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
August 2016

ISSN: 1252-3066

SYRACUSE UNIVERSITY
Maxwell School
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 –Summer 2016

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 and PAIA
Associate Director, Metropolitan Studies

SENIOR RESEARCH ASSOCIATES

Badi Baltagi, Economics
Robert Bifulco, PAIA
Leonard Burman, PAIA
Thomas Dennison, PAIA
Alfonso Flores-Lagunes, Economics
Sarah Hamersma, PAIA
William C. Horrace, Economics
Yilin Hou, PAIA
Hugo Jales, Economics
Duke Kao, Economics
Jeffrey Kubik, Economics
Yoonseok Lee, Economics
Amy Lutz, Sociology
Yingyi Ma, Sociology
Jerry Miner, Economics
Cynthia Morrow, PAIA
Jan Ondrich, Economics
John Palmer, PAIA
David Popp, PAIA
Stuart Rosenthal, Economics
Michah Rothbart, PAIA
Rebecca Schewe, Sociology
Amy Ellen Schwartz, PAIA/Economics
Perry Singleton, Economics
Michael Wasylenko, Economics
Peter Wilcoxen, PAIA

GRADUATE ASSOCIATES

Emily Cardon, PAIA
Carlos Diaz, Economics
Alex Falevich, Economics
Wancong Fu, Economics
Boqian Jiang, Economics
Hyunseok Jung, Economics
Yusun Kim, PAIA
Ling Li, Economics
Michelle Lofton, PAIA
Judson Murchie, PAIA
Brian Ohl, PAIA
Jindong Pang, Economics
Laura Rodriquez-Ortiz, PAIA
Fabio Rueda De Vivero, Economics
David Schwegman, PAIA
Shulin Shen, Economics
Iuliia Shybalkina, PAIA
Kelly Stevens, PAIA
Saied Toossi, PAIA
Rebecca Wang, Sociology
Xirui Zhang, Economics

STAFF

Kelly Bogart, Administrative Specialist
Kathleen Nasto, Administrative Assistant
Candi Patterson, Computer Consultant
Mary Santy, Administrative Assistant
Katrina Wingle, Administrative Assistant

Abstract

As part of the 2010 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 payment 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. This results in point allocation under the payment formula that is random for a large proportion of the hospitals. An alternative payment plan is discussed.

JEL No. H51, I18

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

Acknowledgements: We are grateful for comments by seminar participants at ASHEcon 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.

Authors:

Andrew I. Friedson, Department of Economics, University of Colorado Denver.
Andrew.Friedson@ucdenver.edu (303) 315-2038.

William C. Horrace, Department of Economics, Syracuse University.
whorrace@maxwell.syr.edu (315) 443-9061

Allison F. Marier, Allison.marier@gmail.com

I. Introduction

As part of the 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 this goal, AHRQ's strategy operationalized a value-based purchasing plan that reimburses hospitals based on performance metrics. AHRQ, along with additional stakeholders, designed these metrics to address potential gaps in patient care and coordination that lead to unintended and costly adverse patient outcomes.[1] Hospital Value Based Purchasing (HVBP) is implemented by the Centers for Medicare and Medicaid Services (CMS); CMS redistributes a percentage of the total funds designated for reimbursement to hospitals based on a performance score composed of individual AHRQ metrics. Many of the metrics used in HVBP were already being collected by CMS for their Hospital Compare website, which reports on hospital performance and outcomes.[2]

One prominent part of the formula for calculating the redistribution of funds under HVBP depends on a hospital's ranking within the distribution of all hospitals for any given quality metric. A concern that arises is whether the signal of a hospital's true quality can be differentiated from statistical noise when using the distribution of a given quality metric to generate an ordinal ranking. In other words, do the quality metric distributions have enough information in them to differentiate one hospital's quality from another's? The answer is important to policy makers, as the primary goal of the program is to reward hospitals based on

[1] 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] Hospital Compare has been in operation as of 2005.

their quality. Because the Hospital Compare score is an estimate of a hospital's true quality, the relationship between score variation in the distribution and Medicare payment is particularly relevant to policymakers.

A hospital's true quality is obfuscated by two related factors. The first is that a hospital's true performance is subject to noise that can be a function of statistical risk-adjustments (Dimick, Stager and Birkmeyer 2010, Mathematica 2012), human data entry error, or additional human errors in the data management process (Bowman 2013). The second is that the quality metrics used for HVBP are easily improvable – which leads to bunching near the maximum achievable score over time. This second issue becomes particularly problematic if the metrics do not have a corresponding improvement in their precision (i.e. the scores become a more exact measure of a hospital's true quality). As each metric is effectively an estimate of a hospital's true quality in a given dimension, and as such has a standard error, increased bunching without a corresponding narrowing of standard errors will decrease the ability to statistically differentiate hospitals from one another.

Figures 1a, 1b, and 1c demonstrate the improvement in hospital performance metrics between 2005 and 2014.[3] Hospitals have been steadily improving their scores over time, which is at least partially attributable to pay-for-reporting initiatives (Lindenauer et al. 2014), although the connection between score improvement and pay-for-reporting has been shown to be both limited in scope and modest in size (Ryan, Nallamothu, and Dimick 2012). Further, studies of the impact of HVBP and its pilot programs show that the implementation of the program itself had little to no direct impact on the improvement of quality scores (Ryan, Blustein, and Casalino

[3] The patterns shown in the figures also appear in the distributions of other quality scores reported by Hospital Compare. Those figures are available upon request.

2012; Ryan, Sutton and Doran 2014; Ryan et al. 2015). The observed distributions during the most recent reporting year, 2014, show clearly how the quality metrics have homogenized across hospitals, and underscore the potential problem of an inability to differentiate hospitals based on these metrics.

