AMERICAN COLLEGE OF SURGEONS

NATIONAL SURGICAL QUALITY IMPROVEMENT PROGRAM

The Quality-Profiling "Reliability" of ACS NSQIP Models

Kristopher Huffman, MS

Division of Research and Optimal Patient Care

Continuous Quality Improvement

American College of Surgeons

July 29th, 2014

What is Reliability?

Hospital submits cases to model → **Hospital receives performance metric (odds ratio)**

(1) The variability of the outcome between hospitals (“**Signal**”)

(2) The variability of the outcome within the hospital (“**Noise**”)

2 sources of variability contribute to this metric

- **Reliability** = $\frac{\text{Signal}}{\text{Signal + Noise}}$ → **What percentage of the total variability in the metric is attributable to signal?**

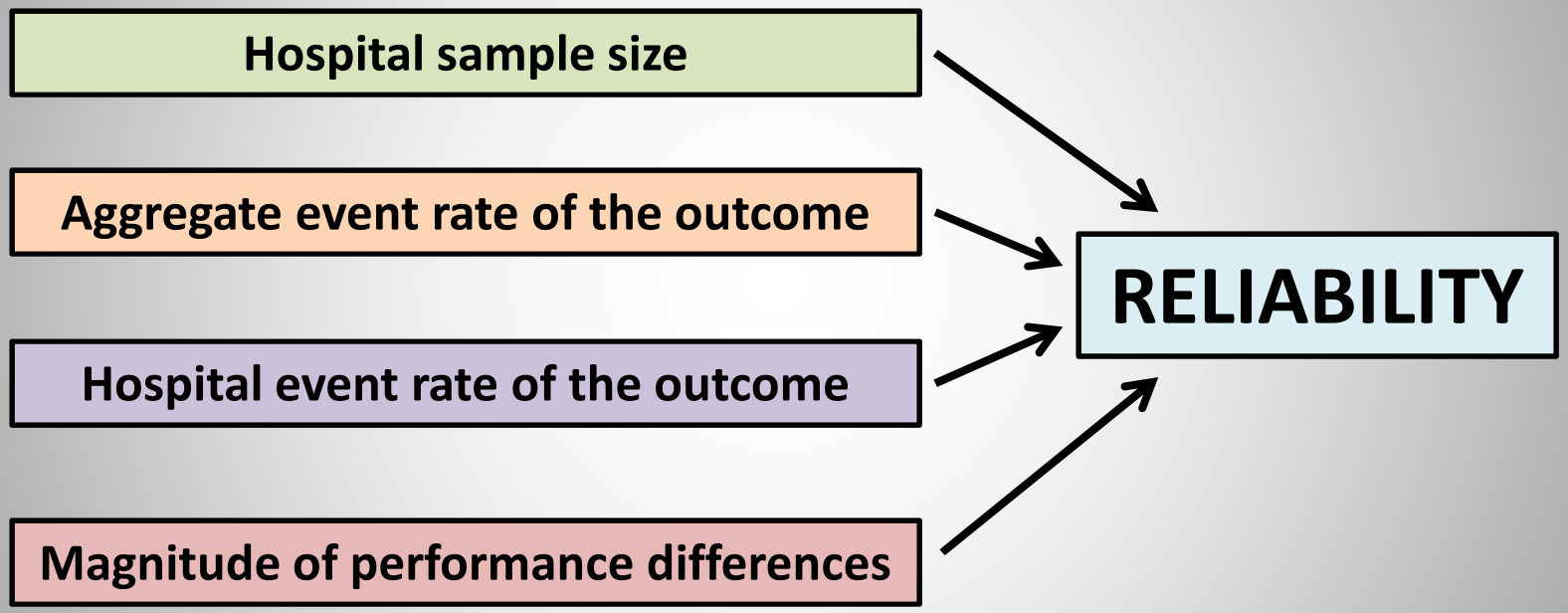
- **Reliability = 0** → **Hospital’s odds ratio is due only to noise**
- **Reliability = 1** → **Hospital’s odds ratio is due only to signal**

→ **Reliability quantifies the degree to which a hospital’s odds ratio is based on true differences in hospital performance.**

What is Reliability?

- What we're really talking about here is ***Statistical Reliability – a technical term***.
- **Statistical Reliability ≠ "Reliability"** ➡ **A term used in everyday speech to characterize the quality of being dependable or trustworthy**
- **Low Statistical Reliability ≠ "Unreliable"**
- Models with low statistical reliability are not "unreliable" or "invalid" – statistical reliability alone does not determine whether performance assessment information is useful.

