CENTER FOR POLICY RESEARCH
THE MAXWELL SCHOOL
WORKING PAPER SERIES

So Many Hospitals, So Little Information: How Hospital Value Based Purchasing is a Game of Chance

Andrew I. Friedson, William C. Horrace, and Allison F. Marier

Paper No. 192
January 2018 (Revised from August 2016)

ISSN: 1525-3066

Maxwell Syracuse University | Center for Policy Research
426 Eggers Hall
Syracuse University
Syracuse, NY 13244-1020
(315) 443-3114/ email: ctrpol@syr.edu
http://www.maxwell.syr.edu/CPR_Working_Papers.aspx

CENTER FOR POLICY RESEARCH –Spring 2018

Leonard M. Lopoo, Director
Professor of Public Administration and International Affairs (PAIA)

Associate Directors

Margaret Austin
Associate Director, Budget and Administration

John Yinger
Trustee Professor of Economics (ECON) and Public Administration and International Affairs (PAIA)
Associate Director, Metropolitan Studies

SENIOR RESEARCH ASSOCIATES

Badi Baltagi, ECON
Robert Bifulco, PAIA
Leonard Burman, PAIA
Thomas Dennison, PAIA
Alfonso Flores-Lagunes**,** ECON
Sarah Hamersma, PAIA
Madonna Harrington Meyer, SOC
Colleen Heflin, PAIA
William Horrace, ECON
Yilin Hou, PAIA
Hugo Jales, ECON
Jeffrey Kubik, ECON
Yoonseok Lee, ECON
Amy Lutz, SOC
Yingyi Ma, SOC
Katherine Michelmore, PAIA
Jerry Miner, ECON
Shannon Monnat, SOC
Jan Ondrich, ECON
David Popp, PAIA
Stuart Rosenthal, ECON
Michah Rothbart, PAIA
Rebecca Schewe, SOC
Amy Ellen Schwartz, PAIA/ECON
Saba Siddiki, PAIA
Perry Singleton, ECON
Yulong Wang, ECON
Michael Wasylenko, ECON
Peter Wilcoxen, PAIA

GRADUATE ASSOCIATES

Rachel Baker, PAIA
Ziqiao Chen, PAIA
Yoon Jung Choi, PAIA
Stephanie Coffey, ECON
Henry Dyer Cruzado, PAIA
Wancong Fu, ECON
Jordana Gilman, PAIA
Emily Gutierrez, PAIA
Jeehee Han, PAIA
Boqian Jiang, ECON
Hyunseok Jung, ECON
Yusun Kim, PAIA
Hyoung Kwon, PAIA
Carolee Lantigua, PAIA
Eunice Lee, PAIA
Ling Li, ECON
Joyce Lin, PAIA
Ibrahim Malik, PAIA
Christopher Markwood, PAIA
Jindong Pang, ECON
Christopher Rick, PAIA
Laura Rodriguez-Ortiz, PAIA
Fabio Rueda de Vivero, ECON
David Schwegman, PAIA
Shulin Shen, ECON
Iuliia Shybalkina, PAIA
Mackenzie Theobald, PAIA
Saied Toossi, PAIA
Rebecca Wang, SOC
Victoria Wright, PAIA
Yimin Yi, ECON
Bo Zheng, PAIA

STAFF

Joanna Bailey, Research Associate
Joseph Boskovski, Maxwell X Lab
Emily Henrich, Administrative Assistant
Kathleen Nasto, Administrative Assistant
Candi Patterson, Computer Consultant
Laura Walsh, Administrative Assistant
Katrina Wingle, Administrative Specialist

Abstract

As part of the Patient Protection and Affordable Care Act, participating Medicare hospitals have part of their Medicare reimbursements withheld and then redistributed based on quality performance. The Hospital Value Based Purchasing reimbursement plan relies partly on ordinal rankings of hospitals to determine how money is distributed. We analyze the quality metric distributions used for payment and show that there is not enough information to reliably differentiate hospitals from one another near the payment cutoffs; and conclude that a large part of the payment formula is driven by sampling variability rather than true quality information. Alternative reimbursement plans are developed.

JEL No. H51, I18

Keywords: Pay-for-Performance, Hospital Value Based Purchasing, Hospital Quality Scores, Ordinal Ranking, Indistinguishability

Authors: Andrew I. Friedson, Department of Economics, University of Colorado Denver, Andrew.Friedson@ucdenver.edu (303) 315-2038; William C. Horrace, Center for Policy Research, Department of Economics, Syracuse University, whorrace@maxwell.syr.edu (315) 443-9061; Allison F. Marier, Department of Economics, Syracuse University, Allison.marier@gmail.com

Acknowledgements

The authors are grateful for comments from Chloe East, Richard Lindrooth, Edward Norton, and Andrew Ryan, as well as by seminar participants at ASHEcon, iHEA, the Colorado School of Public Health, the University of California Santa Barbara and the University of Colorado Boulder/Denver Applied Economics Workshop. We would like to thank George Jacobs and Madia Parker Smith for excellent research assistance.

I. Introduction

As part of the Patient Protection and Affordable Care Act of 2010, the United States Agency for Healthcare Research and Quality (AHRQ) formalized its commitment to improve the quality of care and U.S. population health by publishing the National Quality Strategy (2011). Among other mechanisms to achieve the quality improvement goal, a value-based purchasing plan was developed that reimburses hospitals based on their performance in certain quality metrics.[1] The Hospital Value Based Purchasing (HVBP) program is implemented by the Centers for Medicare and Medicaid Services (CMS); CMS redistributes a percentage of the total funds designated for Medicare reimbursement to hospitals based on a Total Performance Score (TPS) composed of a weighted average of scores on individual quality metrics. There are many metrics used in a given year to calculate the TPS, such as patient survival rates after discharge and rates of timely delivery of appropriate intervention.[2] The ultimate goal of the program is to provide financial incentives for hospitals to improve quality via pay-for-performance.[3]

Pay-for-performance is a concept that has been historically applied within firms,[4] and more recently been applied by governments to incentivize high quality care by health care providers. The literature on pay-for-performance for health care finds modest to no quality gains after implementing

[1] AHRQ, along with additional stakeholders, chose the metrics to address potential gaps in patient care and coordination that lead to unintended and costly adverse patient outcomes. These goals are laid out in the 2011 National Quality Strategy report to congress, which can be found at http://www.ahrq.gov/workingforquality/reports/annual-reports/nqs2011annlrpt.htm

[2] Many of the metrics had been continuously published since 2005 by CMS on their Hospital Compare website, which reports on hospital performance and outcomes.

[3] CMS reports the payment adjustment factor (a percent by which payments are adjusted) for each individual hospital, but only reports actual payment amounts by counts in bins, which are top coded at incentive payments of $1,000,000. Using the minimum value in each bin yields a conservative underestimate of total incentive payments of 1.2 billion dollars in 2015 for the program. These data can be found at https://www.medicare.gov/hospitalcompare/data/value-based-incentive.html

[4] See Prendergast (1999) for a review.

pay-for-performance schemes.[5] Studies of the impact of HVBP and its pilot programs show that the implementation of the programs themselves had little direct impact on quality scores (Ryan, Blustein, and Casalino 2012; Ryan, Sutton and Doran 2014; Ryan et al. 2015; Ryan et al. 2017). The HVBP program has also been shown to generate changes in hospital behavior on margins with the greatest cash returns (Norton et al. 2016), as opposed to margins with the greatest quality benefits.[6] The above findings together suggest that HVBP changes hospital resource allocations, but does not change quality enough to justify the internal resource reallocations found in Norton et al. (2016). Indeed, the previous results in the literature can read as HVBP reassigning resources to the "flat of the curve" where increases in marginal spending do not yield any quality improvements.

While the literature suggests that HVBP has little to no effect on hospital quality performance, this study demonstrates why this may be the case - a lack of useful information in the quality metrics - and recommends ways to improve the program. To do this, let us fix ideas. Along any quality metric, a given hospital has some unobservable latent quality level that each metric is attempting to quantify. Measures are usually rates at which some prescribed medical procedure is correctly followed (a measure of hospital input quality) or at which treated patients survive after discharge (a measure of hospital output quality). For our purposes, we think of a latent quality level as constant in a given period for each

[5] When pay-for-performance does cause quality improvements, they are small in magnitude and only impact the process of care, not health outcomes (Eijkenaar et al. 2013). There is also variation in responses to pay-for-performance based on hospital, physician and patient demographics (Markovitz and Ryan 2016). Additional evidence shows that pay-for-performance changes how hospitals structure their internal incentives, perhaps in an attempt to meet quality goals (Damberg et al. 2009), as well as evidence that pay-for-performance may simply speed up quality improvements that were already happening in the absence of pay-for-performance (Werner et al. 2011).

