

# Rating the Raters: An Evaluation of Publicly Reported Hospital Quality Rating Systems

Article · August 14, 2019

**Karl Y. Bilimoria, MD, MS, John D. Birkmeyer, MD, Helen Burstin, MD, MPH, Justin B. Dimick, MD, MPH, Karen E. Joynt Maddox, MD, MPH, Allison R. Dahlke, MPH, John Oliver DeLancey, MD, MPH & Peter J. Pronovost, MD, PhD**

Northwestern Medicine; Sound Physicians; Council of Medical Specialty Societies; University of Michigan; Washington University in St. Louis; University Hospitals

Originally published by NEJM Catalyst <https://catalyst.nejm.org/evaluation-hospital-quality-rating-systems>




## Introduction

The numerous currently available public hospital quality rating systems frequently offer conflicting results, which may mislead stakeholders relying on the ratings to identify top-performing hospitals. Given that there is no gold standard for how a rating system should be constructed or perform and no objective way to compare the rating systems, we evaluated the strengths and weaknesses of four major public hospital quality rating systems based on our experience as physician scientists with methodological expertise in health care quality measurement.

No rating system received an A or an F. The highest grade received was a B by U.S. News & World Report. The Centers for Medicare and Medicaid Services' (CMS) Star Ratings received a C. The lowest grades were for Leapfrog, C-, and Healthgrades, D+. Each rating system had unique weaknesses that led to potential misclassification of hospital performance, ranging from inclusion of flawed measures, use of proprietary data that are not validated, and methodological decisions.

More broadly, there were several issues that limited all rating systems we examined: limited data and measures, lack of robust data audits, composite measure development, measuring diverse hospital types together, and lack of formal peer review of their methods. Opportunities to advance the field of hospital quality measurement include the need for better data subject to robust audits, more meaningful measures, and development of standards and robust peer review to evaluate rating system methodology.



*Given the importance and need for health care quality transparency, we set out to fill this gap by undertaking a Rating the Raters process to evaluate and compare several major publicly reported hospital quality rating systems in the United States.”*

In this *Rating the Raters* initiative, we found that the current hospital quality rating systems should be used cautiously as they likely often misclassify hospital performance and mislead. These results can offer guidance to stakeholders attempting to select a rating system for identifying top-performing hospitals.

Over the past decade, publicly reported hospital quality rating systems have proliferated. These rating systems may be used in many ways: by patients when selecting where to receive care, by clinicians when deciding where to refer patients for care, by payers and purchasers interested in directing patients to certain hospitals or establishing contracts with high-quality hospitals, by payers in pay-for-

performance programs, and by hospital leaders to identify opportunities for improvement and to market their own performance.

However, it is unclear whether current rating systems are meeting stakeholders' needs. Such rating systems frequently publish conflicting ratings: Hospitals rated highly on one publicly reported hospital quality system are often rated poorly on another. This provides conflicting information for patients seeking care and for hospitals attempting to use the data to identify real targets for improvement.

Though some of the variation may be due to differing goals and measures included, some is likely due to differences in the validity and reliability of the underlying measures and the methods by which performance across measures is summarized. Moreover, there often seems to be a considerable disconnect between the top hospitals identified by the rating systems and those thought by clinicians to be major referral centers. Thus, the potential misclassification of hospital performance is a major concern in need of evaluation. However, to our knowledge, there has been no prior systematic review or evaluation of current rating systems that could help inform patients, clinicians, and policymakers of the various systems' methodologies, strengths, and weaknesses.

Given the importance and need for health care quality transparency, we set out to fill this gap by undertaking a *Rating the Raters* process to evaluate and compare several major publicly reported hospital quality rating systems in the United States: CMS Hospital Compare Overall Star Ratings, Healthgrades Top Hospitals, Leapfrog Safety Grade and Top Hospitals, and U.S. News & World Report Best Hospitals.

Given that there is no gold standard against which to compare these rating systems nor an ability to complete a formal comparative analysis of these rating systems, we sought to fill this void with an assessment of a group of experienced methodologists. Our objective was to provide users of these rating systems with insights into the relative strengths and weaknesses of each rating system, as well as to identify opportunities to improve the rating systems individually and the field as a whole. This is important because an evaluation of public hospital quality rating systems can offer guidance to stakeholders attempting to select a rating system for identifying top-performing hospitals, and the comparative evaluation may also offer the rating systems some insights on how to improve.

### **Approach for Rating the Raters**

The diverse group of six evaluators (Bilimoria, Birkmeyer, Burstin, Dimick, Maddox, Pronovost) includes established physician scientists with methodological expertise in health care quality measurement from both academic centers and the private sector. Given their experience in this field, all evaluators currently or previously have had some relationship with one or more of the rating systems. Thus, each evaluator was required to disclose the nature, timing, and financial arrangement of any current or prior relationships.

The group of evaluators was given the information on conflicts to review independently and then in person to determine when a conflict was perceived and whether the evaluator should be recused. Moreover, each rating system was asked if they had concerns about a conflict with any of the six evaluators. Evaluators were recused from grading a particular rating system if they had a direct current or recent relationship with the rating system itself. We have noted the details of the evaluator relationships in the Acknowledgements section.

*While we did score each system separately for each of our criteria, our goal was to provide an overall grade that represented a holistic evaluation of each of the ratings systems, with a particular focus on that system's potential for misclassifying hospital quality performance."*

We informed the rating systems of our project's intent at its outset, and asked for and received their cooperation throughout the process. This involved providing feedback at four steps of the project (described below and in [Appendix 1](#)) and answering specific questions about their rating systems by email and by phone. The rating systems were asked to review all materials that would eventually be used to grade them. Finally, leadership from each rating system attended an in-person meeting to answer specific questions about their rating system and to discuss the field of public reporting of quality measurement in general, and each rating system was interviewed independently.

Through iterative discussions with our group of coauthors, review of the literature, and discussions with leaders from each of the rating systems examined, we established six major criteria by which to assess these rating systems: Potential for Misclassification of Hospital Performance, Importance/Impact, Scientific Acceptability, Iterative Improvement, Transparency, and Usability (Table 1). We were particularly interested in assessing the potential for misclassification of hospital performance (i.e., incorrectly assessing the performance of a hospital) as we felt this was likely the most critical for avoiding unintended consequences for patients and providers.

#### Criteria Developed to Evaluate Hospital Quality Rating Systems

Domain	Examples of Criteria
Potential for Misclassification of Hospital Performance	Use of measures with serious known flaws Hospitals examined (number and types) Risk adjustment Composite methodology Methodological approach Audit mechanism
Importance/Impact	Unique features that resonate with patients, referring physicians, and hospitals
Scientific Acceptability	Balanced measurement Hospitals examined (number and types) Distribution/assignment of hospital grades/stars Use of available measures Use of unique data Specific methodological concerns Stability of rankings over time Audit mechanism
Iterative Improvement	Response to stakeholder feedback and scientific advances in measurement science Review of methods prior to release Peer review of methods Expert panel level of involvement
Transparency	Detailed methods report available (transparency) Clear rationale for methodological decisions Data availability (replicability) Financial conflicts and details regarding how ratings are monetized
Usability	Ease of overall use Ability to compare hospitals easily Attention to varying health literacy and numeracy

Source: The Authors  
NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society  
Table 1

Next, using information from the rating systems' websites, technical documentation, and available published literature, we created standardized Fact Sheets for each of the rating systems that included information about each rating system in a standardized format according to our evaluation criteria ([Appendix 1](#)). The information used was from the 2017 version of the rating systems. We have monitored since then for methodological changes in the rating systems, and no meaningful changes have occurred in the interim.

The Fact Sheets included objective, factual information (e.g., number of hospitals reviewed, number of elements included, risk-adjustment methodology selected). We gave each rating system the opportunity to review, provide input, and correct issues on their respective Fact Sheets; they all were responsive. These Fact Sheets served as guidance for our evaluation, along with detailed review of the rating system websites, their available technical and methodological documentation, and discussions with the leadership from each rating system. While all six evaluators reviewed all rating systems, two evaluators and two staff were assigned to each rating system to do a detailed review.