Factors Influencing Reliability



An Acceptable Level of Reliability?

- There is no established minimum level of reliability considered necessary for hospital profiling.

- 0.40 = “Moderate” level of reliability
- 0.70 = “Good” level of reliability

Evaluating Reliability in ACS NSQIP – Background

- The ACS NSQIP January 2013 SAR utilized over 503,000 cases from 358 participating institutions.

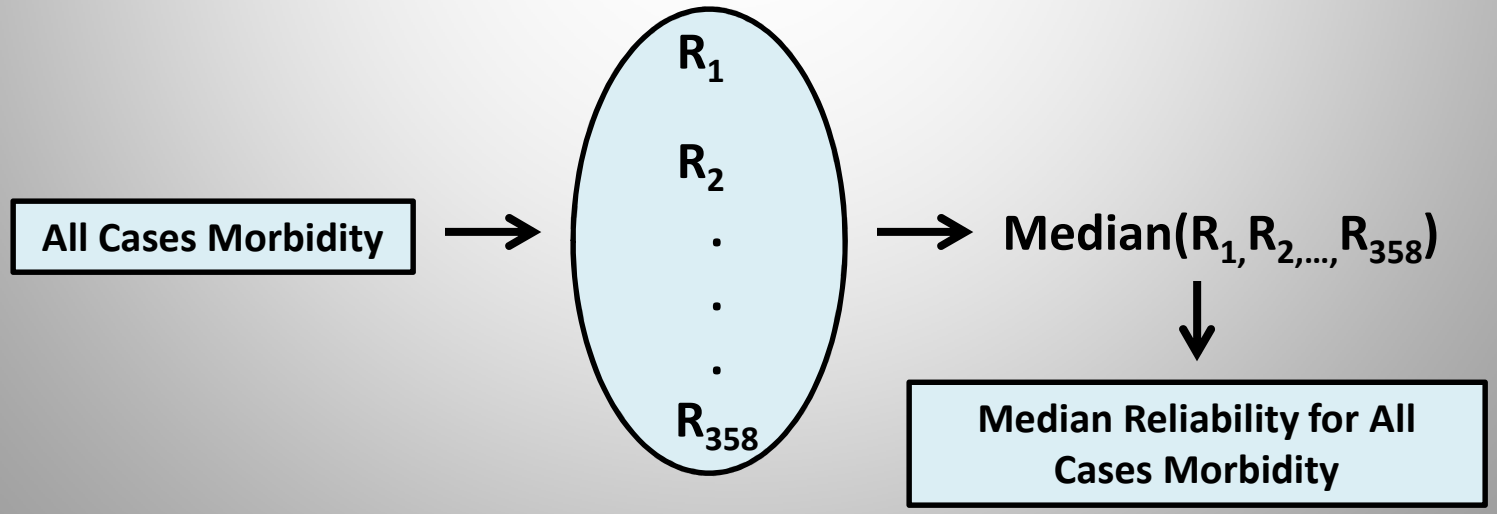
- 137 profiling models were reported on…
  - 72 “Essentials”
  - 65 “Procedure-Targeted”

Evaluating Reliability in ACS NSQIP – Methods

- For each of the 137 profiling models, reliability was computed for each hospital submitting cases to the model.