[6] Norton et al. (2016) show that hospitals focus resources on patients that have high leverage over the marginal dollar of pay-for-performance incentive. For example, a hospital that is close to a quality cutoff influenced by a given patient will on average focus more resources on that patient than a hospital that has a lower chance of attaining that cutoff.

hospital in the universe of participating hospitals (as many at 4,000 hospitals for the most common measures).

For example, one important outcome measure for assessing quality is the 30-day survival rate post-discharge for patients treated for pneumonia. The formula for the TPS largely depends on a hospital's ranking within the sampling distribution of metrics such as 30-day survival rates across all hospitals. HVBP uses sampling techniques (looks at observed patient outcomes) to estimate the latent survival rates for each hospital, and compares the magnitudes of the estimates in the form of an order statistic over all participating hospitals. The program ranks approximately 4,000 hospitals based on their estimated 30-day survival rates for pneumonia, and hospitals with rates less than 88% (a cutoff that at the time of evaluation was pre-determined) receive zero points for this measure in one component of their TPS. Other measures include, for example, the 30-day survival rate for heart failure or the rate at which a given drug is properly administered. Ultimately, our empirical foci are the aforementioned survival rates due to the properties of their data generation process.

A concern that arises when using a pay-for-performance scheme such as HVBP is whether the sampling distributions of the estimates have enough information in them to differentiate one hospital's quality estimate from another's in a statistically meaningful way. In other words, HVBP might compare the estimated pneumonia survival rates of 4,000 hospitals, but ignoring the sampling variances of the 4,000 distributions could lead to false inferences on the relative quality of the ranked hospital. In this regard, the quality comparisons may be statistically meaningless. If the quality comparisons are statistically meaningless, then the assignment of points to the overall quality score of a given hospital is meaningless, causing the overall assessment to be unreliable and (perhaps) simply random.

This inability to make statistically meaningful comparisons will cause the policy to fail at its goal of paying based on quality (regardless of whether pay-for-performance works in general), as funds will be redistributed based on noise rather than the true latent quality of the hospitals. Incomplete or low information causes HVBP to become a "pay-randomly" scheme rather than a quality based payment scheme. A lack of distinguishing information between hospitals would explain the lack of findings of quality metric improvements due to HVBP (and perhaps in other pay-for-performance schemes): the metrics used to implement HVBP may not capture any meaningful quality differences and would thus be incapable of providing any incentives for quality improvement.

A hospital's latent quality relative to others is obfuscated by three main factors. The first is that measurement of quality scores is inherently noisy. A large number of draws from each within-hospital distribution is required to shrink the uncertainty (variance) around the quality estimates to a point where comparisons between hospitals are meaningful.[7] The second factor is that the large number of hospitals adds difficulty due to the multiplicity of the implied inference. That is, we may confidently say that hospital A is better that hospital B, but it is much more difficult to say with any confidence that A is simultaneously better than hospital B and hospital C and hospital D..., even when the quality estimates are not very noisy. The final factor is that many of the quality metrics used for HVBP are improvable - which may lead to bunching near the maximum achievable score over time. If distribution-wide improvements already exist in the absence of HVBP (as shown by Ryan, Blustein, and Casalino 2012, Ryan, Sutton and Doran 2014, and Ryan et al. 2015; Ryan et al. 2017), then increased bunching of the distribution of hospital scores without a corresponding improvement in the precision of the individual

[7] A hospital's true performance is also subject to noise that can be a function of statistical risk-adjustments (Dimick, Staiger and Birkmeyer 2010, Mathematica 2012), human data entry error, or additional human errors in the data management process (Bowman 2013).

scores (i.e. no increases in the size of the within hospital samples) will make it more difficult to statistically differentiate hospitals.[8]

We examine how well HVBP performs at differentiating hospitals given the quality data that is fed into the formula. We use multivariate inference techniques to simultaneously test whether individual hospitals are statistically different from all other hospitals in the sample, paying particular attention to areas around cutoffs for payment under the current HVBP program. The exercises to follow can be seen as a policy evaluation answering the question: how well does HVBP achieve its goal of paying based on true quality? We draw on techniques for multiple comparisons within ordinal rankings (Dunnet 1955) to create groupings in which with a fixed probability, all members of the group are indistinguishable from the relevant payment cutoff. These groupings have the benefit of appropriately accounting for multiplicity in inference. By creating groups that are indistinguishable from the payment cutoff, we are able to say what proportion of the total number of hospitals are essentially subject to a lottery with respect to that payment point.[9] We refer to these groupings as "lottery zones."

Our technique provides a new contribution to the analysis of noise present in quality score distributions used in a policy context. Previous work on this topic such as Chay, McEwan, and Urquiola (2005), and Kane and Staiger (2002) have focused on decomposing how much of an agent's (in our case a hospital, and in their case a school) score is true signal and how much is noise. Their techniques focus on this decomposition at the agent level, whereas our technique can be seen as an analogue for the entire

[8] The improvement over time for the quality estimates is at least partially attributable to public reporting initiatives such as pay-for-reporting (Lindenauer et al. 2014), although the connection between score improvement and public reporting initiatives has been shown to be both limited in scope and modest in size (Ryan, Nallamothu, and Dimick 2012).

[9] We do not suggest that every hospital in the lottery has the exact same probability of meeting a payment cutoff; instead, hospitals subject to the lottery are assigned to a side of the cutoff by a random process.

sampling distribution across agents. Instead of saying how much of an agent's score is due to random assignment of a shock, we provide a measure of how much of a distribution's ranking cannot be differentiated from statistical noise. We are limited because we do not recover a signal to noise ratio for each agent, but have the benefit of needing far less information to operationalize our method while still providing a single policy-relevant estimate. We only need the individual scores that make up an across-individual distribution and their respective errors in a cross-section, whereas previous methods need an individual level panel, or a set of individual level characteristics.[10] Our methods provide a new and useful tool in the policy analysis toolbox for pay-for-performance, an area of current relevance for healthcare as well as other sectors, notably education.

We find that the lottery zones around cutoffs for payment under HVBP are large, in most cases capturing over 60 percent of the hospitals submitting data for the metrics that make up the largest part of the payment formula. This demonstrates that the current HVBP program does a poor job paying based on true performance, as the noise in the sampling distributions makes it impossible to detect most of the true differences in performance. We show that changes in the payment formulae in recent years that phased out older metrics and phased in newer metrics have actually increased the amount of noise in the system, making the problem progressively worse. We also demonstrate that the lottery zones cover a larger percentage of hospitals in regions of the quality distribution where the HVBP formula attempts to make the finest quality distinctions: the points assigned on the intensive margin are indistinguishable from random for the metrics we examine. Lastly, we suggest an alternative, data driven approach for

[10] We do not observe the errors in this case, but are able to recover them from the within hospital sample sizes given the Bernoulli nature of the data generation process.

generating payment cutoffs based on the limited information that is present in the relevant distributions, making broad categories of hospital quality that reflect differences that can be statistically detected.

The implications from our analysis are twofold. From a policy perspective, if the intent of HVBP is to incentivize high effectiveness practices within hospitals, then HVBP fails as a policy. True quality is largely ignored under HVBP, and payment appears to be mostly based on shocks. Second, as hospitals in the lottery zone are assigned plausibly random payments; this source of payment variation could be of use to future researchers who wish to learn the impacts of additional dollars of federal funding on hospital behavior.

II. HVBP Programmatic Details

Under HVBP, hospitals that are eligible for the program receive payment based on their scores on select measures from the Hospital Compare data.[11] The payment formula uses a points system: hospitals earn points based on their performance for each quality metric. A hospital can score up to 9 points based on improvement over their old values, or can score up to 10 points based on their placement within the overall distribution across hospitals for a given metric. These scores are referred to as *improvement* and *achievement* scores respectively. The greater of the two values is then used as the hospital's point value for that quality metric.

