A Causal Tree Approach for Personalized Health Care Outcome Analysis

Guihua Wang
Stephen M. Ross School of Business
University of Michigan

Jun Li
Stephen M. Ross School of Business
University of Michigan

Wallace J. Hopp
Stephen M. Ross School of Business
University of Michigan

Ross School of Business Working Paper
Working Paper No. 1336
December 2016
Submitted to *Management Science*

Submitted to *Management Science*
manuscript

A Causal Tree Approach for Personalized Health Care Outcome Analysis

Guihua Wang
Ross School of Business, University of Michigan, guihuaw@umich.edu

Jun Li
Ross School of Business, University of Michigan, junwli@umich.edu

Wallace J. Hopp
Ross School of Business, University of Michigan, whopp@umich.edu

Using patient-level data from 35 hospitals for 6 cardiovascular surgeries in New York, we provide empirical evidence that outcome differences between health care providers are heterogeneous across different groups of patients. We then use a causal tree approach to identify patient groups that exhibit significant differences in outcome. By quantifying these differences, we demonstrate that a large majority of patients can achieve better expected outcomes by selecting providers based on patient-centric outcome information. We also show how patient-centric outcome information can help providers to improve their processes and payers to design effective pay-for-performance programs.

Key words: Health care, patient-centric, quality information, machine learning

1. Introduction

Choosing a health care provider for a major medical procedure can be literally a life or death decision. However, because they have historically lacked clear quality information about providers, most patients have made these important choices based on proximity or familiarity.[1] Even patients who have relied on physician referrals have been unable to rigorously evaluate their options, because the physicians themselves have lacked objective data and therefore have had to rely on subjective reputation information.

Recognizing the critical need among patients for more and better information about health care providers, government and private organizations have made various efforts to provide patient-oriented hospital ratings. For example, the Center for Medicare & Medicaid Services (CMS) maintains the Hospital Compare web site to compare Medicare-certified hospitals across the country and the US News provides aggregate hospital ratings for broad categories of procedures such as

[1] http://www.infographicsarchive.com/health-and-safety/2014-healthgrades-american-hospital-quality-report-nation/

heart surgery and cancer. These, and other rating systems like them, compare hospitals based on risk-adjusted rates of mortality, complication and/or readmission, and assign scores or star ratings to hospitals based on their outcome measures.

However, a widely overlooked reality is that these ratings are based on population averages (hereinafter referred to as "population-average information"), which imply that the same hospitals are best for all patients. But this is an assumption built into population-average based ratings, rather than an empirical fact. To illustrate how such ratings can be misleading, consider a simple example of three hospitals and two procedure types — Coronary Artery Bypass Grafting (CABG) and Mitral Valve Surgery. The mortality rates of these three hospitals are 1%, 4% and 2% for CABG patients, and 5%, 2% and 3% for mitral patients. If all three hospitals have a 50/50 mix of CABG and mitral patients, the overall mortality rates are 3%, 3% and 2.5%, respectively. If hospitals are ranked according to overall mortality rate, then the third hospital will come out on top, even though it is not the best for either procedure type. Hence a population average ranking on overall mortality rate will misguide patients (and their primary care physicians) in the choice of a hospital. By suggesting the same hospital for everyone, it will also contribute to a capacity imbalance.

In recognition that a hospital may perform well for some procedures and not as well for other procedures, some states such as New York and Pennsylvania have begun publishing hospital quality report cards for individual cardiac surgeries such as coronary artery bypass grafting, aortic valve and mitral valve surgeries. But this still does not provide true patient-centric information, because patients requiring the same procedure differ in their demographics and severity of illness (Huckman and Kelly, 2013). Hospital outcomes may be sensitive to these differences and the best hospital may be different for different patients.[2] In this paper, we focus on how to measure the heterogeneity of patient outcomes using readily available data, and how to use the results to generate patient-centric hospital ratings.

