Safety in Numbers: The Development of Leapfrog's Composite Patient Safety Score for U.S. Hospitals

J. Matthew Austin, PhD, Guy D'Andrea, MBA,† John D. Birkmeyer, MD,‡ Lucian L. Leape, MD,§ Arnold Milstein, MD,|| Peter J. Pronovost, MD, PhD,¶ Patrick S. Romano, MD,** Sara J. Singer, MBA, PhD,†† Timothy J. Vogus, PhD,‡‡ and Robert M. Wachter, MD§§*

Objective: To develop a composite patient safety score that provides patients, health-care providers, and health-care purchasers with a standardized method to evaluate patient safety in general acute care hospitals in the United States.

Methods: The Leapfrog Group sought guidance from a panel of national patient safety experts to develop the composite score. Candidate patient safety performance measures for inclusion in the score were identified from publicly reported national sources. Hospital performance on each measure was converted into a "z-score" and then aggregated using measure-specific weights. A reference mean score was set at 3, with scores interpreted in terms of standard deviations above or below the mean, with above reflecting better than average performance.

Results: Twenty-six measures were included in the score. The mean composite score for 2652 general acute care hospitals in the United States was 2.97 (range by hospital, 0.46–3.94). Safety scores were slightly lower for hospitals that were publicly owned, rural in location, or had a larger percentage of patients with Medicaid as their primary insurance.

Conclusions: The Leapfrog patient safety composite provides a standardized method to evaluate patient safety in general acute care hospitals in the United States. While constrained by available data and publicly reported scores on patient safety measures, the composite score reflects the best available evidence regarding a hospital's efforts and outcomes in patient safety. Additional analyses are needed, but the score did not seem to have a strong bias against hospitals with specific characteristics. The composite score will continue to be refined over time as measures of patient safety evolve.

Key Words: patient safety, hospital, performance measures, composite, score

(*J Patient Saf* 2014;10: 64–71)

Despite the hundreds of thousands of patients who experience preventable harm each year in U.S. hospitals,[1] patients, health-care providers, and health-care purchasers lack a standardized method to evaluate patient safety in U.S. hospitals. This lack of standardized evaluation is especially concerning, given recent evidence that many U.S. hospitals have made small incremental improvements in safety or continue to harm patients in large numbers.[2–5] Although the number of publicly reported patient safety measures has grown over the last decade,[6,7] there is little evidence that patients and providers are using these data to inform their choice of hospital.[8,9] One possible explanation for the lack of use is that safety data have typically been reported on a measure-by-measure basis, requiring patients and providers to synthesize many disparate data points in their decision-making process.

Composite measures, which combine multiple performance measures using a predetermined weighting methodology to produce a single score, provide a more general picture of the safety performance of a hospital, making it more understandable and usable for nonmedical personnel.[10] However, although composite measures are easier for patients to understand, they can mask a hospital's excellent or poor performance on individual measures. Additionally, the predetermined weights assigned to each measure may not reflect the preferences of individual patients.[11] Nevertheless, the potential to motivate patient and provider response, and thereby improve patient utility, makes it worthwhile to develop a composite measure of hospital patient safety.

The Leapfrog Group, a national coalition of health-care purchasers, sought to address the need for a standardized method for evaluating patient safety in U.S. hospitals by creating a composite safety score. The primary goal was a single, publicly available safety score for every U.S. hospital, including hospitals that do and do not voluntarily participate in the Leapfrog Hospital Survey. Other national organizations, such as Consumer Reports and US News and World Report, have developed their own composite measures of hospital performance. These organizations, however, have taken quite different approaches to constructing composites; US News relies heavily on reputational surveys and focuses on major academic medical centers and large sized hospitals, whereas Consumer Reports relies heavily on patient experience surveys and readmission rates that favor small- to medium-sized community hospitals.[12–14] Leapfrog's goal was to balance many different measures of structures, processes, and outcomes that have been linked to patient safety but which may or may not be related to each other, to improve patient and clinician selection of hospitals and to more strongly motivate and better target hospital safety improvement efforts.

The purpose of this paper is to describe the methods used to develop the composite score and to demonstrate its use with recent national data. The Leapfrog Group's Board of Directors has used the composite safety score to assign a letter grade to each hospital. This application of the composite score is beyond the scope of this paper.

