Methodology

U.S. News & World Report 2019-2020 Best Hospitals Procedures & Conditions Ratings

Greta Martin, ScM
Anwesha Majumder, MHS
Zach Adams, ScM
Tavia Binger, MSPH
Ben Harder

July 22nd, 2019

To Whom It May Concern:

U.S. News & World Report's "Best Hospitals: Procedures & Conditions Ratings" study is the sole and exclusive property of U.S. News & World Report, L.P., which owns all rights, including but not limited to copyright, in and to the attached data and material. Any party wishing to cite, reference, publish or otherwise disclose the information contained herein may do so only with the prior written consent of U.S. News. Any U.S. News-approved reference or citation must identify the source as "U.S. News & World Report's Best Hospitals" and must include the following credit line: "Copyright © 2019 U.S. News & World Report, L.P. Data reprinted with permission from U.S. News." For permission to cite or use, contact permissions@usnews.com.

EXECUTIVE SUMMARY

This report describes the methodology underlying U.S. News & World Report's 2019-20 Best Hospitals: Procedures & Conditions, ratings of U.S. medical centers' performance in nine relatively common inpatient procedures and conditions.

The procedures and conditions ratings significantly extend the mission of Best Hospitals: to provide a decision tool that helps the public identify hospitals that best meet their needs. Since 1990, the Best Hospitals rankings have focused on hospitals that excel in treating especially challenging inpatient diagnoses. However, a comparatively small number of patients need such hospitals compared with those who need relatively routine inpatient care. The procedures and conditions in which U.S. News began to rate hospitals in 2015 are much more typical of those needs and represent an integral part of the standard repertoire for most community hospitals. The ratings provide the public with the best possible information, using the best available data sources, for choosing a local source of competent care.

U.S. News is committed to transparency and therefore publishes detailed descriptions of the methodologies used to rank and rate hospitals. Questions and constructive suggestions are welcomed.

The 2019-20 ratings evaluate hospitals in nine procedures and conditions:

Abdominal aortic aneurysm repair (AAA)
Aortic valve surgery (AVR)
Chronic obstructive pulmonary disease (COPD)
Colon cancer surgery
Congestive heart failure (CHF)
Heart bypass surgery (CABG)
Hip replacement
Knee replacement
Lung cancer surgery

Ratings in other procedures and conditions may be added over time.

Unless otherwise noted, the metrics discussed in these pages refer only to the nine cited ratings cohorts.

More than 4,500 hospitals were evaluated in at least one of the nine procedures or conditions, using methods developed by health services researchers at U.S. News & World Report.

Each hospital meeting the rating criteria is assigned to one of three overall performance bands – high performing, average and below average – so patients and families can quickly identify hospitals whose performance meets or exceeds the national norm. Approximately 1,400 hospitals

received a high performing rating in one or more procedures and conditions, and 57 hospitals received a high performing rating in all nine procedures and conditions.

Sources of data included Medicare administrative claims, Medicare Hospital Compare, the American Hospital Association annual survey, publicly available clinical registries, and external designations.

Unless otherwise noted, ratings reflect care received by inpatients age 65 and older. Older patients are at greater risk – they tend to have higher incidence and severity of comorbidities upon admission and illnesses that are more advanced than those of younger patients. The quality of care of over-65 patients is generally regarded as indicative of a hospital's capabilities.

CHANGES IMPLEMENTED IN 2019-2020

- In our hip and knee cohorts, a new measure was incorporated indicating the proportion of procedures at a hospital that were performed by surgeons board certified in orthopedics. The measure accounts for both Medical and Osteopathic Doctors who are board certified in Orthopedic Surgery.
- Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) data has been used in the Procedures & Conditions ratings since its inception. This year, analogous data from the PPS-exempt Cancer Hospital (PCH) HCAHPS dataset were used for cancer specialty hospitals exempt from the CMS Inpatient Prospective Payment System; previously HCAHPS data were treated as missing for these hospitals. For a small number of non-PPS-exempt hospitals for which delivery of cancer care is integrated with a neighboring PPS-exempt cancer hospital, we used the PPS-exempt hospital's PCH HCAHPS score in Colon Cancer Surgery and Lung Cancer Surgery and used the non-PPS-exempt hospital's HCAHPS score in all other procedures and conditions.
- Nurse Magnet recognition was no longer used as a measure in three cohorts: Hip Replacement, Knee Replacement and Heart Failure.
- Length of stay was added as an outcome measure in the AAA and colon cohorts.
- Unplanned readmission was added as an outcome measure in the knee cohort.
- A measure that indicates the percentage of healthcare personnel who received a timely vaccination during flu season (IMM_3_OP_27_FAC_ADHPCT) was added as a candidate process indicator in all cohorts.

TABLE OF CONTENTS

INTRODUCTION

First published in 2015, Best Hospitals: Procedures & Conditions (formerly Best Hospitals for Common Care) is a key component of the U.S. News & World Report suite of healthcare consumer decision-support tools. For 2019-20, hospitals are rated in nine common inpatient procedures and conditions:

Abdominal aortic aneurysm repair (AAA)
Aortic valve surgery (AVR)
Chronic obstructive pulmonary disease (COPD)
Congestive heart failure (CHF)
Colon cancer surgery
Heart bypass surgery (CABG)
Hip replacement
Knee replacement
Lung cancer surgery

Although these procedures and conditions are services common to community hospitals, many studies demonstrate wide variability between hospitals in the quality of the care provided. Access to information about the performance of local hospitals enables patients to better select hospitals that are the most likely to offer better, safer care.

In focusing on large numbers of patients with relatively straightforward needs, these ratings complement the Best Hospitals rankings published annually by U.S. News since 1990. Those rankings identify facilities with demonstrable ability to handle a much smaller but far more challenging patient population of difficult and high-risk cases.

Quality of care has no ready definition or definitive metric, and there is no consensus on the best way to measure it. Some of its aspects are readily quantified while others are more challenging to measure. Moreover, what matters to one patient, such as reported levels of patient satisfaction, may be of little concern to another patient, who might prioritize rates of survival or complications. Offering not only an overall rating but a window into the individual elements that make up the rating recognizes the need for both.

Domains of Quality

Like the Best Hospitals rankings, the procedure and condition ratings use the Donabedian paradigm, which reflects a relationship between structure, process and outcomes. Avedis Donabedian described this now widely accepted dynamic in 1966, which has been applied to

hospital care as follows:

- *Structure* refers to hospital resources connected with patient care, such as the number of nurses, availability of certain specialists and accreditations and certifications by outside organizations.
- *Process* refers to the way in which diagnoses, treatments, practices to avoid harm to patients and other care are rendered – whether steps known to be effective in preventing infections and medical errors, for example, are built into hospital routine.
- *Outcomes* refers to the results of care, including death, harm to patients, preventable readmissions, unusually long hospitalizations, and other consequences.

Failing to acknowledge the influence of random variation in quality metrics can produce results that misleadingly identify one hospital as superior or inferior to another. The methodology for the procedures and conditions ratings takes into account not only how each hospital performed on different measures but also the level of statistical certainty of those performance metrics. Larger sample size produces higher statistical confidence, which can result in a high-volume hospital with modestly above average results to be rated more highly than a low-volume hospital with comparatively better observed results. This is because the second hospital's performance was more likely due to chance.

An important goal of the methodology is to give patients a clear bottom line. Despite the complexity of the measurement issues and the usefulness of particular types of information such as death and readmission rates, patients deserve an overall conclusion: How well does a hospital perform compared to other hospitals in heart bypass surgery or other specific procedure or condition? The ratings aggregate the measures in each cohort of care into an overall assessment by placing a hospital into one of three composite bands: high performing, average and below average.

Data Sources

We used the following data sources:

1. **Publicly available indicators**. Measures of performance in the public domain were obtained from the websites of Hospital Compare maintained by the federal Centers for Medicare & Medicaid Services (CMS), the Society of Thoracic Surgeons, and the National Cancer Institute. The main limitation of many published sources is that data for all hospitals are not available, hampering direct comparison. In some cases where data are very robust, however, including data only for a subset of hospitals can improve the accuracy of relative ratings.

2. **Inpatient Limited Data Set Standard Analytical Files (Inpatient LDS SAF).**

Administered by CMS, the Inpatient LDS SAF dataset contains inpatient hospitalization claims filed on behalf of patients enrolled in traditional Medicare. The LDS SAF provides a thorough administrative record for each patient across all inpatient encounters related to an episode of care. All data were de-identified prior to being provided to U.S. News.