The *Rating the Raters* group then met in person to discuss each rating system in detail. From this discussion, we generated a Strengths and Weaknesses Summary, which described and categorized what we found beneficial and concerning for each of the publicly reported hospital quality rating systems based on our six evaluation criteria (Table 2; [Appendix 2](#)). We reviewed the Summary to ensure that critiques were applied consistently across the rating systems. The Strengths and Weaknesses Summary of our evaluation was then shared with each rating system to clarify certain points, obtain additional details, and engage in discussion on specific areas of concern. Each system responded, and their feedback was again incorporated where appropriate.

### Characteristics of Rating Systems

CMS Hospital Compare Overall Star Ratings	Healthgrades Top Hospitals	U.S. News & World Report Best Hospitals	Leapfrog Hospital Safety Grade and Leapfrog Top Hospitals
<b>Composite Description</b>			
Composite score star rating (1 to 5 stars)	Distinguished Hospital Award for Clinical Excellence (DHACE) and America's 50 and 100 Best Hospitals Award	National "Honor Roll" and "Best Hospitals" (includes 16 specialties and 9 procedures/conditions)	<ul style="list-style-type: none"> <li>Hospital Safety Grade: Composite letter safety grade</li> <li>Top Hospitals: Highest performing hospitals on the Leapfrog Hospital Survey</li> </ul>
<b>Data Source(s)</b>			
CMS claims, NHSN, HCAHPS, CMS reporting system (CASPER), Joint Commission, CMS Abstraction and Reporting Tool, CMS Clinical Data Warehouse	<ul style="list-style-type: none"> <li>CMS inpatient claims</li> <li>All-payer state data for ratings in Obstetrics/ Gynecology, Bariatric Surgery, and Appendectomy</li> </ul>	CMS inpatient claims, AHA Annual Survey, Hospital Compare for HCAHPS, public websites (STS, ACC, Magnet), Proprietary hospital reputation survey	<ul style="list-style-type: none"> <li>Hospital Safety Grade: Leapfrog Hospital Survey, CMS measures publicly reported on Hospital Compare (e.g., AHRQ PSIs, NHSN HAls, HACs, HCAHPS), AHA Annual Survey and Information Technology Supplement</li> <li>Top Hospitals: Leapfrog Hospital Survey</li> </ul>
<b>Measures</b>			
<ul style="list-style-type: none"> <li>Process: Best Practice Measures, Appropriateness of Medical Imaging</li> <li>Outcomes: Hospital-Acquired Infections (NHSN), 30-Day Readmissions and 30-Day Mortality, Surgical Complications, PSIs</li> <li>Patient Experience</li> </ul>	Outcomes: Proprietary 30-day Mortality and Inpatient Complications Measures for Selected Procedures and Conditions	Specialty Rankings: <ul style="list-style-type: none"> <li>Structural: Volume, Hospital Staffing, Clinical Technologies, External Recognition/ Accreditations, Patient Services, Nursing Magnet Status, STS and ACC Transparency</li> <li>Outcomes: Mortality, PSIs</li> <li>Reputation</li> </ul> Procedures and Conditions: <ul style="list-style-type: none"> <li>Structural: Volume, Nurse Staffing, Nursing Magnet Status, Intensivists, STS Clinical Registry Participation and Performance, Cardiac ICU, Heart Failure Program</li> <li>Process: VTE-2, noninvasive ventilation</li> <li>Outcomes: Mortality, Readmissions, Length of Stay, Surgical Complications (SSI, joint revision)</li> <li>Patient Experience</li> </ul>	Hospital Safety Grade: <ul style="list-style-type: none"> <li>Structural: Computerized Physician Order Entry, ICU Physician Staffing</li> <li>Outcomes: Hospital-Acquired Infections (NHSN), PSIs, Hospital-Acquired Conditions (e.g., falls, air embolism, retained foreign object)</li> <li>Patient Experience</li> <li>Safe Practices</li> </ul> Top Hospitals: <ul style="list-style-type: none"> <li>Structural: ICU Physician Staffing, Never Events policy</li> <li>Safe Practices Survey (Maternity Care, Infections and Injuries, Inpatient Care Management, Medication Safety)</li> <li>Value Score</li> <li>Hospital Safety Grade</li> <li>Qualitative component (evaluation of CMS mortality measures)</li> </ul>
<b>Composite Updated</b>			
Twice per year	Annually	Annually	<ul style="list-style-type: none"> <li>Hospital Safety Grade: Twice per year</li> <li>Top Hospitals: Annually</li> </ul>
<b>Measure Selection</b>			
Technical Expert Panel and vetted through public comment	Internal steering committee, external quality advisory board, and review of submitted comments	Measure selection is reviewed annually and determined by U.S. News and its data contractor RTI International with annual latent variable modeling, expert panel, and ongoing discourse with stakeholders	<ul style="list-style-type: none"> <li>Safety Grade: Expert panel</li> <li>Top Hospitals: Committee</li> </ul>

Abbreviations: NHSN: National Healthcare Safety Network, NTSV: Nulliparous Term Singleton Vertex, AHA: American Hospital Association, HCAHPS: Hospital Consumer Assessment of Healthcare Providers and Systems, HAI: Hospital-Acquired Infections, HAC: Hospital-Acquired Condition, PSI: Patient Safety Indicator


Source: The Authors.

NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

Table 2

We then met in person a few months later with leaders and/or methodologists from each of the ratings systems to further clarify issues and learn more about their systems, methodological decisions, and barriers, as well as have a robust discussion about how to advance the field in general. Our group of evaluators agreed unanimously on most of the points noted in the Strengths and Weaknesses Summary; however, the rating systems did not always agree with our opinions or conclusions.

Finally, each eligible evaluator independently assigned letter grades to each of the publicly reported hospital quality rating systems examined, and then the grades were averaged. We discussed the best way to present our findings, which are based on objective and subjective criteria, and determined that a point-by-point analysis of the strengths and weaknesses of each rating system, plus an explicit letter grade, would be most effective for this opinion piece. Those who had a conflict with a particular rating system were recused from assigning a grade for that particular rating system.



*These results offer a guide to stakeholders who are looking to assess hospital quality and wonder which rating system they should use.”*

The grading was done after our initial in-person review of each rating system. Once the Strengths and Weaknesses Summary was compiled, each evaluator then reviewed their grading again for each rating system. Then, the evaluators also had the chance to review their grades again following the in-person interviews with each rating system.

While we did score each system separately for each of our criteria, our goal was to provide an overall grade that represented a holistic evaluation of each of the ratings systems, with a particular focus on that system’s potential for misclassifying hospital quality performance. We likened this to scoring federal grants at a study section. The decision was made to give straight grades without a curve (i.e., we did not require giving any group an A or an F). The following anchors were used: An A would be an ideal rating system with little chance of misclassifying hospital performance, while an F would be a poor rating system that is more likely to misclassify a hospital’s performance than assign it correctly. C would suggest a mediocre rating system where a fair bit of misclassification of hospital performance is thought to occur. While the Consumer Reports hospital quality rating system was included in our evaluation, we have deleted mention of it in this manuscript as Consumer Reports is no longer conducting or posting its hospital quality evaluations.

## Rating the Hospital Quality Rating Systems

There were no hospital quality rating systems meriting an A or A-. The highest grade received was a B by U.S. News. The CMS Star Ratings received a C. The lowest grades were for Leapfrog (C-) and Healthgrades (D+) (Table 3). We qualitatively agreed that the U.S. News rating system had the least chance of misclassifying hospital performance. There was considerable agreement in overall grade assignments among the six individuals who performed the ratings.