This work explores the viability of creating ordinal rankings for U.S. hospitals based on the Hospital Compare metrics. We examine whether individual hospitals can be statistically differentiated from one another, specifically in areas around cutoffs for payment. To do this we draw on techniques for multiple comparisons within ordinal rankings (Dunnet 1955, Gupta 1956, Gupta 1965) to create groupings in which with a fixed probability, all members of the group are indistinguishable from each other in terms of ranking based on a single measure. By creating groups that are indistinguishable from a hospital sitting exactly on or very close to a 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.[4]

We find that the lottery zones around cutoffs for payment under HVBP are large, in most cases capturing over 85 percent of the hospitals submitting data. We show that changes in the payment formulae in future years that phase out older metrics and phase in newer metrics do not appear likely to solve this issue. We also demonstrate that the lottery zones cover a larger percentage of hospitals in regions where HVBP attempts to make the finest quality distinctions. Lastly, we suggest an alternative, data driven approach from generating payment cutoffs.

The implications from our analysis are twofold. From a policy perspective, if AHRQ's intent is to incentivize high effectiveness practices within hospitals, then it would appear that

[4] 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.

payment based on ordinal rank is unnecessary, as hospitals were already improving their scores over time in the absence of the HVBP cash incentives. However, if the goal is simply to reward hospitals for being of high quality, then the payment formula we propose does just that. We demonstrate how to pay hospitals in a way that uses what true differentiating information is contained in the metric distributions. Secondly, we argue that 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 and performance.

II. Background on HVBP

HBVP is a form of pay-for-performance, a concept that has been historically applied within firms (see Prendergast 1999 for a review), and more recently been applied by the government to incentivize high quality care by health care providers. The literature on pay-for-performance for health care finds on average that financial incentives do improve outcomes reported by health care providers (Eijkenaar et al. 2013), although health outcomes vary by hospital, physician and patient demographics (Markovitz and Ryan 2016). There is evidence 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).

Under HVBP, hospitals that qualify receive payment based on their scores in the Hospital Compare data.[5] The payment formula uses a points system: hospitals earn points based on their

[5] 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

performance for each quality metric. A hospital can score up to 9 points based on their improvement over their old metric, or can score up to 10 points based on their placement within the overall distribution 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 from several domains of metrics: Clinical Process of Care, Patient Experience of Care, Efficiency and Outcomes[6]. Within each domain, the total hospital score is the percent of total points earned within the metrics contained by the domain.[7] The domain scores are then averaged together into the TPS, using weights that vary from year to year. In 2013, for example, the TPS was calculated as:

$$TPS_{2013} = 0.70 \times \text{Clinical Process of Care Score} + 0.30 \times \text{Patient Experience of Care Score}$$

and in 2014:

$$TPS_{2014} = 0.45 \times \text{Clinical Process of Care Score} + 0.30 \times \text{Patient Experience of Care Score} + 0.25 \times \text{Outcomes Score}$$

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 will remain onwards) of Medicare reimbursements to

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.

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

[7] The "Patient Experience of Care" domain has an additional source of points, where hospitals earn points for "consistency" by scoring well in multiple categories in the Hospital Consumer Assessment of Healthcare Providers and Systems Survey.

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 scoring points, and then additional data are drawn from a later *performance* to score points. Table 1 describes the relevant time periods used for calculating payments.[8]

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 relevant thresholds and benchmarks. 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

[8] 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.

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, the same as the achievement score, is the 95th percentile of all hospital's performance during the baseline period. 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 Hospital Compare data pulls information for each hospital from a subset of patient visits. Each qualifying visit is entered into the data for the purposes of calculating a quality metric.[9] We start with analysis of the relatively simple metrics that fall under the Clinical Process of Care domain: patients qualify for inclusion in a hospital's data if they have a specific diagnosis and are clinically eligible to receive treatment. Then, the data track whether or not a specific action was taken for that patient 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 aggregate metric reports the percent of total qualifying AMI cases for which patients received the intervention in the appropriate amount of time. In most cases, metrics quantify desirable treatments and higher scores are indicative of better care. Other domains are based on patient survey responses, the rate of

[9] Qualifying visits include patients who both are eligible for a specific treatment, and for which the treatment is clinically appropriate. For instance, a patient who presents with AMI is eligible to receive aspirin upon hospital arrival, but would be excluded from the data if the hospital did not administer aspirin due to a potential allergic reaction.

hospital acquired illnesses, and rates of spending per Medicare beneficiary. The Hospital Compare data reports both the score for each metric for each hospital, and the number of observations that were used to calculate the score. The top panel of Table 2 lists the metrics used in the Clinical Process of Care domain, and briefly describes what they measure.

From a statistical point of view, each observation used to calculate a metric in the Clinical Process of Care domain is a draw from a Bernoulli distribution (one for each hospital). Either the hospital does the desired action or it does not. Each hospital has a true probability of the desired clinical action, 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. Whether those hospitals are distinguishable from one another in a statistical sense also depends upon their standard errors, which given the Bernoulli structure of the underlying data generation process is a function of the reported score and of the number of observations per hospital.[10] If the overall distributions of the metrics (estimates) do not vary greatly, then it becomes difficult to statistically distinguish hospital metric values from one another. CMS understands this and "tops out" metrics. When CMS believes that there is not enough distinguishing information in a distribution of metrics then those metrics are removed from payment calculation; that is why, as shown in table 2, not all metrics are used in all years. An open empirical question that we seek to answer is whether this happens quickly enough. In other words, is there enough distinguishing information in the

[10] 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

distributions in the first place? We accomplish this by directly testing the distinguishability of metric values from one another within their respective distributions.

Using the Clinical Process of Care metrics for analysis has a drawback: over time, these metrics become a smaller and smaller portion of the payment formula. In 2017, they will determine only 5 percent of the TPS. Fortunately, there are also measures within the "Outcomes" domain that follow a Bernoulli data generation process. Hospitals are also scored based on their 30-day survival rate on discharges for AMI, heart failure and pneumonia. The metrics are reported as survival rates, or the percent of patients with such a discharge that survived 30 days after being released.[11] Analysis of these Outcome metrics provides a look into how quality metrics that will remain a large part of the payment formula for the near future perform. We analyze mortality Outcome metrics for 2015, the most recent year that CMS has made the data available for; a description of the metrics can be found in the bottom panel of Table 2.