The final payment depends on the Total Performance Score (TPS), which is calculated as a weighted sum of point values from groupings of metrics, referred to as domains. The domains as of 2017

[11] The entirety of the HVBP payment plan can be found in the Federal Register (Centers for Medicare and Medicaid Services 2011). Eligible hospitals are those that are paid via the prospective payment system, serve a minimum number of patients, do not have payment reductions from the Inpatient Quality Reporting program, and have not been cited for deficiencies that may jeopardize to patient health or safety.

are: Clinical Process of Care, Patient Experience of Care, Efficiency and Outcomes.[12] The relative weighting of the domains in the TPS changes from year to year, so in different years different sets of metrics will have greater or less importance relative to one another. In 2015, for example, the TPS was calculated as:

(1) $TPS_{2015} = 0.45 \times$ Clinical Process of Care Score $+ 0.30 \times$ Patient Experience of Care Score $+ 0.25 \times$ Outcomes Score

and in 2016:

(2) $TPS_{2016} = 0.10 \times$ Clinical Process of Care Score $+ 0.25 \times$ Patient Experience of Care Score $+ 0.25 \times$ Efficiency Score $+ 0.40 \times$ Outcomes Score

CMS funds the incentive payments via reallocation of existing Medicare reimbursement to hospitals. A fixed percentage (1 percent in 2013, ramping up by 0.25 percent a year until it reaches 2 percent in 2017 where it remains onwards) of Medicare reimbursements to hospitals are withheld during the year and reallocated based on the TPS of the hospitals. Data from two time-periods are used to calculate payments for each payment year. Data are drawn from an earlier *baseline* period to set goalposts for point allocation, and then additional data are drawn from a later *performance* to score points. Table 1 describes the relevant time-periods used for calculating payments.[13]

[12] These domains are still evolving, and CMS will apply different weights to each domain in order to calculate the TPS in the coming years.

[13] For the purposes of this study we will assume that during the performance period, the goalposts set by the baseline period data are exogenous as hospitals at that point in time have no way to influence them.

A. How Hospitals Score Points

Each hospital scores points based on their metrics during the performance period. Data-driven cutoffs from the baseline period determine how points are allocated in the performance period. Two relevant values are calculated from the baseline period: the *threshold*, or the minimum value needed to score a single point for a given metric, and the *benchmark*, or the minimum value needed to score the maximum number of points for a given metric.

Points for improvement or achievement are allocated based on cutoffs set in uniform intervals between the threshold and the benchmark. For example, if the achievement threshold is at a value of 60 out of 100 for a metric, and the benchmark is at 100 out of 100, then a hospital with a value of 80 in the performance period would receive 5 out of the 10 possible points for achievement on that metric.

The achievement score threshold is set at the median of all hospitals' performance during the baseline period. The benchmark is set at the 95th percentile of all hospitals' performance during the baseline period. If a hospital that has a value for a metric during the performance period that exceeds the 95th percentile of that overall distribution of scores during the baseline period would receive 10 points for achievement.

For improvement scores, the threshold is set at an individual level. The threshold is each hospital's score on the relevant metric during the baseline period. The benchmark is the 95th percentile of all hospital's performance during the baseline period, which is the same as for the achievement score. Although a maximum of 9 points can be earned for improvement, a hospital that earns 9 points for improvement also earns 10 points for achievement, and is awarded the higher of the two scores (which would be 10 out of 10) for that specific metric.

III. Data

The HVBP data obtains information for each hospital from their CMS records. For most metrics, scores are calculated by dividing the number of patients who had a specific event occur by the number of patients who "qualified" to have that event occur. Patients are included as qualifying if they meet a level of appropriateness (i.e. they have a relevant diagnosis and if the specific event is a treatment, that treatment is not counter-indicated). Visits are also deemed as qualifying or not as a way to homogenize the patient base studied across hospitals. For example, patients that present with symptoms that are far more severe than average will not been deemed qualifying to avoid penalizing hospitals in their quality scores for taking on sicker patients. The metric value for each hospital is reported, along with the number of qualifying patients used to calculate the value.

The main component of our analysis focuses on hospital scores in the "Outcomes" domain: specifically we look at three scores that are the 30-day survival rate after discharge for patients discharged after being admitted with one of three serious diagnoses: acute myocardial infarction (AMI), heart failure, and pneumonia. The score for each hospital is the number of qualifying patients that survived at least 30 days past discharge divided by the total number of qualifying patients discharged. We focus on these metrics for two reasons. The first is that the Outcomes domain makes up the largest part of the payment formula at present, and is scheduled to make up the largest part of the payment formula for future years. The second is that these scores have statistical properties that make our methods tractable: the data generation process for these metrics allows us to recover standard errors when we only observe the value and the sample size.

From a statistical point of view, each observation used to calculate a metric in the Outcomes domain is a draw from a within hospital Bernoulli distribution. Either the patient survives or does not

survive. Each hospital has a true latent survival probability for the patient pool, and the eventual metric is the usual consistent estimator of that true probability. When ranking hospitals, making a claim that one hospital is ranked higher on a given metric than another is a statement about their estimate values relative to one another. Whether those hospitals are distinguishable from one another in a statistical sense depends upon their estimate values and on their standard errors, which given the Bernoulli structure of the underlying data generation process are a function of the reported score and of the number of observations per hospital.[14] Table 2 reports descriptions, mean values and number of patients for the Outcomes metrics used in the 2015 payment year: the most recent year that CMS has made the data available for this set of scores.

To flesh out the story of how well HVBP has performed in the past we also look at scores in the Clinical Process of Care domain. These scores measure how often patients with particular diagnoses receive a recommended intervention in a timely manner. For instance, the metric AMI 8a reports whether or not a patient who arrives with a diagnosis of acute myocardial infarction (AMI) had a stent placed within 90 minutes of arrival. The Clinical Process of Care domain is of interest because it too follows a Bernoulli data generation process, and because it was the largest part of the HVBP formula when the program began. Clinical Process of Care has in the past made up as much of 70 percent of the payment formula, even though it accounts for only 5 percent of the payment formula in 2017. Appendix Table 1 reports descriptions, mean values, and the number of patients for the Clinical Process of Care metrics for 2015 and 2016, the two most recent years with available data.

[14] The analysis to follow does not require an underlying Bernoulli distribution to operationalize, but does require sampling standard errors. As standard errors are not reported by Hospital Compare but sample sizes are, we are able to proceed by relying on the Bernoulli nature of the data generation process to provide the errors.

IV. Methods

To determine the amount of distinguishing information within the distribution across hospitals for a given quality metric, we focus on how distinguishable hospitals' values of the metric are from the cutoff values for payment under HVBP. We estimate what percentage of the overall distribution for a given metric in a given performance period is indistinguishable from the threshold and benchmark payment values for the achievement score. The goal of the exercise is to show what proportion of the distribution has its payment points determined by a process indistinguishable from random assignment; we refer to this grouping as a "lottery zone." If a large proportion of the hospitals fall into the lottery zone, then HVBP is a system that does a poor job of separating true quality differences from noise for the distribution across hospitals with respect to the relevant point in the payment formula.

We focus on achievement scores because they offer a single value for the threshold for all hospitals whereas improvement scores do not. Improvement scores compare hospitals to their past selves, and use a different threshold value for each hospital. Fortunately, a vast majority of hospitals rely of the achievement score rather than the improvement score to determine their point allocation. This is demonstrated in Table 3, which shows that for every metric in the Outcomes domain, achievement scores are the relevant source of points over 75 percent of the time.[15] That said, as the benchmark is the same for achievement and improvement scores, analysis relevant for achievement benchmarks applies to improvement benchmarks as well.

[15] A similar pattern can be seen for the Clinical Process of Care Domain in Appendix Table 2. For this domain, achievement scores are the relevant score for point assignment over 70 percent of the time for all metrics.

A. Creating Lottery Zones

Let $p_i \in (0,1)\ \ i = 1, \ldots n$ be the mean parameter for *n independent* Bernoulli populations. Also, let $p_l^* \in (0,1)$ be a fixed *lower threshold*, such that it is on the closed unit interval [0,1]. Similarly, define a fixed *upper benchmark*, $p_u^* \in (0,1)$, such that $p_l^* < p_u^*$. In what follows, we focus on the lower threshold p_l^*, but everything can be adapted for the upper benchmark, p_u^*, with no modification. Interest centers on estimating each p_i from a sample of events for each hospital $i = 1, \ldots n$ and determining (in a statistical sense) a subset of populations (hospitals) in a neighborhood around the threshold that fall closest to the cutoff.