### Overall and Criteria-Specific Grades for the Hospital Quality Rating Systems

	Average Grade	Grade Range	
		High	Low
<b>CMS Hospital Quality Star Ratings</b>			
<b>Overall Grade</b>	C	B-	C
Potential for Misclassification	D	C	D
Importance/Impact	C+	B	C
Scientific Acceptability	C+	B	C
Iterative Improvement	C-	B	D
Transparency	B	A	B
Usability	B	A	B
<b>Healthgrades Top Hospitals</b>			
<b>Overall Grade</b>	D+	C-	D
Potential for Misclassification	D	C	F
Importance/Impact	B	A	C
Scientific Acceptability	D+	C	D
Iterative Improvement	C	B	D
Transparency	D+	B	D
Usability	C	C	C
<b>U.S. News &amp; World Report Best</b>			
<b>Overall Grade</b>	B	B	B-
Potential for Misclassification	B	B	B
Importance/Impact	B	A	B
Scientific Acceptability	B	B	B
Iterative Improvement	B+	A	B
Transparency	B	A	C
Usability	B	B+	C
<b>Leapfrog Safety Score and Top</b>			
<b>Overall Grade</b>	C-	B-	D
Potential for Misclassification	C-	C	D
Importance/Impact	C+	B	C
Scientific Acceptability	C	B	D
Iterative Improvement	B-	A	C
Transparency	C	B	D
Usability	C+	B-	C

Notes: The "Overall Grade" for a rating system was a separate category assigned by each rater, not an average of the individual criteria for a rating system. "Potential for Misclassification" refers to the likelihood that the rating system incorrectly estimates true hospital performance.

Source: The Authors

NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

Table 3




## COMMON ISSUES ACROSS MOST RATING SYSTEMS

In reviewing the rating systems and discussing the issues with their leaders, we found several limitations to public reporting of hospital quality relevant to the field in general, resulting in a similar critique being applicable to all rating systems examined.

### Data and Measurement Limitations

Many of the rating systems use the same underlying data as the basis for their ratings. Some groups simply use the analyzed data CMS reports on [Hospital Compare](#) to capture already generated process, outcome, and [patient experience](#) metrics and build them into their own composites; others use the raw Medicare claims data to perform their own selected analyses of [patient outcomes](#).

Unfortunately, these administrative data, collected for billing rather than clinical purposes, have notable, [well-described shortcomings](#). The data used are generally limited to those 65 and older who participate in the Medicare Fee-for-Service program. The data often lack adequate granularity to produce valid risk adjustment. Moreover, outcomes reported in administrative data have been shown to have [high false-negative and false-positive rates](#). There are also notable ascertainment or surveillance bias issues that [invalidate some of these measures](#) (e.g., the [PSI-12 VTE outcome measure](#)). Most of the rating systems we reviewed include the Agency for Healthcare Research and Quality (AHRQ) Patient Safety Indicators (PSIs); these were developed for internal hospital surveillance purposes, not to compare hospitals to one another. Several [notable problems have been identified](#) with their use to compare hospital performance.



*Most rating system methodologies smooth or shrink the rates, essentially pushing lower-volume hospitals toward the mean. This results in smaller hospitals being essentially unable to be identified as poor performers or good performers.”*

Similarly, the [National Healthcare Safety Network \(NHSN\) measures](#) from the Centers for Disease Control and Prevention (CDC) address important, clinically relevant, and [potentially preventable hospital-acquired infections](#); however, ascertainment bias impacts these measures when it comes to comparing performance between hospitals. There are differences in identification of events (e.g., technologies available to abstractors, thoroughness and training of abstractors), (i.e., hospitals not reporting events that may not meet the intent of the measure, but technically qualify as an event), and concerns surrounding inadequacy of the risk adjustment (i.e., very few if any patient factors included). Undoubtedly, there is value in hospitals tracking their own rates of hospital-acquired infections, but serious limitations arise when comparing hospitals on the basis of these measures.

## Lack of Robust Data Audits

Many of the items used in the rating systems lack meaningful data audits. For example, there is not a robust audit performed for the data used in the PSI and NHSN measures. The Office of Inspector General has previously stated that CMS needs to improve its audit process to identify gaming (inappropriate intentional data manipulation) and inaccurate reporting of quality data. Similarly, when the rating systems generate their own data through surveys, these data are not always made available publicly for analysis to allow for independent assessment of validity and reliability. These data also need robust audits when self-reported by hospitals, and this is currently lacking.


## Composite Measure Development

The methods the rating systems use for compiling measures into composites to create an overall hospital score or grade vary tremendously, and there is often limited rationale for the selection and weighting of different elements in the composite. Moreover, there are frequently measures or domains that are weighted equally, though they are definitely not equal in the eyes of any stakeholder.

For example, readmissions and mortality were weighted equally in some rating systems. There is no question that experiencing a readmission is better than a death for all involved stakeholders. Mortality should be weighted much more heavily than a resource-use measure. Similarly, other composites inappropriately equate serious complications (e.g., mortality, hospital-acquired infections) with structural measures (e.g., computerized physician order entry), patient experience measures (e.g., communication about medicines), and CT-scan appropriateness measures.

## Handling Diverse Hospitals Together

A related issue is that of small hospitals. Rating systems often have difficulty handling outcomes measurement at smaller hospitals, which have lower volumes and therefore less reliable performance estimates. Currently, most rating system methodologies smooth or shrink the rates, essentially pushing lower-volume hospitals toward the mean. This results in smaller hospitals being essentially unable to be identified as poor performers or good performers. This masks hospital performance, may not reveal opportunities for improvement, and may mislead patients. Conversely, unsmoothed rates may be noisy and unfairly reward or penalize low-volume hospitals (i.e., if only four patients qualify for a measure, one death results in the highest mortality rate in the country, whereas no deaths results in the best mortality rate).



*Many rating sites claimed they were transparent by publishing their methods, but that is not the same as the transparency of reproducibility.”*

One straightforward solution would be stratifying hospitals into similar groups (e.g., large academic centers or small community hospitals). Stratifying ratings by hospital type is a reasonable approach, but patients may want to identify the best hospitals in their geographic area or nationally, not just the best critical access hospital or best community hospital. Though not used in any of the current systems, there is a potential opportunity to advance the field by shrinking a hospital’s performance estimates to that particular hospital’s

peer volume group or subtype’s prior average performance. This allows incorporation of information when there is a volume-outcome relationship and does not assume that small hospitals are average. Another option to handle the small numbers problem is to rely more heavily on process and patient experience measures as they are less affected by lower volumes than outcome measures.

### **Lack of Formal Peer Review**

The rating systems examined generally do not publish research on their measure testing and composite measure development methods, nor do they submit their methods to journals for formal, rigorous external peer review. All the groups use expert panels to varying degrees, but typically expert panels provide input intermittently and without detailed methodological review. This is quite different than the rigorous peer review of well-respected journals. The field as a whole would benefit from this type of rigor.

### **Potential Financial Conflicts**

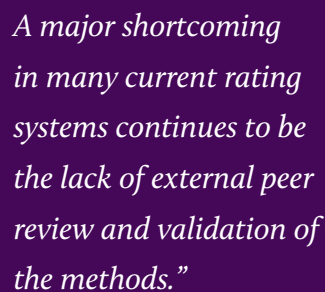
Monetizing ratings — that is, having hospitals pay the rating systems to be able to display their performance or to allow use of their ratings for hospital marketing or advertising — may create unfortunate incentives. Specifically, there is a concern that the business of selling these ratings leads to a model that encourages multiple rating systems to intentionally identify different “best hospitals.” By using different measures, analytic methods, and composite development approaches, the rating systems can identify very different hospitals as the top hospitals. Thus, most hospitals in the country can purchase a “best hospital” emblem of some type from at least one of these rating systems.

## ISSUES SPECIFIC TO INDIVIDUAL RATING SYSTEMS

While there are several issues that affect the field of hospital quality measurement in general, we need to separate these from specific methodological decisions made by individual rating systems ([Appendix 2](#)). Where there is heterogeneity among the four rating systems in use of a measure (e.g., the PSI-12 VTE outcome measure) or in the methodological approach to composites, then that is a decision the rating systems have made and could elect to do differently.

### CMS Star Ratings

The CMS Star Ratings have some unique strengths. The ratings carry considerable weight as they are put forth by a federal agency and the largest payer in the country. Other rating systems noted that they often included measures or methods “because CMS does,” so the CMS Star Ratings have an important influence on other rating systems. CMS has assembled multiple Technical Expert Panels to guide their decision-making. They have made their code available to allow replication of much of their analyses. They do not monetize their rating system. Their website is also to be commended for usability and facilitating comparisons between hospitals.



