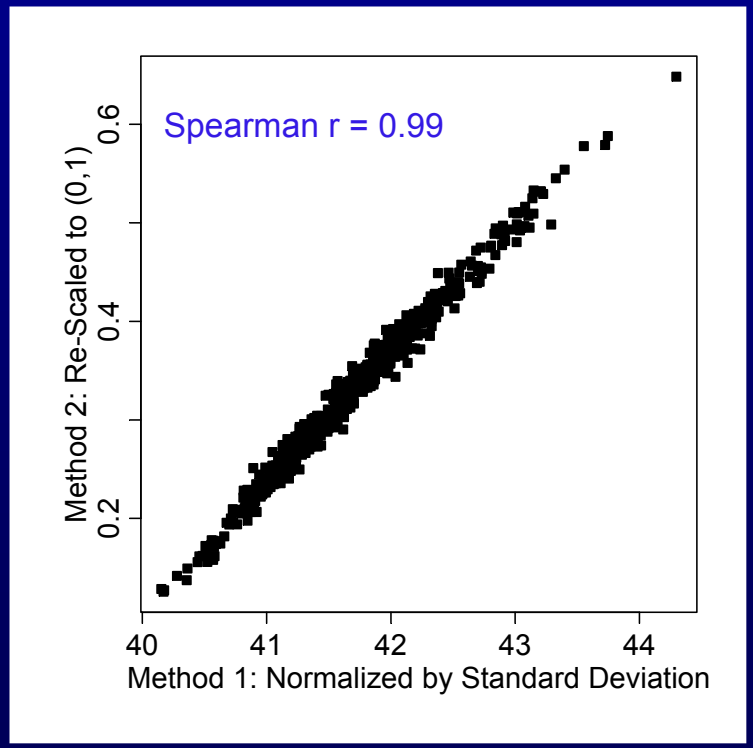


Sensitivity Analysis: Method of Standardization

Divide by the range instead of the standard deviation

$$\text{Composite} = \frac{\hat{\pi}_{\text{mort}}}{\text{range}_{\text{mort}}} + \frac{\hat{\pi}_{\text{morb}}}{\text{range}_{\text{morb}}} + \frac{\hat{\pi}_{\text{IMA}}}{\text{range}_{\text{IMA}}} + \frac{\hat{\pi}_{\text{meds}}}{\text{range}_{\text{meds}}}$$

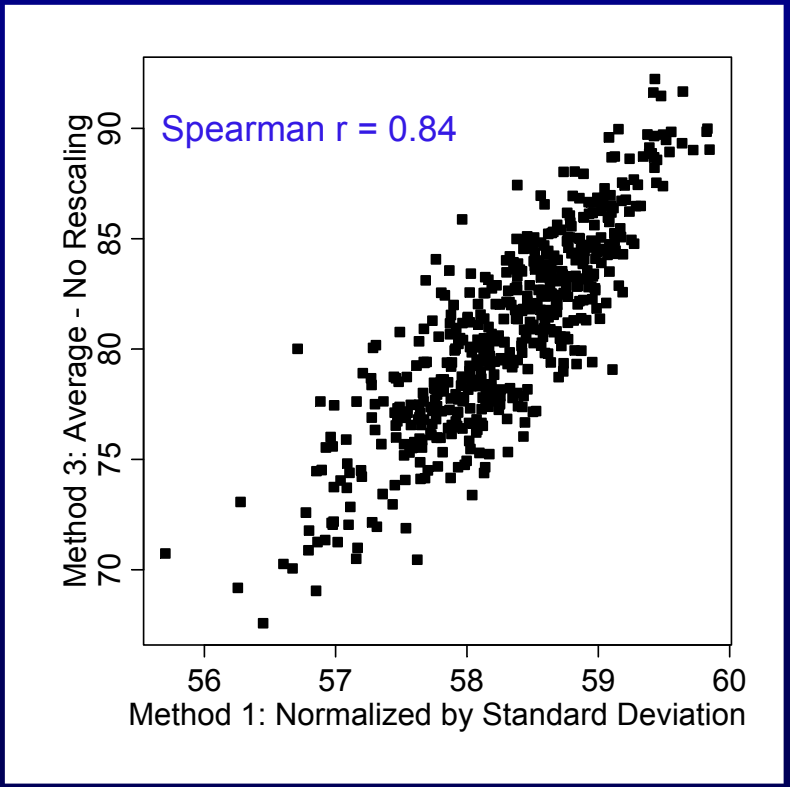
where range denotes the maximum minus the minimum (across hospitals)



Sensitivity Analysis: Method of Standardization

Don’t standardize

$$\text{Composite} = \hat{\pi}_{\text{mort}} + \hat{\pi}_{\text{morb}} + \hat{\pi}_{\text{IMA}} + \hat{\pi}_{\text{meds}}$$