Let $x_{it}\ t = 1, \ldots, T_i$ be a random sample of size T_i from independent Bernoulli p_i populations, $i = 1, \ldots, n$. Define the usual consistent (as $T_i \to \infty$) estimators:

$$\hat{p}_i = T_i^{-1} \sum_{i=1}^{T_i} x_{it} \tag{3}$$

$$s_i^2 = T_i^{-1}\, \hat{p}_i\, (1 - \hat{p}_i) \tag{4}$$

In what follows, we assume that the sampling distribution of $\hat{p}_i$ is normal (or asymptotically so). It is important to stress that our focus is not on inference for the underlying Bernoulli population but for the sampling distributions of the statistics $\hat{p}_i$, which for large T_i may be treated as normal for the purpose of statistical inference. Therefore, if we are willing to assume normality for inference (or at least asymptotic normality), then the underlying population can be from *any* non-degenerate family of distributions with finite mean and variance.

Let the set of all hospital indices be $N = \{1, \ldots, n\}$. We consider simultaneous confidence intervals for all p_i:

$$(5) \qquad p_i \in [L_i, U_i], i \in N,$$

where $L_i = \hat{p}_i - d_\alpha s,\ U_i = \hat{p}_i + d_\alpha s, d_\alpha$ is an appropriately chosen critical value such that:

$$(6) \qquad \Pr\{\, p_i \in [L_i, U_i], i \in N\} \geq 1 - \alpha,$$

and $\alpha < 0.50$, is a pre-selected error rate.[16] That is, we would like to construct simultaneous confidence intervals, $[L_i, U_i]$, for all p_i with a confidence level of at least $(1 - \alpha) \times 100\%$. Then, determining the lottery zone around the threshold p_l^* (or the benchmark p_u^*), is simply a matter creating a subset of hospital indices based on the selection rule:

$$(7) \qquad S_{l,\alpha} = \{k: p_l^* \in [L_k, U_k]\ \forall\ k \in N\} \subseteq N.$$

Then, the hospital indices in the subset $S_{l,\alpha}$ or $S_{u,\alpha}$ have estimated scores, $\hat{p}_i$, that are statistically indistinguishable from p_l^* or p_u^*, respectively. Therefore, the hospital indices s in each subset form a lottery around each cutoff at the pre-specified error rate α.

All the difficulty in calculating (7) is determining the critical value d_α in (5). Following standard multiple comparison procedures (e.g., Horrace and Schmidt, 2000), exact $(1 - \alpha) \times 100\%$ confidence intervals can be found by letting $d_\alpha = z_{\alpha,n}$, a two-sided critical value from an n dimensional standard normal distribution, such that $\Pr(max_{1 \leq i \leq n} |z_i| \leq z_{\alpha,n})$. See Horrace and Schmidt (2000) for a precise definition of the critical values, which are easily simulated (Horrace, 1998). The simulation algorithm provides intuition for the procedure.

[16] As the usual error rates used are $\alpha = 0.1$ or less, this restriction is usually non-binding.

1. Draw independent standard normal random variables $z_i, i = 1, \ldots, n$.

2. Scale each draw *i* by the individually estimated s_i, respectively.

3. Find $y = \max |z_i|$.

4. Perform steps 1-3 many times.

5. Then $d = z_{\alpha,n}$ is the $(1-\alpha) \times 100$ percentile from the sample distribution of the *y*.

Clearly, the 'max' operator in step 3 ensures the coverage probability for the multivariate confidence intervals in (5). When n =1 and $\alpha = 0.05$, the simulated critical value will be the usual univariate $z_{0.05,1} = 1.96$. The critical value is increasing in *n* and decreasing in α. Also, the cardinality of $S_{l,\alpha}$ is increasing in α and in the extent to which the estimated $\hat{p}_i$ are bunched (tightly distributed) around the cut-off. For our empirical analysis we let $n_{l,\alpha} \leq n$ be the cardinality of $S_{l,\alpha}$, so that $n_{l,\alpha}/n$ is the lottery share of the hospitals. The confidence intervals in (5) are easily adapted from the Multiple Comparisons with a Control (MCC) procedure of Dunnett (1955), which we now explain.

B. Relation to the MCC procedure of Dunnett (1955)

Our procedure is a special case of the MCC of Dunnett (1955). Let $k \in N$ be the index of a pre-specified control, then $(1-\alpha) \times 100\%$ multiple comparisons with a control (MCC) confidence intervals of Dunnett (1955) are:

$$(8) \qquad \left[L_i^k, U_i^k\right], i \neq k, i \in N$$

$$(9) \qquad L_i^k = \hat{p}_k - \hat{p}_i - z_{k,\alpha,n}(s_k^2 + s_i^2)^{\frac{1}{2}}$$

$$(10) \qquad U_i^k = \hat{p}_k - \hat{p}_i + z_{k,\alpha,n}(s_k^2 + s_i^2)^{\frac{1}{2}},$$

where $z_{k,\alpha,n}$ is a two-sided critical value from an $n-1$ dimensional standard normal distribution such that $\Pr(max_{1\leq i\leq n-1}|z_i| \leq z_{k,\alpha,n-1})$. Notice the similarities between the intervals in (5) and in (8). In fact, the *n* confidence intervals in (5) can be adapted from the $n-1$ intervals in (8), but with an n^{th} component added as a non-random control: $\hat{p}_k = p_l^*$, $s_k^2 = 0$ and $n = n+1$. In other words, any standard computer-program that can calculate the MCC intervals of (8) can be used to calculate those of (5) and conduct the inference. Additionally, (8) can be used in place of (5) even if the threshold is considered random ($p_l^* = \hat{p}_k$, with, for example, $s_k^2 > 0$). In other words, our selection rule for determining the lottery around the cutoff is a special case of MCC (but with a non-random control), which has been extensively studied (e.g., Dunnett, 1955), so the coverage probability in (6) is assured. We can then use these subsets to conduct evaluation of the policy objective of paying hospitals based on latent quality rather than based on noise.

Another useful application of the above tools is to empirically determine threshold and benchmark values that have better properties in terms of distinguishing hospitals based on quality. We will ultimately use the confidence intervals in (8) to estimate this class of cutoffs using the subset selection procedure due to Gupta (1956, 1965). This procedure is adaptable from the MCC intervals in (8), so they are important in the sequel. Let the ranked population parameters be $p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[n]}$. Then Gupta (1956, 1965) defines the "subset of the best" (the subset of population indices with the largest parameters) as:

$$(11) \quad S_\alpha^* = \left\{k: U_i^k > 0 \; \forall \; i \in N\right\} \subseteq N.$$

That is, our decision rule is to select populations that have all positive MCC upper bounds. The selected populations will simultaneously dominate all other populations in term of their outcome scores. Then, the selected subset based on the decision rule satisfies the probability:

$$(12) \quad Pr\{[n] \in S_\alpha^*\} \geq 1 - \alpha.$$

That is, an implication of the MCC intervals in (8) is that we can also make the probability statement in (12) to identify a subset of populations that contains the best population with probability at least $1 - \alpha$. In other words, our ability to identify the indices of the best hospitals is confounded by the statistical noise in the sample rank statistic, $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \cdots \leq \hat{p}_{(n)}$, and it limits our ability to infer the extent to which (*n*) in the sample may equal [*n*] in the population. Similar to the lottery subsets $S_{l,\alpha}$ and $S_{u,\alpha}$, which identify hospitals that are indistinguishable from a predetermined threshold and benchmark (respectively), the subset of the best (S_α^*) identifies hospitals that are indistinguishable from an unknown best hospital at a pre-specified error rate, α. Later, we will propose an alternative method for estimating a threshold and benchmark that better fits the policy objective of paying based on detectible quality. We do this by using these well-known inferential results.