METHODS

The Leapfrog Group invited nine national experts to serve on a panel to develop a composite score to evaluate patient safety in U.S. hospitals (see side bar on page 65 for the list of

From the *Armstrong Institute for Patient Safety and Quality, The Johns Hopkins University School of Medicine, Baltimore, MD; †Discern Consulting, LLC, Baltimore, MD; ‡George D. Zuidema Professor of Surgery, University of Michigan Medical School, Ann Arbor, MI; §Department of Health Policy and Management, Harvard School of Public Health Boston, MA; ||Medicine, Stanford University School of Medicine, Stanford, CA; ¶Departments of Anesthesiology/Critical Care Medicine and Surgery, The Johns Hopkins University School of Medicine, Baltimore, MD; **Medicine and Pediatrics, University of California (UC) Davis School of Medicine, Sacramento, CA; ††Department of Health Policy and Management, Harvard School of Public Health, Boston, MA; ‡‡Management, Vanderbilt University Owen Graduate School of Management, Nashville, TN; and §§Department of Medicine, University of California (UC) San Francisco, San Francisco, CA.

Correspondence: J. Matthew Austin, PhD, Instructor, Armstrong Institute for Patient Safety and Quality, The Johns Hopkins University School of Medicine; 750 E. Pratt St, 15th Floor, Baltimore, MD 21202 (e-mail: jmaustin@jhmi.edu).

The authors disclose no conflict of interest.

Members of Safety Score Expert Panel

- John Birkmeyer, MD (University of Michigan)
- Ashish Jha, MD, MPH (Harvard University)
- Lucian Leape, MD (Harvard University)
- Arnold Milstein. MD, MPH (Stanford University)
- Peter Pronovost, MD, PhD(Johns Hopkins University)
- Patrick Romano, MD, MPH (University of California, Davis)
- Sara Singer, MBA, PhD (Harvard University)
- Timothy Vogus, PhD (Vanderbilt University)
- Robert Wachter, MD (University of California,

panelists). The work was done in late 2011 and involved defining a conceptual framework for the score, assigning relative weights to each measure, standardizing scores across different measure types, and identifying methods for dealing with missing data.

Defining a Conceptual Framework for the Score

Initially, the panel discussed what the composite score should be designed to measure, and the consensus was patient safety, defined as "freedom from harm." The panel agreed that this focus, which aligns with the Institute of Medicine's improvement aim of "safe" care, was narrower than hospital quality.

One goal was to develop a composite score suitable for assessing the largest number of general acute care hospitals across the country. Thus, the panel recommended that Leapfrog include publicly reported measures from national data sources in the score. The panel excluded state-reported and regionally reported measures because of variations in measure specifications, data collection, and availability that would prevent a consistent comparison across hospitals. Leapfrog staff (J.M.A., G.D.) scanned publicly reported patient safety measures from the Centers for Medicare & Medicaid Services (CMS), The Joint Commission, The Leapfrog Group, and The Commonwealth Fund to compile a list of candidate measures for the expert panel to review and consider for inclusion in the composite score. Forty-five candidate measures, representing a mix of structural, process, and outcome measures were identified for review.

The panel debated whether the safety score should recognize hospital efforts toward patient safety (process and structural measures), achievements in patient safety (outcome measures), or both. Process and structural measures signal what hospitals have done to improve safety (e.g., adopting specified protocols to reduce infections, implement computerized order entry) and fill gaps where valid outcome data are not yet available. However, outcome measures (e.g., infection rates) reflect whether a hospital has actually achieved safety goals. The panel ultimately recommended that process/structural measures and outcome measures carry equal weights of 50% in the composite score but recognized that this choice was arbitrary and should be re-evaluated periodically as new outcome measures become available to capture the impact of safety-enhancing activities that are currently represented by structure and process measures.

The expert panel reviewed the candidate measures and selected the final set through a process of discussion and repeated voting to maximize consensus. The panel eliminated measures that were not clear safety measures, measures supported by weak research evidence, and process measures with a corresponding outcome measure that better represents patient utility. Table 1 provides the final list of 26 measures included in the composite score.

Weighting the Individual Measures

The expert panel next developed a framework to assign relative weights to each measure using 3 criteria: strength of evidence, opportunity for improvement, and impact. These criteria track closely with criteria used by the National Quality Forum in its Consensus Development Process for measure endorsement.[15] These criteria were defined as follows:

- **Strength of Evidence**. Measures with stronger research evidence carry higher weight than measures with weaker evidence. Because the measures selected for the composite score already met high standards of research evidence, any variation in strength of the evidence was assessed using the following scoring method:
  1 = Supported by either suggestive clinical or epidemiological studies or theoretical rationale
  2 = Supported by experimental, clinical, or epidemiological studies and strong theoretical rationale
- **Opportunity for Improvement**. The panel agreed that measures with greater variation in hospital performance are more valuable for decision making because there is greater opportunity for performance improvement. For each measure, the coefficient of variation (standard deviation/mean) was calculated using the measure's national performance data.[16] The coefficient of variation was then converted into an opportunity score by adding 1.0 to the coefficient of variation, capping the maximum score at 3.0 to limit distortions in the final weights that could result from very high coefficients of variation. Thus, the opportunity scores ranged from 1.0 to 3.0, with higher values denoting greater variation and more opportunity for improvement.
- **Impact**. Some measures affect a greater percentage of patients than others. For example, computerized medication orders are applicable to nearly every patient in the hospital, whereas an intensivist manages only patients in the intensive care unit (ICU). Some measures are associated with clinical events that indicate greater severity of harm (e.g., mortality or serious morbidity). Therefore, impact reflects both the