*A major shortcoming in many current rating systems continues to be the lack of external peer review and validation of the methods.”*

However, there were several weaknesses regarding the CMS Overall Hospital Quality Star Ratings. First, there is likely a high rate of misclassifying hospital performance given the inclusion and comparison of a [heterogeneous collection of hospitals into a single group](#). Large academic medical centers that report nearly all component measures are compared to critical access hospitals and specialty hospitals reporting less than half of the component measures.

Second, the weighting appears fairly arbitrary, so there would be value in bringing some rigor to weighting of measure domains and component measures. Third, there are few, if any, diagnosis- or procedure-specific measures for elective conditions. Many of the disease-specific measures are for nonelective admissions like myocardial infarction, for which patients do not have the liberty to compare hospitals ahead of time. Fourth, one of the most heavily weighted measures in the Star Rating composite is [PSI-90](#), the AHRQ patient safety and adverse events composite. This measure has been improved recently, but it still has several flaws that should preclude such heavy emphasis in the rating systems.

Finally, CMS continues to use several measures that other rating systems have deemed not valid for comparing hospital quality and excluded from their rating system (e.g., PSIs). CMS has statutory limitations that preclude changing some aspects of its Star Ratings, but there is considerable opportunity and need to improve multiple aspects of this highly visible and influential federal rating system from the largest payer in the country.

## Healthgrades

Healthgrades had some unique strengths. They include procedure- and condition-specific rankings that offer more granular information to patients in selecting a hospital and to hospitals seeking to identify improvement targets. Also, Healthgrades does not use the CDC's NHSN measures; that is a strength because Healthgrades insists on being able to run its own analyses rather than reusing aggregated data from others.

There were some notable weakness regarding the Healthgrades composite measure. First, the composite only contains outcome measures. Our group emphasized the need for balanced measurement with the inclusion of all domains of quality. This is particularly important here given the concerns with administrative data for outcomes measurement. Moreover, some of the data they use in their composite are only available for certain states, which precludes equitable comparisons nationally.

Importantly, their methods are not sufficiently described to allow replication and evaluation. An arbitrary 90% confidence interval is employed to identify outliers on individual measures. They still include a number of flawed PSIs. Healthgrades also evaluates all hospital types together leading to misclassification concerns. There also appear to be some important inconsistencies in the codes they count as complications, as many of these would be unrelated to the primary diagnosis or procedure. We also noted that their Expert Panel does not include any methodologists with expertise in the science quality ratings.


## Leapfrog

Leapfrog has some important strengths. They take a balanced measurement approach that includes all domains of quality (structure, process, outcomes, and patient experience). Moreover, no other rating system includes an assessment of the culture of safety, and the coauthors believed that was an important, unique feature. While many rating systems had fairly arbitrary weighting approaches, Leapfrog uses an approach that incorporates expert evaluation of measure impact, opportunity, and evidence basis. Hospitals also receive a calculator to replicate or predict scores.

However, there were several important weaknesses for the Leapfrog rating systems, specifically with the Safety Score and the Top Hospitals designation. The greatest concerns were with their internally developed and modified Safety Survey. The survey is self-reported and there is not a robust audit in place; the audits are done on very few hospitals. During our discussion, Leapfrog leadership stated that they had only done a formal audit for approximately five hospitals of about 2,600 in the past year, and only 72 hospitals underwent an electronic audit. Concerns were also raised about the value of many of the items on the survey as they may not truly reflect patient safety efforts or be currently meaningful to stakeholders (e.g., computerized physician order entry).

In addition, many other important aspects of quality are excluded from their rating system. For example, the Leapfrog rating system excludes mortality, as their team believes it is “not a safety metric.” The coauthors disagreed with this assessment and noted that the exclusion of mortality was a notable oversight.

The assessment of all hospital types together for the Safety Score was an issue, but when Leapfrog denotes “Top Hospitals,” they do collect separate hospitals into peer groups (e.g., general, academic, rural). Leapfrog continues to use most PSIs; whereas, other groups have dropped certain PSIs based on evidence and validity concerns.



*PROs, which are distinct from patient experience measures, are important measures of successful health care management, but they are completely absent from all current publicly available hospital quality rating systems.”*

Another major area of concern was how Leapfrog assesses hospitals that do not answer their Safety Survey. When hospitals do not report the survey, the missing data are filled in from other secondary sources. However, hospitals that answer the survey and those that do not are assessed in the same way despite not having the same measures upon which to base the ratings. Approximately 50% of hospitals answer their survey, so a good deal of the rankings are based on missing or inconsistent data.

Leapfrog uses unadjusted, internally developed central line infection and urinary tract infection measures rather than other more standard measures. While flawed for hospital

quality comparisons, the NHSN measures are at least somewhat standardized and have some minimal risk adjustment. Finally, when Leapfrog assembles the list of Top Hospitals, they noted that they have a subjective component that allows elimination of the hospital from the Top Hospitals list if they are poor performers on the mortality rates reported by CMS (even though mortality is excluded from their Safety Grade), but no defined criteria were articulated (e.g., poor performance on 1 of 6 vs. all 6 mortality metrics). While we acknowledge that mortality is a very important measure, this subjective component seemed antithetical to an objective hospital quality rating.


## U.S. News

This rating system was thought by the authors to be the most responsive to changes in measurement science and feedback from stakeholders. They revise their rating system annually to address measurement issues that have come to their attention from experts, the literature, hospitals, or their internal investigations. This is largely beneficial, although it limits year-to-year comparisons, as hospitals that shift ranks may be unable to determine whether the change was due to change in performance or ranking methodology.

Some notable recent changes for the U.S. News rating system include eliminating all NHSN measures and most PSIs, weighting volume for proportion of Medicare Advantage patients, improving outcome measures with exclusion of external transfers, and adding risk adjustment for sociodemographic factors. One important needed improvement would be to release these changes and supporting evidence further in advance of releasing the new rankings to allow time for more widespread input and peer review. Although we liked the incorporation of patient experience into the procedure/condition-specific rankings, the patient experience scores are not specific to that specific group; they reflect all specialties combined and only come from the inpatient survey.

A unique feature of the U.S. News ranking that is in some ways both a strength and a weakness is the inclusion of the “reputation” domain. This component is ascertained through surveys of practicing providers in 16 specialties by asking them where they would send their most challenging cases. Respondents can select up to five hospitals, including their own. The Reputation Survey is worth 27.5% of the score for overall ranking for 12 specialties for which most of the score is based on data, and 100% for the four specialties (ophthalmology, psychiatry, rehabilitation, and rheumatology) that are assessed solely on the reputational surveys.

The inclusion of reputation in the U.S. News rankings is widely debated, but our group of coauthors concluded that this is generally a beneficial component. It serves as the question that we would ask as physicians when referring patients or that patients would ask us when deciding where to go for complex care. The Reputation Survey is also a surrogate for measures that are currently not readily available or collected, such as availability of expert services and specialists, unique technologies and innovation, and clinical trials.



*There is no acknowledged gold standard for misclassification of hospital performance, nor for assigning grades to hospital rating systems.”*

There are, however, some important concerns about how U.S. News conducts the survey, and we believe there may be opportunities for bias and gaming of the survey. Specialties where reputation is the only metric are not a balanced representation of quality care (i.e., no inclusion of outcomes or patient experience), and additional metrics in those specialties are needed. Finally, U.S. News does not make any of their Reputation Survey data available publicly for external analysis and validation. Until they do, reputation data will

be a source of controversy. That said, as a group, we felt that the Reputation Survey offers valuable information to patients until better measures are available to reflect this concept.

Another feature that was beneficial about the U.S. News system was its inclusion of volume as a quality measure, which none of the other rating systems include. While a surrogate, volume remains an important and valuable measure of quality, particularly where more specific quality assessments are not available. This is an easy measure that others should consider including in their rating systems.

U.S. News also includes specialty- and procedure-specific rankings, so patients can ascertain information about their particular needs, rather than just a global overall hospital evaluation. High-acuity, high-complexity conditions and procedures are included, as well as common procedures. However, side-by-side comparisons of hospitals on overall and specific components of the rating system were not easily done. U.S. News incorporates Society of Thoracic Surgeons registry data into its rankings; the database includes components that focus on cardiothoracic surgery. This strengthens the U.S. News rating system, but there are many other registries that should be considered for inclusion.