The data imperfectly mirror the overall hospital inpatient population because other than those with disabilities and end-stage kidney disease, Medicare members are age 65 and older. How older inpatients fare, on the other hand, represents a test of hospital performance that for many procedures and conditions is more revealing than results would be from a population that includes younger and healthier patients. Broad "all-payer" data that would permit such an evaluation for all hospitals, moreover, is unavailable.

Consequently, inpatient LDS SAF data are widely used in academic literature to permit meaningful comparisons of rates of death, complications, readmission, infection and other outcomes on a like-to-like basis across most hospitals.

3. **American Hospital Association Annual Survey**. Through its Health Forum arm, the AHA surveys all U.S. hospitals annually (including AHA nonmembers) to obtain operational and clinically relevant information, such as types and levels of staffing. The collected data is the most complete such information available on U.S. hospitals.

4. **Hospital Consumer Assessment of Healthcare Providers and Systems Survey (HCAHPS)**. The federal government releases quarterly results of ongoing surveys of recently discharged inpatients conducted by more than 4,000 hospitals. The results comprise a rolling 12-month assessment of inpatients' opinions about their stay in various respects such as staff communication, treatment of pain and overall opinion of the hospital. The procedure and condition ratings incorporate overall patient opinions into the methodology. Other HCAHPS survey results are displayed but not integrated into the ratings. Because the government aggregated HCAHPS data across each hospital, patients' opinions about their care in specific departments cannot be determined.

5. **Doximity.** With a membership of over 70 percent of all doctors in the US, Doximity is currently the largest social network for physicians and clinicians in the country. Doximity's database of doctor profiles is pre-populated using various sources, and can be modified once the profile is claimed by the verified physician. For the two orthopedic procedures, the current analysis incorporates information on board certifications and specialties from this database.

Selection of Conditions and Procedures

Procedures and conditions were selected based on the significance of their Medicare inpatient volumes, the availability of data allowing for hospital-to-hospital comparisons, and the presence of a sufficient degree of risk or complexity that the quality of a hospital's performance could be important.

The procedures and conditions evaluated for publication are listed in Table 1 along with total numbers of Medicare inpatients at rated hospitals (those with 15 or more patients over five years) and at all hospitals, both rated and unrated, during the five-year period from January 2013 through December 2017 (for simplicity, referred to as 2013-2017 throughout this document). As the table shows, most Medicare patients in these procedure and condition cohorts received care at rated rather than unrated hospitals.

NB: A revised version of the 2017 Inpatient Standard Analytical File was released in March of 2019, which included approximately 150,000 additional beneficiaries that were not contained in the original file from November 2018. The inclusion of this file added marginal volume across all cohorts, with the most volume being in the orthopedic and medical cohorts.

Table 1: *Number of Patients 2013-2017, by Cohort**

	Estimated Medicare Volume	
	Rated Hospitals	*All Hospitals*
Abdominal aortic aneurysm repair	121,183	123,547
Aortic valve surgery	507,028	508,252
Chronic obstructive pulmonary disease (COPD)	1,880,606	1,881,763
Colon cancer surgery	215,107	221,391
Heart bypass surgery	507,028	508,252
Heart failure	3,234,353	3,235,466
Hip replacement	951,604	954,866
Knee replacement	2,024,101	2,025,493
Lung cancer surgery	130,129	133,256

*Estimates include fee-for-service and Medicare Advantage visits

The definitions for the procedures and conditions were created for this project and are not identical to those used by CMS for its performance indicators. In defining patients to include or exclude for a given condition or procedures, three aims were paramount for maximizing statistical and clinical accuracy:

1. **Maximal homogeneity:** a group of patients as alike as possible other than with regard

to factors that could be adequately managed through risk adjustment.

2. **Maximal sample size:** a sufficiently large number of patients for statistical robustness. The CMS data contain all traditional Medicare patients ages 65 and above. Selection of procedures and conditions was therefore limited to those involving sufficient numbers of such patients to provide statistically meaningful measures.
3. **Minimal coding variation:** code definitions that are relatively immune to large variations due to differences in coding practices. In considering this issue, it was particularly important to try to avoid systematic coding biases that might benefit particular organizations and encourage gaming, as opposed to random coding variations that would simply add noise and reduce precision.

These three goals are not in harmony. While (1) argues for narrowly defined patient cohorts, (2) and (3) argue for broader inclusion criteria. This dynamic factored into the decisions regarding which specific procedures and conditions are evaluated in the ratings.

Procedures

When we rate procedures or conditions for which CMS has also developed quality measures, our case inclusion requirements are generally broader than the CMS definitions. The rationale is that outcome measures should not be distorted by decisions about the way in which patients are treated or procedures are coded. Using procedure codes to exclude patients from a cohort or to risk-adjust may be inappropriate if the choice of code and/or procedure is within a doctor's or hospital's discretion. In such cases, exclusion or risk adjustment by procedure code could encourage upcoding, or perversely reward a hospital for performing a higher-risk procedure when a lower-risk alternative may be indicated, such as selection of open surgery over a minimally invasive procedure.

To the extent that a hospital's use of different interventions and associated procedure codes is a reliable indicator of a patient's risk, the desire for homogeneity suggests using procedure codes for risk adjustment or to define exclusion criteria. However, to the extent that the use of different procedures represents a hospital's decisions in treating an otherwise homogeneous group of patients, procedure codes should not be used in this way. This last issue was of particular concern, since using procedure codes in this way could encourage manipulation of data. With these considerations in mind, the procedures were defined as follows.

Abdominal aortic aneurysm repair. Two surgical approaches -- endovascular and open -- are employed to repair aneurysms. As the risk profiles differ for these two approaches, the open approach was included in the risk-adjustment model for the survival outcome in this cohort. Open aneurysm repair represents about a tenth of all repairs in the Medicare fee-for-service population, and that number is falling. Accordingly, we rate hospitals primarily on endovascular repair, and include one risk-adjusted outcome (survival) in the ratings that includes open surgeries. We limited

the cohort to abdominal repairs because repairs in other locations pose different levels of risk. We excluded patients with ruptured aneurysm, indicating an emergent procedure. Outcomes from elective surgery may differ from those for emergent surgery, and patients undergoing emergent surgery typically are unable to choose which hospital they visit.

Aortic valve surgery. The cohort includes only procedures with ICD-9 and -10 codes typically used for open surgical valve replacement. It excludes patients undergoing concurrent coronary artery bypass. Transcatheter aortic valve therapies, which have become increasingly common since the time period covered by this analysis, were not included. As in our CABG cohort, each hospital's relevant volume was based on the sum of isolated CABG and isolated aortic valve surgery cases.

The transcatheter aortic valve replacement (TAVR) procedure has emerged in recent years as a feasible alternative to surgical aortic valve replacement (AVR). No methodology changes regarding TAVR have been implemented for this year's analysis; however, an indication of which hospitals perform TAVR will be published on the Best Hospitals website. The flag on the website will reflect hospitals that either: 1) participate in the Transcatheter Valve Therapy (TVT) registry, as indicated by ACC NCDR data published on February 1st, 2019, or 2) performed 20 or more Medicare fee-for-service TAVR procedures in the past two years. As the volume of TAVR procedures continues to increase, the incorporation of TAVR in future analytic strategies is also increasingly probable.

Colon cancer surgery. We limited the cohort to patients diagnosed with this cancer because outcome profiles for colon cancer patients differ from those of patients undergoing colon surgery for other conditions, and because there may be differences in hospital quality depending on the condition giving rise to the surgery.

Heart bypass surgery. Only patients undergoing isolated open coronary artery bypass graft (CABG) were included in the outcome measures; patients who had valve replacement or repair or other significant cardiac procedures at the same time were excluded, as they have a different risk profile. However, to estimate each hospital's relevant volume in heart surgery, the volume measure for this cohort was based on the sum of isolated CABG and isolated aortic valve surgery cases.

Hip replacement, knee replacement. These two cohorts focus on patients receiving elective primary arthroplasty of the hip or knee for osteoarthritis. We excluded patients who lacked a principal diagnosis of osteoarthritis, and those receiving partial joint replacements, revisions, and concurrent hip and knee surgeries.

Lung cancer surgery. The cohort included patients undergoing lobectomy and pneumonectomy, as well as patients undergoing sublobar resection. We limited the cohort to those

with a related cancer diagnosis, as outcomes for cancer patients may differ from those undergoing lung resection for other reasons.