TABLE 1. List of Measures Included in Composite Safety Score

Measure	Primary Data Source (Secondary Data Source, if Applicable)	Evidence Score*	Opportunity Score†	Impact Score‡	Relative Weight (%)
Process and Structural Measures					
Computerized Physician Order Entry (CPOE)	Leapfrog Hospital Survey (AHA Annual Survey)	2	3.00	3	7.7
ICU Physician Staffing (IPS)	Leapfrog Hospital Survey (AHA Annual Survey)	2	2.89	3	7.5
Safe Practice 1: Leadership Structures and Systems	Leapfrog Hospital Survey	1	1.16	2	2.3
Safe Practice 2: Culture Measurement	Leapfrog Hospital Survey	1	1.32	2	2.5
Safe Practice 3: Teamwork Training and Skill Building	Leapfrog Hospital Survey	1	1.36	2	2.6
Safe Practice 4: Identification and Mitigation of Risks and Hazards	Leapfrog Hospital Survey	1	1.22	2	2.4
Safe Practice 9: Nursing Workforce	Leapfrog Hospital Survey	1	1.25	3	3.3
Safe Practice 17: Medication Reconciliation	Leapfrog Hospital Survey	1	1.19	2	2.4
Safe Practice 19: Hand Hygiene	Leapfrog Hospital Survey	2	1.21	2	3.1
Safe Practice 23: Care of the Ventilated Patient	Leapfrog Hospital Survey	1	1.23	2	2.4
SCIP INF 1: Antibiotic within 1 Hour	CMS Hospital Compare	2	1.05	2	2.9
SCIP INF 2: Antibiotic Selection	CMS Hospital Compare	1	1.04	2	2.2
SCIP INF 3: Antibiotic Discontinued After 24 Hours	CMS Hospital Compare	1	1.06	2	2.2
SCIP INF 9: Catheter Removal	CMS Hospital Compare	2	1.12	2	3.0
SCIP VTE 2: VTE Prophylaxis	CMS Hospital Compare	2	1.09	3	3.7
Outcome Measures					
Foreign Object Retained	CMS Hospital Compare	1	3.00	2	5.6
Air Embolism	CMS Hospital Compare	1	3.00	2	5.6
Pressure Ulcer – Stages 3 and 4	CMS Hospital Compare	1	2.75	3	7.4
Falls and Trauma	CMS Hospital Compare	2	1.82	3	6.0
Central Line Associated Bloodstream Infections (CLABSI)	Leapfrog Hospital Survey (CMS Hospital Compare)	2	2.00	3	6.4
PSI 4: Death Among Surgical Inpatients	CMS Hospital Compare	1	1.17	2	2.7
PSI 6: Iatrogenic Pneumothorax	CMS Hospital Compare	1	1.40	2	3.1
PSI 11: Postoperative Respiratory Failure	CMS Hospital Compare	1	1.37	2	3.0
PSI 12: Postoperative PE/DVT	CMS Hospital Compare	1	1.50	2	3.2
PSI 14: Postoperative Wound Dehiscence	CMS Hospital Compare	1	1.19	2	2.7
PSI 15: Accidental Puncture or Laceration	CMS Hospital Compare	1	1.46	3	4.3

*A binary score reflecting the strength of evidence of the measure (1 or 2).

‡The coefficient of variation for the measure plus 1.0, capped at a maximum score of 3.0 reflects the variation in performance between hospitals.

‡Measure of the number of patients possibly affected by the safety event and on the severity of harm for individual patients.

number of patients possibly affected by the event and the severity of harm for individual patients. We defined and scored the dimensions as follows:

- Number of patients possibly affected by the event:
  - 1 = Rare event (e.g., foreign object retained after surgery)
  - 2 = Some patients in hospital affected (e.g., ICU physician staffing)
  - 3 = All patients in hospital affected (e.g., hand hygiene safe practice)
- Severity of harm for individual patients:
  - 1 = Limited direct evidence of harm or harm reduction (e.g., culture measurement safe practice)
  - 2 = Clear documentation of harm or harm reduction; adverse events (e.g., foreign object retained after surgery)
  - 3 = Significant mortality reduction (>1000 deaths across the US annually or a 10% reduction in hospital-wide mortality) (e.g., ICU physician staffing)

The values from each dimension were added together and an overall Impact Score given as follows:

1 = Summed value of 2 (low impact)
2 = Summed value of 3 to 4 (medium impact)
3 = Summed value of 5 to 6 (high impact)

Through an iterative process, the panel reviewed different schemes to assign weights until they reached consensus on a scheme that seemed fair, simple, and intuitively sound. The panel chose a calculation that emphasized the opportunity and the impact of each measure because all selected measures already met a high standard of evidence. Each measure received an overall weight score as follows:

$$\text{Weight Score} = \text{Evidence} + (\text{Opportunity} \times \text{Impact})$$

The possible weight score for each measure ranged between 2 and 11. By design, this wide scoring range provided

substantial differences in weights across measures so that more informative, higher impact measures had a stronger influence on the overall composite score.