IV. Methods

To determine the amount of relevant ordinal information within the distribution of a given metric, we focus on how distinguishable values of the metric are from the cutoff values for payment under HVBP. We seek to discover 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. We focus on achievement scores because they offer a single value for the threshold for all hospitals whereas improvement scores do not. Additionally, a vast majority of hospitals rely of the achievement score rather than the improvement score to determine

[11] Adjustments are made to account for patients who die of causes unrelated to their discharge.

their point allocation. This is demonstrated in Table 3, which shows that for every metric, achievement scores are the relevant source of points over 70 percent of the time. That said, as the benchmark is the same for achievement and improvement scores, analysis relevant for achievement benchmarks applies to improvement benchmarks as well.

The grouping of hospitals which are indistinguishable from a hospital on the cutoff represent areas in which whether a particular hospital exceeds the payment cutoff or not is random in a statistical sense. In other words, hospitals that fall into these groupings are essentially subject to a lottery for scoring points and subsequently for payment under HVBP. If these groupings are large (i.e. cover a large proportion of the overall distribution of the metric), then HBVP assigns a larger number of points to hospitals as a result of sampling variability, rather than a hospital's true metric performance.

Let $p_i \in (0,1)\ \ i = 1, \dots n$ be the mean parameter for *n independent* Bernoulli populations. Also, let $min_i p_i < p_l^* < max_i p_i$ be a fixed *lower threshold*, such that it is strictly less than the largest value of p_i and strictly larger than the smallest value of p_i. Similarly, define a fixed *upper benchmark*, $min_i p_i < p_u^* < max_i p_i$, 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 $i = 1, \dots n$ and determining (in a statistical sense) a subset of populations (hospitals) around the threshold that are closest to the cutoff. To do this we fold those populations with $p_i > p_l^*$ around p_l^*.[12] Consider the folding of populations above p_l^*:

[12] In situations where the cutoff being used is the maximum values (as is the case for many metric benchmarks), no folding is necessary, and the standard subset of the best groupings are used.

$$r_{l,i} = r(p_l^*, p_i) = \begin{cases} p_i & p_i \leq p_l^* \\ 2p_l^* - p_i & p_i > p_l^* \end{cases}, i = 1, \dots n.$$

The function $r(p_l^*, p_i)$ is an *affine* transformation of the population parameters that will be important for the inference that follows. Notice that $r_{l,i} \leq p_l^*$, so the largest values of $r_{l,i}$ are those closest or equal to p_l^*. If $p_i > p_l^*$ is in a small neighborhood above the lower threshold, then after the fold $r_{l,i} \leq p_l^*$ is in the same small neighborhood reflected below the threshold. Our focus is on identifying such a neighborhood. Let the ranked folded values be:

$$r_{l,[1]} \leq r_{l,[2]} \leq \cdots \leq r_{l,[n]} \leq p_l^*$$

Then our selection procedure will determine a non-empty subset $S_{l,\alpha} \subset N = \{1, \dots, n\}$ at a pre-specified error rate $\alpha \in (0,1)$ such that $\Pr\{[n] \in S_{l,\alpha}\} \geq 1 - \alpha$. Simply put, $S_{l,\alpha}$ will contain the index of the population with parameter $r_{l,i}$ closest to the threshold, p_l^*, with probability at least $1 - \alpha$. Therefore, $S_{l,\alpha}$ contains those populations in the neighborhood of the threshold that are indistinguishable. We now consider inference on $S_{l,\alpha}$. In doing so, our concern is only to select those populations, $r_{l,i}$, that are closest to the threshold. In particular, the folding confounds our ability to infer to which side of the threshold the underlying population parameters, p_i , lie.

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

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

Let $\hat{r}_{l,i} = r(p_l^*, \hat{p}_i)$ Notice that,

$$V(\hat{r}_{l,i}) = \begin{cases} V(\hat{p}_i) & \hat{p}_i \leq p_l^* \\ V(2p_l^* - \hat{p}_i) & \hat{p}_i > p_l^* \end{cases} = \begin{cases} V(\hat{p}_i) & \hat{p}_i \leq p_l^* \\ V(\hat{p}_i) & \hat{p}_i \leq p_l^* \end{cases} = V(\hat{p}_i)$$

Not surprisingly, independence of the $\hat{p}_i$ ensures independence of the $\hat{r}_{l,i}$. Hence, the usual unbiased estimator of the variance of $\hat{r}_{l,i}$ is,

$$\hat{V}(\hat{r}_{l,i}) = \hat{V}(\hat{p}_i) = T_i^{-1}\hat{p}_i(1 - \hat{p}_i).$$

Therefore, the sampling distributions of $\hat{r}_{l,i}$ and $\hat{p}_i$ have the same variance, greatly simplifying inference. In what follows, we assume that the sampling distribution of $\hat{p}_i$ is normal (or asymptotically so), so that the sampling distribution of $\hat{r}_{l,i}$ is normal, as the normal family of distributions in closed to affine transformations. It is important to stress that our focus is not on inference for the underlying Bernoulli population but for the sampling distribution of the folded statistic $\hat{r}_{l,i}$. 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. Moreover, all our results hold for the benchmark p_u^* by simply substituting u (upper) for l (lower) everywhere.

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:

$$\left[L_{l,i}^j, U_{l,i}^j\right], i \neq k, i \in N$$

$$L_{l,i}^k = \hat{r}_{l,k} - \hat{r}_{l,i} - z_{k,\alpha,n}\left(\hat{V}(\hat{p}_k) + \hat{V}(\hat{p}_i)\right)^{1/2}$$

$$U_{l,i}^k = \hat{r}_{l,i} - \hat{r}_{l,i} + z_{k,\alpha,n}\left(\hat{V}(\hat{p}_k) + \hat{V}(\hat{p}_i)\right)^{1/2}$$

where $z_{k,\alpha,n}$ is a two-sided critical value from an $n - 1$ dimensional standard normal distribution with the k^{th} population treated as control, such that $\Pr(max_{1 \leq i \leq n-1}|z_i| \leq z_{k,\alpha})$. See Horrace and Schmidt (2000) for a precise definition of the critical values, which are easily simulated

(Horrace, 1998). MCC is due to Dunnett (1955), and the confidence intervals account for the multiplicity inherent in the order statistic, and are therefore wider that the usual univariate confidence intervals associated with the difference of two means. This will be reflected in the fact that $z_{k,\alpha,n} > 1.96$ for $n > 2$, in our application. Define the subset:

$$S_{l,\alpha} = \left\{k: U_{l,i}^{k} > 0 \; \forall \; i \in N\right\} \subseteq N$$