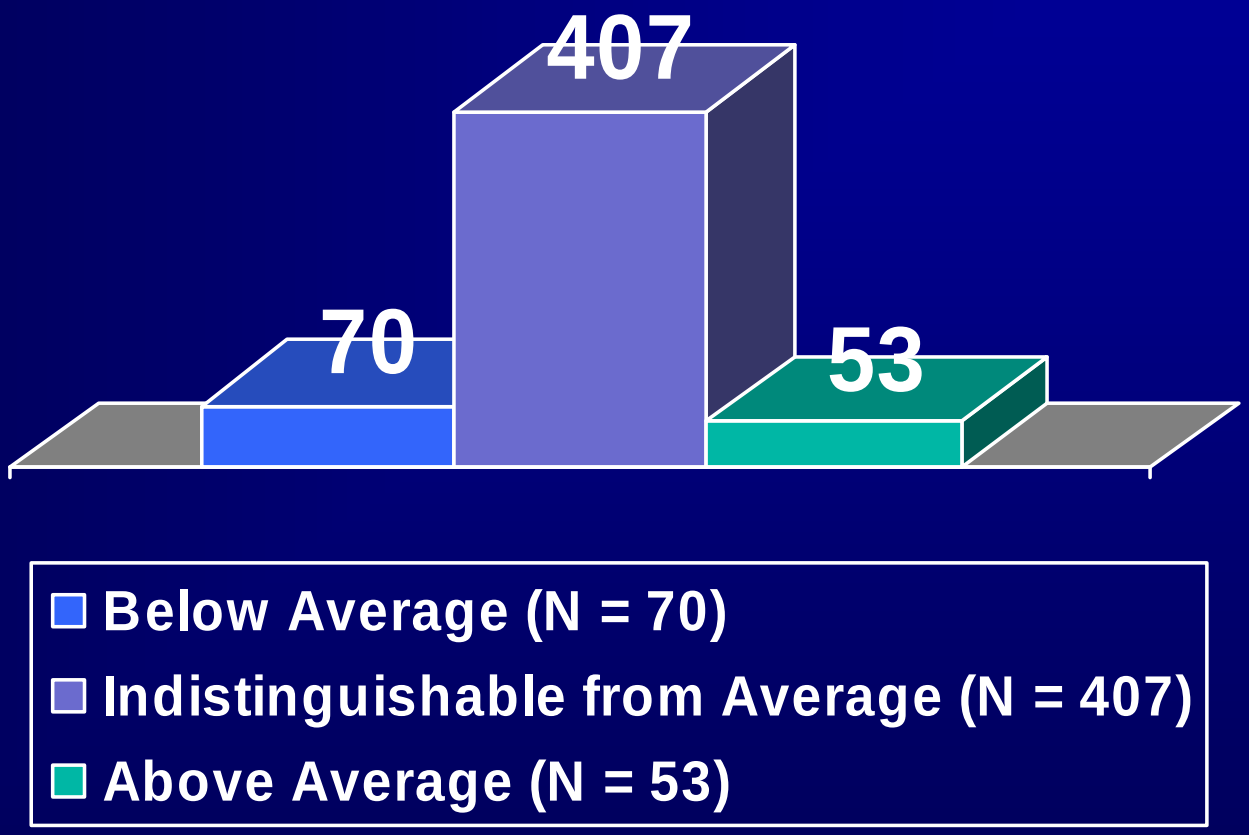
Sensitivity Analysis: Summary

- Inferences about hospital quality are generally robust to minor variations in the methodology
- However, standardizing vs. not standardizing has a large impact on hospital rankings

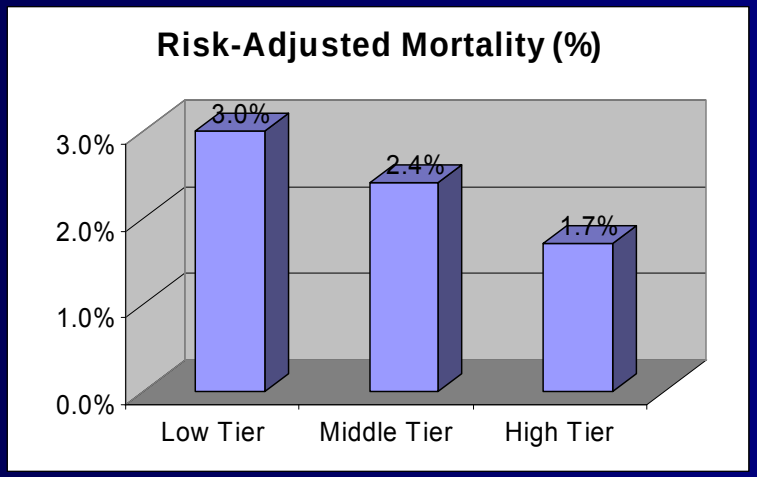
Performance of Hospital Classifications Based on the STS Composite Score

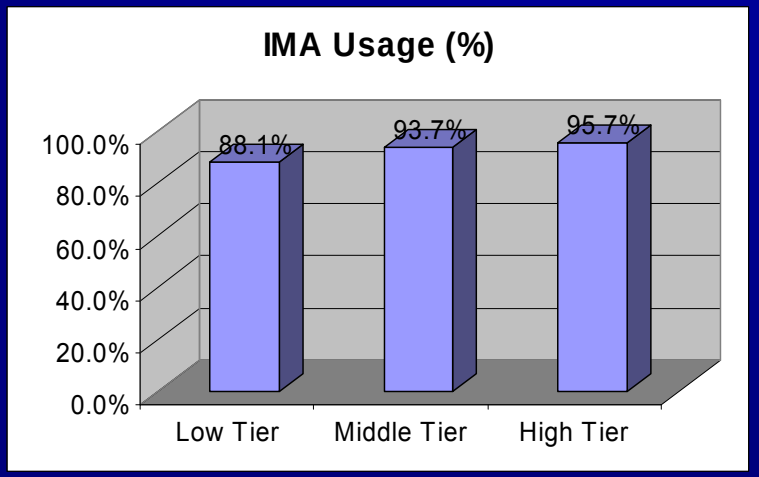
- **Bottom Tier**
  - ≥ 99% Bayesian probability that provider's true score is lower than STS average
- **Top Tier**
  - ≥ 99% Bayesian probability that provider's true score is higher than STS average
- **Middle Tier**
  - < 99% certain whether provider's true score is lower or higher than STS average.

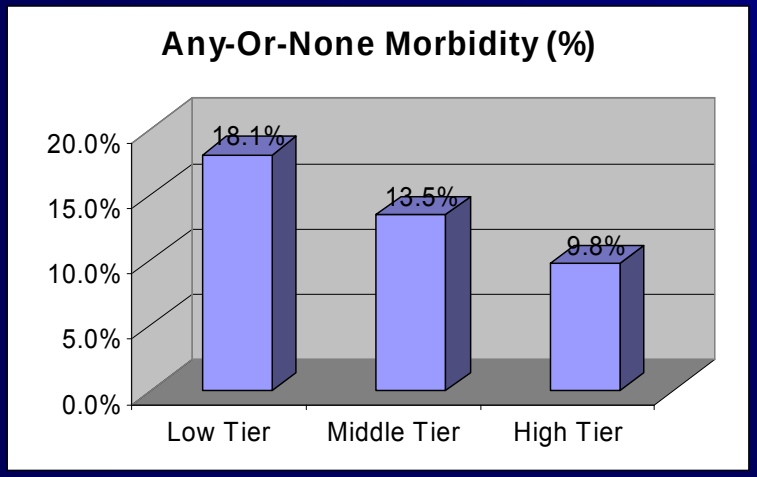
Results of Hypothetical Tier System in 2004 Data