In the final stage of the weighting process, each measure's weight score was converted to a percentage weight. To ensure equal overall weights for the process/structural domain and the outcomes domain, we divided the weight score for each measure by the sum of all of the weight scores from measures in that domain and then multiplied by 50%.

Standardizing Scores Across Disparate Measures

One challenge in creating the composite score was combining measures that were expressed in different ways. Some measures were expressed as a point system (the hospital earned x points out of a possible y points), whereas other measures were expressed as proportions or rates (x occurrences out of a relevant population of y patients).

The panelists agreed that the scores had to be standardized to ensure the measure weights are the primary driver of the composite score and the measure values do not distort the score. Three approaches were considered for standardizing scores: 1) assign points based on a provider's performance relative to the mean or median, 2) assign a score based on the provider's ranking—expressed as simple percentile or categorical value (e.g., quartiles), or 3) divide the difference in the hospital's score from the national mean by the standard deviation.

The panel recommended option 3 to standardize a hospital's performance on each measure. This option is often called a z-score and is based on the following equation that expresses the hospital's performance as a number of standard deviations from the national mean:

$$\text{Hospital's Z-Score} = (\text{Hospital's Measure Performance} - \text{Mean Performance for All Hospitals}) \div \text{Standard Deviation for All Hospitals}$$

Hospitals that perform close to the national mean earn a z-score near zero. Better hospitals earn a positive z-score, reflecting measure values above the national mean. Worse hospitals earn a negative z-score, reflecting measure values below the national mean.

The z-score allowed us to compare and combine individual scores from different types of data. Besides being simple to understand and explain, a z-score has precedent in health-care measurement. It has been used in the Multiple Sclerosis Functional Composite, which is a measure of disability in patients with Multiple Sclerosis (MS).[17] Research has shown that the z-score method produces results comparable to alternative rescaling approaches.[18,19] One potential disadvantage of the z-score, or any rescaling approach, is the exaggeration of clinically meaningless differences when measures have a small standard deviation; this problem was minimized by giving more weight to measures with higher coefficients of variation, as described previously.

Dealing With Missing Data

Although the expert panel selected nationally reported measures, some data were missing for hospitals that did not report to The Leapfrog Group, and other data were missing because hospitals did not meet the CMS' minimum sample size requirements to publicly report their results.

The panelists discussed several possible approaches for assessing hospital performance when data were missing, including the following:

1. Impute the hospital's performance at the national median/mean.
2. Impute the hospital's performance on the missing measures using other available data.
3. Impute a score above or below the median/mean (e.g., -1 standard deviation).
4. Give the hospital zero credit for the measure (e.g., when the hospital chose not to report data to Leapfrog).
5. Exclude the measure and recalibrate the weights for the affected hospital(s), using only those measures for which data are available.

After conducting sensitivity analyses using each of the above approaches, the panel recommended approach number 5 to address missing data (excluding the measure for a hospital and recalibrating the weights for those measures for which performance data were available) for most measures. In this way, we avoided making questionable assumptions about the performance of hospitals with missing data, and we also maintained variation in the final calculated scores by not imputing identical values for all missing observations. Second, with the 2 exceptions described in the next paragraph, imputation using other variables was deemed inappropriate, as hospital scores on one measure did not predict performance on other measures. Finally, the third and fourth approaches were rejected to avoid creating a structural bias in favor of Leapfrog-reporting hospitals. By scoring hospitals on the available data, the composite scores were based on what was known about the hospital's performance, with no penalties for missing data.

Imputation using alternative data was used for two Leapfrog survey-based measures. Data on Computerized Physician Order Entry (CPOE) and ICU Physician Staffing (IPS) from the American Hospital Association's (AHA) annual survey were used to impute measure scores for hospitals that did not respond to the Leapfrog Hospital Survey. The AHA survey does not collect the same level of detail as the Leapfrog Hospital Survey. Thus, for this first round of scores, hospitals were given a score that was equivalent to what they would have earned had they reported that same limited data to the Leapfrog Hospital Survey. After considerable feedback on the initial imputation approach, the panel conducted additional analyses using data for hospitals that had reported to both data sets and has since revised the imputation method to better reflect an "expected" value concept. With this change, a hospital's imputed value now reflects an empirically estimated probability of the hospital scoring in each Leapfrog category given their AHA response.