Patient-centric ratings have obvious use in helping individual patients choose a hospital. But they have other important uses as well. The US government is devoting considerable energy to designing payment structures that incentivize hospitals to improve quality. Most prominently, CMS has developed programs to link Medicare payments to hospital performance. For example, it

[2] For example, diabetic patients in need of coronary bypass surgery have generally not been treated using the Bilateral Internal Thoracic Artery (BITA) grafting technique, because of concerns that they are at higher risk of infection involving the breast bone. However, the Cleveland Clinic found recently that BITA grafting can work very well for diabetic patients, except for those that are very overweight with diffuse atherosclerosis or widespread hardening of the arteries (see https://health.clevelandclinic.org/2014/11/the-best-bypass-surgery-option-for-diabetic-patients/ for more details). Similarly, surgeons at the Greenville Health System have found that patients with end stage renal disease (ESRD) require special care because they are at a higher risk for complications and death after surgical procedures including bypass grafting (Schneider et al., 2009).

launched the Readmission Reduction Program (RRP) in 2013 to penalize hospitals with excessive 30 day readmission rates and the Hospital Acquired Conditions Reduction Program (HACRP) in 2015 to penalize low performers with regard to hospital acquired infections.[3] In both programs, if a hospital's performance is below a threshold, the hospital is penalized for all its Diagnosis-Related Groups (DRGs). In 2015, more than 2,000 hospitals were penalized under RRP and more than 700 hospitals were penalized under HACRP.

A problem with both RRP and HACRP is that they rely on population average data. As a result, they penalize some hospitals for all their procedures and do not penalize other hospitals for any procedure. As we noted above, low average performance does not necessarily mean that the hospital is poor at treating all patients. It is possible that some of the penalized hospitals have good or even excellent performance for some patients. Likewise, hospitals that are not penalized at all may be providing poor performance to some patients. The result is a misalignment between the penalties (or lack of them) and hospital performance, and hence misalignment in the incentives to improve. Using patient-centric ratings allows payers such as CMS to assess hospital quality by patient group and thereby direct penalties more accurately at areas of poor performance.

In this paper, we examine six cardiovascular surgeries at thirty-five NY hospitals and address four key questions: (1) Are the outcome differences between hospitals heterogenous across different patient groups? (2) How can we identify groups of patients that exhibit significant differences in outcome? (3) How can we quantify the differences in patient outcomes between hospitals in a (patient-centric) manner that is useful to individual patients? and (4) What are the benefits of patient-centric ratings to patients, payers and providers?

To answer the first question, we can partition patients into different groups according to their medical condition/procedure, as well as various patient characteristics such as age and comorbidities. For each group, we compare the outcomes of different hospitals and estimate the outcome differences between hospitals.

As mentioned earlier, various consumer-based hospital rating systems such as the New York State Cardiac Surgery Reporting System attempt to do this by comparing hospitals across different procedures. Table 1 summarizes the risk-adjusted mortality rates and the relative ranking of six hospitals for three cardiovascular surgeries based on New York Cardiovascular Surgery Quality Report Cards 2011-2013.[4] The results show clearly that outcome differences are indeed heterogenous across procedures. To see whether outcome differences are also heterogenous across other dimensions of patient characteristics, we need a way to group patients to generate patient-centric outcomes. This presents us with the second question.

[3] https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/index.html

[4] https://www.health.ny.gov/statistics/diseases/cardiovascular/

The standard approach for partitioning patients into groups would be to include interaction terms between hospital indicators and patient groups as covariates in a multivariate regression model. This method works well when there is a small number of groups, but quickly breaks down when, as is the case here, the number of patient characteristics is large. Methods such as LASSO can reduce the dimensionality of the problem, but rely on assumptions of sparsity and linear additivity, and impose distributions on the error term.

Table 1 Relative Performance of Hospitals for Different Procedures

Procedures		Lenox Hill Hospital	Mount Sinai	NYP-Columbia	NYP-Weill Cornell	Rochester General	St. Francis Hospital
Coronary Artery Bypass Grafting	Count	256	385	419	176	306	658
	Mortality	2.23%	1.80%	1.10%	1.74%	1.65%	1.54%
	Rank	6	5	1	4	3	2
Valve-Related Surgeries	Count	479	1820	2228	1303	1025	1831
	Mortality	3.30%	3.10%	2.88%	2.63%	4.91%	3.28%
	Rank	5	3	2	1	6	4
Percutaneous Coronary Intervention	Count	1551	4522	2541	1298	1569	2289
	Mortality	0.59%	0.92%	1.05%	1.50%	0.99%	0.82%
	Rank	1	3	5	6	4	2

Source: New York Cardiovascular Surgery Quality Report Cards 2011-2013.

These issues can be addressed by a nonparametric method that partitions patients into groups such that patients within the same group have similar outcome differences between providers. Unfortunately, while simple to state, it is not straightforward to find the best way to group patients. First, there are many patient characteristics to consider, so we need to identify those that affect provider outcome differences. Second, for a given set of patient characteristics, there are many different ways to group patients. To see this, consider a simple example with patients of two genders {male, female} and two races {white, black}. These can be grouped into {male, white}, {male, black}, {female, white} and {female, black}. Since the number of patient groups increases exponentially with the number of patient characteristics, real world settings will have too many groups to evaluate each one individually with statistically significant results.