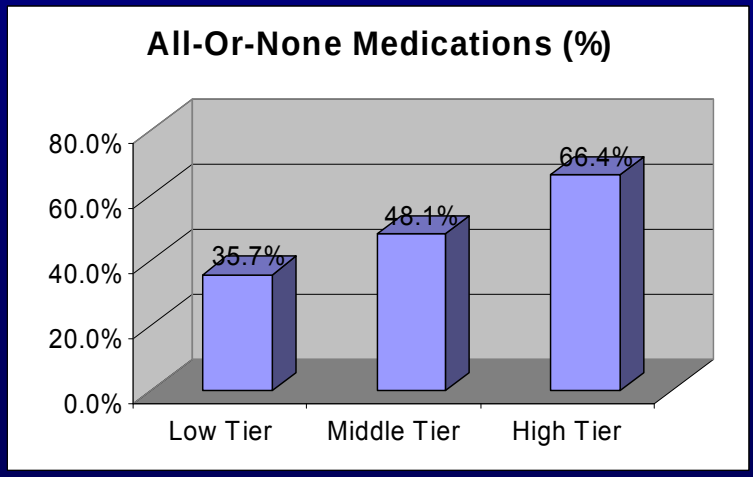


Ability of Composite Score to Discriminate Performance on Individual Domains









Summary of STS Composite Method

- Use of all-or-none composite for combining items within domains
- Combining items was based on rescaling and adding
- Estimation via Bayesian hierarchical models
- Hospital classifications based on Bayesian probabilities

Advantages

- **Rescaling and averaging is relatively simple**
  - Even if estimation method is not
- **Hierarchical models help separate true quality differences from random noise**
- **Bayesian probabilities provide a rigorous approach to accounting for uncertainty when classifying hospitals**
  - Control false-positives, etc.

Limitations

- **Validity depends on the collection of individual measures**
  - Choice of measures was limited by practical considerations (e.g. available in STS)

    Measures were endorsed by NQF

- **Weak correlation between measures**
  - Reporting a single composite score entails some loss of information
  - Results will depend on choice of methodology

    We made these features transparent

    - Examined implications of our choices

    - Performed sensitivity analyses

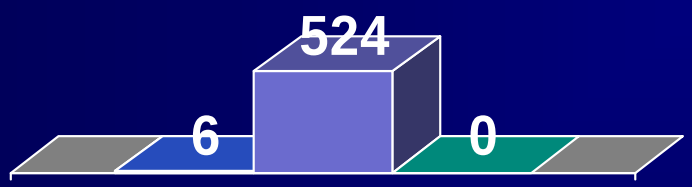
Summary

- **Composite scores have inherent limitations**
- **The implications of the weighting method is not always obvious**
- **Empirical testing & sensitivity analyses can help elucidate the behavior and limitations of a composite score**
- **The validity of a composite score depends on its fitness for a particular purpose**
  - Possibly different considerations for P4P vs. public reporting

Extra Slides

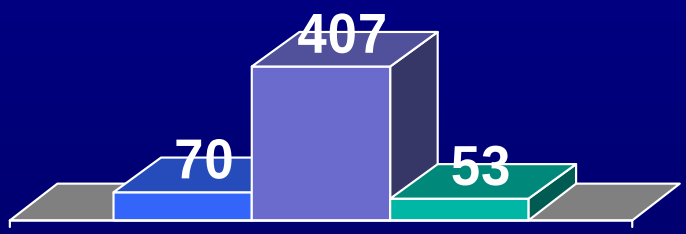
Comparison of Tier Assignments Based on Composite Score Vs. Mortality Alone

Mortality Only



- Worse Than Average (N = 6)
- Indistinguishable from Average (N = 524)
- Better Than Average (N = 0)