Scores could not be calculated for certain types of hospitals because of the systemic lack of publicly reported safety data; these include the following:

- Critical access hospitals (CAH)
- Long-term care facilities
- Mental health facilities
- Federal hospitals (e.g., Veterans Affairs, Indian Health Services)
- Specialty hospitals, including surgical centers and cancer hospitals
- Free-standing pediatric hospitals
- Hospitals located in U.S. territories (e.g., Puerto Rico, Guam)
- Hospitals in Maryland, since they did not report data through CMS's Inpatient Prospective Payment System (IPPS)

In addition, some hospitals had so much missing data that the panel determined it was not possible to reliably calculate a safety score. Hospitals missing more than 9 of 15 process/structural measures or more than 3 of 11 outcome measures were not scored; 605 hospitals were excluded from our analysis for

TABLE 2. Descriptive Statistics of the Composite Safety Scores by Hospital Characteristics

Characteristic	No. Of Hospitals*	Mean Score	Median Score	Standard Deviation	$P^{\dagger}$
All hospitals	2,573	2.97	2.98	0.32	
Region of country‡					0.001
Northeast	485	3.00	2.98	0.30	
Midwest	596	3.00	3.01	0.31	
South	987	2.94	2.98	0.31	
West	505	2.97	2.98	0.35	
Size					0.001
Small (1–99 beds)	496	2.93	2.98	0.36	
Medium (100–399 beds)	1,583	2.99	3.00	0.30	
Large (400+ beds)	493	2.95	2.93	0.32	
Ownership type					0.000
For-profit	475	2.97	3.04	0.31	
Private nonprofit	1,757	3.00	2.99	0.31	
Public	341	2.83	2.88	0.34	
Teaching status					0.043
Major	296	2.97	2.93	0.35	
Minor	683	2.99	2.98	0.31	
Not Teaching	1,594	2.96	2.99	0.31	
Location					0.047
Rural	64	2.89	2.97	0.38	
Urban	2,509	2.97	2.98	0.31	
Member of hospital system					0.000
No	882	2.90	2.94	0.34	
Yes	1,691	3.01	3.01	0.30	
% Medicare patients					0.000
Lowest	643	2.93	2.94	0.34	
Quartile 2	643	3.01	3.02	0.30	
Quartile 3	643	2.98	3.00	0.32	
Highest	643	2.96	2.97	0.30	
% Medicaid patients					0.000
Lowest	644	3.02	3.03	0.33	
Quartile 2	642	2.96	2.97	0.31	
Quartile 3	643	3.00	3.02	0.31	
Highest	643	2.89	2.89	0.31	

*Characteristics for 76 hospitals could not be identified from the 2011 AHA annual survey.

†One-way ANOVA was used to compare characteristics within each subgroup.

‡Regions are based on the U.S. Census Bureau's definitions and available at http://www.census.gov/geo/www/us_regdiv.pdf.

having too little available data. These minimum measure thresholds were chosen to balance the reliability of the calculated score with the number of hospitals receiving a score.

Final Scoring

A composite safety score for each hospital was calculated by multiplying the weight for each measure by the hospital's z-score on that measure. We added 3 to each hospital's final score to avoid possible confusion with interpreting negative patient safety scores. The final calculation of the safety score was as follows:

$$\text{Safety Score} = 3 + \text{Weight}_{\text{Measure1}} \times \text{Z} - \text{Score}_{\text{Measure1}} + \text{Weight}_{\text{Measure2}} \times \text{Z} - \text{Score}_{\text{Measure2}} + \text{Weight}_{\text{Measure3}} \times \text{Z} - \text{Score}_{\text{Measure3}} + \ldots + \text{Weight}_{\text{Measure n}} \times \text{Z} - \text{Score}_{\text{Measure n}}$$

In April 2012, Leapfrog applied the final scoring methodology to the most recent available data.[i]

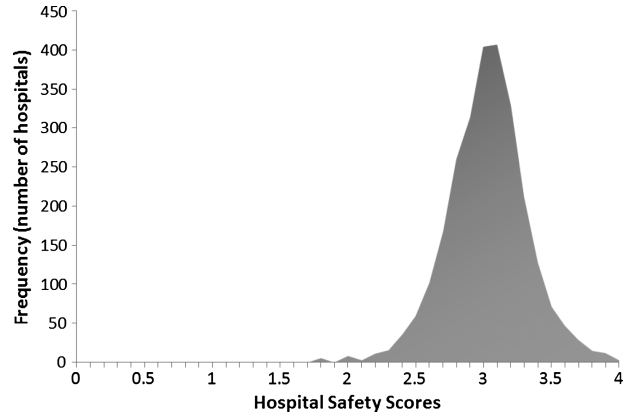

FIGURE 1. Illustrates the distribution of composite safety scores for the 2652 general acute care hospitals. The scores ranged from 0.46 to 3.94. Scores were calculated by multiplying the weight for each measure by the hospital's *z*-score on the measure. Three was added to each hospital's final score to avoid possible confusion with interpreting negative patient safety scores.