NOTE: Reliability varies across hospitals in a profiling model

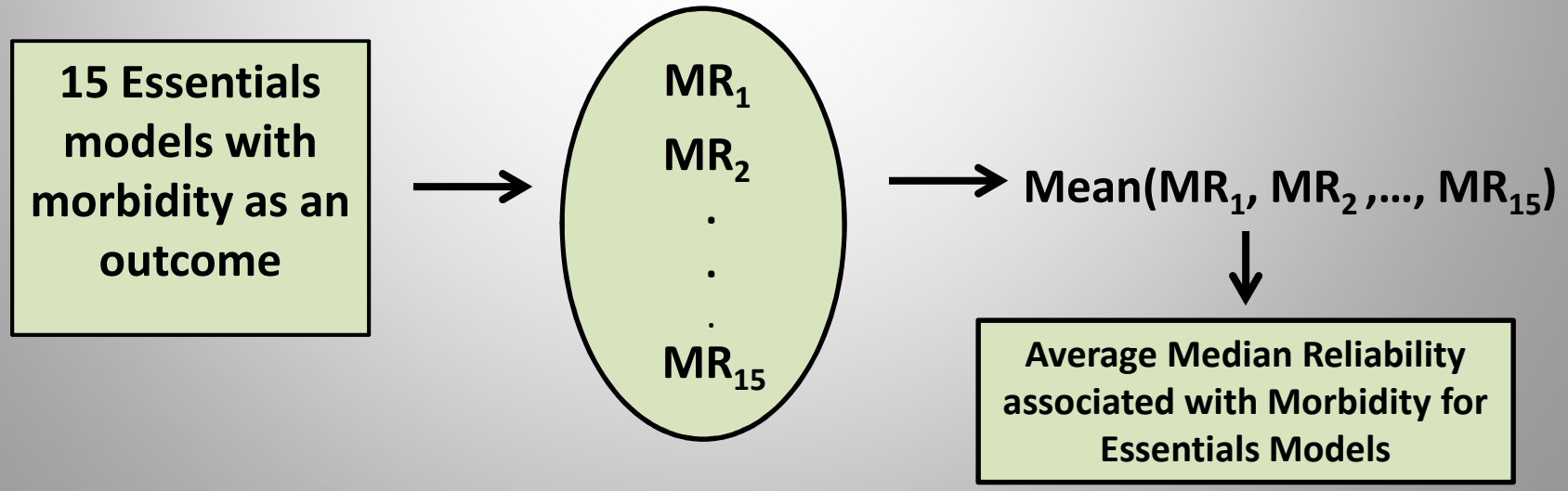
- The median reliability attained over all hospitals submitting cases to the model is then computed



Evaluating Reliability in ACS NSQIP – Methods

- To assess reliability associated with particular outcomes, we computed the average median reliability across all models with common outcomes.

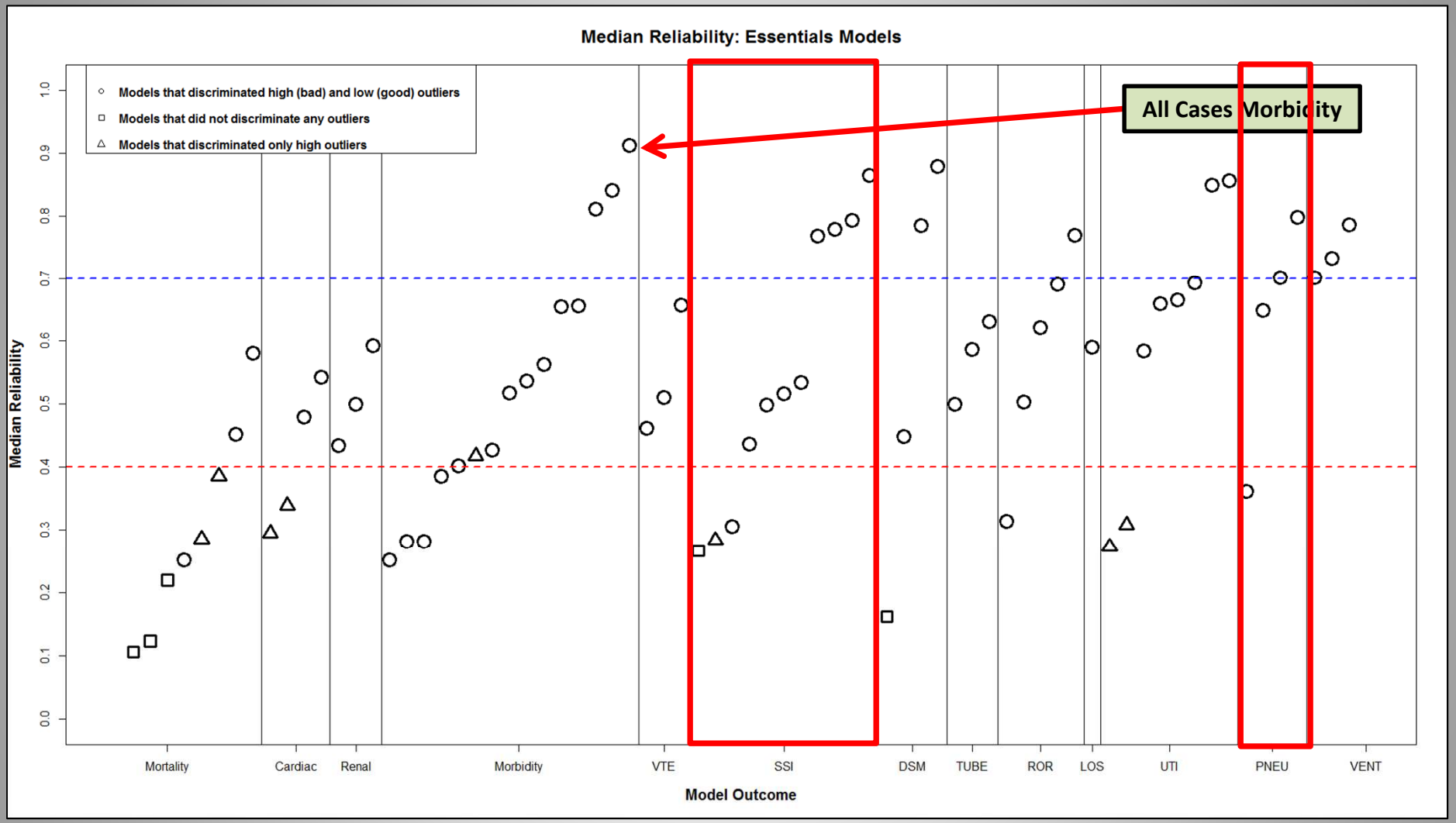
- This was done for the Essentials and Procedure-Targeted programs.