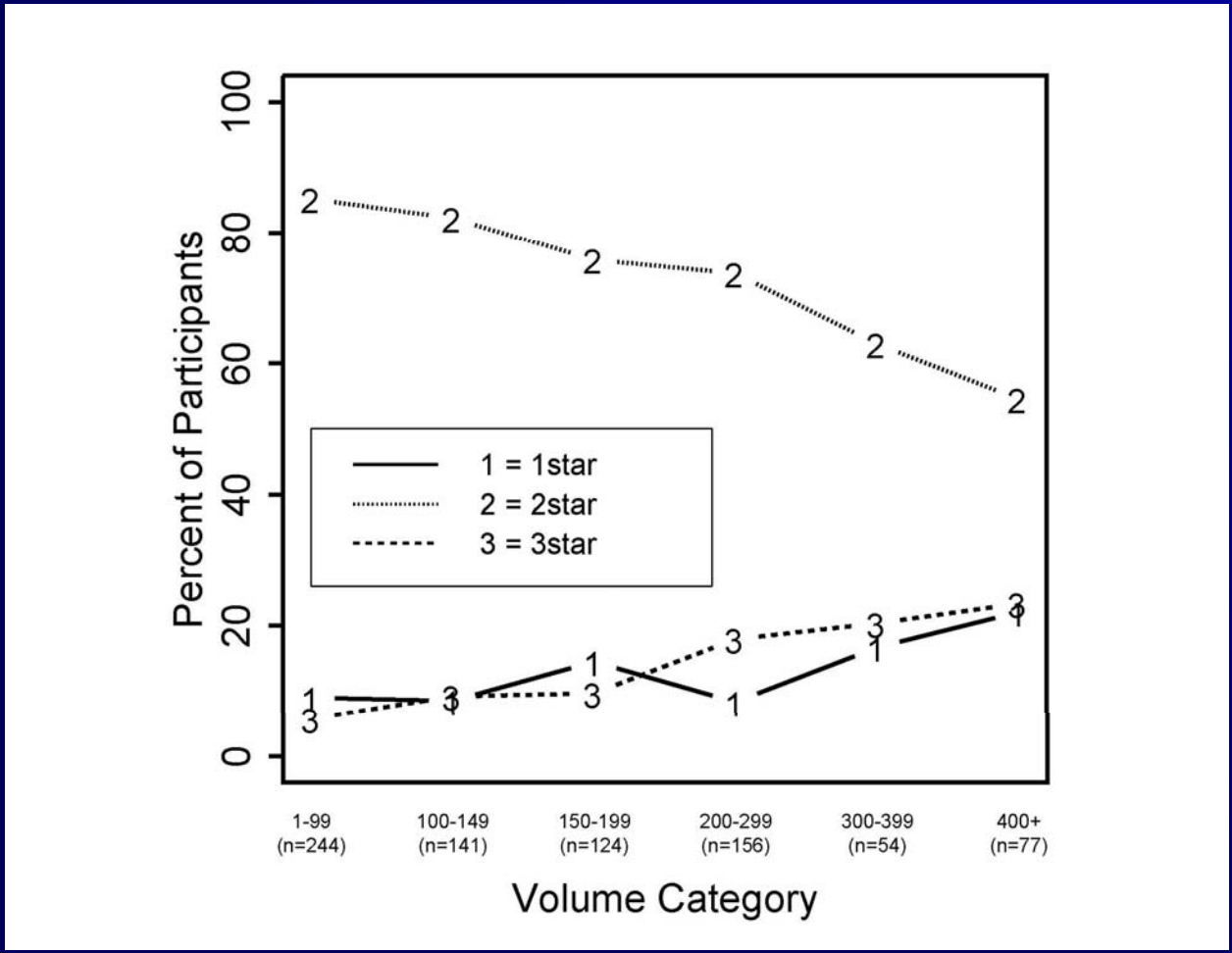
Composite Score



- Worse Than Average (N = 70)
- Indistinguishable from Average (N = 407)
- Better Than Average (N = 53)