Data Analysis

To better understand the associations between calculated composite scores and hospital characteristics, descriptive statistics of the composite safety scores were calculated by characteristic (Table 2). A *P* value from a 1-way ANOVA was calculated for each characteristic. For each characteristic, the *P* value was less than the critical value of 0.05, indicating that the subgroup scores within each characteristic are not all the same.

In addition, we compared the safety of the 605 unscored hospitals with those of the scored hospitals by calculating the average performance of each group on the structural/process measures and the outcome measures.

RESULTS

Composite safety scores were calculated for 2652 general acute care hospitals; 605 hospitals were excluded because of excessive missing data. The final composite safety scores for hospitals ranged from 0.46 to 3.94 (Fig. 1). For interpretation, the top score of 3.94 represents a hospital that averages almost 1.0 standard deviation above (better than) the national mean across all measures for which data were available. The bottom score of 0.46 represents a hospital that averages almost 2.5

[i]The reporting time periods from each data source used to calculate a hospital's score were a follows: 2011 Leapfrog Hospital Survey (April 2011–March 2012); CMS HospitalCompare SCIP process measures (April 2010–March 2011); CMS HospitalCompare No-Pay Events and AHRQ PSIs (October 2008–June 2010); CMS HospitalCompare CLABSI rates (2006–2008); AHA Annual Survey (January 2010–December 2010).

TABLE 3. Mean Composite Safety Scores by State

State	Hospital Count	Mean Score	State	Hospital Count	Mean Score	State	Hospital Count	Mean Score
MA	62	3.33	FL	178	3.00	AR	32	2.89
ME	20	3.17	CT	29	2.99	LA	56	2.89
MI	78	3.11	RI	11	2.98	AZ	47	2.89
MN	46	3.07	CO	34	2.97	ND	6	2.88
VT	6	3.07	IA	27	2.97	NM	21	2.88
VA	58	3.06	WI	45	2.96	AK	7	2.88
NJ	73	3.05	IN	67	2.96	WY	4	2.88
IL	113	3.05	PA	130	2.95	OK	42	2.88
DE	5	3.05	KY	49	2.94	KS	37	2.87
TN	69	3.04	UT	21	2.94	OR	32	2.85
MT	10	3.02	ID	12	2.92	WV	24	2.85
SC	39	3.02	GA	78	2.92	HI	12	2.85
CA	264	3.01	NE	17	2.91	NY	150	2.84
OH	101	3.00	NV	19	2.90	SD	10	2.82
NC	70	3.00	MO	68	2.90	DC	7	2.80
WA	41	3.00	TX	221	2.90	AL	54	2.78
NH	12	3.00	MS	38	2.89			

standard deviations below (worse than) the national mean for all measures for which data were available.

The range of safety scores for hospitals with a large number of measures that were not reported or were not available[ii] was similar to those hospitals for which data were available for most measures (0.46–3.82 versus 1.72–3.93), reflecting poor and excellent performance in both subgroups. However, the scores for hospitals with a large number of missing measures did cluster more toward the middle than those hospitals for which data were available for most measures (standard deviation of scores: 0.29 versus 0.32). Of the 605 hospitals that were not scored because of excessive missing data, their performance averaged 3% worse on those structural/process measures for which data were available and 11% better on those outcome measures for which data were available (listed in Table 1) than the hospitals that received a score. The unscored hospitals were 1% to 73% better on hospital-acquired conditions than the scored hospitals.

Composite Scores by Hospital Characteristic

Table 2 describes the mean composite safety score by hospital characteristic. Hospitals in the Midwest and Northeast had higher mean composite scores than hospitals in the South and West ($P = 0.001$). Publicly owned hospitals had lower mean composite scores than either for-profit or private nonprofit hospitals ($P = 0.000$). Rural hospitals had lower mean composite scores than their urban counterparts ($P = 0.047$). Mean composite scores were lower for hospitals with the largest percentage of patients with Medicaid as the primary payer ($P = 0.000$). The teaching status of the hospital did not seem to be correlated with a hospital's composite score, as the mean composite scores for all teaching statuses were quite similar ($P = 0.043$). Also, the mean composite scores did not vary substantially by the percentage of patients with Medicare as the primary payer ($P = 0.000$).

Table 3 shows the total number of hospitals and the mean composite score for hospitals in each state. Hospitals in Massachusetts had the highest mean score (3.33) and those in Alabama had the lowest mean score (2.78). There was a 0.55 difference in mean scores between Massachusetts and Alabama, representing a spread of about 1.75 standard deviations.