Conditions

Heart Failure and COPD. We based these cohorts on the Clinical Classification Software developed for the federal Agency for Healthcare Research and Quality. The Heart failure cohort includes ICD-9 and -10 principal diagnosis codes in the nonhypertensive congestive heart failure, congestive heart failure, and heart failure subgroupings. The COPD cohort includes principal diagnosis codes in the chronic obstructive pulmonary disease and bronchiectasis grouping.

ICD-9 and -10 inclusion and exclusion codes for the nine cohorts are presented in Appendix A, which is a separate, downloadable spreadsheet.

Inclusion of Providers and Cases

All hospitals represented in the 2017 American Hospital Association Annual Survey were initially considered for inclusion in the ratings analysis, unless categorized on the AHA survey by a CNTRL code (41-48) indicating government ownership.

Hospitals also were excluded if they lacked a valid six-digit Medicare provider number (MPN) associated with their AHA profile. In some cases, we attributed outcomes from multiple MPNs to a single AHA profile. This occurred when, in the judgment of U.S. News, the AHA profile encompassed the operations of two or more clinically integrated facilities or campuses that maintained separate MPNs during any portion of the 2013-2017 analytical period.

In the Heart failure and COPD cohorts only, we excluded hospitals with SERV codes indicating service type other than general acute-care from rating eligibility.

Further, we excluded a small number of hospitals from the ratings in individual cohorts when their volume, although otherwise sufficient, was not large enough to allow estimation for at least one outcome used in that cohort. This occurred, for instance, with hospitals that began offering knee replacement near the end of our analytical period, but performed no surgeries during the surveillance period for postoperative infection in that cohort.

Table 2: *Number of Hospitals, by Cohort*

	Rated	All
Abdominal aortic aneurysm repair	1,302	1,683
Aortic valve surgery	984	1,134
Chronic obstructive pulmonary disease (COPD)	4,242	4,410
Colon cancer surgery	2,482	3,493
Heart bypass surgery	1,132	1,160
Heart failure	4,245	4,396
Hip replacement	2,798	3,264
Knee replacement	3,213	3,410
Lung cancer surgery	1,233	1,789

Ratings are displayed on usnews.com for all other hospitals with estimated volume of at least 15 Medicare patients (fee for service cases plus estimated managed care cases) for a particular procedure or condition. For hospitals with volumes of lower than 15, we display information on selected metrics, but not overall composite ratings or claims-based outcome measures. The number of hospitals rated in each of the nine procedures and conditions is shown in Table 2.

We excluded patients under 65 from our analysis, as well as patients who left against medical advice. Further, we excluded records from our analysis of the CMS Inpatient SAF files when the admission was missing key information for modelling purposes or contained data that were logically inconsistent or otherwise indicative of data-entry errors. Specifically:

- The admission was identified as a duplicate admission record
- The patient did not appear in the denominator file
- Admissions where patient sex is not identified
- Admissions with length of stay greater than 1 year
- Admissions involving patients with date of death prior to admission date
- Admissions involving patients with multiple dates of death

Outcomes

Outcomes were primarily derived from the 2012-2017 Limited Data Set (LDS) Standard Analytic File (SAF) published by the Centers for Medicare and Medicaid Services (CMS). This data set enabled us to capture and attribute outcomes, such as death and readmission, to the index hospital even if a patient experienced the outcome out of hospital or at a different facility. We constructed measure definitions so that they encompassed five years of data and used the most recent data available. The time periods from which index visits are drawn vary, depending on the

post-admission surveillance requirements specific to each measure.

The following CMS-based outcomes were used in our final composite models to evaluate hospital performance in each cohort. The relative contribution of each outcome to the overall rating is depicted under "Indicators and Correlations With Scores." Dates during which a patient was considered to be eligible for inclusion as an index case for each measure are included in parentheses after the measure description.

1. **Deaths within 30 days**. Reflects mortality within 30 days of the procedure date for surgical cohorts, or 30 days of admission for medical cohorts.
   (12/2/2012-12/1/2017)
2. **Unplanned readmission within 30 days.** Reflects the relative number of patients who had an unplanned readmission for any cause within 30 days of discharge. Unplanned readmission is determined according to the algorithm developed for the CMS hospital-wide 30-day unplanned readmission measure.[1]
   (12/2/2012-12/1/2017)
3. **Surgical site infection derived from claims data**. Reflects the relative number of patients who developed a surgical site infection after the index procedure. Recently published literature[2,3,4,5,6] indicates that a careful approach to constructing claims-based infection measures can accurately identify hospitals with unusually low or high infection rates.
   (1/1/2012-1/1/2017 for hip and knee replacement, 11/1/2012-10/31/2017 for aortic aneurysm, aortic valve surgery, and heart bypass surgery)

---

[1] Horwitz, L. I., Partovian, C., Lin, Z., Grady, J. N., Herrin, J., Conover, M., … Drye, E. E. (2014). Development and use of an administrative claims measure for profiling hospital-wide performance on 30-day unplanned readmission. *Annals of Internal Medicine*, *161*(0), S66–S75. http://doi.org/10.7326/M13-3000

[2] Calderwood, M. S., A. Ma, Y. M. Khan, M. A. Olsen, D. W. Bratzler, D. S. Yokoe, D. C. Hooper*, et al.* "Use of Medicare Diagnosis and Procedure Codes to Improve Detection of Surgical Site Infections Following Hip Arthroplasty, Knee Arthroplasty, and Vascular Surgery." *Infect Control Hosp Epidemiol* 33, no. 1 (Jan 2012): 40-9.

[3] Letourneau, A. R., M. S. Calderwood, S. S. Huang, D. W. Bratzler, A. Ma, and D. S. Yokoe. "Harnessing Claims to Improve Detection of Surgical Site Infections Following Hysterectomy and Colorectal Surgery." *Infect Control Hosp Epidemiol* 34, no. 12 (Dec 2013): 1321-3.

[4] Calderwood, M. S., K. Kleinman, D. W. Bratzler, A. Ma, R. E. Kaganov, C. B. Bruce, E. C. Balaconis*, et al.* "Medicare Claims Can Be Used to Identify Us Hospitals with Higher Rates of Surgical Site Infection Following Vascular Surgery." *Med Care* 52, no. 10 (Oct 2014): 918-25.

[5] Calderwood, M. S., K. Kleinman, D. W. Bratzler, A. Ma, C. B. Bruce, R. E. Kaganov, C. Canning*, et al.* "Use of Medicare Claims to Identify Us Hospitals with a High Rate of Surgical Site Infection after Hip Arthroplasty." *Infect Control Hosp Epidemiol* 34, no. 1 (Jan 2013): 31-9.

6 Calderwood, M. S., Kleinman, K., Murphy, M. V., Platt, R., Huang, S. S. "Improving Public Reporting and Data Validation for Complex Surgical Site Infections After Coronary Artery Bypass Graft Surgery and Hip Arthroplasty." Open Forum Infectious Diseases 1, no. 3 (Dec 2014).

4. **Revision within 1 year, total joint replacement cohorts**. Reflects the relative number of patients who had a procedure to address a problem with a replacement joint within 1 year of the original operation.
   (1/1/2012-1/1/2017)
5. **Length of stay**. Reflects the relative number of patients who experienced length of stay in the highest quartile for that cohort.
   (12/2/2012-12/1/2017)
6. **Discharge to a location other than the patient's home.** Reflects the relative number of patients discharged to a location other than home, such as a skilled nursing facility, a long-term acute care facility, or a different hospital. Details are provided in Appendix B.
   (1/1/2013-12/31/2017)

Other claims-based outcome measures were risk-adjusted and included in the model selection process, but were not included in the final composite measure. We evaluated complications of total joint replacement, which reflects the relative number of hip or knee patients who developed one or more of the complications detailed in the NQF 1550 quality measure. These include heart attack, pneumonia, blood infections, death, or other complications in patients undergoing joint replacement surgery. In the joint cohorts, we looked at revision within a 5 year period using a time-to-event regression model.

Readmission for cancer cohorts was considered -- criteria for this indicator closely follow those developed by the Alliance of Dedicated Cancer Centers. The measure excludes elective readmissions, those preceded by development of metastatic cancer, and those for chemotherapy or radiation. We also looked at 7-day all-cause readmission, which reflects the relative number of patients who returned to the hospital for any reason (planned or unplanned) within 7 days of discharge. In the COPD cohort, we looked at an extended readmission measure that evaluated readmission in a 60-day period post-discharge, as a longer window may more accurately assess readmissions for COPD exacerbation.