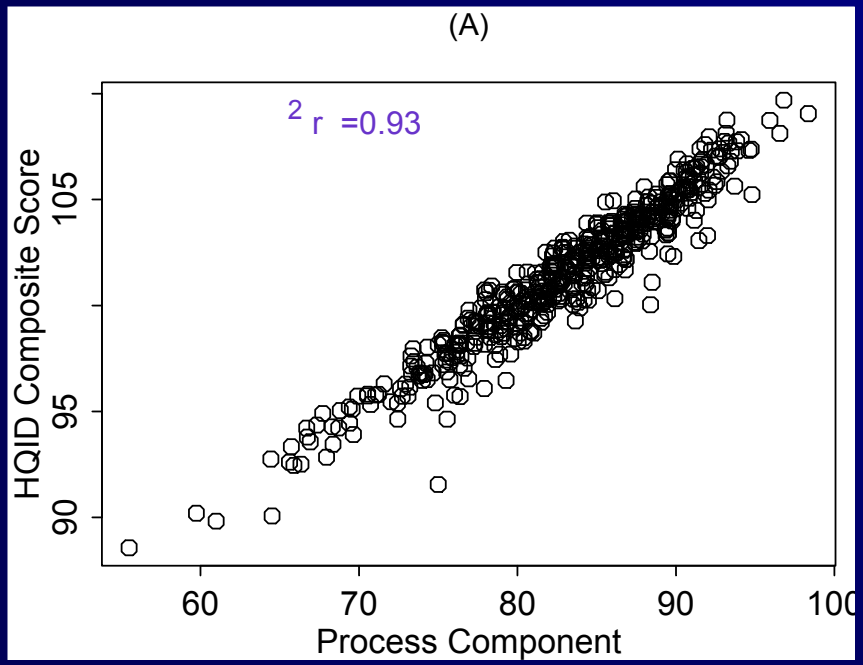
EXTRA SLIDES – STAR RATINGS VS VOLUME

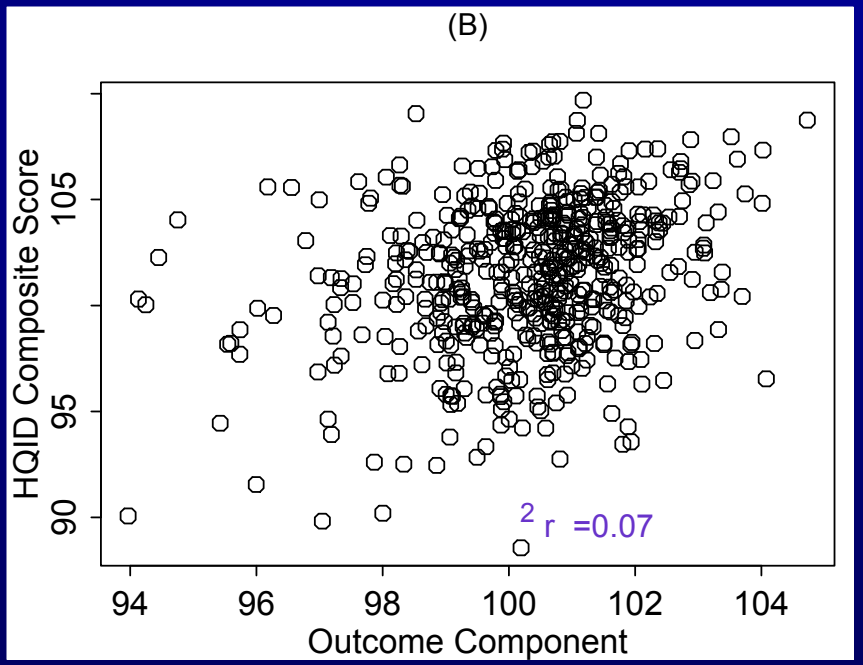
Frequency of Star Categories By Volume



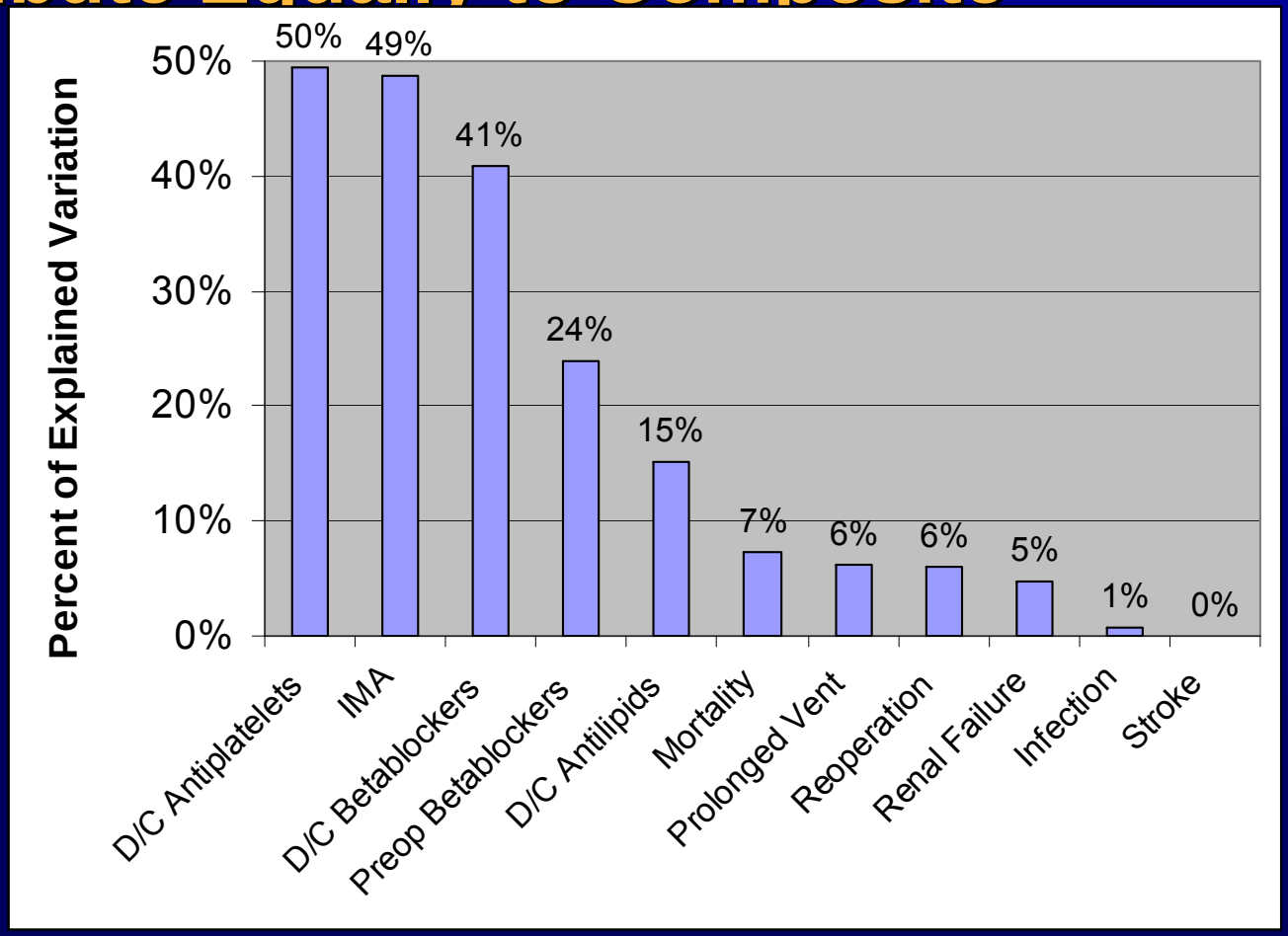
EXTRA SLIDES – HQID METHOD APPLIED TO STS MEASURES