DISCUSSION

The Leapfrog patient safety composite provides a standardized method to evaluate overall patient safety in those U.S. acute care hospitals that have publicly reported data on a sufficient number of individual safety measures. The composite score could be estimated for 2652 hospitals using a methodology that is transparent, reproducible, and has face validity, as determined by the national expert panel that guided its development.

The reporting of the composite safety score can serve multiple functions. First, it could raise consumer awareness about patient safety in hospitals and prompt patients to engage their providers in discussing variations in safety across hospitals. Second, by serving as a national reference point, it could encourage improvement efforts and assist hospital leaders in identifying and improving drivers of safety and quality.[20] Third, the score could be used by purchasers and payers to set minimum performance thresholds for inclusion in provider networks or recognizing and rewarding performance through value-based purchasing (VBP) initiatives. One example of a payer using a composite measure is CMS, which plans to use the Agency for Healthcare Research and Quality's patient safety indicator (PSI) composite measure (PSI 90) in its 2015 Inpatient Prospective Payment System VBP program.[21]

A strength of the composite patient safety score is the inclusion of measures that reflect both a hospital's efforts toward safety (process and structural measures) and its safety-related results (outcome measures). Another strength of the score was the minimal bias from individual hospital characteristics. Hospitals that were publicly owned, rural, not part of a larger health system, or had a larger number of patients with Medicaid as

[ii]"Large number" of missing measures is defined as 8 or more missing process/structural measures. Most of the measures in which data were not reported or not available were those reported through voluntary data sources and are almost exclusively in the process/structural domain. Hospitals with 8 or more missing process/structural measures are generally hospitals that choose not to report to the Safe Practices section of the Leapfrog Hospital Survey.

their primary insurance did have slightly lower mean composite scores compared with their counterparts; however, many hospitals within these subgroups outperformed hospitals that did not share these characteristics.

There are several limitations to this composite score. The main limitation is the small pool of publicly reported national measures available to assess hospital safety, which resulted in composite scores that only cover a subset of patient safety. The publicly reported scores are also limited by a relatively small number of valid outcome measures and by flaws in measures derived from administrative rather than clinical data. Some elements of the score come from voluntary hospital surveys. Voluntary reporting may be a limitation to the extent that hospitals differently interpret the survey questions or report inaccurately. This concern is somewhat mitigated by having hospital senior executives affirm the accuracy of their survey responses and having the reporting organizations critically review submitted responses. In addition, our decision to exclude a measure to address missing data may have overstated or understated a hospital's safety performance, depending on how the hospital performed on the measures that remained. Finally, because of the limited reporting of the data underlying each measure (i.e., the numerators and denominators), the weights in the composite score could not be reliability adjusted for the 'signal' of each measure, a practice commonly used in other composite scores.[22,23]

Future research should examine the relationship between hospitals' composite scores and outcomes, specifically risk-adjusted mortality, readmission rates, and infection rates. Although previous studies did not find any relationship between hospitals' mortality rates and their Leapfrog National Quality Forum Safe Practices Scores,[24,25] future research should examine whether the Leapfrog composite score, which includes nonmortality outcome measures plus structural and process measures, is a better correlate of patient mortality than previously developed safety composites. It would also be worthwhile comparing hospitals' performance on the Leapfrog patient safety composite score with performance on other safety composite scores.

Those involved in developing the Leapfrog composite score believe it reflects the best patient safety measures and data currently available for a large number of U.S. hospitals. Nonetheless, 2 challenges we faced in developing a score were the small number of safety measures that are publicly reported at a national level and deficiencies with many of the existing measures. These problems could be addressed with a more formal mechanism to support and encourage the development of robust patient safety measures. The authors hope that this paper will catalyze the entire health-care community—patients, providers, purchasers, and policymakers—to work together to develop better and more meaningful patient safety measures and to publicly report those data nationally. As new measures become available, the panel is committed to revisiting the measures and weights that comprise the composite score.

ACKNOWLEDGMENTS

The authors thank Ashish Jha, M.D., M.P.H., for leadership and contributions in helping guide the development of the composite score. The authors also thank Leah Binder, M.A., M.G.A., and Missy Danforth, B.A., from The Leapfrog Group, Barbara Rudolph, Ph.D., M.S.S.W., from the University of Wisconsin–Madison, and Morgan Fishbein, M.B.A., from Discern Consulting, LLC for their leadership in the actual implementation of the composite scoring methodology. The authors also thank Sidney Le, B.S., from the Harvard School of Public Health for assistance with data analysis, and Christine G. Holzmueller, B.L.A., from the Johns Hopkins University Department of Anesthesiology and Critical Care Medicine, for thoughtful review and editing of this paper.