All claims-based outcomes were risk-adjusted using a hierarchical logistic regression model that controlled for potential confounders, with a random intercept for hospital identity. Details on the model specified for each cohort are listed under "Indicators and Correlations With Scores". In all instances, continuous variables were treated as such in our composite modeling in order to make maximum use of the information contained in the variable, and to minimize the risk of measurement error due to categorization.

In addition to SAF-derived outcomes, the heart bypass surgery and aortic valve surgery cohorts used data from the Society of Thoracic Surgeons. STS publishes ratings for approximately 600 hospitals that evaluate hospital performance in mortality, major morbidity, and adherence to certain best practices. STS did not provide data for the U.S. News analysis, and their use by U.S. News in public reporting does not represent an endorsement by STS for this purpose. Data for heart

bypass surgery and aortic valve surgery were retrieved from https://publicreporting.sts.org/acsd on February 1, 2019.

Process Measures

We evaluated a variety of process measures, which were obtained primarily from the CMS Hospital Compare website as well as the inpatient claims datasets. Most were not suitable for inclusion due to issues with missing data or other data validity concerns, and others were excluded after going through our modeling process because they did not demonstrate good empirical model fit. The following measures were included in the composite model for one or more cohorts:

- **Patient flu immunization.** Percentage of patients who received a timely vaccination during flu season. Derived from the CMS Hospital Compare Database.
- **Worker flu immunization.** Percentage of healthcare personnel who received a timely vaccination during flu season. Derived from the CMS Hospital Compare Database.
- **Noninvasive ventilation.** Whether the hospital treated at least 20% of ventilated patients noninvasively. Derived from CMS claims and used in the heart failure and COPD cohorts.
- **HCAHPS.** Measure reflecting patient experience, as reported by the HCAHPS survey of recently discharged patients.[7] Indicative of the overall score from the HCAHPS survey and used in all cohorts. We used the linear mean score rather than the HCAHPS star rating because the former is a continuous measure and provides more information for our analysis. We used the HCAHPS dataset for 1/1/2017-12/31/2017.
- **Transfusion.** Percentage of patients who did not need to undergo transfusion with donor blood, which can be necessary if unexpected blood loss occurs during surgery. Derived from CMS claims and used in the surgical cohorts.
- **Board certification.** Percentage of surgeries in the hip and knee replacement cohorts that were performed by board-certified surgeons. The measure accounts for both MDs and DOs who are board certified with either the American Board of Orthopaedic Surgery, the American Osteopathic Board of Orthopedic Surgery, or the National Board of Physicians and Surgeons (orthopedic specialty) as represented in Doximity on March 5th, 2019. The American Osteopathic Association provided additional information on board certification of its members, in order to supplement data on DOs from the Doximity database.

Structural Measures

Structural measures of health care evaluate staff, services, equipment and other resources used to deliver care. Structural indicators that have been associated with good outcomes for patients

[7] The current version of the survey is available at http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalHCAHPS.html.

were included. A provider with the right structures will not necessarily have good outcomes, but it is more likely.

It is known, for example, that hospitals that operate more frequently tend to have better outcomes. For this reason, volume – number of operations or patients treated for the targeted medical conditions – was one of the metrics used. Evidence similarly suggests that higher nurse and specialty staffing levels contribute to a better outcome and a better experience for patients. Therefore measures of nurse staffing and presence of specialized intensive care physicians were included. In addition to volume, five structural indicators were employed, two of them derived from data collected for the 2017 American Hospital Association Annual Survey.

- **Volume.** There is widespread evidence that hospitals performing higher numbers of common procedures get better outcomes. Volume derived from CMS claims was therefore included as an indicator. In the heart bypass surgery and aortic valve surgery ratings, volume of isolated CABG procedures was combined with the volume of valve replacement or repair procedures to get a more accurate picture of the total volume of cardiothoracic procedures provided by a hospital. Volume totals were adjusted to account for Medicare Advantage cases. (Hospitals with very low volumes – defined as fewer than 15 cases over five years – were not rated because their numbers were too low to establish whether the quality of care was different from average.)
- **Nurse staffing**. The number of nurses involved in direct patient care at a hospital is known to play a major role in the quality of care[8,9,10,11,12,13]. Various ways to calculate a measure of nursing intensity are available. For this project, it is a ratio that reflects inpatient and outpatient care. The numerator is the total number of staff registered nurses (RNs), converted to full-time equivalents (FTEs) – two half-time nurses add up to one FTE, for example. Only nurses with an RN degree from an approved nursing school and a current state registration are included. Supervisory positions are excluded.

  Making sense of nurse staffing requires comparing the number of staff to the total workload. The two most commonly used approaches are:

  - Total inpatient days

[8] Unruh, L. (2003) Licensed Nurse Staffing and Adverse Events in Hospitals. Medical Care. 41(1) (pp142-152)

[9] Stanton MW, Rutherford MK. (2004) Hospital nurse staffing and quality of care. Agency for Healthcare Research and Quality. Research in Action Issue 14. ARHQ Pub. No. 04-0029.

[10] Spetz J, Donaldson N, Aydin C, Brown DS. (2008) How Many Nurses per Patient? Measurements of Nurse Staffing in Health Services Research. Health Services Research. 43(5) (pp1674-1692)

[11] Lankshear AJ, Sheldon TA, Maynard A. (2005) Nurse Staffing and Healthcare Outcomes. Advances in Nursing Science. 28(2) (pp163-174)

[12] Hickham DH, Severance S, Feldstein A. (2003) The Effect of Health Care Working Conditions on Patient Safety. AHRQ Evidence Report/Technology Assessment (74)

[13] Needleman J, Buerhaus P, Pankratz VS, Leibson CL, Stevens SR, Harris M. (2011) Nurse Staffing and Inpatient Hospital Mortality. New England Journal of Medicine 364(11) (pp1037-1045)

- Adjusted patient days

The denominator used in the ratings is adjusted patient days, an aggregate figure reflecting the number of days of inpatient care plus an estimate of the volume of outpatient services, expressed in units equivalent to an inpatient day in terms of level of effort. The latter was derived by first multiplying the number of outpatient visits by the ratio of outpatient revenue per outpatient visit to inpatient revenue per inpatient day. The product (which represents the number of patient days attributable to outpatient services) was then added to the number of inpatient days.

We chose to use the number of FTE registered nurses divided by adjusted patient days; numerous published studies have validated this metric as an indicator of quality.

- **Intensivists**. Intensivists are board-certified physicians with subspecialty or fellowship training in critical-care medicine. They specialize in managing critically ill patients in hospital intensive care units (ICUs). A hospital received credit if it reported having at least one full-time equivalent intensivist available, on staff or from another source, in any ICU other than neonatal or pediatric. Research indicates that better outcomes are associated with the presence of intensivists[14, 15] This measure was derived from the 2017 AHA Annual Survey.
- **Heart failure programs.** Indicates whether a hospital provided CHF patients with LVAD implantation, heart transplantation, or both. Derived from CMS claims.
- **National Cancer Institute status.** Whether hospital is designated a cancer center by the National Cancer Institute, which funds clinical trials and other advances in cancer care.
- **Transparency.** The heart-surgery cohorts also included a U.S. News "transparency indicator" that rewarded hospitals that permit STS to publicly report their performance data. This was done in part to encourage all hospitals, regardless of performance, to release their data and by doing so expand the data universe. Published research by STS-affiliated researchers[16] and independent analysis by U.S. News found that hospitals that do not publicly report via STS performed worse than STS reporters on quality measures such as risk-adjusted mortality, morbidity and readmissions following heart surgery. While not establishing the direction of causality, these observed correlations between STS-mediated transparency and better outcomes support the use of transparency as an indicator of higher quality of care.

[14] Pronovost PJ, Holzmueller CG, Clattenburg L, Berenholtz S, Martinez EA, Paz JR, Needham DM. "Team care: beyond open and closed intensive care units." *Current Opinion in Critical Care*. 2006; 12(6):604-8.

[15] Sapirstein A, Needham DM, Pronovost PJ. "24-hour intensivist staffing: balancing benefits and costs." *Critical Care Medicine*. 2008; 36(1):367-8.

[16] Shahian, David M., et al. "The Society of Thoracic Surgeons voluntary public reporting initiative: the first 4 years." Annals of surgery 262.3 (2015): 526-535.

Risk Adjustment for Medicare Inpatient Claims-Based Outcomes

When comparing outcomes between hospitals, adjusting for differences in the patients treated at each hospital is critical. A hospital with a 50 percent death rate might be superior to a hospital with a 10 percent death rate, if most of the patients at the first hospital are expected to die and most of the patients at the second hospital are low risk.