V. Results

Results for the Outcomes metrics are presented in Table 4. The first panel of the table (the first two columns) contains the metric under consideration and the relevant payment year, which for the

Outcomes metrics are always 2015.[17] For example, in the first row the metric is "30-Day AMI" (the 30-Day mortality rate for an acute myocardial infarction discharge) as used for payment year 2015 (therefore the metrics were measured in 2013 as laid out in Table 1). The next panel of Table 4 (the 3rd, - 5th columns), contains information on the entire sample of hospitals for that metric in that year. The 3rd column has the total number of hospitals that reported scores, the 4th column has maximum and minimum values of $\hat{p}_i$ across all *n* hospitals in the set *N*, and the 5th column has the average value of the simulated critical values over all *n* hospitals in *N*. For example, in the first row of the table, for the metric 30-Day AMI, there were *n* = 2,502 hospitals reporting data for payment year 2015. The maximal value of the score is 0.9915 and the minimal value is 0.3000. Then, we simulated multivariate critical values to perform the analysis, one for each measure. Notice that down 3rd column in the table as *n* increases, so does the critical value. That is, a large number of hospitals means more (multiple) comparisons to consider, leading to larger critical values and less sharp inference. This demonstrates how the inference controls for multiplicity in the probability statement of (6).

The next panel of the Table 4 (columns 6-9) contains our results for the subset of hospitals in the neighborhood of the threshold p_l^*. The sixth column contains the value of threshold for each metric (which is set by CMS based on the previous performance of the hospitals), the seventh contains the cardinality of the subset in the neighborhood of the threshold $(n_{l,\alpha})$, the eighth contains that extreme values of $\hat{p}_i$ in the lottery zone, and the ninth contains the lottery share of hospitals$(n_{l,\alpha}/n)$. Again, the sharpness of the inference is decreasing in the cardinality and the lottery share of hospitals. In all cases in the table, the inference around the threshold is not sharp at all. For example, for metric 30-Day AMI, we

[17] For simplicity, we limited our analyses to metrics that had performance periods corresponding to calendar years, and in which the data was readily available for public use on the CMS website.

see that there are $n_{l,\alpha} = 1{,}549$ in the lottery, so the lottery share around the threshold of 0.84747 at the 5% error rate is 0.6191. That is, a majority of the hospitals are indistinguishable from 0.84747, and their estimated scores range from 0.4280 to 0.9302. The inference is just as poor for the 30-Day HF metric, which has a lottery share of 0.5861 around the threshold of 0.88151, and for the 30-Day PN metric, which has a share of 0.6192 in the lottery zone around the threshold of 0.88165. What is most notable about the analysis, is that even though the thresholds are fairly large across the three measures (0.84747, 0.88151, and 0.88165), the scores of the worst hospitals in the lottery are quite low (0.4280, 0.46000, and 0.4577, respectively) relatively speaking.

The final panel of the table (columns 10-13) contains results for the benchmark analysis. The tenth column contains the value of benchmark for each metric, the eleventh contains the cardinality of the subset in the neighborhood of the benchmark, the twelfth contains that extreme values of $\hat{p}_{u,i}$ in the lottery (after unfolding the $\hat{r}_{u,i}$), and the thirteenth contains the lottery share of hospitals. Similar to the threshold results, the benchmark inference is not very sharp. For example, for the 30-Day AMI metric, we see that there are 1,628 hospitals in the lottery, so the lottery share around the benchmark of 0.86237 is 0.6507 at the 5% error rate. Again, the majority of hospitals are in the lottery zone for all of the outcome metrics. In general, neither the threshold nor the benchmark is useful in differentiating hospital performance. There is simply too much uncertainty and multiplicity in the order statistics for the HVBP formula to allocate funds in a way that rewards true quality.

A. Clinical Process of Care

Results for Clinical Process of Care scores are reported in Appendix Table 3, which follows the same layout as Table 4. There are two patterns of note in this table. The first is that the lottery zones for

Clinical Process of Care imply less sharp inference than what is possible for Outcomes. With two exceptions, all of the metrics have lottery zones that capture over 50 percent of hospitals, and most lottery zones capture over 80 percent of hospitals. The second pattern is that generally, the inference on these metrics gets worse over time. For instance, for SCIP VTE 2 in the last two rows of the table, the lottery share goes from 0.6816 for 2015 to 0.9646 for 2016. The same is true for the metric SCIP Card 2 (which goes from a lottery share of 0.8790 to 0.9899), and for SCIP INF 9 (which goes from a lottery share of 0.5670 to 0.9192). This result is consistent with hospitals improving their scores over time, bunching together, and making it more difficult to differentiate their scores.

The lottery zone sizes for Clinical Process of Care show that inference was poor in the older payment formulas and would have worsened over time if the same metrics were kept.

B. Is Quality Differentiation Feasible with the Current Formula?

Given that the current payment formula generates large lottery zones, a relevant question is under what circumstances would the current formula be able to differentiate a large enough number of hospitals from payment cutoffs to be useful in achieving its policy goals? This could be achieved through limiting the multiplicity of inference by using a smaller number of hospitals (perhaps by generating separate regional distributions), or by having each hospital generate more data, which would mechanically shrink each hospital's standard error for their metrics.

Increasing the sample size for each hospital could potentially help sharpen inference. As hospitals report all of their eligible Medicare data, the sample size increases would need to come from other segments of the population being reported (or an overall population increase). To examine the effect of larger sample sizes, we artificially rescale the number of observations for each hospital and calculate new

lottery zones. This is analogous to assuming a fixed proportion of the population uses the hospital system in a given year and that the hospital system works through these patients without changing scale or efficiency, and then artificially increasing the population of the United States. We increase each hospital's sample size by a factor of 3.93 (analogous to using the population of India rather than that of the United States), a factor of 13.88 (inflating to the population of Asia), and a factor of 22.34 (inflating to the global population).

Results of this exercise for 30-Day AMI scores with respect to the threshold are reported in Table 5. Increasing the sample size by a factor of 3.93 (inflating the U.S. population to that of India), decreases the lottery zone from approximately 62 percent of hospitals to approximately 34 percent of hospitals. The payment formula is still subjecting about a third of the hospitals to a lottery when the amount of data is almost quadrupled. Increasing the sample size by a factor of 13.88 (inflating the U.S. population to that of Asia) still subjects 17 percent of hospitals to a lottery, and even an absurdly large increase of a factor of 22.34 (inflating the U.S. population to the global population) leaves over 13 percent of the hospitals in the lottery zone. Based on these results, it seems unlikely that any feasible increase in sample size will overcome the shortcomings of the payment formula.

VI. An Alternative Points System

Based on the above analysis, the distributions of quality metrics do not have enough information within them to allow the current HVBP payment formulae detect true quality differences between hospitals. Rather, the cutoffs, though well intentioned, appear to create arbitrary point assignments. Here we suggest an alternative point system. In broad strokes, our proposed system uses the above methods to construct groupings in which the little information that is contained within the distributions is appropriately leveraged. We make fewer distinctions between hospitals, avoiding fine point assignment.

There is not enough information in the distributions of the quality metrics to merit assigning between 0 and 10 points to hospitals for each metric. We propose a system in which a hospital can earn 0, 1 or 2 points for each metric. A hospital would receive 1 point for reaching some estimated threshold, and 2 points for reaching some estimated benchmark, with 0 points awarded as before for not reaching the threshold as before. Hence, hospitals that achieve the estimated threshold could be considered to be "among the best" and hospitals that achieve the estimated benchmark could be considered "among the best of the best." Improvement scores could keep their previous threshold value, and reward hospitals a single point if they manage to surpass their previous year's score. This proposed points system would preserve the intent of HVBP, while at the same time allocating points to hospitals based on statistically relevant quality distinctions.

We can use the entire set of hospitals to generate our alternative estimates for the threshold and the benchmark. To develop a data-driven estimate of the threshold, we construct MCC intervals in (8) allowing construction of the subset of the best hospitals in (11), satisfying the multivariate probability statement in (12). Using this subset of hospital indices, $S_\alpha^* \subseteq N$, we estimate the threshold as the lowest hospital score therein contained. Our estimated threshold is:

$$(13) \quad \hat{p}_l^* = min_{i \in S_\alpha^*} \hat{p}_i.$$

That is, the threshold estimate is pegged to the worst performing hospital index in the subset of the best. Let the cardinality of the subset of the best hospital indices be n_*. Based on our proposed points method, hospitals that do not make it into S_α^* get zero points.