REFERENCES

1. Corrigan JM, Kohn LT, Donaldson MS, eds. *To Err is Human: Building a Safer Health System*. Washington, DC: National Academies Press; 1999.
2. Leape LL, Berwick DM. Five years after To Err Is Human: what have we learned? *JAMA*. 2005;293:2384–2390.
3. Pronovost PJ, Miller MR, Wachter RM. Tracking progress in patient safety: an elusive target. *JAMA*. 2006;296:696–699.
4. Landrigan CP, Parry GJ, Bones CB, et al. Temporal trends in rates of patient harm resulting from medical care. *N Engl J Med*. 2010;363:2124–2134.
5. Longo DR, Hewett JE, Ge B, et al. The long road to patient safety: a status report on patient safety systems. *JAMA*. 2005;294:2858–2865.
6. Ross JS, Bernheim SM, Drye ED. Expanding the frontier of outcomes measurement for public reporting. *Circ Cardiovasc Qual Outcomes*. 2011;4:11–13.
7. Meyer GS, Nelson EC, Pryor DB, et al. More quality measures versus measuring what matters: a call for balance and parsimony. *BMJ Qual Saf*. 2012;21:964–968.
8. The Henry J. Kaiser Family Foundation. 2008 Update on Consumers' Views of Patient Safety and Quality Information 2008, 2008 October [cited 27 July 2012]. Available at: http://www.kff.org/kaiserpolls/upload/7819.pdf.
9. Tu HT, Lauer J. Word of Mouth and Physician Referrals Still Drive Health Care Provider Choice, Research Brief: Findings from HSC, Center for Studying Health System Change. December 2008. Available at: http://www.hschange.com/CONTENT/1028/.
10. Peters E, Dieckmann N, Dixon A, et al. Less is more in presenting quality information to consumers. *Med Care Res Rev.* 2007;64:169–190.
11. Peterson ED, Delong ER, Masoudi FA, et al. ACCF/AHA 2010 position statement on composite measures for healthcare performance assessment: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures (Writing Committee to Develop a Position Statement on Composite Measures). *Circulation*. 2010;121:1780–1791.
12. Best Hospitals 2012-13: How They Were Ranked. Available at: http://health.usnews.com/health-news/best-hospitals/articles/2012/07/16/best-hospitals-2012-13-how-they-were-ranked. Accessed October 22, 2012.
13. How we rate hospitals. http://www.consumerreports.org/cro/2012/10/how-we-rate-hospitals/index.htm. Accessed October 22, 2012.
14. Hospital Rankings Get Serious. http://thehealthcareblog.com/blog/2012/08/15/hospital-rankings-get-serious/#more-50270. Accessed October 22, 2012.
15. National Quality Forum Measurement Evaluation Criteria (January 2011). Available at: http://www.qualityforum.org/docs/measure_evaluation_criteria.aspx. Accessed July 27, 2012.
16. Allison PD. Measures of Inequalit. *Am Sociol Rev.* 1978;43:865–880.
17. Polman CH, Rudick RA. The multiple sclerosis functional composite: a clinically meaningful measure of disability. *Neurology.* 2010;74(Suppl 3):S8–S15.
18. DeVellis RF. *Scale Development: Theory and applications*. 2nd ed. Thousand Oaks, CA: Sage Publications, Inc; 2003.
19. O'Brien SM, DeLong ER, Dokholyan RS, et al. Exploring the behavior of hospital composite performance measures: an example from coronary artery bypass surgery. *Circulation*. 2007;116:2969–2975.

20. AHRQ Quality Indicators: Inpatient Quality Indicators Composite Measure Workgroup Final Report. Available at: http://www.qualityindicators.ahrq.gov/Downloads/Modules/IQI/IQI%20Composite%20Development.pdf. Accessed on July 15, 2012.
21. CMS Releases FY 2013 IPPS Proposed Rule: 12 Points to Know. Available at: http://www.beckershospitalreview.com/racs-/-icd-9-/-icd-10/cms-releases-fy-2013-ipps-proposed-rule-12-points-to-know.html. Accessed July 15, 2012.
22. Dimick JB, Staiger DO, Baser O, et al. Composite measures for predicting surgical mortality in the hospital. *Health Aff*. 2009;28:1189–1198.
23. AHRQ Quality Indicators: Composite Measures User Guide for the Patient Safety Indicators (PSI). Available at: http://www.qualityindicators.ahrq.gov/Downloads/Modules/PSI/V42/Composite_User_Technical_Specification_PSI.pdf. Accessed on September 15, 2012.
24. Kernisan LP, Lee SJ, Boscardin WJ, et al. Association between hospital-reported Leapfrog safe practices scores and inpatient mortality. *JAMA*. 2009;301:1341–1348.
25. Qian F, Lustik SJ, Diachun CA, et al. Association between Leapfrog safe practices score and hospital mortality in major surgery. *Med Care*. 2011;49:1082–1088.