Evaluating Reliability in ACS NSQIP – Methods

- To provide program-wide assessments, we computed the percentage of profiling models in each program attaining a median reliability of ≥ 0.40 and ≥ 0.70

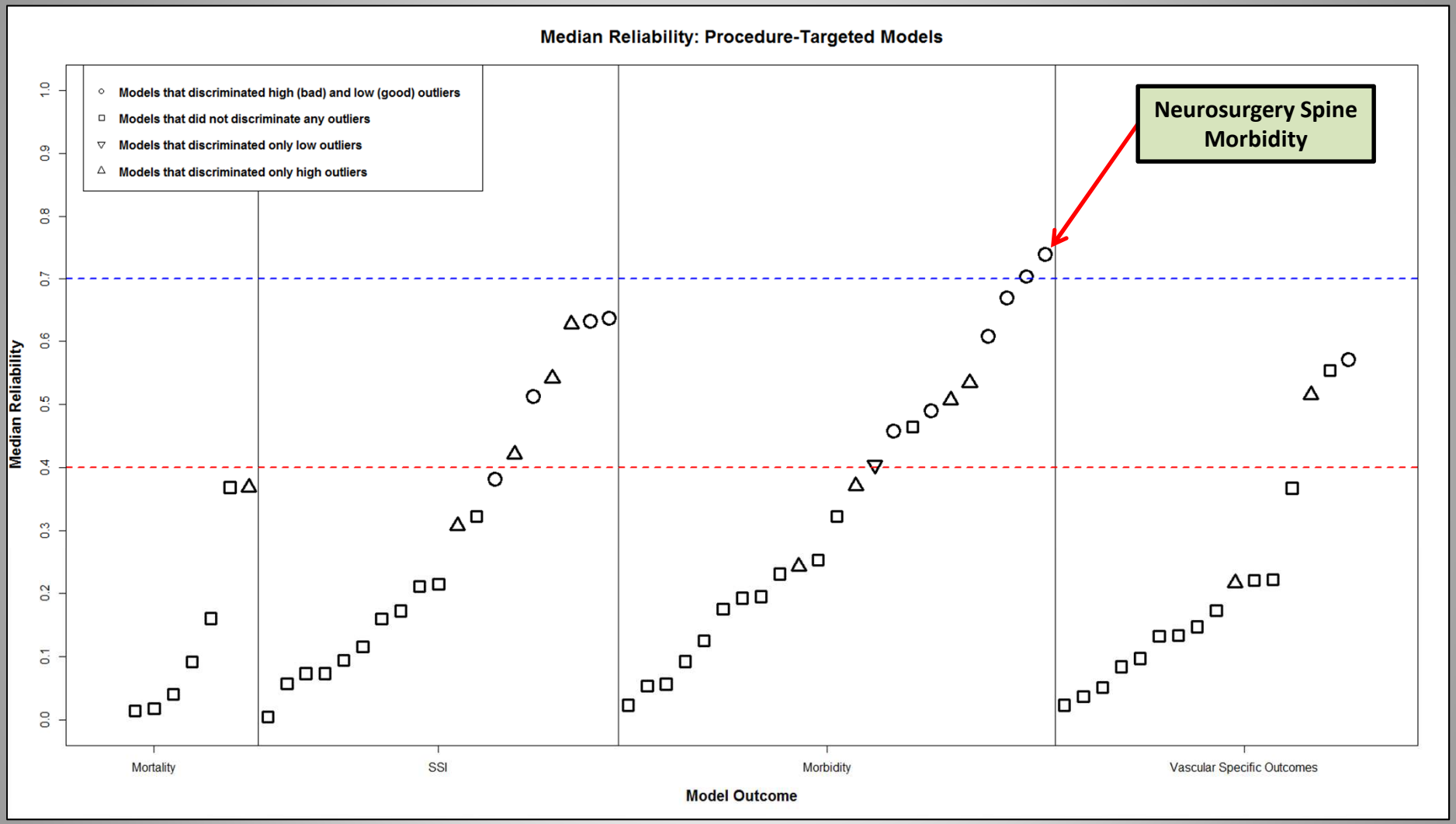
Median Reliability: Essentials



72% (52/72) of Essentials models achieved "moderate" median reliability

24% (17/72) of Essentials models achieved "good" median reliability