We used multilevel logistic regression models to adjust for differences in case mix between hospitals. Multilevel models are a form of regression that allocates variance between variables on two or more levels. We used the empirical Bayes estimate of the hospital intercept as an estimate of each hospital's value for a given outcome. Multilevel modeling accounts for clustering of patient observations within hospitals and allows for more precise rating of hospitals with lower patient volume and fewer outcomes.

We selected covariates for inclusion in risk-adjustment models based on the literature, discussions with clinicians in relevant specialties and a causal-inference model aimed at achieving unbiased estimation of the effect of treatment at a particular hospital on a given outcome.

The model (Appendix C) indicates that an unbiased estimate of the effect of treatment at a given hospital as compared to a hospital selected at random from among those eligible for rating in a cohort, requires adjustment for age, sex, comorbidities, severity of index condition, socioeconomic status (SES), admission urgency, inbound transfer status, and year of admission. In other instances, we have controlled for severity of index condition via restriction (for example, by limiting evaluation of aneurysm repair outcomes to unruptured cases). Because severity of the index condition is correlated with many of the other covariates for which we adjusted, we suspect residual confounding is negligible. "Strengths and Limitations" contains further discussion of this issue.

Risk-Adjustment Variables

- **Age at admission**. Age in years as a continuous variable, obtained from the denominator or Master Beneficiary Summary file.
- **Sex**. Male or female.
- **Inbound transfer status.** Transfer from the initial receiving hospital may indicate complex procedures or conditions. Patients were classified as an inbound transfer if they had been treated at another acute-care hospital on the day of admission, or if the claim admission source variable indicated they were transferred, or if the claim was preceded by another claim indicating outbound transfer status.
- **Year of hospital admission**. Quality of care tends to improve over time. This means the risk of adverse outcomes is less year to year. For that reason, year of admission is included as a risk factor.

- **Elixhauser comorbidities.** We controlled for the comorbidities identified by Elixhauser et al[17] as being predictive of mortality.
- **Medicare status code**. The reason or reasons why the patient is eligible for Medicare: age, disability or end-stage renal failure. This is a proxy for comorbidities.
- **Socioeconomic status**. Patients with lower incomes are typically sicker when they arrive at the hospital, and may face more challenges in obtaining or managing their care after they are discharged. This can affect their risk of death, readmission and complications. When hospitals differ by the socioeconomic status of their patients, this can create bias in comparing outcomes. Our risk models include "dual eligibility" as a measure of socioeconomic background. Patients who are eligible for both Medicare and Medicaid are treated as a separate risk group.
- **ICD version.** We controlled for which ICD version each visit was coded under. Visits with claims dated October 1, 2015 or later have procedures and diagnoses coded in ICD-10, and visits with claims dated September 30, 2015 or earlier are coded in ICD-9.
- **Medical cohort risk adjusters.** Binary flags indicating whether a patient ever left against medical advice, had ever been admitted for the same condition, or had a history of mechanical ventilation were included in the CHF and COPD risk adjustment models.
- **Surgical cohort risk adjusters.** A binary bilateral flag was included in the hip and knee risk adjustment models, indicating whether the operation was performed on both joints simultaneously. A flag for approach (open or endoscopic) was included in the abdominal aortic aneurysm repair mortality risk adjustment model.
- **Source of admission.** In our discharge not home outcome measure, we controlled for whether a patient came from a skilled nursing facility.

EVALUATION OF RISK-ADJUSTMENT MODELS

The accuracy of risk-adjustment models is measured by two statistics, the C-statistic and the Hosmer-Lemeshow goodness of fit statistic. The C-statistic estimates the probability that if one subject who experienced an outcome (death, for example) and another who did not are drawn randomly from the data, the model will assign a higher probability of death to the person who died. A C-statistic of .5 indicates the model has no better than random chance at predicting the outcome. A C-statistic in the .60-.69 range indicates limited discrimination, .70-.79 indicates acceptable discrimination and above .8 indicates good discrimination.

Typically, the C-statistic for mortality models implemented using clinical data range from approximately .75-.85[18]. Our models for some outcomes were generally of similar predictive quality

[17] Elixhauser, Anne, et al. Comorbidity measures for use with administrative data. *Medical care* 36.1 (1998): 8-27.

[18] e.g.: Kozower, Benjamin D., et al. "STS database risk models: predictors of mortality and major morbidity for lung cancer resection." *The Annals of Thoracic Surgery* 90.3 (2010): 875-883; Hamel, Mary Beth, et al. "Surgical outcomes for patients aged 80 and older: morbidity and mortality from major noncardiac surgery." *Journal of the American Geriatrics Society* 53.3 (2005): 424-429.

as those based on clinical data. Our models for readmission and other outcomes had lower predictive power, with C-statistics similar to those in the published literature drawing on claims data. The Hosmer-Lemeshow goodness of fit statistic looks at whether the observed number of outcomes matches the expected number predicted by the model in samples of the population. As this test is not informative for samples over 25,000, we used a procedure designed to evaluate Hosmer-Lemeshow fit in large samples, in which multiple Hosmer-Lemeshow tests are conducted on small samples of the data. A Hosmer-Lemeshow test results in a p-value, which below 0.05 indicates a bad fit. The closer to 1 the mean p-value across all of the sample Hosmer-Lemeshow tests, the better fit.

Table 3: *Predictive Accuracy of Risk-adjustment Models*

Cohort	Outcome	C-statistic	Mean (min, max) of Large-sample Hosmer-Lemeshow Tests
Abdominal aortic aneurysm repair	Readmission prevention	0.681	0.36 (0.06,0.78)
	Infection prevention	0.773	0.60 (0.00,0.97)
	Prevention of prolonged hospitalizations	0.819	0.50 (0.02,0.99)
	Discharging patients to home	0.844	0.66 (0.06,0.99)
	Survival	0.853	0.64 (0.32,0.97)
Aortic valve surgery	Readmission prevention	0.664	0.29 (0.10,0.67)
	Prevention of prolonged hospitalizations	0.801	0.49 (0.03,0.94)
	Discharging patients to home	0.805	0.40 (0.05,0.74)
	Survival	0.788	0.47 (0.01,0.96)
Chronic obstructive pulmonary disease (COPD)	Discharging patients to home	0.753	0.34 (0.04,0.96)
	Survival	0.736	0.58 (0.17,0.99)
Colon cancer surgery	Readmission prevention	0.934	0.83 (0.00,1.00)
	Prevention of prolonged hospitalizations	0.837	0.21 (0.03,0.58)
	Discharging patients to home	0.844	0.45 (0.01,0.98)
	Survival	0.829	0.52 (0.19,0.95)
Heart bypass surgery	Readmission prevention	0.688	0.68 (0.26,0.96)
	Infection prevention	0.771	0.49 (0.07,0.85)
	Prevention of prolonged hospitalizations	0.782	0.60 (0.17,0.89)
	Discharging patients to home	0.801	0.47 (0.01,0.99)
	Survival	0.786	0.47 (0.00,0.80)
Heart failure	Discharging patients to home	0.739	0.57 (0.02,0.93)
	Survival	0.704	0.56 (0.11,0.85)
Hip replacement	Infection prevention	0.707	0.52 (0.04,0.98)
	Prevention of revision surgery	0.641	0.55 (0.01,0.92)
	Prevention of prolonged hospitalizations	0.843	0.64 (0.04,1.00)
	Survival	0.694	0.79 (0.00,1.00)
Knee replacement	Readmission prevention	0.707	0.50 (0.17,0.86)
	Prevention of revision surgery	0.615	0.39 (0.05,0.83)
	Prevention of prolonged hospitalizations	0.839	0.39 (0.15,0.75)
	Survival	0.663	0.71 (0.01,1.00)
Lung cancer surgery	Readmission prevention	0.955	0.92 (0.19,1.00)
	Prevention of prolonged hospitalizations	0.790	0.47 (0.12,0.90)
	Discharging patients to home	0.813	0.42 (0.11,0.96)
	Survival	0.843	0.63 (0.10,0.98)

Construction of Composite Ratings

There are two major issues in constructing a composite rating of quality of medical or surgical care: Determining how much weight each indicator should receive, and accounting for measurement error. Some approaches, such as averaging a set of indicators, weight each equally, and do not address measurement error. More sophisticated statistical procedures can determine empirically how much weight each indicator should be assigned. They can also account for the degree to which an indicator is measured inaccurately due to incomplete risk adjustment, random

variation due to low sample size, and other factors.