This is the standard Gupta (1956, 1965) subset selection procedure that ensures that $\Pr\{[n] \in S_{l,\alpha}\} \geq 1 - \alpha$. The subset is non-empty and contains all indices k that have all positive MCC upper bounds. That is, $\hat{r}_{l,k}$ is simultaneously larger (in a statistical sense) than all $\hat{r}_{l,i}$, $i \neq k$. The indices in the subset constitute our lottery around p_l^*. Populations (hospitals) in $S_{l,\alpha}$ are statistically indistinguishable and are in the neighborhood of the threshold. The size of the neighborhood (the cardinality of $S_{l,\alpha}$) is strictly decreasing in α, but the convention is to pre-select $\alpha = 0.05$ or $\alpha = 0.10$ so inference is at the 95% or 90% level. 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 cardinality and the lottery share are both inversely related to the "sharpness" of the inference. When $n_{l,\alpha} = 1$, the subset is a singleton and inference is sharpest. In this case, the inference has identified the hospital i that has $\hat{r}_{l,i}$ and p_i closest to p_l^*. When $n_{l,\alpha} = n$, the subset contains all the hospital indices and inference is least sharp. In this case, the data tell us nothing about the hospital's relative proximity to the threshold: all hospitals are indistinguishable.

Since the number of hospitals for each measure can be quite large, our critical values can be quite large. To reduce the level of multiplicity in our inference, we can also focus on just those hospitals between the threshold and the benchmark. Let n^* be the cardinality of the set $N^* = \{i: p_l^* < \hat{p}_i < p_u^*, i = 1, \ldots, n\}$. Notice that for this analysis we will not be reflecting the

estimated scores around the cutoff. The number of hospitals between the threshold and benchmark, n^*, will be considerably smaller than the total number of hospitals, *n*, so that multiplicity will have less of an effect on the inference that follows. Then we can calculate MCC intervals:

$$\left[L_{*i}^{j}, U_{*i}^{j}\right], i \neq k, i \in N^*$$

$$L_{*i}^{k} = \hat{p}_k - \hat{p}_i - z_{k,\alpha,n^*}\left(\hat{V}(\hat{p}_k) + \hat{V}(\hat{p}_i)\right)^{1/2}$$

$$U_{*i}^{k} = \hat{p}_k - \hat{p}_i + z_{k,\alpha,n^*}\left(\hat{V}(\hat{p}_k) + \hat{V}(\hat{p}_i)\right)^{1/2},$$

leading to the subset of the best hospitals between the threshold and benchmark:

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

V. Results

Clinical Process of Care Lottery Zones

Results for the Clinical Process of Care 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.[13] For example, in the first row the metric is "AMI 8a" 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 $\hat{p}_i \in (0,1)$, 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 critical values over all *n* hospitals in *N*. For example, in the first row of the table, for the metric AMI 8a, there were *n* = 1,192 hospitals

[13] 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.

reporting data for payment year 2015 that had scores of $\hat{p}_i \in (0,1)$. The maximal value of the score is 0.9967 and the minimal values is 0.1581.[14] Then, we simulated critical value of each hospital (per Horrace 1998) to perform the MCC procedure. There were *n*-1 critical values generated for each $k \in N$ as the control. The average of these critical values for AMI 8a, was 3.83. Notice that 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.

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, the seventh contains the cardinality of the subset in the neighborhood of the threshold $(n_{l,\alpha})$, the eighth contains that extreme values of $\hat{p}_{l,i}$ in the lottery (after unfolding the $\hat{r}_{l,i}$), 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 AMI 8a for 2015, we see that there are $n_{l,\alpha} = 1{,}127$ in the lottery, so the lottery share around the threshold at the 5% error rate is 0.9455. That is, most of the hospitals are indistinguishable, and their estimated scores range from 0.6814 to 0.9967. The inference is only slightly sharper for the HF 1 metric for 2015, which has a lottery share of 0.5949, but this is still a large portion of the hospitals in the lottery. Looking down column nine, we see that it is never the case that less than half of the hospitals are in the lottery around the threshold. Also, in cases when we have metrics for two consecutive years (2015 and 2016), the lottery is generally getting worse over time. For

[14] It should be noted that the data that we received were truncated after the 2nd decimal point. Nonetheless, when we aggregated truncated scores, we allowed for four places after the decimal point.

instance, for SCIP VTE 2 in the last two rows of the table, the lottery share goes from 0.7563 for 2015 to 0.9597 for 2016. The same is true for the metric SCIP Card 2 (which goes from a lottery share of 0.8459 to 0.9813), and for SCIP INF 9 (which goes from a lottery share of 0.5859 to 0.9323). This implies that hospital performance is increasingly indistinguishable over time. The only case where this was not true of the threshold lottery was for metric PN 6, where the lottery share stayed virtually unchanged (compare 0.8763 for 2015 to 0.8761 for 2016).

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. If the benchmark p_u^* equals 1, then all the hospitals in the sample are below it, and inference is on $r_{u,i}$ reduces to inference on p_i for all n. Otherwise, there may be some populations that get folded from above the benchmark to the neighborhood below it. Similar to the threshold results, the benchmark inference is not very sharp. For example, for metric AMI 8a for 2015, we see that there are 1,002 hospitals in the lottery, so the lottery share around the benchmark at the 5% error rate is 0.8460. Most of the hospitals around the benchmark are indistinguishable, and their estimated scores range from 0.6900 to 0.9967. The lottery around the benchmark is always smaller than the lottery around the threshold, and in two cases (HF1 for 2015 and SCIP 9 for 2015) the lottery share of hospitals around the benchmark is below 50%. 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 used to allocate HVBP funds.

Outcome Lottery Zones

The Clinical Process of Care metrics become a smaller and smaller part of the payment formula over time. Though the Clinical Process of Care metrics used in past years may have undesirable distributions for ordinal ranking, this may not be the case for metrics in other domains that are slated to remain a large part of the payment formula in upcoming years. To assess this we carry out the prior analyses for scores in the Outcomes domain that also follow a Bernoulli data generation process: 30-day survival rate for discharges related to AMI, heart failure, and pneumonia.