To estimate the benchmark, we perform MCC on only those n_* hospitals in the subset of the best hospitals, S_α^*. That is, define

$$(14) \quad \left[L_{*i}^{k}, U_{*i}^{k}\right], i \neq k, i \in S_{\alpha}^{*}$$

$$(15) \quad L_{*i}^{k} = \hat{p}_k - \hat{p}_i - z_{k,\alpha,n_*}(s_k^2 + s_i^2)^{\frac{1}{2}}$$

$$(16) \quad U_{*i}^{k} = \hat{p}_k - \hat{p}_i + z_{k,\alpha,n_*}(s_k^2 + s_i^2)^{\frac{1}{2}},$$

Leading to the Gupta subset:

$$(17) \quad S_{\alpha}^{**} = \left\{k: U_{*i}^{k} > 0 \; \forall \; i \in S_{\alpha}^{*}\right\} \subseteq S_{\alpha}^{*} \subseteq N.$$

This is the subset of the best of the best hospitals. Hospitals in S_{α}^{**} would receive two points, and those hospitals in S_{α}^{*} but not in S_{α}^{**} would receive 1 point. Let the cardinality of S_{α}^{**} be $n_{**} \leq n_* \leq n$. Then, our estimated benchmark is:

$$(18) \quad \hat{p}_u^* = min_{i \in S_{\alpha}^{**}} \hat{p}_i.$$

This produces a sequential inference procedure, and we can use the Bonferroni inequality to bound the overall error rate of the procedure. For example, if we set $\alpha = 0.05$, then the overall error rate of the sequential procedure is 0.10.

One complication with the sequential procedure is that there may be hospitals with estimated scores that lay slightly above the estimated threshold, $\hat{p}_l^*$, yet they may not be contained in S_{α}^{*}. Even though it may lead to more conservative inference, we may want to treat these hospitals as if they are in the subset of the best hospitals. After all if the threshold is set at $\hat{p}_l^*$, a hospital may complain if it has $\hat{p}_i > \hat{p}_l^*$, but it is not considered to be among the best hospitals contained in S_{α}^{*} (and receives zero points in our scoring scheme). Therefore, define the alternative subset of hospitals indices that have scores above the estimated threshold:

$$\widetilde{N} = \{i: \hat{p}_l^* < \hat{p}_i, i = 1, \dots, n\},$$

with cardinality $\tilde{n}$. By design $S_\alpha^* \subseteq \widetilde{N}$ is so we have $\tilde{n} \geq n_*$. Then we can perform the MCC and selection in (14)-(17) on the set $\widetilde{N}$ (as opposed to on the set S_α^*) with sample size $\tilde{n}$ (as opposed to n_*). This leads to an alternative estimate of the benchmark:

$$(19) \quad \tilde{p}_u^* = min_{i \in \widetilde{N}} \hat{p}_i.$$

The subset $\widetilde{N}$ relaxes the requirement that a hospital must be in S_α^* to be part of the second step of the inference procedure. Let the cardinality of the alternative subset of the best of the best be $\tilde{n}_*$. By definition $\tilde{p}_u^* \leq \hat{p}_u^*$.

Estimated thresholds and benchmarks for Outcome metrics for fiscal year 2015 using our proposed methods are presented in Table 6. The first panel of the table consists of three columns: the outcome metric, the year (always 2015), and the total number of hospitals. The second panel (columns 4 - 8) contains the results for estimating the threshold. Column 4 contain the HVPB threshold (p_l^*), and column 5 has our estimated threshold ($\hat{p}_l^*$). In all cases our estimated threshold is much lower that the HVPB threshold. For example, compare 0.84747 to 0.7692 for the 30-Day AMI, indicating that the inference determined that the worst hospital in the subset of the best hospitals has a considerably lower score than the HVPB threshold. Under the original HVPB scheme this hospital would have received 0 points, but under our proposed scheme it would have received 1 point. Column 6 (n_*) has the number of hospitals that are in the subset of the best (S_α^*): those that are indistinguishable from the unknown best hospital in the sample at the 95% level. The 7th column ($\tilde{n}$) contains the number of hospitals that were above the estimated threshold. These hospitals are contained in the set $\widetilde{N}$. For example, for the 30-Day

AMI we have 1,622 hospital in the subset of the best, but if we include all hospitals above $\hat{p}_l^*$ = 0.7692, then that number grows to the 1,091 hospitals in $\widetilde{N}$. We can use either set ($\widetilde{N}$ or S_α^*) as the basis for our second round of inference to estimate the benchmark.

Panel three of Table 6 (columns 9-13) contains our estimated benchmark analysis. Column 9 has the HVPB benchmark (p_u^*), while column 10 has our benchmark estimate ($\hat{p}_u^*$) based on analysis of the subset of the best, S_α^*. For the 30-Day AMI, our estimate is not much lower than the HVPB benchmark. Compare our 0.8586 to 0.86237. This difference is much starker for the 30-day HF and PN measures. Continuing with the 30-Day AMI, of the 1,662 hospitals in the subset of the best, 1,360 of them were in the subset of the best of the best, and the lowest AMI value in this subset of the best of the best provides our estimated benchmark, 0.8586. Under our proposed scoring scheme, these 1,360 hospitals would receive another point in addition to the point they received for being in the subset of the best (S_α^*). For completeness column 12 indicates that there are 1,511 hospitals above the estimated benchmark ($\hat{p}_u^*$), so there are hospitals above the benchmark that did not make it into the subset of the best of the best, S_2.

Panel four of Table 6 (columns 14-17) contains our alternative estimated benchmark analysis. Column 14 has our alternative benchmark estimate ($\tilde{p}_u^*$) based on analysis of the subset of the best, $\widetilde{N}$, with consists of all hospitals above our estimated threshold. For the 30-day AMI, there were 1,901 hospitals above $\hat{p}_l^*$ (who would receive 1 point under our scheme). Of these, 1,449 hospitals were in our best of the best subset and would receive and additional point in our scheme. For completeness column 16 indicates that there are 1,675 hospitals above the estimated benchmark ($\tilde{p}_u^*$), so there are hospitals

above the benchmark that did not make it into the subset of the best hospitals in $\widetilde{N}$; this subset which only contains $\tilde{n}_* = 1{,}449$ hospitals.

VII. Discussion and Conclusions

Though the intent of HVBP is to pay hospitals based on their true latent quality, it appears that most hospitals are indistinguishable from one another on the metrics used for evaluation. In summary, CMS is effectively paying hospitals based on shocks that bump their metrics to one side or another of payment thresholds, rather than for truly distinguishable differences in quality. Although CMS "tops out" and removes metrics as their distributions collapse towards the maximum attainable values, the above analyses show that the metrics that are being used for HVBP even after the removal of "topped out" metrics still do not contain enough ordinal information in them to meet the goal of creating cash incentives for quality.

Fund redistribution may not be necessary to prompt quality improvement. Quality scores have been improving over time to reasons that appear to be unrelated to the HVBP program (Ryan, Blustein, and Casalino 2012; Ryan, Sutton and Doran 2014; Ryan et al. 2015; Ryan et al. 2017). However, a possibility exists that the program creates perverse incentives: a hospital that enacts a useful program to try to meet quality thresholds may be adversely impacted by the program's imprecision and receive a smaller payment. Similarly, a hospital that enacts a wasteful program to try to meet quality thresholds may benefit from the variability and receive a larger payment. Fortunately, there is no reason to think that this would be systematically the case. It is also possible that hospitals could receive payments that correspond with good practices. In the current world of HBVP, statistical noise dominates the true quality performance signal for most of the hospitals participating in the program. As a result, the program will not

likely generate payments that consistently reward hospitals for effective performance in administering the desired treatments.

Potentially inconsistent reimbursement for hospital quality improvement efforts is supported by Norton et al. (2016), who show that hospitals respond to the incentives presented by HVBP based on their marginal future reimbursement from a given outcome for a given patient. The calculated marginal future reimbursements demonstrate a large amount of heterogeneity across hospitals and metrics. If the cutoffs that generate these marginal future reimbursements do not divide hospitals based on statistically useful differences in quality, then the widely varying incentives imposed on hospitals found by Norton et al. (2016) can be seen as a product of statistical noise. In other words, the inability of the HBVP formula to adequately recognize true underlying quality could be creating incentives and provoking hospital behaviors that do not correspond with the government's stated goal of addressing potential gaps in patient care and coordination that lead to adverse patient outcomes.

For researchers, HVBP may be an untapped opportunity. To the extent that payments under HVBP are a random redistribution of funds to hospitals, which is especially the case for hospitals that scored between the threshold and the benchmark, HVBP offers a new identification strategy for researchers studying the impact of the marginal dollar of government transfers to hospitals on any of an array of hospital behaviors and outcomes.