Best Hospitals: Procedures & Conditions relies on a statistical method known as confirmatory factor analysis, which assigns empirical weights to the indicators. This approach has been previously used to evaluate provider quality of care[19]. Confirmatory factor analysis is based on the statistical principle that variables sharing a common cause will be correlated. Here, we hypothesize that the various candidate indicators for a given condition or procedure are caused by an underlying, or latent, variable that represents quality of medical or surgical care rendered by a hospital. Thus, for each indicator the model can estimate the extent to which the values are the result of a relationship with quality of care. The remaining variance in the indicator is attributed to measurement error. The degree to which an indicator is correlated with other indicators helps to determine its weight in the equation for the composite scores.

We developed models by evaluating model statistics for all possible combinations of indicators that included at least one indicator from each of the three domains of quality (structure, process and outcomes). From the resulting list of candidate models exhibiting acceptable fit statistics, we selected a final model offering an optimal combination of number of indicators (models with more indicators produce more accurate factor scores), number of outcomes, model fit, and consistency with models in related cohorts. The selected models showed acceptable fit statistics in the majority of the bootstrapped samples in all cohorts.

We evaluated our confirmatory factor analysis models using three measures: the comparative fit index (CFI), the Tucker Lewis Index (TLI), and the root-mean-square error of association (RMSEA). The literature provides a variety of standards for acceptable model fit using these statistics. We sought final models with a CFI and TLI of .9 or greater, and RMSEA of .1 or lower, while also considering our theoretical understanding of the factors that are relevant for quality of care. Most models displayed fit characteristics better than the cutoff value.

We estimated fit statistics with the WLSMV estimator after multiply imputing missing data. We did not assign quality scores to hospitals based on imputed data. To avoid using imputed data for that purpose, we fit final models separately using Full Information Maximum Likelihood with empirical Bayes estimation of hospital factor scores and standard errors. These models are appropriate for use with missing data, but do not provide the fit statistics necessary to guide model development. Fit statistics can change depending on the estimator used, so there is no assurance that the fit statistics estimated apply directly to the models used for score estimation. However, we found the models, including factor loadings, fit statistics, and factor scores, to be consistent across a variety of estimators and software packages.

We assigned each rated hospital in a cohort to one of three bands: below average, average, or

[19] e.g. Keller, S., A. J. O'Malley, R. D. Hays, R. A. Matthew, A. M. Zaslavsky, K. A. Hepner, and P. D. Cleary. "Methods Used to Streamline the CAHPS Hospital Survey." *Health Serv Res* 40, no. 6 Pt 2 (Dec 2005): 2057-77.

high performing. Inference that a hospital was below average or high performing was made at the 75% confidence level. Health researchers more commonly use a 95% confidence level, an approach that is geared toward minimizing the number of false positive results (in this context, incorrectly identifying average hospitals as better or worse than the mean). However, because false negatives (identifying poor-performing hospitals as average) can have serious consequences for patients, we sought to strike a balance between minimizing false positive and false negative results.

Table 4: *Confirmatory Factor Analysis Fit Statistics, by Cohort*

	CFI	TLI	RMSEA
Abdominal aortic aneurysm repair	0.966	0.956	0.095
Aortic valve surgery	0.996	0.995	0.112
Chronic obstructive pulmonary disease (COPD)	0.955	0.932	0.087
Colon cancer surgery	0.965	0.954	0.095
Heart bypass surgery	0.997	0.997	0.099
Heart failure	0.939	0.919	0.101
Hip replacement	0.975	0.967	0.072
Knee replacement	0.943	0.924	0.077
Lung cancer surgery	0.990	0.986	0.082

Indicators and Correlations With Scores

The following tables list the indicators that were included in each cohort's final composite model. The quality score correlation, or standardized factor loading, indicates the relative strength of the relationship in a cohort between a given indicator and hospitals' quality scores. The quality score correlation is determined by the statistical model; it is not a weight and is not applied as a factor of a summative formula. Instead it is applied to a maximum likelihood estimation algorithm that produces the overall quality score for each hospital. The greater the value of the correlation, the stronger the relationship to the quality score. It may be noted that outcome measures in some cohorts are relatively weakly correlated with quality scores. That is to be expected if the incidence of negative outcomes is very low, as it is, for example, for mortality in the hip and knee replacement cohorts, or if there is little variation in the measure from one hospital to another.

Table 5: *Indicator Correlations, Abdominal Aortic Aneurysm Repair*

Indicator	Quality Correlation
Discharging patients to home	0.175
Infection prevention	0.133
Intensivists	0.568
Number of patients	0.500
Nurse staffing	0.461
Patient experience	0.550
Prevention of prolonged hospitalizations	0.186
Readmission prevention	0.192
Survival	0.183
Worker influenza immunization	0.301

Table 6: *Hospital Distribution by Performance Band, Abdominal Aortic Aneurysm Repair*

Band	Description	Number of Hospitals
1	Below average	210
2	Average	890
3	High performing	202

Table 7: *Indicator Correlations, Aortic Valve Surgery*

Indicator	Quality Correlation
Discharging patients to home	0.344
Intensivists	0.540
Number of heart surgery patients	0.540
Nurse staffing	0.473
Patient experience	0.468
Prevention of prolonged hospitalizations	0.362
Public transparency	0.490
Readmission prevention	0.337
STS score	0.383
Survival	0.495

Table 8: *Hospital Distribution by Performance Band, Aortic Valve Surgery*

Band	Description	Number of Hospitals
1	Below average	214
2	Average	581
3	High performing	189

Table 9: *Indicator Correlations, Heart Bypass Surgery*

Indicator	Quality Correlation
Discharging patients to home	0.308
Infection prevention	0.134
Intensivists	0.544
Number of heart surgery patients	0.593
Nurse staffing	0.439
Patient experience	0.471
Prevention of prolonged hospitalizations	0.373
Public transparency	0.557
Readmission prevention	0.464
STS score	0.295
Survival	0.502

Table 10: *Hospital Distribution by Performance Band, Heart Bypass Surgery*

Band	Description	Number of Hospitals
1	Below average	288
2	Average	603
3	High performing	241

Table 11: *Indicator Correlations, Heart Failure*

Indicator	Quality Correlation
Advanced heart program	0.783
Cardiac ICU	0.749
Discharging patients to home	0.348
Intensivists	0.834
Noninvasive breathing aid (≥20%)	0.734
Number of patients	0.808
Nurse staffing	0.344
Patient influenza immunization	0.285
Survival	0.380

Table 12: *Hospital Distribution by Performance Band, Heart Failure*

Band	Description	Number of Hospitals
1	Below average	1,356
2	Average	1,782
3	High performing	1,107

Table 13: *Indicator Correlations, Colon Cancer Surgery*

Indicator	Quality Correlation
Discharging patients to home	0.326
Intensivists	0.675
NCI cancer center	0.722
Number of patients	0.687
Nurse staffing	0.390
Prevention of prolonged hospitalizations	0.136
Readmission prevention	0.126
Survival	0.418
Worker influenza immunization	0.115

Table 14: *Hospital Distribution by Performance Band, Colon Cancer Surgery*

Band	Description	Number of Hospitals
1	Below average	487
2	Average	1,605
3	High performing	390

Table 15: *Indicator Correlations, Chronic Obstructive Pulmonary Disease (COPD)*

Indicator	Quality Correlation
Cardiac ICU	0.705
Discharging patients to home	0.341
Noninvasive breathing aid (≥20%)	0.691
Number of patients	0.730
Nurse staffing	0.288
Patient influenza immunization	0.330
Survival	0.253

Table 16: *Hospital Distribution by Performance Band, Chronic Obstructive Pulmonary Disease (COPD)*

Band	Description	Number of Hospitals
1	Below average	1,026
2	Average	2,348
3	High performing	868

Table 17: *Indicator Correlations, Hip Replacement*

Indicator	Quality Correlation
Board certified physicians	0.212
Infection prevention	0.306
Intensivists	0.454
Number of patients	0.669
Nurse staffing	0.317
Prevention of blood transfusion	0.415
Prevention of prolonged hospitalizations	0.517
Prevention of revision surgery	0.211
Survival	0.262

Table 18: *Hospital Distribution by Performance Band, Hip Replacement*

Band	Description	Number of Hospitals
1	Below average	516
2	Average	1,822
3	High performing	460

Table 19: *Indicator Correlations, Knee Replacement*

Indicator	Quality Correlation
Board certified physicians	0.259
Intensivists	0.486
Number of patients	0.634
Nurse staffing	0.360
Prevention of blood transfusion	0.394
Prevention of prolonged hospitalizations	0.495
Prevention of revision surgery	0.174
Readmission prevention	0.151
Survival	0.209