### **Opportunities to Advance the Field**

As an individual hospital's performance can be rated quite differently between seemingly similar public hospital quality rating systems, we sought to assess the four major rating systems to provide insights about their relative strengths and weaknesses. Given that an empirical comparison is not possible due to lack of a gold standard and availability of all underlying data and methods, a thorough evaluation by experienced methodologists may offer the best way to compare these rating systems. We were particularly focused on the potential for misclassification of hospital performance as this could have important adverse consequences and mislead stakeholders seeking to use a rating system to assess hospital quality. We found that there were many limitations to the field of public reporting, but there were also several methodological decisions that rating systems made that resulted in better or worse assessments of hospital quality.



These results offer a guide to stakeholders who are looking to assess hospital quality and wonder which rating system they should use. This is certainly important for patients and referring doctors, but payers and purchasers are increasingly using these rating systems to direct their patients or establish contracts. Finally, hospitals themselves often use one of these rating systems to assess their performance and set goals. Thus, providing a comparison of these rating systems may be beneficial to those seeking to use one of these ratings systems.


We have identified several opportunities that can advance the field of hospital quality rating systems.

### **BETTER DATA**

The field definitely needs better data: All rating systems rely on administrative data or self-reported data, which have numerous limitations. All rely heavily on Medicare claims data, which represents an important part of the population, but using all-payer data would present a more accurate and complete representation of quality. Incorporation of registry data would be helpful, understanding that this requires a partnership with and permission from the registry owners. Nonetheless, while registries would in many ways be an ideal alternative to overcome the limitations of administrative data, the abstraction required is laborious and expensive, and this results in only a fraction of hospitals participating in most registries with only a fraction of a hospital's relevant cases being captured by the registry. Thus, we need to make major strides in moving toward new methods of obtaining data directly from the electronic health record to support valid, meaningful quality metrics. While there is much discussion of interoperability, little has been translated to meaningful national quality measures.

### **BETTER MEASURES**

We need to leverage the data to create quality measures that are valid, valuable, and timely. The currently available measures fall far short in many domains and suffer from inadequate risk adjustment, questionable relationship to outcomes, and unacceptable lag times. There are also areas of quality that are entirely missing from the rating systems, such as collection, analysis, and patient-reported outcomes (PROs). PROs, which are distinct from patient experience measures, are important measures of successful health care management, but they are completely absent from all current publicly available hospital quality rating systems. Similarly, no long-term measures are included, such as cancer recurrence and survival, undoubtedly important measures for assessing health care quality. Finally, rating systems must become nimble and adapt more rapidly to changes in measurement science.



*Stakeholders want, need, and deserve comparative public data on hospital quality, and rating systems play an important role in providing such information. However, there are improvements needed to advance the field as a whole and opportunities to improve.”*

## MEANINGFUL AUDITS

We need to ensure that all data used in these ratings systems are subject to a strong audit program to ensure the data are valid and allow for fair comparisons of providers. In no other industry undertaking high-stakes public reporting and pay-for-performance would the lack of a meaningful audit be tolerated. Detailed audits should be required before the data and resulting rating systems will be widely accepted. The results of the audits must be made public and available for analysis. Until this is addressed, there will continue to be real questions about the validity of these rating systems. Moreover, all data sources and methods used in rating systems must be transparent and available publicly. Many rating sites claimed they were transparent by publishing their methods, but that is not the same as the transparency of

reproducibility. The actual underlying data need to be available for there to be reproducibility allowing external analysis and validation.

## EXTERNAL PEER REVIEW

A major shortcoming in many current rating systems continues to be the lack of external peer review and validation of the methods. In this *Rating the Raters* initiative, our group of coauthors served as an unsolicited peer review of the methodology of these rating systems, but if we were reviewing a manuscript of a rating system’s methodological changes for publication in a journal, we would require much more description of the statistical methods, better validation of underlying data, many more sensitivity analyses, and more justification of the rationale for methodological decisions before we could even begin to render a decision on whether to accept or reject the study.

We encourage all rating systems to submit studies of their analytical approach, decisions, and periodic modifications for real peer review and publication, preferably well ahead of employing them in public ratings. This would encourage innovation and improvement over time. Moreover, an independent standing group could be set up to establish standards and evaluate public rating systems based on requirements and criteria, similar to the approach the National Quality Forum (NQF) uses for individual measures or as in the [financial sector by the Financial Accounting Standard Board](#).

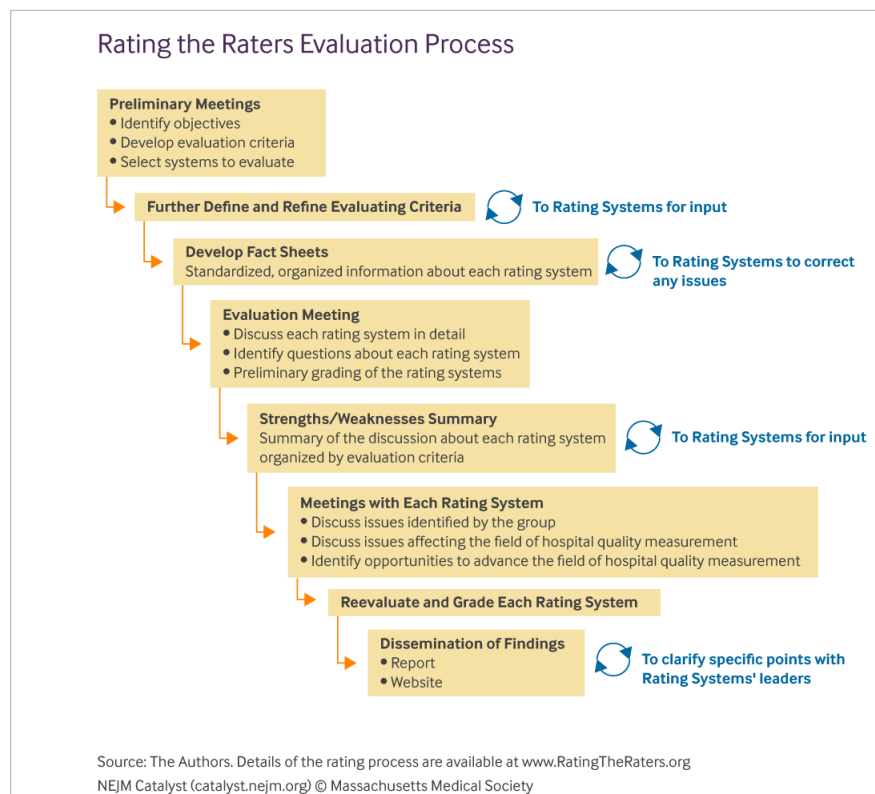
## Limitations

This evaluation has some important limitations. First, there is no acknowledged gold standard for misclassification of hospital performance, nor for assigning grades to hospital rating systems — this was the reason we took on the process in the first place, but we consider these findings to be a starting point rather than a definitive last word on the matter. Because stakeholders' priorities and preferences may differ, a different group of people could have come to alternative conclusions. Second, we did not evaluate every publicly available rating, nor did we evaluate every rating put out by these four rating systems. Finally, most of the evaluators have worked at academic medical centers, which may influence our grading. However, many of us have or previously had positions where we also have some responsibility for community hospitals.

## Next Steps for Hospital Quality Rating Systems

Stakeholders want, need, and deserve comparative public data on hospital quality, and rating systems play an important role in providing such information. However, there are improvements needed to advance the field as a whole and opportunities to improve each of these four hospital quality rating systems before these ratings will truly meet the needs of stakeholders. Until then, these rating systems should be interpreted very cautiously, as most seem likely to misclassify hospital performance and may mislead patients, referring doctors, payers, purchasers, and hospitals themselves.

## Appendix 1



## Appendix 2

### Rating the Raters – Strengths and Weaknesses Assessment of the Four Public Hospital Quality Rating Systems

The comments in the table below reflect the discussion that the Rating-the-Raters group had about each rating system. These comments for each rating system were provided to the leaders of that rating system to solicit feedback.