References

Bowman, Sue. 2013. "Impact of Electronic Health Record Systems on Information Integrity: Quality and Safety Implications." *Perspectives in Health Information Management* 10: 1c.

Centers for Medicare and Medicaid Services. 2011. "Medicare Program: Hospital Inpatient Value-Based Purchasing Program: Final Rule" *Federal Register* 76(88): 26,490-26,547.

Chay, Kenneth Y., McEwan, Patrick J., and Miguel Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools" *The American Economic Review* 95(4): 1237-1258.

Damberg, Cheryl L., Raube, Kristina, Teleki, Stephanie S., and Erin dela Cruz. 2009. "Taking Stock of Pay-for-Performance: A Candid Assessment from the Front Lines." *Health Affairs*, 28(2): 517-525.

Dimick, Justin B., Staiger, Douglas O., and John D. Birkmeyer. 2010. "Ranking Hospitals on Surgical Mortality: The Importance of Reliability Adjustment." *Health Services Research* 45(6p1): 1614-1629.

Dunnett, C. W. 1955. "A multiple comparison procedure for comparing several treatments with a control." *Journal of the American Statistical Association* 50: 1096-1121.

Eijkenaar, Frank, Emmert, Martin, Scheppach, Manfred and Oliver Schöffski. 2013. "Effects of Pay for Performance in Health Care: A Systematic Review of Systematic Reviews" *Health Policy* 110(2-3): 115-130.

Gupta, S. S. 1956. "On a decision rule for a problem of ranking means." Institute of Statistics Mimeo Series No. 150.

Gupta, S. S. 1965. "On some multiple decision (selection and ranking) rules." *Technometrics* 7: 225-245.

Horrace, William C. 1998. "Tables of percentage points of the k-variate normal distribution for large values of k." *Communications in Statistics: Simulation & Computation* 27: 823-31.

Horrace, William C. and Peter Schmidt. 2000. "Multiple comparisons with the best, with economic applications." *Journal of Applied Econometrics* 15: 1-26.

Horrace, William C. and K.E. Schnier. 2010. "Fixed effect estimation of highly-mobile production technologies." *American Journal of Agricultural Economics* 92:1432-1445.

Kane, Thomas J., and Douglas O. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures" *Journal of Economic Perspectives* 16(4): 91-114.

Lindenauer, Peter K., Lagu, Tara, Ross, Joseph S., Pekow, Penelope S., Shatz, Amy, Hannon, Nicholas, Rothberg, Michael B., and Evan M. Benjamin. 2014. "Attitudes of Hospital Leaders Towards Publicly Reported Measures of Health Care Quality" *JAMA Internal Medicine* 174(12): 1904-1911.

Markovitz, Adam A., and Andrew M. Ryan. 2016. "Pay-for-Performance: Disappointing Results or Masked Heterogeneity?" *Medical Care Research and Review* forthcoming.

Mathematica Policy Research. 2012. *Results of Reliability Analysis from Mathematica Policy Research*. Memorandum to Centers for Medicare and Medicaid Research, February 13, 2012. [Accessed online on April 11, 2016 at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Downloads/HVBP_Measure_Reliability-.pdf]

Norton, Edward C., Li, Jun, Das, Anup, and Lena M. Chen. 2016. "Moneyball in Medicare." NBER Working Paper No. 22371.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms" *Journal of Economic Literature* 37(1): 7-63.

Ryan, Andrew M., Blustein, Jan, and Lawrence P. Casalino. 2012. "Medicare's Flagship Test of Pay-for-Performance did not Spur more Rapid Quality Improvement Among Low-Performing Hospitals." *Health Affairs* 31(4): 797-805.

Ryan, Andrew M., Burgess Jr., James F., Pesko, Michael F., Borden, William B., and Justin B. Dimick. 2015. "The Early Effects of Medicare's Mandatory Hospital Pay-for-Performance Program." *Health Services Research* 50(1): 81-97.

Ryan, Andrew M., Krinsky, Sam, Maurer, Kristin A., and Justin B. Dimick. 2017. "Changes in Hospital Quality Associated with Hospital Value-Based Purchasing" *The New England Journal of Medicine* 376: 2358-2366.

Ryan, Andrew M., Nallamothu, Brahmajee K., and Justin B. Dimick. 2012. "Medicare's Public Reporting Initiative on Hospital Quality had Modest or no Impact on Mortality from Three Key Conditions." *Health Affairs* 31(3): 585-592.

Ryan, Andrew M., Sutton, Matthew, and Tim Doran. 2014. "Does Winning a Pay-for-Performance Bonus Improve Subsequent Quality Performance? Evidence from the Hospital Quality Incentive Demonstration." *Health Services Research* 49(2): 568-587.

Werner, Rachel M., Kolstad, Jonathan T., Stuart, Elizabeth A., and Daniel Polsky. 2011. "The Effect of Pay-for-Performance in Hospitals: Lessons for Quality Improvement." *Health Affairs* 30(4): 690-698.

Table 1. Relevant Time Periods for Payment Calculation under HVBP

	Baseline Period		Performance Period	
Payment Year	**Start Date**	**End Date**	**Start Date**	**End Date**
FY 2013	7/1/2009	3/31/2010	7/1/2011	3/31/2012
FY 2014	4/1/2010	12/31/2010	4/1/2012	12/31/2012
FY 2015	1/1/2011	12/31/2011	1/1/2013	12/31/2013
FY 2016	1/1/2012	12/31/2012	1/1/2014	12/31/2014
FY 2017	1/1/2013	12/31/2013	1/1/2015	12/31/2015

Table 2. HVBP Outcomes Metrics Descriptions

Metric	Fiscal Years Used for HVBP	Description	Mean Value	Mean Observations
30-Day AMI	2014-current	30 day survival rate for AMI discharges	83.980	191.360
30-Day HF	2014-current	30 day survival rate for heart failure discharges	87.977	256.906
30-Day PN	2014-current	30 day survival rate for pneumonia discharges	89.293	228.770

Table 3. Percent of Hospitals using Achievement Score

Metric	Year	Percent of Hospitals using Achievement Score
30-Day AMI	2015	0.8042
30-Day HF	2015	0.7612
30-Day PN	2015	0.7535

Table 4. Outcome Metric Lottery Around the Threshold and Benchmark with α = 0.05 Error Rate

Metric	Year	Total *Hospitals* n	Total Extrema $min_{i\in N}\hat{p}_i$, $max_{i\in N}\hat{p}_i$,	$z_{\alpha,n}$	p_l^*	Lottery Count $n_{l,\alpha}$	Lottery Extrema $min_{i\in S_{l,\alpha}}\hat{p}_i$, $max_{i\in S_{l,\alpha}}\hat{p}_i$	Share in Lottery	p_u^*	Lottery Count $n_{u,\alpha}$	Lottery Extrema $min_{i\in S_{u,\alpha}}\hat{p}_i$, $max_{i\in S_{u,\alpha}}\hat{p}_i$	Share In Lottery
30-Day AMI	2015	2,502	0.3000, 0.9915	4.27	0.84747	1,549	0.4280, 0.9302	**0.6191**	0.86237	1,628	0.4400, 0.9370	**0.6507**
30-Day HF	2015	3,781	0.4200, 0.9961	4.34	0.88151	2,216	0.4600, 0.9490	**0.5861**	0.90032	2,462	0.4760, 0.9578	**0.6416**
30-Day PN	2015	4,191	0.3920, 0.9955	4.39	0.88165	2,595	0.4577, 0.9512	**0.6192**	0.90418	2,923	0.4808, 0.9610	**0.6975**

Simulation sample size 10,000.

Table 5. AMI Lottery Around the Threshold with α = 0.05 Error Rate - Rescaled Sample Sizes

Metric	Year	Total *Hospitals* n	Sample Size Rescaling	Total Extrema $min_{i\in N}\hat{p}_i$, $max_{i\in N}\hat{p}_i$,	$z_{k,\alpha,n}$	p_l^*	Lottery Count $n_{l,\alpha}$	Lottery Extrema $min_{i\in S_{l,\alpha}}\hat{p}_i$, $max_{i\in S_{l,\alpha}}\hat{p}_i$	Share in Lottery
30-Day AMI	2015	2,502	×1	0.3000, 0.9915	4.27	0.84747	1,549	0.4280, 0.9302	**0.6191**
30-Day AMI	2015	2,502	×3.93	0.3000, 0.9915	4.27	0.84747	859	0.6976, 0.9031	**0.3433**
30-Day AMI	2015	2,502	×13.88	0.3000, 0.9915	4.27	0.84747	430	0.7862, 0.8816	**0.1719**
30-Day AMI	2015	2,502	×22.34	0.3000, 0.9915	4.27	0.84747	330	0.8057, 0.8757	**0.1319**

Simulation sample size 10,000.