Table 20: *Hospital Distribution by Performance Band, Knee Replacement*

Band	Description	Number of Hospitals
1	Below average	585
2	Average	2,092
3	High performing	536

Table 21: *Indicator Correlations, Lung Cancer Surgery*

Indicator	Quality Correlation
Discharging patients to home	0.454
Intensivists	0.619
Number of patients	0.641
Nurse staffing	0.445
Patient experience	0.498
Prevention of prolonged hospitalizations	0.451
Readmission prevention	0.236
Survival	0.421
Worker influenza immunization	0.280

Table 22: *Hospital Distribution by Performance Band, Lung Cancer Surgery*

Band	Description	Number of Hospitals
1	Below average	264
2	Average	763
3	High performing	206

VALIDATION OF FACTOR SCORES

The primary means of evaluating construct validity of our measurement models and resulting factor scores was a multi-trait matrix, by which we compared the relative correlations of hospital ratings across cohorts. Specifically, we hypothesized that hospital factor scores for heart bypass surgery and aortic valve surgery would be more closely correlated with each other than with those for other surgeries, and that the two cardiac surgeries would be least correlated with the medical cohorts, CHF and COPD. Similarly, we hypothesized that hip and knee ratings would be highly intercorrelated, that these two ratings would be less well correlated with other surgical procedures, and that they, like the cardiac surgeries, would be least correlated with the medical conditions. Finally we hypothesized that CHF and COPD would be strongly intercorrelated, and less well correlated with the surgical procedure ratings. The correlations, shown in Table 23, provide strong evidence of construct validity. We also examined how ranked hospitals in the Best Hospitals Specialties rankings performed in the Procedures and Conditions ratings, hypothesizing that hospitals who were ranked in specialty care would more often be rated high performing in related P&C cohorts.

Table 23: *Multi-Trait Correlation Table, Hospital Cohort Scores*

	AAA	AVR	CABG	COLON	LUNG	HIP	KNEE	CHF	COPD
AAA	1.000	0.806	0.806	0.755	0.795	0.620	0.664	0.562	0.216
AVR	0.806	1.000	0.914	0.721	0.785	0.585	0.593	0.586	0.238
CABG	0.806	0.914	1.000	0.736	0.773	0.615	0.640	0.624	0.297
COLON	0.755	0.721	0.736	1.000	0.786	0.662	0.684	0.810	0.551
LUNG	0.795	0.785	0.773	0.786	1.000	0.581	0.586	0.578	0.230
HIP	0.620	0.585	0.615	0.662	0.581	1.000	0.913	0.565	0.383
KNEE	0.664	0.593	0.640	0.684	0.586	0.913	1.000	0.620	0.459
CHF	0.562	0.586	0.624	0.810	0.578	0.565	0.620	1.000	0.883
COPD	0.216	0.238	0.297	0.551	0.230	0.383	0.459	0.883	1.000

We further investigated validity by examining concordance of the heart bypass and aortic valve surgery ratings with ratings published by the Society of Thoracic Surgeons. Our intent was not to demonstrate strong concordance with the STS ratings. The STS ratings and the U.S. News ratings cover different time periods and patient populations. The U.S. News ratings are based on three domains of quality measurement (outcomes, process and structure), while the STS ratings do not use structural indicators. Further, the two sets of ratings used different standards for statistical inference. U.S. News employed statistical testing at the $p<.25$ level, while STS ratings employ a standard of $p<.05$. Because of this difference, one would expect that the U.S. News ratings would identify more hospitals as performing above or below average. We expected to find modest agreement between the two sets of ratings, and very few instances of marked disagreement, in which a hospital received the lowest rating from one organization and the highest from the other. Tables 24 and 25 show findings consistent with this hypothesis.

Table 24: *Concordance with STS Aortic Valve Surgery Star Rating*

	STS Star Rating		
	1	2	3
Below Average	3	47	0
Average	7	330	5
High Performing	0	138	29

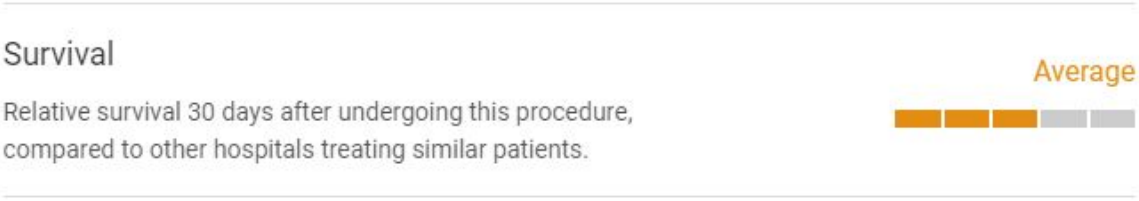
Table 25: *Concordance with STS Heart Bypass Surgery Star Rating*

	STS Star Rating		
	1	2	3
Below Average	8	52	2
Average	10	326	17
High Performing	0	178	39

CATEGORICAL DISPLAY

In our confirmatory factor analysis, we use the continuous form of each measure when possible. For the sole purpose of making information more accessible for patients, we display categorical descriptions of each continuous outcome or process measure on hospital scorecards. See an example of the survival rating in Figure 1.

Figure 1. *Display of survival outcome on U.S. News website.*



A band is displayed on the website for each hospital for every outcome used in a particular cohort's CFA. Our approach to estimating each hospital's band falls under the general rubric of statistical significance testing. The band cutoffs are ***different for each hospital*** and for each measure. This band is reflective of a hospital's estimated risk-adjusted value on the outcome compared to other hospitals, as well as its Medicare claims volume and the incidence of that outcome. We compare each hospital's risk-adjusted outcome value to a normal distribution, taking into account precision as well as how a hospital compares to other hospitals—the greater a hospital's volume, the more certain we are of its estimated outcome value. For rare outcomes, such as death after joint

replacement, relatively few hospitals will have a rate that would designate it as above or below average. It is important to keep in mind that the bands displayed provide a heuristic for the underlying continuous metric that is used to evaluate each hospital's performance.

STRENGTHS AND LIMITATIONS

The Best Hospitals: Procedures & Conditions ratings benefit from our use of the Medicare claims dataset, which tracks a large and clearly defined population. Quality measures derived from this population are generally believed to be representative of those that would emerge from the overall population. The Medicare claims dataset affords statistical power to distinguish between providers, even in some cases when procedures are performed only rarely. Further, the SAF allows researchers to accurately characterize exposures (treatment at a particular hospital) and outcomes (e.g., death, readmission). Our study makes use of multiple datasets, which allowed us to consider indicators from most, if not all, domains relevant to the measurement of hospital quality. We employed statistical procedures that allowed us to simultaneously mitigate the effect of measurement error and empirically determine which combination of indicators would yield valid quality measurement. We conducted extensive research on the validity of our results, which included construction of a multitrait matrix and comparison with external datasets. Our work benefited from the input from a panel of leading health services researchers and clinicians.

A noteworthy limitation of the ratings is that the outcome indicators rely on administrative data, which could lead to bias in several ways. As previously discussed, controlling for severity of the index condition is required to achieve adequate case-mix adjustment. We believe we have largely mitigated this problem by adjusting for a number of variables that are correlated with severity of the index condition, such as transfer status and urgency of admission, and by using other statistical procedures that account for measurement error.

It is possible, however, that our results are biased by residual confounding. Similarly, ascertainment of some outcomes, such as surgical-site infection, requires accurate coding across hospitals. If hospitals differ in the way they code conditions related to SSI, this could result in bias. Error in coding of covariates used in risk adjustment would have the same effect.

Another issue is our use of datasets with incomplete data for all hospitals. Not all hospitals, for example, report process-of-care measures displayed on the Hospital Compare website. We used two methods to deal with incomplete data. To build and evaluate composite models, we imputed data for missing indicators. To calculate factor scores, we relied on a full information maximum likelihood estimator. Both of these approaches assume that the data are missing at random. If the data are missing dependent on values of the process measures themselves, or on other unmeasured variables, the missing data could result in biased estimates of a particular indicator. There is no way to guarantee that this assumption has not been violated. However, we determined that missing Hospital Compare process measures are primarily associated with hospital size, so we do not suspect

that the data are missing conditional on levels of the process variables.

To evaluate the fit of our composite models, we used the WLSMV estimator in the Mplus statistical package. This estimator does not provide full information maximum likelihood estimation of factor scores, so we used a different estimator (MLR) to specify models that produced the actual procedure and condition ratings. We verified that loading coefficients were similar for the two estimators; however, it is possible that if the software was able to calculate fit statistics for the MLR models used to estimate the factor scores, those statistics would differ from the fit statistics published above.