Results of analysis around the threshold and the benchmark are presented in Table 5, which follows the same layout as Table 4. Just as with the Clinical Process of Care metrics, the lottery zones a large, and inference is not very sharp. In fact, the lottery zones for the outcome scores are much larger than most of the lottery zones for Clinical Process of Care, implying that HVBP may be exchanging metrics with bad properties for metrics with worse properties.

One exercise that helps to further illustrate importance of noise in the outcome metrics used for HVBP is to calculate the subset of hospitals that are indistinguishable among the hospitals that scored between the threshold and the benchmark values. This is useful for two reasons. The first is that the critical values used in the subset selection procedure are sensitive to the amount of multiplicity of inference. Increasing the number of hospitals increases the size of the confidence intervals around each estimate, so if the problem of large lottery zones persists with a drastically smaller number of hospitals, the implications of our first analysis will be strongly reinforced.

The second reason is that the region between the threshold and the benchmark is of particular importance for the payment formula: this is a region where HVBP attempts to make fine quality distinctions between hospitals. At even intervals between the threshold and the benchmark are cutoffs for each of the different point values that can be obtained. If a lottery zone for the benchmark contains any of these values, then whether a hospital is given that value of the benchmark is essentially random. For example, if the lottery zone around the benchmark contains the cutoff for obtaining 7 points, then whether a hospital is assigned 7, 8, 9, or 10 achievement points is statistically random.

Our analysis of only those hospitals that scored between the benchmark and the threshold is reported in Table 6. In the area between the threshold and the benchmark, all of the hospitals are indistinguishable for all of the Outcome metrics. This implies that the assignment of point values between 1 and 10 for all hospitals that did not exceed the benchmark was random.

Relevance of Multiple Comparisons

Appropriate accounting for multiplicity of inference gives large lottery zones. This is a feature of the underlying data that would likely have been missed had multiple inference not been accounted for. Table 7 recreates the analysis reported in Table 5 using single inference techniques. We construct standard confidence intervals around each estimate, using critical values calculated from a t-distribution with degrees of freedom equal to the number of observations for the given hospital. Hospitals are considered to be in a lottery zone for a given metric if the 95 percent confidence interval for that hospital contains the threshold or the benchmark respectively.

The lottery zones reported in Table 7 are much smaller in magnitude than the lottery zones reported in Table 5, although in some cases they are still quite large. For example, the threshold lottery zone for 30-day heart failure survival 2015 when correctly adjusting for multiple inferences contains 99.15 percent of all hospitals, whereas the naïve lottery zone that uses single inference contains only 26.53 percent of hospitals. In most cases, the lottery zones created using single inference are approximately 70 percentage points smaller than those that account for multiple inferences. This underscores the importance of correctly accounting for the multiple comparisons that are made when creating an ordinal ranking of hospitals.

VI. An Alternative Points System

Based on the above analysis, the distributions of quality metrics do not have enough information within them to have the current HVBP payment formulae detect true quality differences between hospitals. Rather, the cutoffs, though well intentioned, appear to create arbitrary point assignment. Here we suggest an alternative point system that uses a data-driven methodology to estimate the threshold and benchmark cutoffs.

There is simply 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 N estimated threshold ($\hat{p}_l^*$), and 2 points for reaching an estimated benchmark (either $\tilde{p}_u^*$ or $\hat{p}_u^*$), 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 for the non-reflected estimates:

$$\left[L_{1i}^{j}, U_{1i}^{j}\right], i \neq k, i \in N$$

$$L_{1i}^{k} = \hat{p}_k - \hat{p}_i - z_{k,\alpha,n}\left(\hat{V}(\hat{p}_k) + \hat{V}(\hat{p}_i)\right)^{1/2}$$

$$U_{1i}^{k} = \hat{p}_k - \hat{p}_i + z_{k,\alpha,n}\left(\hat{V}(\hat{p}_k) + \hat{V}(\hat{p}_i)\right)^{1/2}.$$

Estimation in this setting is a two-step (nested) inference procedure, so the subscript "1" simply indicates the first step. The MCC intervals leads to the subset of the best hospitals in *N*:

$$S_1 = \left\{k: U_{1i}^{k} > 0 \; \forall \; i \in N\right\} \subseteq N.$$

Then an estimated threshold is $\hat{p}_l^* = \min_{i \in S_1} \hat{p}_i$. The estimated threshold is the minimal value for the scores contained in the subset of the best hospitals. That is, the threshold estimate is pegged to the worst performing hospital in the subset of the best. Let the cardinality of S_1 be n_1. To estimate the benchmark, perform MCC again on only the best hospitals, those contained in S_1. That is:

$$\left[L_{2i}^{j}, U_{2i}^{j}\right], i \neq k, i \in S_1$$

$$L_{2i}^{k} = \hat{p}_k - \hat{p}_i - z_{k,\alpha,n_1}\left(\hat{V}(\hat{p}_k) + \hat{V}(\hat{p}_i)\right)^{1/2}$$

$$U_{2i}^{k} = \hat{p}_k - \hat{p}_i + z_{k,\alpha,n_1}\left(\hat{V}(\hat{p}_k) + \hat{V}(\hat{p}_i)\right)^{1/2},$$

leading to the s*ubset of the best of the best* hospitals in S_α:

$$S_2 = \{k: U_2^k > 0 \; \forall \; i \in S_1\} \subseteq S_1.$$

Then an estimated benchmark is $\hat{p}_u^* = \min_{i \in S_2} \hat{p}_i$. The estimated benchmark is the minimal value for the scores contained in the "subset of the best of the best" hospitals. That is the benchmark estimate is pegged to the worst performing hospital in the subset of the best of the best. Then the cardinality of S_2 is n_2. Therefore, we have $S_2 \subseteq S_1 \subseteq \; N$, and $n_2 \leq n_1 \leq n$.