#### Importance/Impact

(CMS) Hospital Compare Overall Star Ratings	Healthgrades Top Hospitals	USNWR (U.S. News & World Report) Best Hospitals	Leapfrog Hospital Safety Grade and Leapfrog Top Hospitals
<b>Pro</b>			
<ul style="list-style-type: none"> <li>• Compiled by largest payer in U.S.</li> </ul>	<ul style="list-style-type: none"> <li>• Procedure- and condition-specific rankings</li> </ul>	<ul style="list-style-type: none"> <li>• Overall, specialty and procedure/condition rankings are helpful</li> <li>• Useful and rigorous rating system overall</li> <li>• High-complexity and high-acuity measures where quality tends to vary, but also focuses on more common procedure areas</li> <li>• Measures generally have high face validity</li> <li>• Inclusion of registry data (STS data for CABG and AVR) in procedure/condition rankings</li> </ul>	<ul style="list-style-type: none"> <li>• Focus on safety</li> <li>• Includes assessment of culture of safety</li> </ul>
<b>Con</b>			
<ul style="list-style-type: none"> <li>• Very few elective condition- or procedure-specific measures (most are less common conditions or not elective admissions)</li> </ul>		<ul style="list-style-type: none"> <li>• Paucity of clinically meaningful and rigorous outcomes data (e.g., mortality considered quality metric and not safety metric)</li> <li>• No information for individual common elective procedures/conditions</li> <li>• Safety focused, but misses a lot of other important measures for quality and should be more balanced</li> </ul>	

Source: The Authors. Details of the rating process are available at [www.RatingTheRaters.org](http://www.RatingTheRaters.org)  
 NEJM Catalyst ([catalyst.nejm.org](http://catalyst.nejm.org)) © Massachusetts Medical Society

### Rating the Raters – Strengths and Weaknesses Assessment of the Four Public Hospital Quality Rating Systems

The comments in the table below reflect the discussion that the Rating-the-Raters group had about each rating system. These comments for each rating system were provided to the leaders of that rating system to solicit feedback.

Scientific Acceptability			
(CMS) Hospital Compare Overall Star Ratings	Healthgrades Top Hospitals	USNWR (U.S. News & World Report) Best Hospitals	Leapfrog Hospital Safety Grade and Leapfrog Top Hospitals
<b>Pro</b>			
<ul style="list-style-type: none"> <li>Incorporates process, outcomes, and Patient Experience measures</li> <li>The star rating is typically updated twice/year</li> </ul>	<ul style="list-style-type: none"> <li>Does not use NHSN measures</li> <li>Data updated annually</li> <li>Volume-based weighting of performance</li> </ul>	<ul style="list-style-type: none"> <li>Incorporates structure, process, outcomes, reputation</li> <li>Compares hospitals in different ways to allow grouping of similar types of hospitals</li> <li>Eliminated most PSIs</li> <li>Eliminated all NHSN measures</li> <li>Structural measures (e.g., volume) mitigate some of the measurement issues with outcome metrics</li> <li>Reputation measure offers some information where there is a lack of more granular measures capturing the same concept</li> <li>Risk adjustment for socioeconomic status in 12 mortality rates</li> <li>Excluded external transfers from outcomes measures</li> <li>Adjusts volume in each specialty to account for regional differences in Medicare Advantage Enrollment</li> </ul>	<ul style="list-style-type: none"> <li>Incorporates process, structural, outcomes, and Patient Experience measures</li> <li>Scientifically rigorous composite methodology</li> <li>Impact Score for weighting approach is available and equally applied based on expert evaluation of impact, opportunity, and evidence base of measure</li> <li>Z-scores are used to standardize data from measures with different scales and are applied to measures available for each hospital</li> <li>Calculator available to hospitals to replicate measures scores, weights, and total score</li> <li>The Safety Grade rating is updated twice/year</li> <li>Hospitals must complete all applicable sections of survey to be included in Top Hospitals</li> <li>Top Hospitals ratings separates peer groups (General, Rural, Teaching)</li> </ul>
<b>Con</b>			
<ul style="list-style-type: none"> <li>Some concerns regarding measure weighting approach</li> <li>Attempts to measure diverse hospital types together (major issue)</li> <li>Concerns regarding approach for assigning stars (e.g., k-means clustering)</li> <li>Use of PSIs, particularly PSI-90 and PSI-4</li> <li>PSI-90 and readmission weighted too heavily</li> <li>Use of NHSN infection measures</li> <li>No inclusion of clinical registry measures</li> <li>Inclusion relatively unimportant imaging measures</li> <li>No robust data audit process</li> <li>Data lag can be up to 3 years from collection to release</li> <li>Administrative data and NHSN data are not rigorously audited</li> </ul>	<ul style="list-style-type: none"> <li>Not balanced measurement. No process measures or patient experience; composite relies only on outcome measures.</li> <li>Patient experience scores not included – shown as a separate rating.</li> <li>Measures vary by state availability</li> <li>Evaluating all hospital types together is a major issue</li> <li>Proprietary methodology is not transparent and cannot be replicated and thoroughly evaluated</li> <li>Assigning best hospitals based on percentiles of raw scores is a major concern (arbitrary threshold, top 5%)</li> <li>No statistical testing done in “overall ranking” – just top 5% of hospitals</li> <li>Inclusion of PSIs, particularly PSI-3, 4, 7, 12, 13</li> <li>Outcome measures and clinical cohorts (e.g., small bowel obstruction) are not conditions where patients typically can use the comparisons to decide where to go for care</li> <li>No registry data included</li> <li>Some complications are not related or relevant to the procedure (e.g., Legionnaire’s disease is a complication for AAA repair)</li> <li>Administrative data are not rigorously audited</li> </ul>	<ul style="list-style-type: none"> <li>Still uses some flawed PSIs</li> <li>Patient experience measures/Patient Experience used only for Procedures and Conditions rankings</li> <li>Rankings done on hospitals of different types together (e.g., critical access, teaching)</li> <li>Some concerns in their “reputation” measurement methodology with respect to sampling and ranking own institution highly (i.e., gaming)</li> <li>Some specialties are ranked on reputation alone without any other measures of quality (Ophthalmology, Psychiatry, Rehabilitation, Rheumatology)</li> <li>Limited use of registry data, except to give credit for participation in certain registries (but missing other major registries)</li> <li>Hospitals missing patient safety data are assigned median patient safety score (PSI) for all hospitals</li> <li>Administrative data are not rigorously audited</li> <li>Risk adjustment different for procedures/conditions</li> <li>Some metrics developed in-house have not been scientifically vetted in a robust fashion</li> </ul>	<ul style="list-style-type: none"> <li>Voluntary, self-reported survey data account for 35% of score, and the survey data are not rigorously validated. This is a major concern.</li> <li>Concerns about handling of hospitals not reporting (~50%) and about the corresponding missing data</li> <li>Evaluating all hospital types together is a major issue</li> <li>Very few clinically meaningful and rigorous outcomes data (exclusion of mortality is a major issue)</li> <li>Inclusion of PSIs</li> <li>Some particularly weak component measures (e.g., CPOE)</li> <li>Hospitals still included if they don’t answer Leapfrog survey</li> <li>Methodology report does not comment on how reliable the primary and secondary data sources are with one another (used when primary not reported by hospital).</li> <li>Use of legacy NQF Safe Practices (not maintained anymore), not intended to be scored, but rather intended to be used as an improvement tool</li> <li>Administrative data are not rigorously audited</li> <li>Use of NHSN infection measures</li> <li>NHSN data are not rigorously audited</li> <li>Leapfrog audits sample a very small number of hospitals annually</li> <li>Audit results not released publicly</li> <li>Voluntary, self-reported survey data account for 100% of score for Top Hospitals, but the survey data are not rigorously validated or relevant. This is a major concern.</li> <li>Use of non-risk adjusted infection measures</li> </ul>

Source: The Authors. Details of the rating process are available at [www.RatingTheRaters.org](http://www.RatingTheRaters.org)  
 NEJM Catalyst ([catalyst.nejm.org](http://catalyst.nejm.org)) © Massachusetts Medical Society

## Rating the Raters – Strengths and Weaknesses Assessment of the Four Public Hospital Quality Rating Systems

The comments in the table below reflect the discussion that the Rating-the-Raters group had about each rating system. These comments for each rating system were provided to the leaders of that rating system to solicit feedback.