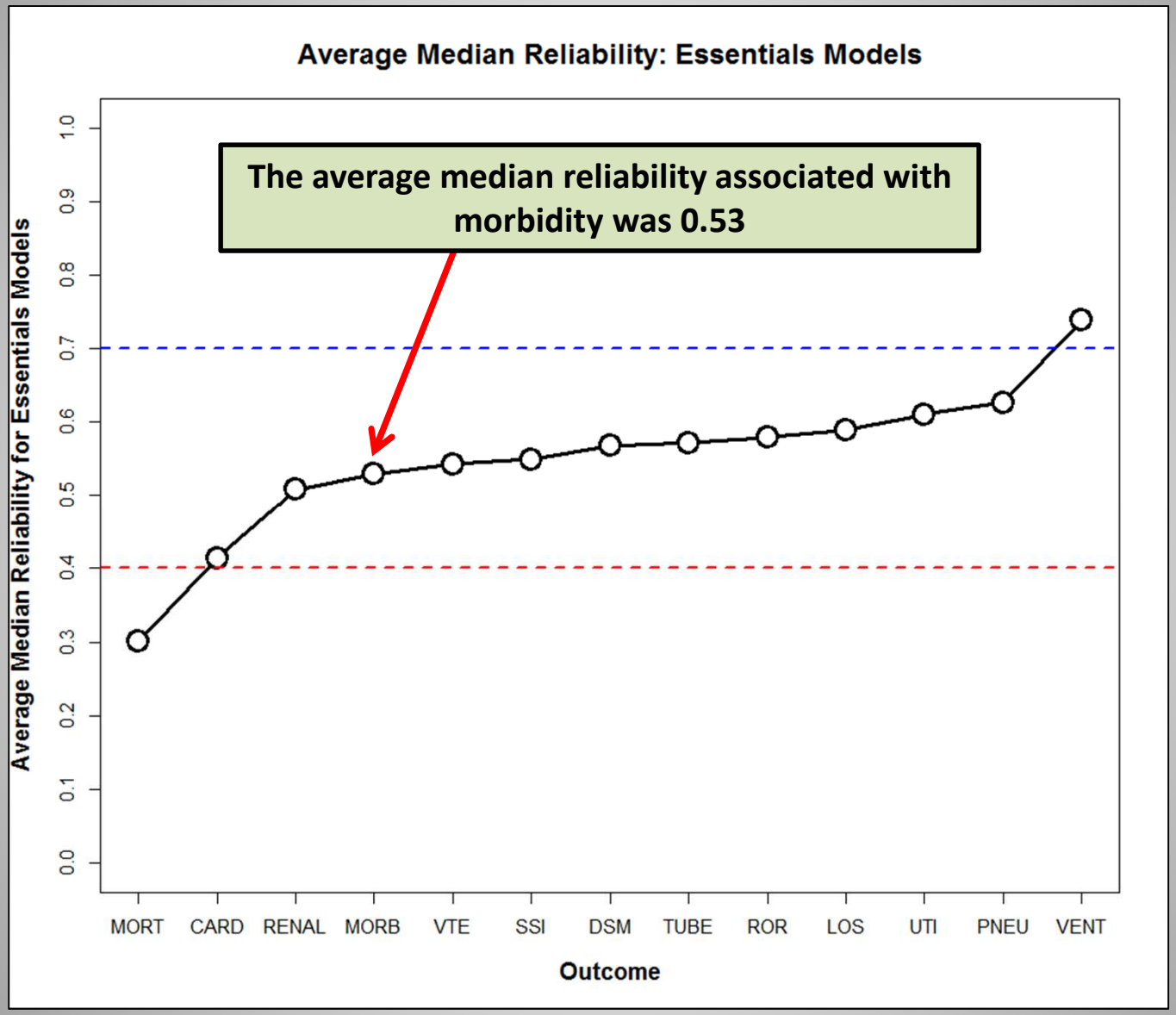
Median Reliability: Procedure-Targeted



29% (19/65) of Procedure-Targeted models achieved "moderate" median reliability

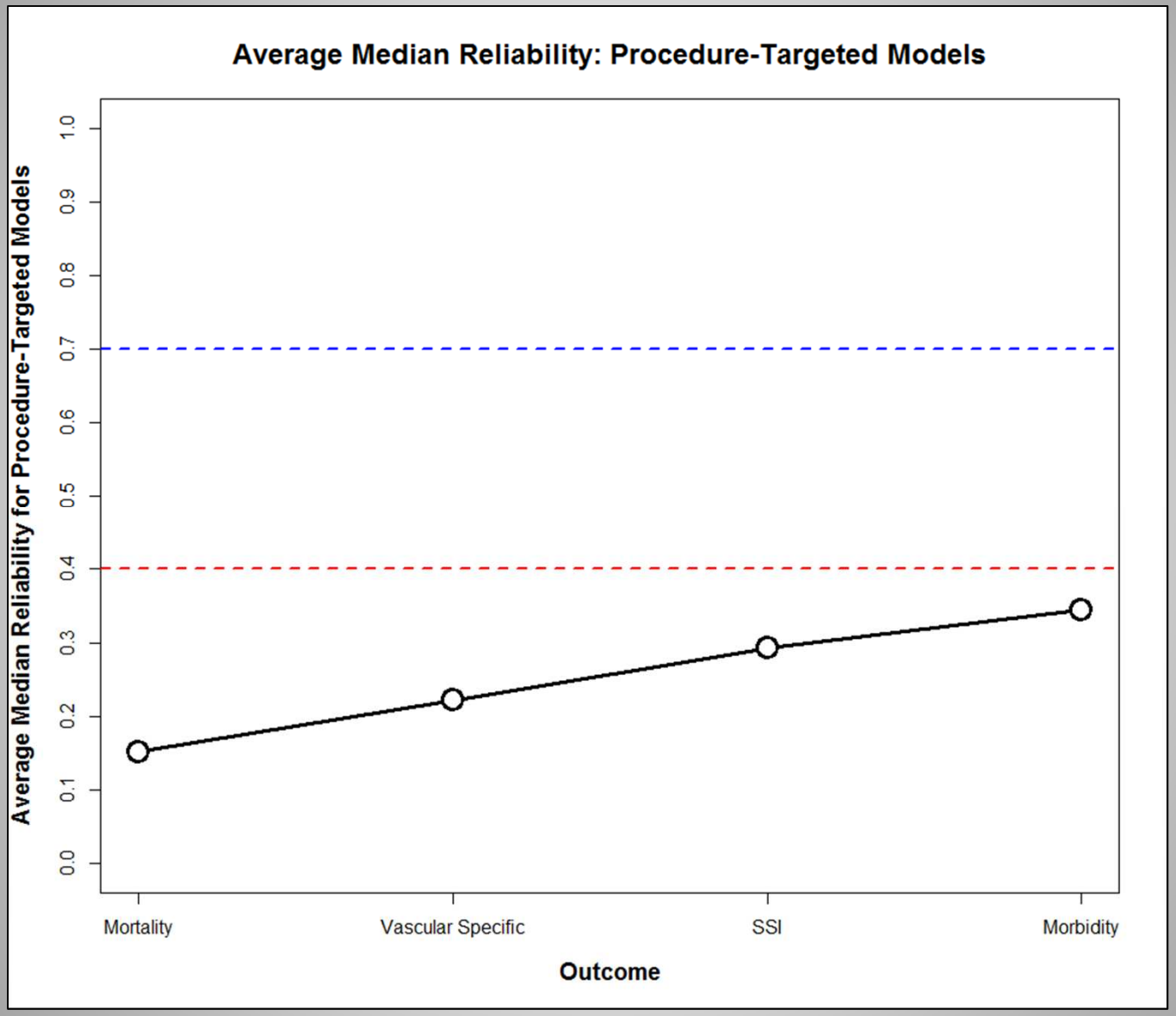
3% (2/65) of Procedure-Targeted models achieved "good" median reliability

Average Median Reliability: Essentials



All outcomes, excluding mortality, achieved “moderate” average median reliability

Average Median Reliability: Procedure-Targeted



All outcomes failed to achieved "moderate" average median reliability

Conclusions: Essentials

- Across most hospitals and models, the reliability of Essentials models is reasonable for assessing surgical quality and differentiating hospital performance.

- From a programmatic perspective, while room for improvement exists, these results are quite encouraging.