Due to differences in the precision of the estimated scores across hospitals, it may be the case that there are hospital with scores above the estimated threshold, $\hat{p}_l^*$, that are not contained in S_1. Therefore, define the set of hospitals 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 the cardinality of $S_1 \subseteq \widetilde{N}$ is so that $\tilde{n} \geq n_1$. Then we can perform MCC on the set $\widetilde{N}$ (as opposed to on the set S_1):

$$\left[\tilde{L}_{2i}^j, \widetilde{U}_{2i}^j\right], i \neq k, i \in \widetilde{N}$$

$$\tilde{L}_{2i}^k = \hat{p}_k - \hat{p}_i - z_{k,\alpha,\tilde{n}} \left(\hat{V}(\hat{p}_k) + \hat{V}(\hat{p}_i)\right)^{1/2}$$

$$\widetilde{U}_{2i}^k = \hat{p}_k - \hat{p}_i + z_{k,\alpha,\tilde{n}} \left(\hat{V}(\hat{p}_k) + \hat{V}(\hat{p}_i)\right)^{1/2},$$

leading to the alternative s*ubset of the best of the best* hospitals in $\widetilde{N}$:

$$\tilde{S}_2 = \left\{k: \widetilde{U}_2^k > 0 \; \forall \; i \in \widetilde{N}\right\} \subseteq \widetilde{N}.$$

Then an alternative estimated benchmark is $\tilde{p}_u^* = \min_{i \in \tilde{S}_2} \hat{p}_i$. The subset $\widetilde{N}$ relaxes the requirement that a hospital must be in S_1 to be part of the second step of the inference

procedure. In fact we have $S_2 \subseteq S_1 \subseteq \tilde{N}$. Let $\tilde{n}_2$ be the cardinality of $\tilde{S}_2$. 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 8. 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_1) has the number of hospitals that are in the subset of the best (S_1): 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 $\tilde{N}$. 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 $\tilde{N}$. We can use either set ($\tilde{N}$ or S_1) as the basis for our second round of inference to estimate the benchmark.

Panel three of table 8 (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_1. 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_1). 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 8 (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 $\tilde{S}_2$, which only contains 1,449 hospitals.

VII. Discussion and Conclusion

Though the intent of HVBP is to pay hospitals based on their true underlying 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" metrics as their distributions collapse towards the maximum attainable values, the above analyses show that the metrics that are being used for HVBP 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 meet quality goals. 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), and the amounts that are redistributed by HBVP in practice are relatively small. Simulated payments suggest that on average hospitals give up around 1 percent of their earnings and then receive back approximately 1 percent, plus or minus a small amount (Werner and Dudley 2012).

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 statistical 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 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 hospital behaviors that do not correspond with the federal 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.

VIII. 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.

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. J. Amer. Statist. Assoc. 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, University of North Carolina.

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

Horrace, W.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, W.C. and P. Schmidt. 2000. Multiple comparisons with the best, with economic applications. Journal of Applied Econometrics 15: 1-26.

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

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., 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.

Werner, Rachel M., and R. Adams Dudley. (2012) "Medicare's New Hospital Value-Based Purchasing Program is Likely to Have Only a Small Impact on Hospital Payments" *Health Affairs,* 31(9): 1932-1940.

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 Metrics that follow a Bernoulli Data Generation Process 2013 to 2017

Metric	Fiscal Years Used for HVBP	Description
	Clinical Process of Care	
AMI 8a	2013-2015	Percutaneous coronary intervention (stent placement) performed within 90 minutes of arrival for heart attack patients
HF 1	2013-2015	Discharge instructions given to heart failure patients
PN 3b	2013-2015	Blood culture performed before 1st antibiotic given to pneumonia patients
PN 6	2013-2016	Most appropriate initial antibiotic given to pneumonia patients
SCIP 1	2013-2015	Antibiotics given within 1 hour before surgery (within 2 hours if certain drugs are used)
SCIP 2	2013-2016	Received recommended prophylactic antibiotics with surgery
SCIP 3	2013-2016	Prophylactic antibiotics discontinued within 24 hours of surgery (48 hours for cardiac surgery)
SCIP 4	2013-2015	
SCIP 9	2014-2016	Post-operative catheter removed within two days of surgery
SCIP VTE 1	2013-2014	Patients for venous thromboembolism (blood clots in veins) surgery received correct prophylactics
SCIP VTE 2	2013-2016	Patients for venous thromboembolism surgeries received anti-clotting treatment
SCIP Card 2	2013-2016	Surgery patients on beta-blockers pre-hospitalization given beta blockers during hospitalization
	Outcomes	
30-Day AMI	2014-2017	30 day survival rate for AMI discharges
30-Day HF	2014-2017	30 day survival rate for heart failure discharges
30-Day PN	2014-2017	30 day survival rate for pneumonia discharges

***** Denotes a clinical process that is considered unhealthy, better scores are lower.

Table 3. Percent of Hospitals using Achievement Score

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

Table 4. Clinical Process of Care Metric Lottery Around the Threshold and Benchmark with α = 0.05 Error Rate