### Potential for Misclassification of Hospital Performance

(CMS) Hospital Compare Overall Star Ratings	Healthgrades Top Hospitals	USNWR (U.S. News & World Report) Best Hospitals	Leapfrog Hospital Safety Grade and Leapfrog Top Hospitals
<b>Pro</b>			
<ul style="list-style-type: none"> <li>Some data integrity checks in place to determine anomalous data</li> </ul>		<ul style="list-style-type: none"> <li>Least likely of the major rating systems to misclassify hospital performance</li> </ul>	<ul style="list-style-type: none"> <li>Some data integrity checks in place to determine anomalous data</li> </ul>
<b>Con</b>			
<ul style="list-style-type: none"> <li>High potential for misclassification</li> <li>Inclusion of PSIs</li> <li>Use of NHSN measures</li> <li>High risk of misclassification due to the inclusion and comparison of heterogeneous hospital types that do not report the same numbers and types of measures. (e.g., Critical Access Hospitals and Specialty hospitals do not report most of the measures used)</li> <li>Concerns regarding adequacy of risk adjustment with administrative data</li> <li>Many measures lead to paradoxical misclassification, and thus likely demonstrate the inverse of quality (PSI-3, PSI-12)</li> <li>NHSN data are not rigorously audited</li> </ul>	<ul style="list-style-type: none"> <li>High potential for misclassification</li> <li>Inclusion of PSIs</li> <li>Evaluating all hospital types together is a major issue</li> <li>A lot of potential misclassification and noise in codes included in outcomes measures</li> <li>Concerns regarding adequacy of risk adjustment with administrative data</li> <li>Administrative data are not rigorously audited</li> <li>Many measures lead to paradoxical misclassification, and thus likely demonstrate the inverse of quality (PSI-3, PSI-12)</li> </ul>	<ul style="list-style-type: none"> <li>Lower likelihood of misclassification</li> <li>Inclusion of some PSIs</li> <li>Some rankings based on “reputation” only</li> <li>Administrative data are not rigorously audited</li> <li>Concerns regarding adequacy of risk adjustment with administrative data</li> </ul>	<ul style="list-style-type: none"> <li>High potential for misclassification based on issues with self-reported Leapfrog survey (gaming, lack of robust audit) and some outcomes subject to surveillance bias and ascertainment issues</li> <li>Inclusion of PSIs, particularly PSI-12</li> <li>Use of NHSN measures</li> <li>Concerns about self-report of process measures</li> <li>Administrative data are not rigorously audited</li> <li>Concerns regarding adequacy of risk adjustment with administrative data</li> <li>Many measures lead to paradoxical misclassification, and thus likely demonstrate the inverse of quality (PSI-3, PSI-12)</li> <li>NHSN data are not rigorously audited</li> <li>Use of non-risk adjusted infection measures</li> </ul>

Source: The Authors. Details of the rating process are available at [www.RatingTheRaters.org](http://www.RatingTheRaters.org)  
 NEJM Catalyst ([catalyst.nejm.org](http://catalyst.nejm.org)) © Massachusetts Medical Society

## Rating the Raters – Strengths and Weaknesses Assessment of the Four Public Hospital Quality Rating Systems

The comments in the table below reflect the discussion that the Rating-the-Raters group had about each rating system. These comments for each rating system were provided to the leaders of that rating system to solicit feedback.

### Iterative Improvement

(CMS) Hospital Compare Overall Star Ratings	Healthgrades Top Hospitals	USNWR (U.S. News & World Report) Best Hospitals	Leapfrog Hospital Safety Grade and Leapfrog Top Hospitals
<b>Pro</b>			
<ul style="list-style-type: none"> <li>• Have assembled multiple Technical Expert Panels of diverse stakeholders</li> </ul>	<ul style="list-style-type: none"> <li>• They have become more transparent over time</li> </ul>	<ul style="list-style-type: none"> <li>• Has been very responsive generally to feedback from stakeholders and scientific advances</li> <li>• Most responsive of the various rating systems to changes in measurement science, recent literature, and stakeholder feedback</li> <li>• Uses expert panel effectively</li> <li>• Ranking for specific specialty or condition will not be shown if data are missing</li> </ul>	<ul style="list-style-type: none"> <li>• Safety Grade: Uses national expert panel</li> <li>• Top Hospitals: Uses Technical and Content Experts for measure selection</li> </ul>
<b>Con</b>			
<ul style="list-style-type: none"> <li>• No robust peer review of methods prior to release</li> <li>• Some concerns regarding incorporation of feedback in recent literature and from Technical Expert Panels</li> </ul>	<ul style="list-style-type: none"> <li>• No robust public or peer review/comment of methods prior to release</li> <li>• Some concerns about lack of incorporation of stakeholder feedback and advances in science</li> <li>• Lack of expert methodologies to Advisory Panel</li> </ul>	<ul style="list-style-type: none"> <li>• No robust public or peer review/comment of methods prior to release</li> <li>• Frequently releases changes without opportunity for public comment well in advance</li> </ul>	<ul style="list-style-type: none"> <li>• No robust public or peer review/comment of methods prior to release</li> <li>• Concerns regarding lack of responsiveness to the issues raised by their Expert Panel, hospitals, scientific advancements, and other stakeholders</li> </ul>

Source: The Authors. Details of the rating process are available at [www.RatingTheRaters.org](http://www.RatingTheRaters.org)  
 NEJM Catalyst ([catalyst.nejm.org](http://catalyst.nejm.org)) © Massachusetts Medical Society

## Rating the Raters – Strengths and Weaknesses Assessment of the Four Public Hospital Quality Rating Systems

The comments in the table below reflect the discussion that the Rating-the-Raters group had about each rating system. These comments for each rating system were provided to the leaders of that rating system to solicit feedback.

### Transparency

(CMS) Hospital Compare Overall Star Ratings	Healthgrades Top Hospitals	USNWR (U.S. News & World Report) Best Hospitals	Leapfrog Hospital Safety Grade and Leapfrog Top Hospitals
<b>Pro</b>			
<ul style="list-style-type: none"> <li>• Data and statistical code made available for some ability to replicate analyses</li> <li>• Extensive methodology description</li> <li>• No monetization of ratings</li> </ul>	<ul style="list-style-type: none"> <li>• None discussed</li> </ul>	<ul style="list-style-type: none"> <li>• Extensive methodology description</li> <li>• Detailed scores and rankings are published in methodology report for top hospitals</li> </ul>	<ul style="list-style-type: none"> <li>• Public information on website about how they monetize their product (costs of using emblems, promotional materials, etc.)</li> </ul>
<b>Con</b>			
<ul style="list-style-type: none"> <li>• Important details missing from methodology report (e.g., details on weighting approach)</li> <li>• Not transparent how all measures are weighted; weighting is vetted by Technical Expert Panel and stakeholders, but no further details</li> <li>• Unclear rationale for some methodological decisions</li> </ul>	<ul style="list-style-type: none"> <li>• Much less transparent than other systems</li> <li>• Proprietary models that are not transparent</li> <li>• Inadequate information to judge validity and appropriateness of methodological decisions</li> <li>• No public information on how they monetize their product (e.g., costs of using emblems, promotional materials, etc.)</li> </ul>	<ul style="list-style-type: none"> <li>• Reputation survey data are not made available for analysis and verification</li> <li>• No public information how they monetize their product (e.g., costs of using emblems, promotional materials, etc.)</li> </ul>	<ul style="list-style-type: none"> <li>• None discussed</li> </ul>

Source: The Authors. Details of the rating process are available at [www.RatingTheRaters.org](http://www.RatingTheRaters.org)  
 NEJM Catalyst ([catalyst.nejm.org](http://catalyst.nejm.org)) © Massachusetts Medical Society



## Rating the Raters – Strengths and Weaknesses Assessment of the Four Public Hospital Quality Rating Systems

The comments in the table below reflect the discussion that the Rating-the-Raters group had about each rating system. These comments for each rating system were provided to the leaders of that rating system to solicit feedback.