Conclusions: Procedure-Targeted

- 29% of Procedure-Targeted models achieved “moderate” median reliability.
- These results are promising, **especially** when considering that the Procedure-Targeted program was in existence for only one year when this analysis was performed.
- There are clear advantages of Procedure-Targeted.
- **THESE ADVANTAGES MERIT CONTINUED DEVELOPMENT OF THE PROCEDURE-TARGETED APPROACH, RATHER THAN ITS ABANDONMENT.**

Low Reliability and Performance Differences

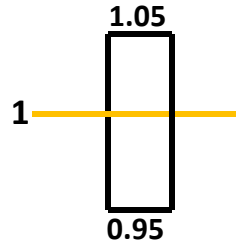
Model

Low Reliability → Lack of Value

Performance differences ≈ 0

Narrow Range of Odds Ratios

Statistical outliers are not identified

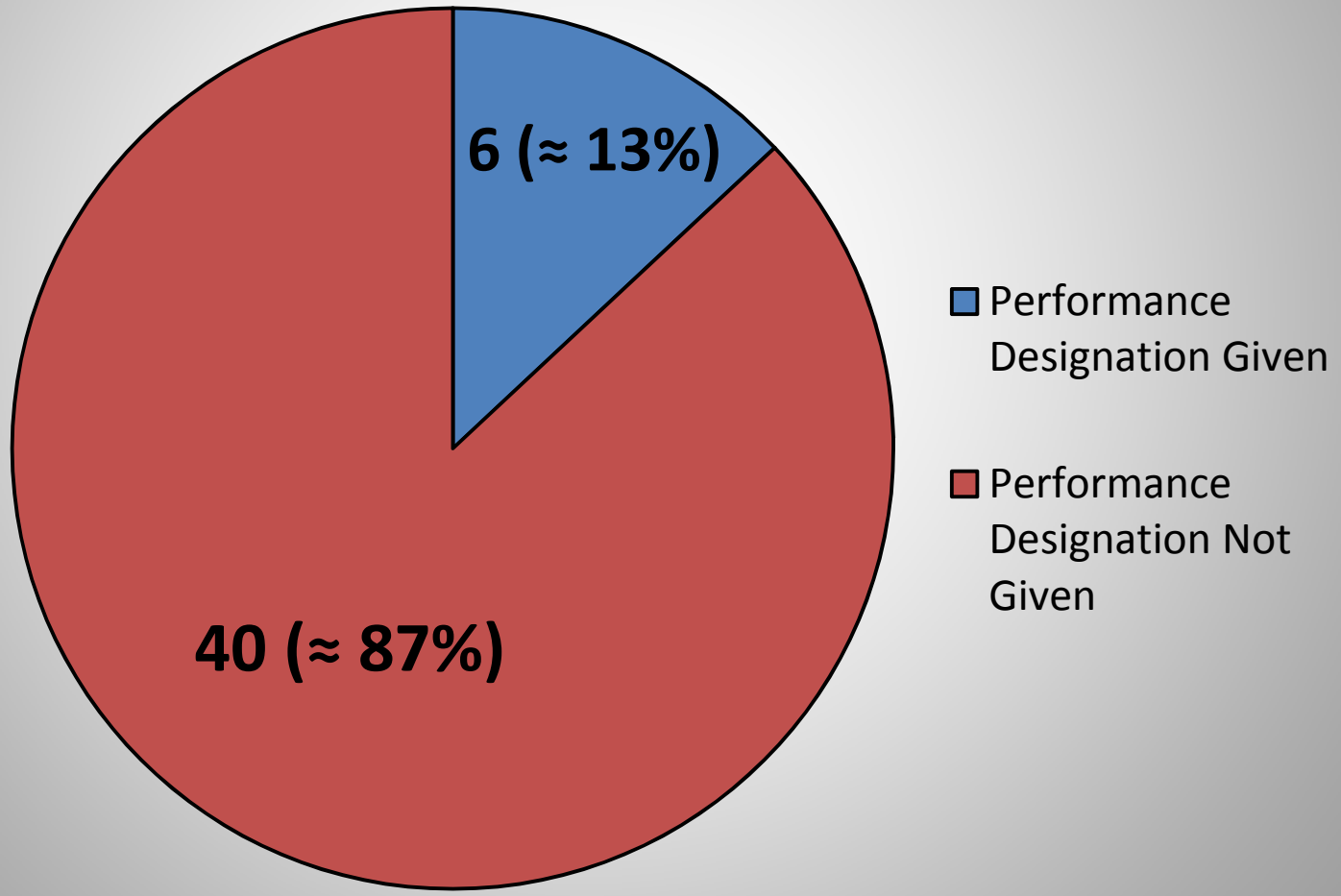


Performance Designation is NOT assigned in PT models

Performance uniformity is useful information

Provision of baseline performance data to a hospital is still exceptionally valuable