Metric	Year	Total *Hospitals* n	Total Extrema $min_{i\in N}\hat{p}_i$, $max_{i\in N}\hat{p}_i$,	Average $z_{k,\alpha,n}$	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	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	3.83	0.95349	1,127	0.6814, 0.9967	**0.9455**	1	1,002	0.6900, 0.9967	**0.8460**
HF 1	2015	3,194	0.0239, 0.9980	4.06	0.94118	1,900	0.5000, 0.9892	**0.5949**	1	1,498	0.5500, 0.9980	**0.4690**
PN 3b	2015	3,290	0.1800, 0.9971	4.06	0.97783	3,002	0.5500, 0.9971	**0.9186**	1	2,647	0.6200, 0.9971	**0.8046**
PN 6	2015	3,655	0.1513, 0.9972	4.08	0.95918	3,203	0.5400, 0.9941	**0.8763**	1	2,706	0.5500, 0.9922	**0.7404**
PN 6	2016	3,304	0.1100, 0.9900	4.06	0.96552	3,225	0.5300, 0.9900	**0.8761**	1	3,165	0.5400, 0.9900	**0.9579**
SCIP 1	2015	2,793	0.1021, 0.9976	4.03	0.98639	2,436	0.7100, 0.9973	**0.8722**	1	2,086	0.7100, 0.9976	**0.7469**
SCIP 2	2015	2,777	0.0846, 0.9979	4.03	0.98637	2,457	0.7300, 0.9973	**0.8848**	1	2,280	0.7300, 0.9979	**0.8210**
SCIP 2	2016	2,176	0.5600, 0.9900	3.97	0.99074	2,146	0.6400, 0.9900	**0.9862**	1	2,146	0.6400, 0.9900	**0.9862**
SCIP 3	2015	3,054	0.1332, 0.9973	4.04	0.97494	2,406	0.6400, 0.9967	**0.7878**	1	1,872	0.6400, 0.9973	**0.6130**
SCIP 3	2016	2,717	0.1900, 0.9900	4.02	0.98086	2,645	0.5500, 0.9900	**0.9735**	1	2,574	0.5500, 0.9900	**0.9474**
SCIP 4	2015	1,137	0.6161, 0.9969	3.82	0.95798	938	0.7967, 0.9916	**0.8250**	0.99767	718	0.8152, 0.9969	**0.6315**
SCIP 9	2015	2,956	0.1317, 0.9973	4.04	0.94891	1,732	0.6400, 0.9903	**0.5859**	0.99991	1,447	0.6400, 0.9973	**0.4895**
SCIP 9	2016	2,586	0.2500, 0.9900	4.01	0.97059	2,411	0.5700, 0.9900	**0.9323**	1	2,249	0.5700, 0.9900	**0.8697**
SCIP Card 2	2015	2,771	0.1700, 0.9972	4.03	0.97175	2,344	0.5500, 0.9941	**0.8459**	1	2,027	0.5500, 0.9972	**0.7315**
SCIP Card 2	2016	2,347	0.0500, 0.9900	3.99	0.97727	2,303	0.5700, 0.9900	**0.9813**	1	2,265	0.5700, 0.9900	**0.9651**
SCIP VTE 2	2015	3,090	0.1694, 0.9977	4.05	0.97403	2,337	0.5800, 0.9946	**0.7563**	0.99998	1,761	0.5800, 0.9977	**0.5700**
SCIP VTE 2	2016	2,683	0.0800, 0.9900	4.01	0.98225	2.575	0.6400, 0.9900	**0.9597**	1	2,500	0.6400, 0.9900	**0.9318**

Simulation sample size 10,000.

Table 5. Outcome Metric Lottery Around the Threshold and Benchmark with α = 0.05 Error Rate

Metric	Year	Total *Hospitals* n	Total Extrema $min_{i\in N}\hat{p}_i$, $max_{i\in N}\hat{p}_i$,	Average $z_{k,\alpha,n}$	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	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	4.00	0.84747	2,404	0.4808, 0.9910	**0.9608**	0.86237	2,377	0.5037, 0.9910	**0.9500**
30-Day HF	2015	3,781	0.4200, 0.9961	4.09	0.88151	3,749	0.5160, 0.9915	**0.9915**	0.90032	3,724	0.5360, 0.9961	**0.9849**
30-Day PN	2015	4,191	0.3920, 0.9955	4.11	0.88165	4,176	0.5200, 0.9953	**0.9943**	0.90418	4,152	0.5480, 0.9955	**0.9907**

Simulation sample size 10,000.

Table 6. Outcome Metric Lottery Between the Threshold and Benchmark with α = 0.05 Error Rate

Metric	Year	p_l^*	p_u^*	Hospitals Between $n^* < n$	Between Extrema $min_{i\in N^*}\hat{p}_i$, $max_{i\in N^*}\hat{p}_i$,	Average z_{k,α,n^*}	Lottery Count	Lottery Extrema $min_{i\in S_{*\alpha}}\hat{p}_i$, $max_{i\in S_{*\alpha}}\hat{p}_i$,	Share in Lottery
30-Day AMI	2015	0.84747	0.86237	70	0.8476, 0.8620	3.21	70	0.8476, 0.8620	**1.0000**
30-Day HF	2015	0.88151	0.90032	184	0.8815, 0.9900	3.43	184	0.8815, 0.9900	**1.0000**
30-Day PN	2015	0.88165	0.90418	322	0.8817, 0.9041	3.55	322	0.8817, 0.9041	**1.0000**

Simulation sample size 10,000.

Table 7. Outcome Metric Naïve Lottery Around the Threshold and Benchmark with α = 0.05 Error Rate

Metric	Year	Total *Hospitals* n	Total Extrema $min_{i\in N}\hat{p}_i$, $max_{i\in N}\hat{p}_i$,	Average $z_{k,\alpha,n}$	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	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	1.6622	0.84747	650	0.7426, 0.8941	**0.2600**	0.86237	692	0.7679, 0.9046	**0.2768**
30-Day HF	2015	3,781	0.4200, 0.9961	1.6610	0.88151	1,003	0.7833, 0.9206	**0.2653**	0.90032	1,014	0.7217, 0.9345	**0.2683**
30-Day PN	2015	4,191	0.3920, 0.9955	1.6590	0.88165	1,292	0.7795, 0.9223	**0.3084**	0.90418	1,413	0.8157, 0.9371	**0.3372**

Critical values generated from t-distribution.

Table 8. Estimated Threshold and Benchmark with a α = 0.05 Error Rate

Metric	Year	Total n	p_l^*	$\hat{p}_l^*$	n_1	$\tilde{n}$	Average $z_{k,\alpha,n}$	p_u^*	$\hat{p}_u^*$	n_2	Above $\tilde{p}_u^*$	Average z_{k,α,n_1}	$\tilde{p}_u^*$	$\tilde{n}_2$	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.