Finding #1. Composite Is Primarily Determined by Outcome Component

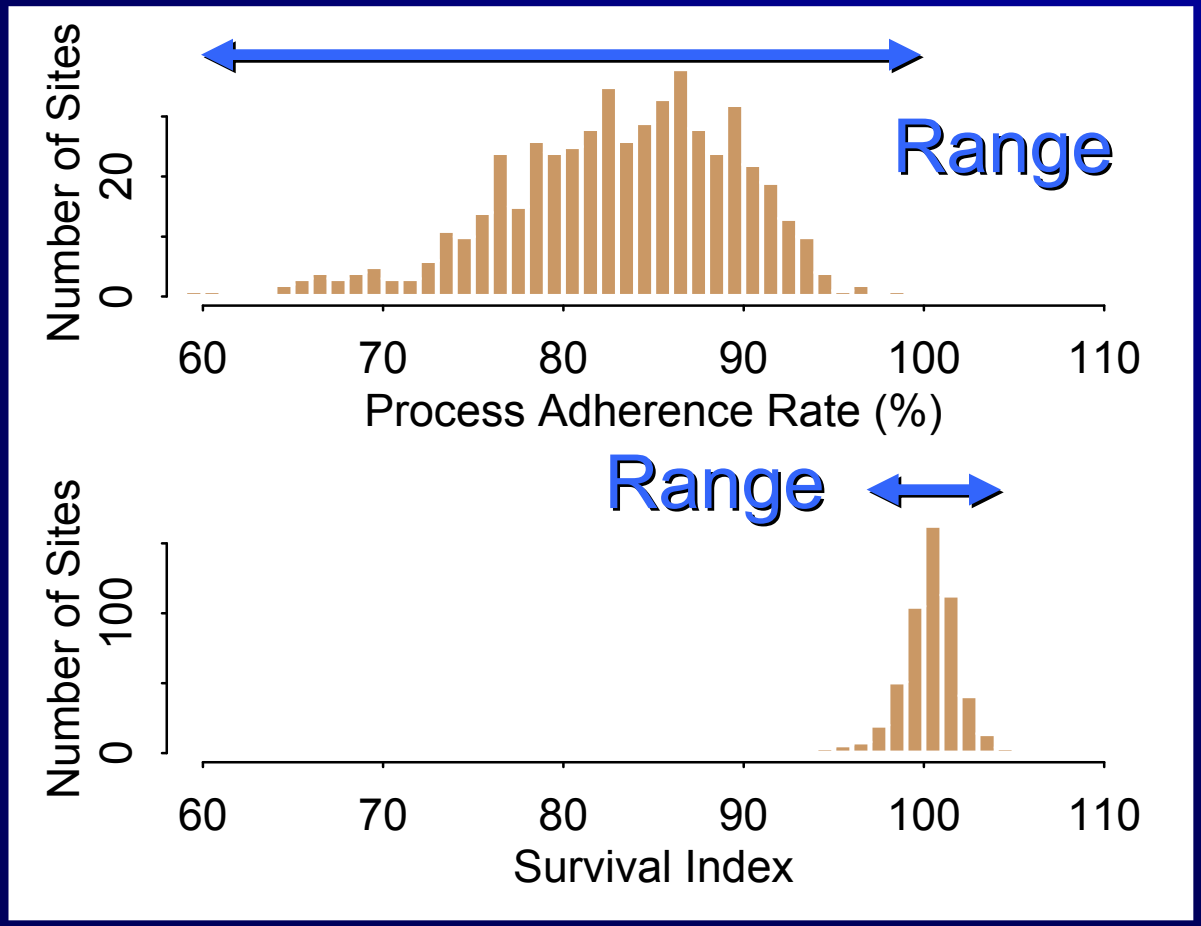




Finding #2. Individual Measures Do Not Contribute Equally to Composite

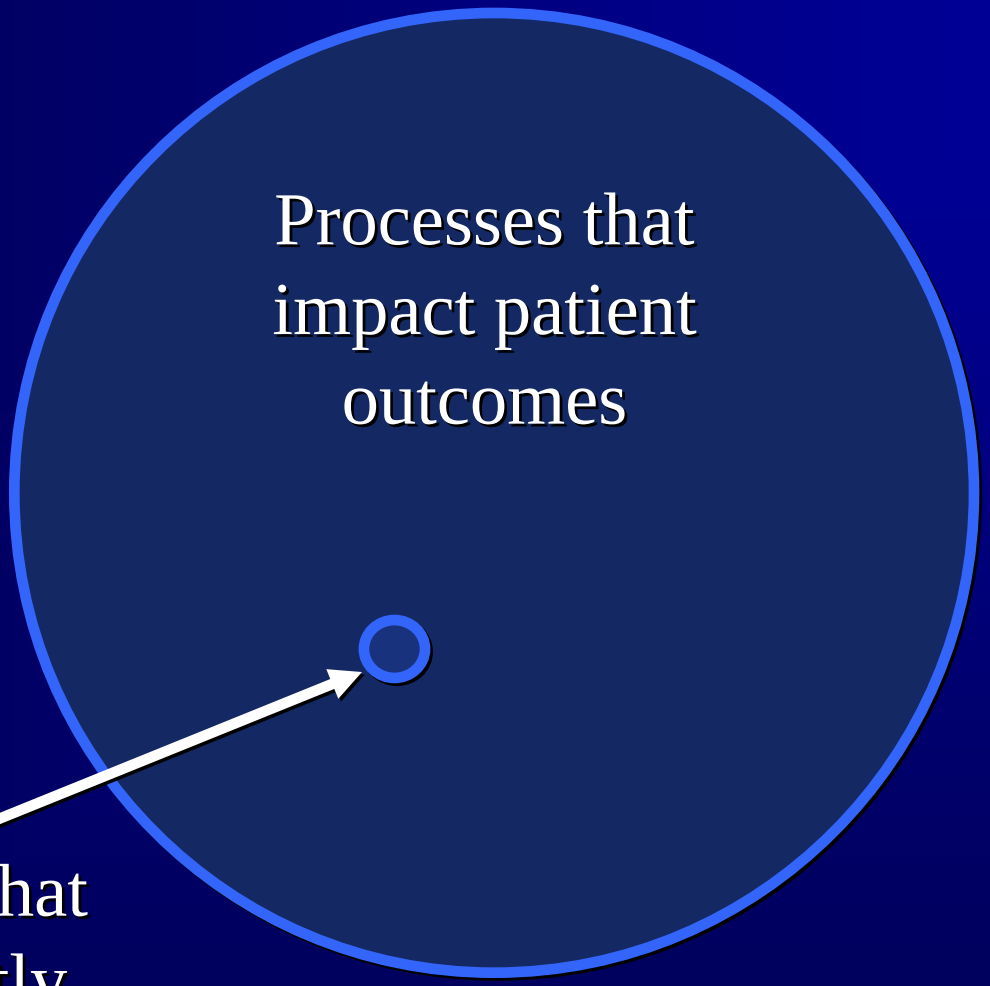


Explanation: Process & Survival Components Have Measurement Unequal Scales

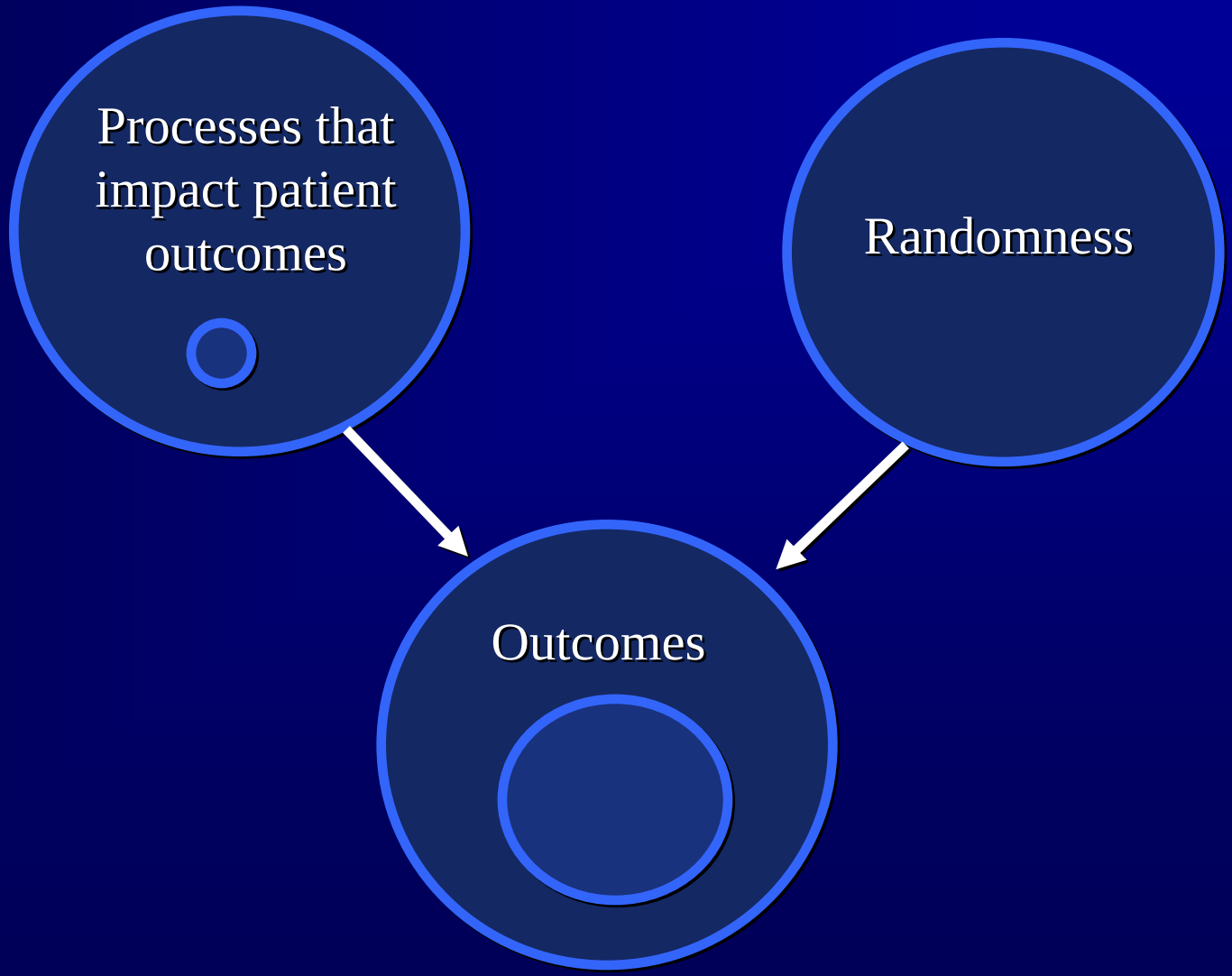


EXTRA SLIDES – CHOOSING MEASURES

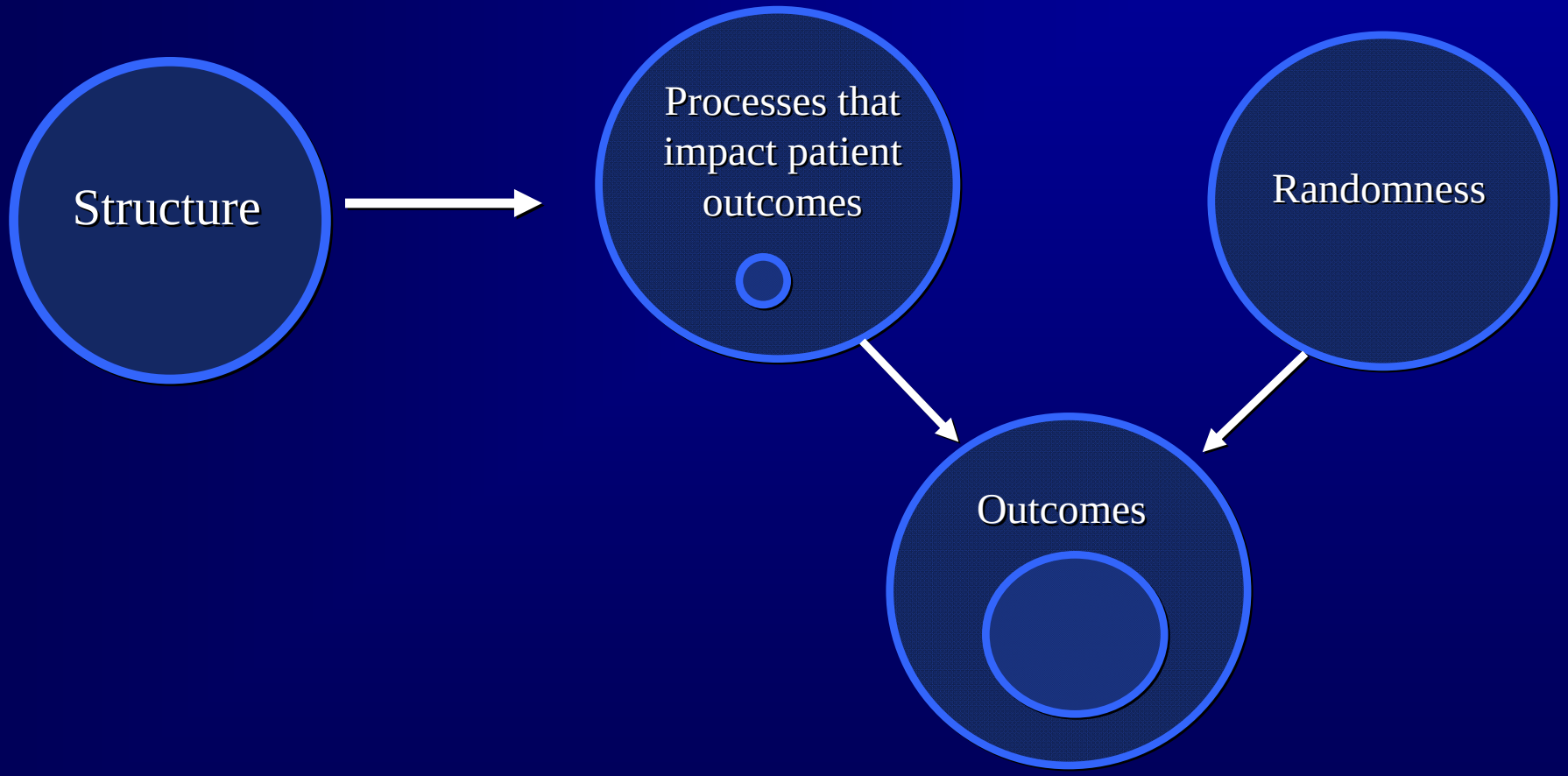
Process or Outcomes?



Process or Outcomes?



Structural Measures?



EXTRA SLIDES – ALTERNATE PERSPECTIVES FOR DEVELOPING COMPOSITE SCORES

Perspectives for Developing Composites

- **Normative Perspective**
  - Concept being measured is defined by the choice of measures and their weighting

    Not vice versa
  - Weighting different aspects of quality is inherently normative

    - Weights reflect a set of values

    - Whose values?