In this study, we use tree-based methods from the machine learning literature to recursively partition patients into smaller groups such that patients within each group have similar characteristics. We compare the traditional regression tree method with the recently proposed causal tree method and explain why the causal tree method is better able to find heterogenous outcome differences between providers. However, we also note that the causal tree method was originally developed to identify binary treatment effects. To extend this approach to identify heterogeneous provider effects when there are multiple providers, we have to overcome two challenges. First, in addition to grouping patients, we also need to group providers because there may not be sufficient data to detect significant differences between all pairs of providers. Second, we need to derive from our groupings easy-to-understand outcome information for use by individual patients. Accomplishing the latter addresses the third key question of deriving patient-centric information by comparing outcomes of different providers.

To address the fourth key question of how patient-centric information can be used to improve patient outcomes, we compare scenarios in which patients use patient-centric and population-average information to select the best provider for them. This characterizes the magnitude of benefit to individual patients of having patient-centric, instead of population-average, data. We also illustrate the potential impact of patient-centric information on hospitals and payers to show how hospitals can use such information to target quality improvements and how Medicare can use it to better align payments with hospital performance.

2. Literature Review

There is growing interest in hospital quality from both the medical and operations management communities. The medical literature has focused primarily on identifying hospital characteristics that indicate better performance. For example, Keeler et al. (1992) compared 197 hospitals and found that teaching, large and urban hospitals are generally better than non-teaching, small, and rural hospitals for congestive heart failure, acute myocardial infraction, pneumonia, stroke or hip replacement. Birkmeyer et al. (2003), Gammie et al. (2009) and Vassileva et al. (2012) found high-volume hospitals tend to perform better than low-volume hospitals. Tsai et al. (2015) found that hospitals with boards that pay greater attention to clinical quality and use clinical quality metrics have more effective management practices and provide higher-quality care.

The operation management literature has taken a more detailed perspective by focusing on the impact of specific provider practices on performance. For example, Barro, Huckman and Kessler (2006), Clark and Huckman (2012), Huckman and Zinner (2008), and KC and Terwisch (2011) analyzed the impact of hospital specialization/focus on productivity and patient outcome; Clark, Huckman and Staats (2013), Huckman and Pisano (2006), KC and Staats (2012), KC et al. (2013) and Ramdas et al. (2014) analyzed the impact of related experiences on surgeon performance; Freeman et al. (2015), Jaeker and Tucker (2015) and Kim et al. (2015) analyzed the impact of workload on quality and patient outcome; Bavafa et al. (2013), Lu and Lu (2016), and Song et al. (2015) analyzed the impact of patient-physician communication, mandatory overtime laws and queue management on productivity and patient outcome.

A common assumption in both literatures is that the effects of quality driver are homogeneous across patient groups. Any study that gives a single ranking of providers or a single estimate of the impact of a practice on quality, regardless of patient group, is implicitly making this assumption. But a number of scholars have recognized the potential for this assumption to lead to inaccurate information to patients and have called for heterogeneous effect analysis in both patient care and quality assessment (see for example, FDA, 2013, Gerteis, 1993, IOM, 2011, Kattan and Vickers, 2004, Kent and Hayward, 2007, Kravitz et al., 2004). Wang et al. (2016) compared medical outcomes of mitral valve patients treated by surgeons at different hospitals and found heterogeous outcome differences across different patient groups.

Existing models that incorporate heterogeneity usually assume latent classes of consumers with different tastes or that consumer tastes are random draws from a known distribution. For example, Xu et al. (2016) used a random coefficient multinomial logit model to characterize heterogeneous patient preferences in choose doctors. Guajardo, Cohen and Netessine (2016) also used a random coefficient multinomial logit model to study the impact of service attributes on consumer demand in the US automobile industry. Lu et al. (2013) used a similar model to analyze how waiting in queue in the context of a retail store affects customers' purchasing behavior. While such modeling framework is useful in incorporating heterogeneous consumer preferences, they cannot systematically identify different combinations of characteristics that define heterogeneous consumer groups. As a result, it offers litter practical guidance to individual consumers.