We also did not conduct extensive subgroup analysis to determine if there was any difference in measurement and fit of our model between specific types of hospitals, such as those with a higher proportion of Medicare Advantage patients, or teaching hospitals. There is a chance that our measurement model may have performed differently among these subgroups.

Finally, the statistical procedures used to estimate composite scores cannot assure that the label a researcher applies to the composite score (quality of care, in this case), is in fact germane to the content of the score itself. The factor scores we estimated might measure a latent variable different from the one we sought to measure. We addressed this possibility through extensive evaluation of construct validity. As illustrated above, those efforts were strongly supportive of our conceptualization of the factor scores as a measure of hospital quality.

Future Opportunities

Like healthcare delivery itself, quality measurement warrants continuous improvement. Among the opportunities we recognize to improve this methodology, those that stand out include: the incorporation of outpatient claims data, particularly for patient populations who may be treated in either inpatient or outpatient settings (such as total joint replacement patients); analysis of additional procedure and conditions, both to provide decision support to more patients and to account for innovations in care delivery (such as the adoption of transcatheter approaches to aortic valve replacement); and the development of additional candidate measures, including a larger portfolio of risk-adjusted outcome measures and additional measures of process, appropriateness and value.

Best Regional Hospitals

U.S. News first published Best Regional Hospitals in 2011 as a way to offer information on community hospitals that are highly rated but may not be nationally ranked. A Best Regional Hospital is a hospital that offers a full range of services (as opposed to a specialty hospital) and that either was nationally ranked in one of the 12 data-driven Best Hospitals specialties or had three or more ratings of high performing in the nine Best Hospitals procedures and conditions. This year, high performing recognitions in the specialties were not counted toward the required minimum of

three; in our view, a hospital must perform at a high level in a variety of common procedures and conditions in order to warrant recognition as one of the best hospitals in its state, metro area or region.

Geographical definitions

In a given region (state or major metro area), a hospital on the Best Hospitals Honor Roll outranked all other hospitals that were not on the Honor Roll regardless of point totals. Other hospitals located in each region were ranked according to the number of points they earned: Hospitals earned two points for each of the 12 data-driven Best Hospitals specialties in which they were nationally ranked and one point for each specialty and each of the nine procedures and conditions in which they were rated high performing. In addition, they lost one point for each procedure or condition in which they were rated below average.

Regional rankings are displayed for every state and for the 100 metro areas with the largest populations in the 2010 census, provided there is at least one Best Regional Hospital located in the state or metro area. In 2019-20, 569 hospitals were recognized as Best Regional Hospitals. Several states had no Best Regional Hospitals. In all, hospitals were ranked in 89 metro areas.

U.S. News departed from the U.S. Census Bureau list of Metropolitan Statistical Areas in three cases by using larger Combined Statistical Areas to include nearby smaller cities with nationally ranked hospitals. The three CSAs are Detroit (by adding Ann Arbor); Raleigh-Cary, North Carolina (adding Durham and Chapel Hill and renaming the expanded area Raleigh-Durham); and Salt Lake City (adding Ogden).

Some metropolitan areas, such as Cincinnati and New York, cross state lines. That is also true for Washington, D.C., which was included as a metro area but not a state. Rankings were not published for U.S. territories.

U.S. News has grouped counties and county equivalents like parishes into approximately 200 regions that reflect geography, local customs, and regional health care markets. Best Regional Hospitals were recognized but not numerically ranked in regions that are not major metro areas.

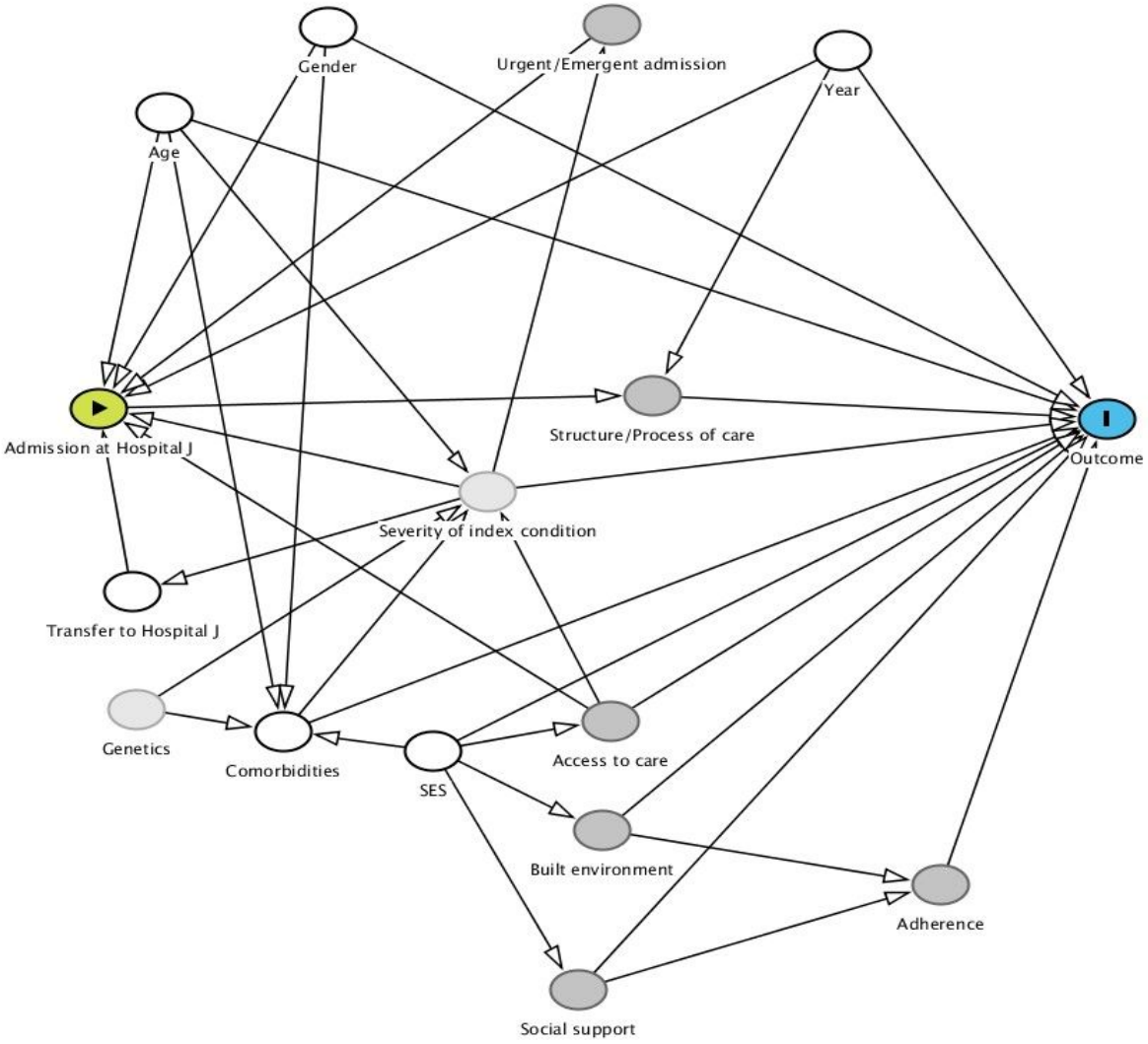
Appendix B: Discharge to a location other than home

The denominator for this measure includes only patients who have been discharged following a visit qualifying as an index visit in one of the nine Procedures and Conditions cohorts. Discharge status codes of 07 (left against medical advice or discontinued care), 20 (expired, did not recover - Christian Science), 30 (still a patient), 40 (expired at home, hospice claim), 41 (expired in facility, hospice claim), or 42 (expired place unknown, hospice claim) are excluded, as are visits with a missing or invalid discharge status code.

Discharge to a location other than home is indicated by one of the following patient discharge status codes: 0, 02, 03, 04, 05, 08, 09, 21, 43, 50, 51, 61, 62, 63, 64, 65, 66, 69, 70, 71, 72, 81, 82, 83, 84,85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95.

Appendix C: Causal model for risk adjustment

The following directed acyclic graph[20] shows the hypothesized relationship between covariates, hospital selection and outcomes.



[20] Johannes Textor, Juliane Hardt, and Sven Knuppel. Dagitty: A graphical tool for analyzing causal diagrams. Epidemiology, 22(5):745, 2011.