### Usability

(CMS) Hospital Compare Overall Star Ratings	Healthgrades Top Hospitals	USNWR (U.S. News & World Report) Best Hospitals	Leapfrog Hospital Safety Grade and Leapfrog Top Hospitals
<b>Pro</b>			
<ul style="list-style-type: none"> <li>• User friendly</li> <li>• Easy to find a hospital</li> <li>• Allows for comparisons of multiple hospitals</li> <li>• Website indicates how current data are</li> <li>• Uses graphical displays</li> </ul>	<ul style="list-style-type: none"> <li>• Easy to find a hospital</li> <li>• Ratings searchable based on specific procedures and conditions</li> <li>• Video explaining what stars mean</li> </ul>	<ul style="list-style-type: none"> <li>• Patients can readily identify which procedure /specialty-specific ranking are most applicable to their clinical circumstance</li> </ul>	<ul style="list-style-type: none"> <li>• Easy to use</li> <li>• Color coding is helpful</li> <li>• Multiple ways to sort and compare hospitals</li> <li>• Grade is displayed prominently and is very apparent to user</li> <li>• Has a "how to use Leapfrog hospital Safety Grade" video tool posted to website</li> <li>• Top Hospitals ratings breaks up hospitals' rankings by hospital type</li> </ul>
<b>Con</b>			
<ul style="list-style-type: none"> <li>• None discussed</li> </ul>	<ul style="list-style-type: none"> <li>• Not easy to compare hospitals</li> <li>• Uses 3 options only on a 5-star scale (1, 3, 5)</li> <li>• Each measure category is displayed differently (e.g., Stars, percentage better/average/worse, or overall "patient safety" score)</li> </ul>	<ul style="list-style-type: none"> <li>• No comparison tool; difficult to compare hospitals</li> <li>• A lot of information, many clicks needed</li> <li>• Underlying hospital data not shown in user-friendly format</li> <li>• Detailed display tables were helpful to understanding and were removed in recent iteration</li> <li>• Filled with distracting hospital advertisements</li> <li>• Only summary data ("Average," "Good," "Very High," "Best") provided on public website. Would prefer layering of information for those interested in more detail.</li> </ul>	<ul style="list-style-type: none"> <li>• No feature to navigate to detailed measure scores from list of Top Hospitals</li> <li>• Does not rank hospitals within state</li> </ul>

Source: The Authors. Details of the rating process are available at [www.RatingTheRaters.org](http://www.RatingTheRaters.org)  
 NEJM Catalyst ([catalyst.nejm.org](http://catalyst.nejm.org)) © Massachusetts Medical Society

## Rating the Raters – Strengths and Weaknesses Assessment of the Four Public Hospital Quality Rating Systems

The comments in the table below reflect the discussion that the Rating-the-Raters group had about each rating system. These comments for each rating system were provided to the leaders of that rating system to solicit feedback.

### Audit Information

<b>(CMS) Hospital Compare Overall Star Ratings</b>	<b>Healthgrades Top Hospitals</b>	<b>USNWR (U.S. News &amp; World Report) Best Hospitals</b>	<b>Leapfrog Hospital Safety Grade and Leapfrog Top Hospitals</b>
<ul style="list-style-type: none"> <li>• CMS will validate up to eight cases for clinical process of care measures per quarter per selected hospital. Cases are randomly selected from data submitted to the warehouse by the hospital</li> <li>• CMS will validate up to 10 candidate hospital acquired infection (HAI) cases total per quarter per selected hospital</li> <li>• Each quarter, the Clinical Data Abstraction Center (CDAC) will send hospitals a written request to Medical Records Director to submit a patient medical record for each case and candidate case that CMS selected for validation</li> <li>• If a hospital does not meet the overall validation requirement, the hospital will not receive full credit</li> </ul>	<ul style="list-style-type: none"> <li>• Not applicable (uses claims data from CMS)</li> </ul>	<ul style="list-style-type: none"> <li>• Not applicable (uses data from other sources)</li> </ul>	<ul style="list-style-type: none"> <li>• Leapfrog gives hospitals opportunity to review data, but assumes no responsibility for accuracy of data from elsewhere (CMS, AHA)</li> <li>• Data review/check for erroneous values completed</li> <li>• An internal Leapfrog data review is conducted monthly after submission from hospitals</li> <li>• Hospitals are randomly selected to submit documentation demonstrating adherence to NQF safe practices and Never Events policy</li> <li>• Contracted with DHG Healthcare to conduct in-hospital data verification for selected hospitals</li> <li>• Very few hospitals are audited annually and audit results are not released</li> </ul>

Source: The Authors. Details of the rating process are available at [www.RatingTheRaters.org](http://www.RatingTheRaters.org)  
 NEJM Catalyst ([catalyst.nejm.org](http://catalyst.nejm.org)) © Massachusetts Medical Society

## Acknowledgments and Disclosures

**Acknowledgments:** *The authors would like to acknowledge the following people for their assistance with the initiative: Carol Cronin, Informed Patient Institute; Mary Dixon-Woods, MSc, DPhil, University of Cambridge; Elizabeth McGlynn, PhD, Kaiser Permanente; Ryan Ellis, MD, Northwestern University; and Cindy Barnard, PhD, MBA, MSJS, Northwestern Medicine. We acknowledge Jeanette Chung, PhD, Northwestern University, for assistance with statistical analyses. We also appreciate the participation of the hospital rating systems' leadership and methodologists in various stages of this initiative.*

**Author contributions:** *Dr. Bilimoria had full access to all data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design: All authors. Acquisition, analysis, or interpretation of data: All authors. Drafting of the manuscript: All authors. Critical revision of the manuscript for important intellectual content: All authors. Statistical analysis: All authors. Supervision: All authors.*

**Funding/support:** *None.*

**Disclosure of any current or prior relationships with a hospital rating system:** *Karl Bilimoria, U.S. News (unpaid; Advisory Panel in the past), CMS (unpaid; Technical Expert Panel); John Birkmeyer, Leapfrog (unpaid; in the past); Helen Burstin, CMS (NQF supported by CMS; in the past); Justin Dimick, Leapfrog (unpaid; Volume Standards Committee Chair), U.S. News (unpaid; Advisory Panel); Karen Joynt Maddox, Department of Health and Human Services (contract work); Peter Pronovost, Leapfrog (unpaid; Advisory Panel), U.S. News (unpaid; Advisory Panel), Health and Human Services (Co-Chair, HHS Summit).*



### **Karl Y. Bilimoria, MD, MS**

John Benjamin Murphy Professor of Surgery, Director, Surgical Outcomes and Quality Improvement Center, and Vice Chair for Quality, Department of Surgery, Feinberg School of Medicine, Northwestern Medicine; Vice President, Quality and Clinical Integration, Northwestern Medicine Health System

### **John D. Birkmeyer, MD**

Chief Clinical Officer, Sound Physicians; Adjunct Professor, Dartmouth Institute for Health Policy & Clinical Practice

### **Helen Burstin, MD, MPH**

Executive Vice President and Chief Executive Officer, Council of Medical Specialty Societies

### **Justin B. Dimick, MD, MPH**

George D. Zuidema Professor of Surgery, Director, Center for Healthcare Outcomes & Policy, and Associate Chair for Strategy and Finance, University of Michigan

**Karen E. Joynt Maddox, MD, MPH**

Assistant Professor of Medicine, Cardiovascular Division, School of Medicine and Assistant Professor, Brown School of Social Work, Washington University in St. Louis

**Allison R. Dahlke, MPH**

Assistant Director, Population Sciences, University of Wisconsin Carbone Cancer Center; Former Administrative Director, Surgical Outcomes and Quality Improvement Center, Department of Surgery, Feinberg School of Medicine, Northwestern Medicine

**John Oliver DeLancey, MD, MPH**

Research Fellow, Surgical Outcomes and Quality Improvement Center, Department of Surgery, Feinberg School of Medicine, Northwestern Medicine

**Peter J. Pronovost, MD, PhD**

Chief Clinical Transformation Officer and Professor, Case Western Reserve University Schools of Medicine and Nursing, University Hospitals