The machine learning literature, on the other hand, offers several useful frameworks to measure heterogeneity and to identify heterogeneous groups. For example, a few studies have proposed methods to analyze the heterogeneous treatment effects. Evaluating patient differences in the effect of a single treatment (e.g., a clinical trial of a new drug) is similar, although not identical, to evaluating patient differences in the relative outcomes across a set of providers. Hence, we discuss the literature on identifying heterogenous treatment effects as a guide to addressing heterogenous provider effects.

In two separate studies of biological markers in high-dimensional genomic data, Signovitch (2007) and Tian et al. (2014) applied the standard LASSO procedure with modified outcomes or covariates to determine from a large set of biological markers the subset of patients that can potentially benefit from a treatment. Imai and Ratkovic (2013) modified the standard LASSO procedure using different penalty factors for the covariates and treatment effects to distinguish the effect of treatment from that of covariates and to allow for the possibility of treatment effects with small magnitudes. Since they do not systematically partition patients into groups, these methods require users to define patient groups a priori. All of them apply a single global model to all observations, and assume that effects are linearly additive and errors follow some distribution.

Realizing that a single global model can not be applied to all observations, Zeileis, Hothorn and Hornik (2008) proposed to partition the observations into groups and apply separate local models such as linear regression or maximum-likelihood based models to individual groups. They proposed using a tree-based method to partition observations, where the feature with the highest instability is used to split groups, with a fluctuation test to analyze the parameter stability at a node. Su et al. (2009) modified the regression tree method to split the predictor space in a way that maximizes the square of the t-statistic for testing the null hypothesis that the average treatment effect is the same in the two potential groups. A tuning parameter is used to penalize complex trees with many terminal nodes, where the value of the parameter is determined through cross-validation based on the sum of squares of the split t-statistics. These methods split the predictor space based on model fit or a test-statistic, and do not use cross-validation to select the tuning parameter or to assess the goodness of fit of the estimated model. Furthermore, by their design these methods are better suited to outcome prediction than to heterogenous treatment effect analysis.

Recently, Athey and Imbens (2016) proposed a causal tree method to analyze heterogeneous treatment effects in studies with binary treatments. This method effectively partitions subjects into groups with either large or small treatment effects. The same concept can be applied to analyze the heterogenous provider effect when there are two providers by interpreting one hospital as “treatment” and the other hospital as “control”. However, the causal tree method cannot be used directly when there are multiple providers, because it is unclear which provider or providers should be designated as the treatment or control groups. Moreover, while the causal tree method can be applied to each pair of providers, presenting such pairwise comparisons directly to patients is likely to be confusing since there may be hundreds of comparisons for a patient to process to come to a conclusion. In this study, we address all these issues in order to derive easy-to-understand patient-centric information on a set of providers.

We are not the first to apply machine learning techniques to the field of operations management. Exiting studies have developed and applied machine learning techniques for better prediction or decision-making. For example, Ang et al. (2015) developed a new method that combines queueing theory and the LASSO procedure to improve the prediction of emergency department waiting time. Bertsimas et al. (2016) used several machine learning methods (LASSO, random forest and support vector machines) to predict the outcomes of clinical trials and optimize the test regimes. Bastani and Bayati (2016) developed a new efficient multi-armed bandit algorithm based on the LASSO estimator to tailor decision-making at individual levels. They illustrated that superior performance of this algorithm in warfarin dosing. Ban et al. (2016) introduced performance-based regularization to improve portfolio performance. Ferreira et al. (2015) used a regression tree approach to predict demand and to optimized price, which led to 9.7% revenue increase in a field experiment implemented at an online retailer.

3. The Model

In this section, we first describe the needs and the challenges of generating patient-centric outcome information. We then introduce the regression tree and causal tree methods from the machine learning literature and discuss how to extend them to identify heterogeneous outcome differences between providers across patient groups.

3.1. Problem Description

The basic problem in which we are interested is identifying the provider, or set of providers, with the highest likelihood of providing a good outcome for a given patient. The data to us are the outcomes of prior patients at the various providers. However, because it is possible that outcomes are influenced by patient characteristics (e.g., age, comorbidities, etc.), prior patient outcomes are not equally relevant to the given patient. Patients with characteristics that match those of the given patient are more likely to be representative, than are patients with radically different characteristics. For instance, a 48-year old black woman with mitral valve disease and hypertension will probably get better information from outcomes of other middle aged mitral valve patients than she would from patients in their 90s with coronary artery disease.