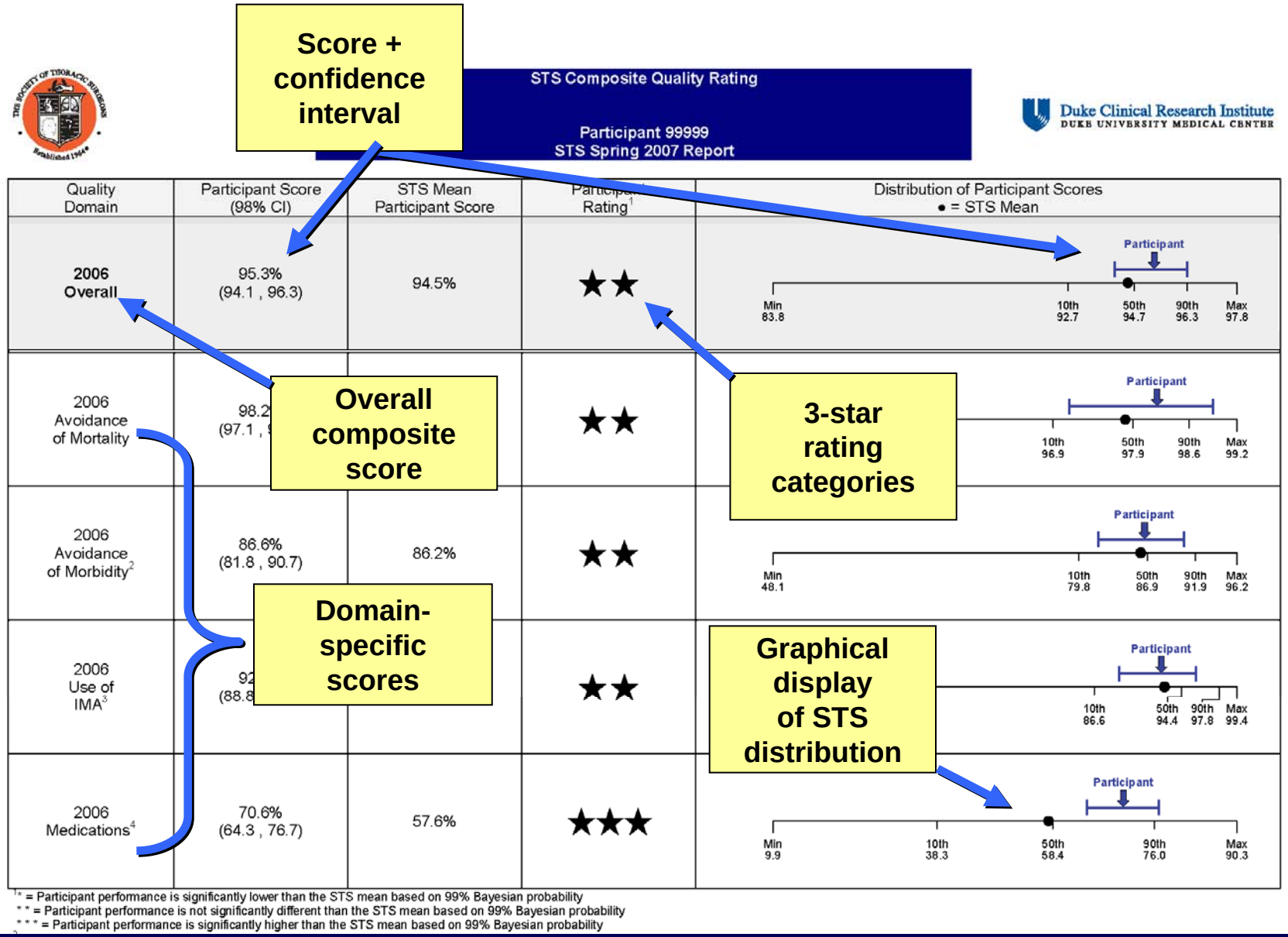
Perspectives for Developing Composites

- **Behavioral Perspective**
  - Primary goal is to provide an incentive
  - Optimal weights are ones that will cause the desired behavior among providers
  - Issues:

    Reward outcomes or processes?

    Rewarding X while hoping for Y

SCRAP

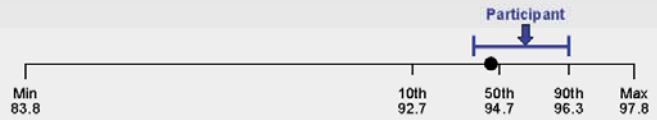
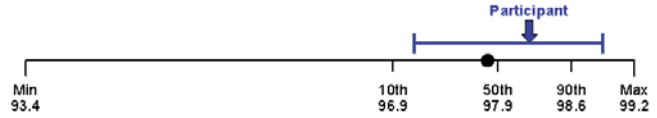
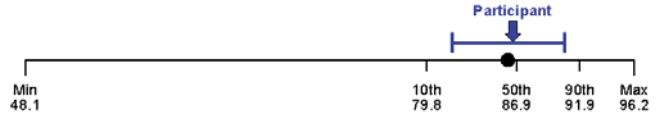
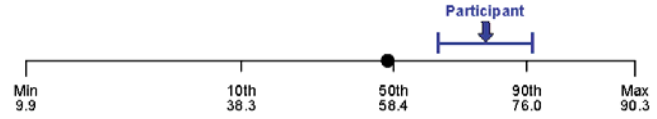


[1]* = Participant performance is significantly lower than the STS mean based on 99% Bayesian probability
** = Participant performance is not significantly different than the STS mean based on 99% Bayesian probability
*** = Participant performance is significantly higher than the STS mean based on 99% Bayesian probability

STS Composite Quality Rating

Participant 99999
STS Spring 2007 Report

Duke Clinical Research Institute
DUKE UNIVERSITY MEDICAL CENTER

Quality Domain	Participant Score (98% CI)	STS Mean Participant Score	Participant Rating[1]	Distribution of Participant Scores ● = STS Mean
2006 Overall	95.3% (94.1 , 96.3)	94.5%	★★	
2006 Avoidance of Mortality	98.2% (97.1 , 98.9)	97.8%	★★	
2006 Avoidance of Morbidity[2]	86.6% (81.8 , 90.7)	86.2%	★★	
2006 Use of IMA[3]	92.6% (88.8 , 95.7)	92.9%	★★	
2006 Medications[4]	70.6% (64.3 , 76.7)	57.6%	★★★	

[1] * = Participant performance is significantly lower than the STS mean based on 99% Bayesian probability
** = Participant performance is not significantly different than the STS mean based on 99% Bayesian probability
*** = Participant performance is significantly higher than the STS mean based on 99% Bayesian probability
[2] Includes Reoperations, Renal Failure, Deep Sternal Wound Infection, Prolonged Ventilation, and CVA
[3] Excludes patients with prior CABG surgery
[4] Includes Preoperative Beta Blockade, Discharge Beta Blockade, Discharge Anti-Lipids, and Discharge Anti-Platelets