Figure 1a. Distribution of Score AMI-8a (percutaneous coronary intervention Received within 90 minutes of arrival for heart attack patients) Over Time

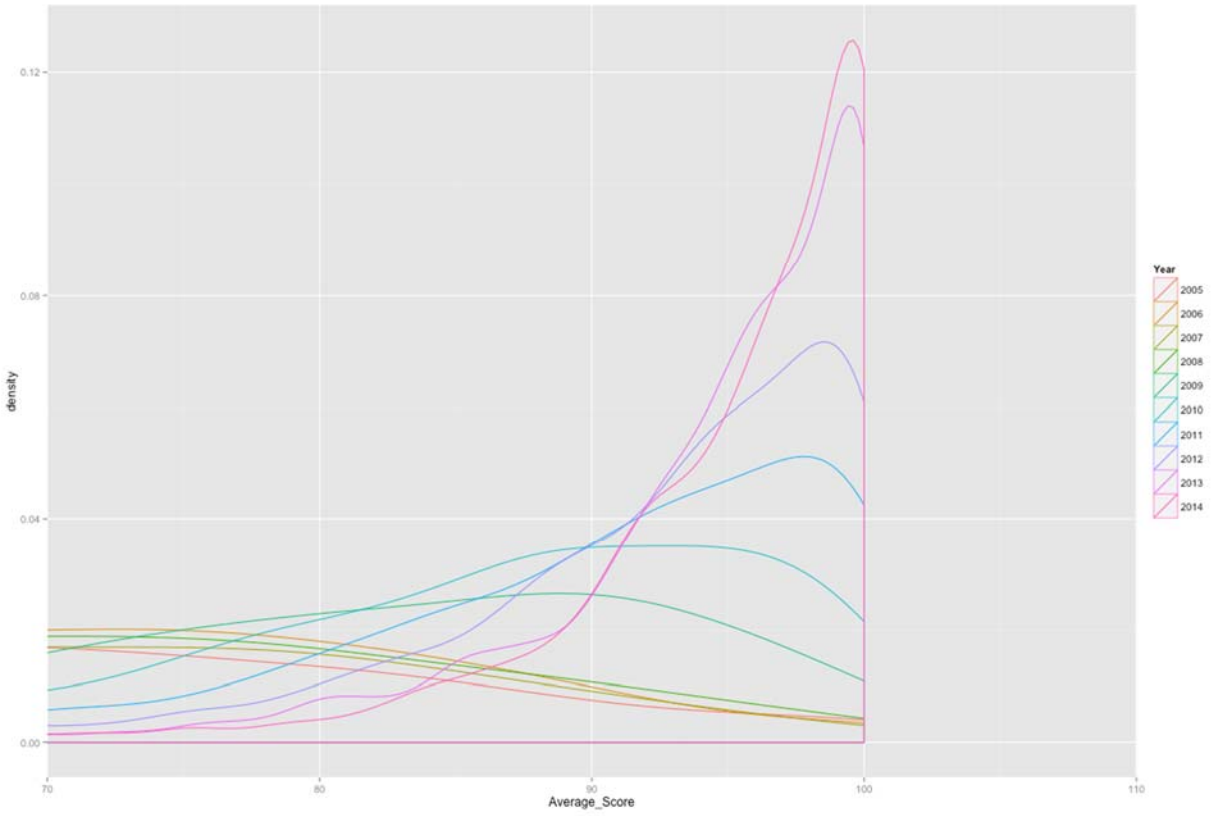

Figure 1b. Distribution of Score PN-3b (blood culture before first antibiotic given to pneumonia patients) Over Time

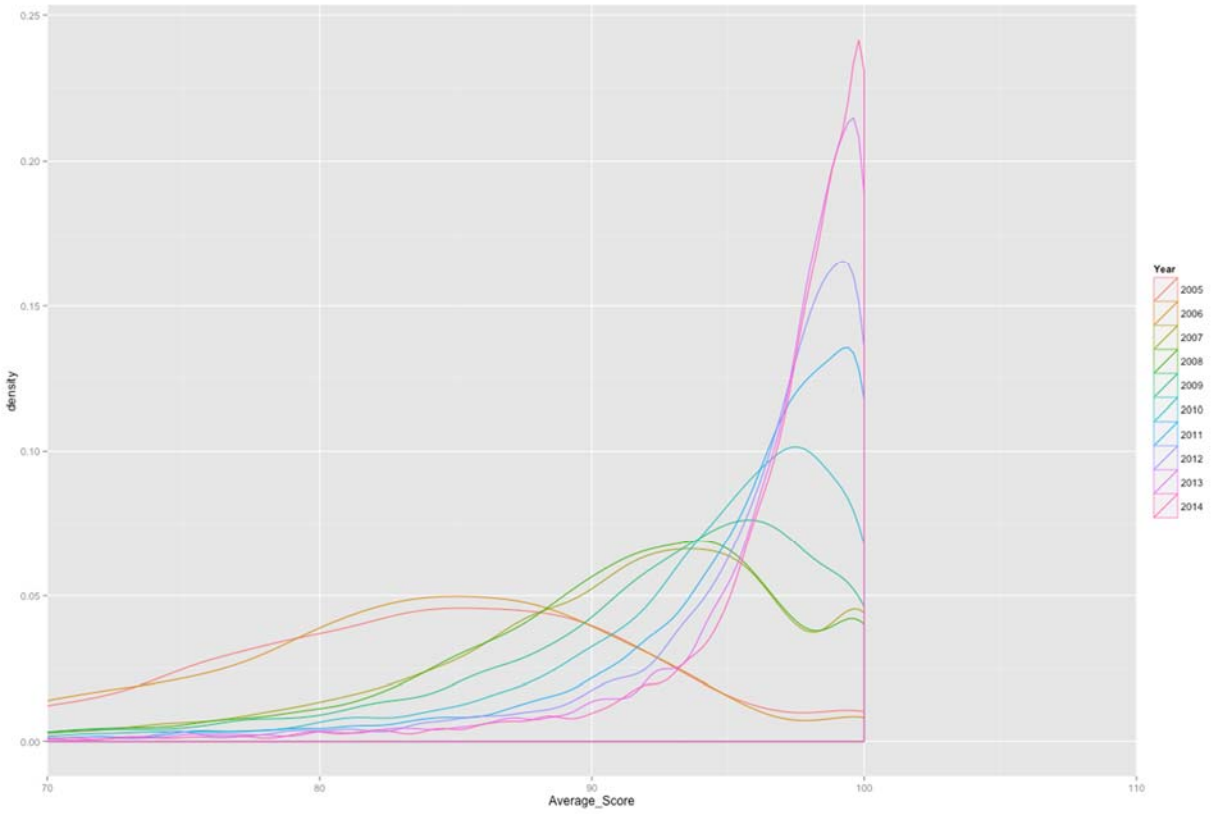

Figure 1c. Distribution of Score HF-1 (discharge instructions given to heart failure patients) Over Time