Table 6. Estimated Threshold and Benchmark with α = 0.05 Error Rate

Metric	Year	Total *n*	p_l^*	$\hat{p}_l^*$	n_*	$\tilde{n}$	Average $z_{k,\alpha,n}$	p_u^*	$\hat{p}_u^*$	n_{**}	Above $\tilde{p}_u^*$	Average z_{k,α,n_1}	$\tilde{p}_u^*$	$\tilde{n}_*$	Above $\tilde{p}_u^*$	Average $z_{k,\alpha,\tilde{n}}$
30-Day AMI	2015	2,502	0.84747	0.7692	1,662	1,901	4.00	0.86237	0.8586	1,360	1,511	3.90	0.8275	1,449	1,675	3.93
30-Day HF	2015	3,781	0.88151	0.7114	3,075	3,344	4.09	0.90032	0.7750	2,787	3,026	4.04	0.7595	2,916	3,155	4.06
30-Day PN	2015	4,191	0.88165	0.7243	3,594	3,858	4.11	0.90418	0.7714	3,348	3,556	4.07	0.7486	3,437	3,785	4.09

Simulation sample size 10,000.

APPENDIX

Table A1. HVBP Clinical Process of Care Metrics Descriptions

Metric	Fiscal Years Used for HVBP	Description	Mean Value	Mean Observations per Hospital
AMI 8a	2013-2015	Percutaneous coronary intervention (stent placement) performed within 90 minutes of arrival for heart attack patients	75.857	128.231
HF 1	2013-2015	Discharge instructions given to heart failure patients	72.935	581.787
PN 3b	2013-2015	Blood culture performed before 1st antibiotic given to pneumonia patients	90.140	478.994
PN 6	2013-2016	Most appropriate initial antibiotic given to pneumonia patients	86.315	336.343
SCIP 1	2013-2015	Antibiotics given within 1 hour before surgery (within 2 hours if certain drugs are used)	88.173	1,074.331
SCIP 2	2013-2016	Received recommended prophylactic antibiotics with surgery	93.871	1,086.577
SCIP 3	2013-2016	Prophylactic antibiotics discontinued within 24 hours of surgery (48 hours for cardiac surgery)	85.689	1,050.943
SCIP 4	2013-2015	Post-operative serum glucose for cardiac surgery	92.405	528.329
SCIP 9	2014-2016	Post-operative catheter removed within two days of surgery	89.229	609.090
SCIP VTE 1	2013-2014	Patients for venous thromboembolism (blood clots in veins) surgery received correct prophylactics	87.555	805.662
SCIP VTE 2	2013-2016	Patients for venous thromboembolism surgeries received anti-clotting treatment	86.636	761.496
SCIP Card 2	2013-2016	Surgery patients on beta-blockers pre-hospitalization given beta blockers during hospitalization	91.125	494.769

Table A2. Percent of Hospitals using Achievement Score for Clinical Process of Care

Metric	Year	Percent of Hospitals using Achievement Score
AMI 8a	2015	0.7480
HF 1	2015	0.7060
PN 3b	2015	0.7467
PN 6	2015	0.7603
PN 6	2016	0.7480
SCIP 1	2015	0.7478
SCIP 2	2015	0.7491
SCIP 2	2016	0.7917
SCIP 3	2015	0.7244
SCIP 3	2016	0.7348
SCIP 4	2015	0.7204
SCIP 9	2015	0.7035
SCIP 9	2016	0.7114
SCIP Card 2	2015	0.7250
SCIP Card 2	2016	0.7558
SCIP VTE 2	2015	0.7033
SCIP VTE 2	2016	0.9091

Table A3. Clinical Process of Care Metric Lottery Around the Threshold and Benchmark with α = 0.05 Error Rate

Metric	Year	Total *Hospitals* n	Total Extrema $min_{i\in N}\hat{p}_i$, $max_{i\in N}\hat{p}_i$,	$z_{\alpha,n}$	p_l^*	Lottery Count $n_{l,\alpha}$	Lottery Extrema $min_{i\in S_{l,\alpha}}\hat{p}_i$, $max_{i\in S_{l,\alpha}}\hat{p}_i$	Share in Lottery	p_u^*	Lottery Count $n_{u,\alpha}$	Lottery Extrema $min_{i\in S_{u,\alpha}}\hat{p}_i$, $max_{i\in S_{u,\alpha}}\hat{p}_i$	Share In Lottery
AMI 8a	2015	1,192	0.1598, 0.9967	4.08	0.95349	1,061	0.4500, 0.9913	**0.8901**	1	1,021	0.4500, 0.9967	**0.8565**
HF 1	2015	3,194	0.0239, 0.9980	4.32	0.94118	1,860	0.3600, 0.9884	**0.5823**	1	1,491	0.4200, 0.9980	**0.4668**
PN 3b	2015	3,290	0.1800, 0.9971	4.33	0.97783	2836	0.5500, 0.9948	**0.8620**	1	2,613	0.5500, 0.9972	**0.7942**
PN 6	2015	3,655	0.1513, 0.9972	4.35	0.95918	3,057	0.4300, 0.9928	**0.8364**	1	2,720	0.4300, 0.9972	**0.7442**
PN 6	2016	3,304	0.1100, 0.9900	4.32	0.96552	3,243	0.4400, 0.9900	**0.9815**	1	3,183	0.4500, 0.9900	**0.9634**
SCIP 1	2015	2,793	0.1021, 0.9976	4.28	0.98639	2,408	0.4200, 0.9971	**0.8622**	1	1,915	0.4200, 0.9976	**0.6856**
SCIP 2	2015	2,777	0.0846, 0.9979	4.27	0.98637	2,441	0.5400, 0.9973	**0.8790**	1	2,074	0.5400, 0.9979	**0.7469**
SCIP 2	2016	2,176	0.5600, 0.9900	4.22	0.99074	2,154	0.6400, 0.9900	**0.9899**	1	2,108	0.6400, 0.9900	**0.9688**
SCIP 3	2015	3,054	0.1332, 0.9973	4.25	0.97494	2,296	0.5616, 0.9958	**0.7518**	1	1,652	0.6400, 0.9973	**0.5409**
SCIP 3	2016	2,717	0.1900, 0.9900	4.27	0.98086	2,646	0.5500, 0.9900	**0.9739**	1	2,435	0.5300, 0.9900	**0.8962**
SCIP 4	2015	1,137	0.6161, 0.9969	4.05	0.95798	871	0.7967, 0.9909	**0.7661**	0.99767	745	0.7967, 0.9969	**0.6552**
SCIP 9	2015	2,956	0.1317, 0.9973	4.29	0.94891	1,676	0.5200, 0.9903	**0.5670**	0.99991	1,349	0.6400, 0.9973	**0.4564**
SCIP 9	2016	2,586	0.2500, 0.9900	4.28	0.97059	2,377	0.5700, 0.9900	**0.9192**	1	2,111	0.5700, 0.9900	**0.8163**
SCIP Card 2	2015	2,771	0.1700, 0.9972	4.27	0.97175	2,256	0.4700, 0.9941	**0.8142**	1	1,917	0.4700, 0.9972	**0.6918**
SCIP Card 2	2016	2,347	0.0500, 0.9900	4.23	0.97727	2,308	0.5700, 0.9900	**0.9834**	1	2,246	0.5700, 0.9900	**0.9570**
SCIP VTE 2	2015	3,090	0.1694, 0.9977	4.32	0.97403	2,106	0.5300, 0.9941	**0.6816**	0.99998	1,583	0.5300, 0.9977	**0.5123**
SCIP VTE 2	2016	2,683	0.0800, 0.9900	4.28	0.98225	2,588	0.5400, 0.9900	**0.9646**	1	2,347	0.5400, 0.9900	**0.8748**

Simulation sample size 10,000.