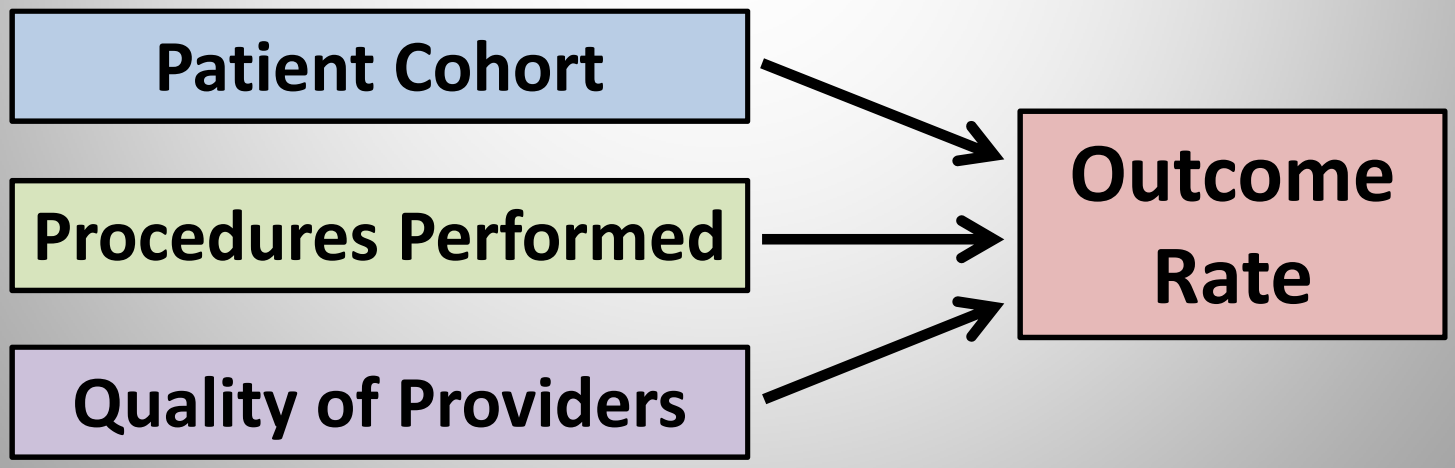
46 Procedure Targeted Models Failing to Achieve "Moderate" Median Reliability – Jan 2013 SAR



The "Reliability of an Outcome"

- Our computation of average median reliability across models with identical outcomes attempts to assess reliability associated with particular outcomes.

- But this is simply conceptual, since a particular outcome does not have an inherent reliability.



The "Reliability of an Outcome"

- Given the strong relationship between sample size and reliability, any subset of cases selected from a larger cohort will likely result in a model with lower median reliability

Model	Median Reliability
General/Vascular Morbidity	**0.84**
Vascular Morbidity	**0.52**
Vascular CEA Morbidity	**0.18**

- An outcome with low reliability for a particular patient cohort across a particular set of providers **does not** imply that the same outcome will have low reliability for a different patient cohort or set of providers. **Nor does it mean that the outcome is fundamentally flawed for provider profiling.**

A Note of Caution

- If performance evaluations are required to reach arbitrarily high levels of reliability, two unintended consequences will follow:

(1) Many hospitals will be excluded from participation in useful performance evaluations

(2) Much useful performance information will be lost

Top Ten Takeaways

(1) Statistical reliability is a critical concept in assessing hospital performance.

(2) Statistical reliability quantifies the degree to which hospitals' odds ratios reflect true differences in hospital performance.

(3) Statistical reliability is not "reliability" as it is commonly understood. Low statistical reliability does not equate to "unreliable" in the pejorative sense.

(4) Statistical reliability alone does not determine whether performance assessment is useful.

(5) There is no minimum level of statistical reliability considered necessary for hospital profiling – an "acceptable" level of statistical reliability will depend on programmatic context.

(6) The statistical reliability of ACS NSQIP Essentials models is sufficient for differentiating hospital performance.

(7) The Procedure-Targeted program is evolving in terms of statistical reliability, with promising results to date.

(8) Low statistical reliability may be informative, as it may reflect consistent surgical mastery (or failure) across hospitals.

(9) A particular outcome does not have inherent statistical reliability – it will depend on how the outcome is modeled.

(10) If performance evaluations are required to reach arbitrarily high levels of statistical reliability, unintended consequences will follow.