While this insight is intuitive, it raises the important question of how similar a patient must be to provide useful information about likely outcomes. For example, are gender or race important? Or could the black female patient use outcomes from white male patients to help evaluate her options? Are only mitral valve patients relevant, or are patients with aortic valve disease also representative? Does hypertension matter? Or are outcomes from patients with other comorbidities, or no comorbidities, good indicators for our patient with hypertension? How much does age matter? Should our patient look only to outcomes for other 48 year olds, or should she consider patients within some wider window of ages? And so on. Ideally, a method for generating outcome information for a specific patient should also identify the cohort of patients from which this information should come.

The basic tradeoff involved in selecting a cohort is one of precision versus power. A very narrow cohort that closely matches the patient in question along all dimensions will be highly representative and hence precise in characterizing outcomes, but may be too small to offer statistical power needed to detect real and important differences between providers. A very broad cohort, which contains patients that may not resemble the patient in question, will be less precise in estimating outcomes but will have more power due to the larger sample size. The balance between precision and power should be struck endogenously by making use of the data itself.

Finally, a key characteristic of our problem is that we are seeking to characterize differences between provider outcomes. In contrast, most analyses focus on outcome prediction. The latter is relevant if a patient is choosing whether or not to receive a procedure. For example, to decide whether the risk of heart surgery is justified by the benefits, we need an estimate of the mortality rate from the procedure. However, once we have decided to receive a procedure and must decide on a provider, it is the difference in the mortality rates between the candidate providers that matters. In a deterministic world, where we know the absolute mortality rates, we can compute the differences via simple subtraction. But in a statistical world, where we can only estimate the rates, a method that focuses on prediction of the absolute rates may not yield the most accurate estimate of the differences between rates. We focus explicitly on estimating differences between providers, in the following discussion of regression and causal trees, and in the subsequent empirical analysis.

3.2. Regression Trees

The regression tree method partitions observations into smaller groups such that the outcomes within each group are similar to each other. A typical algorithm starts at the top of the tree, which consists of a single group called "parent group", and successively makes binary splits of groups based on the most important predictor. The process is repeated until a stopping criterion is met (for example, the incremental improvement in prediction accuracy or the number of observations in a group reaches a specified minimal level). The terminal nodes of a tree represent the final groups of observations that are expected to have similar outcomes.

Obviously, there is a tradeoff between prediction accuracy and tree complexity (number of terminal nodes). It is easy to see that a complex tree (e.g., each observation has its own group) will closely represent the data used to create the tree. Therefore, regression trees are generally evaluated according to their ability to predict a separate out-of-sample set of data. A complex tree will produce highly accurate in-sample predictions but may lead to poor out-of-sample predictions due to over-fitting. To formulate the process for creating a regression tree, we let N^{train} denote the number of observations in the training sample. We let M denote the number of terminal nodes, which are exhaustive and non-lapping. Finally, we let $L = \{l_1, .., l_M\}$ denote the M terminal nodes, and Y_i denote the observed outcome of patient i. A regression tree solves

$$
\begin{aligned}
&min \ \tfrac{1}{N^{train}}[\textstyle\sum_{j=1}^{M}\sum_{i\in l_j}(Y_i - \bar{Y}_{l_j})^2] + \alpha M \\
&s.t. \ L = \{l_1, ..., l_M\}; \\
&\quad\ \ l_i \cap l_j = \emptyset, \forall i \neq j.
\end{aligned}
$$

where $\bar{Y}_{l_j}$ is the average outcome of the jth terminal node, and α is the tuning parameter which penalizes complex trees. Given any α, one can solve the above optimization problem to minimize in-sample prediction error. One can also vary the value of α to minimize the out-of-sample prediction error over a number of cross-validation test samples: $\frac{1}{N^{test}}\sum_{i=1}(Y_i^{test} - \hat{Y}_i^{test})^2$, where Y_i^{test} and $\hat{Y}_i^{test}$ denote the true and predicted outcomes for patient i in the test samples.

The regression tree method is well-suited to estimation of absolute outcomes because it identifies important predictors of outcomes and partitions observations into groups with similar characteristics. However, a regression tree designed to achieve the best average out-of-sample predictions may not accurately characterize the relative differences between providers for different patient groups.

There are two potential approaches to modify the regression tree method to better serve our goal of providing patient-centric information for comparing providers. One approach is to include providers as predictors in the regression tree. If the tree does not split on any of the providers in a group, it means that all of the providers in that group have the same outcome. However, a shortcoming of this approach is that, if one or more patient characteristics (e.g., age or comorbidities) has a strong effect on outcomes, the tree will split first on these patient characteristics, leading to smaller groups in successive stages. Small samples may prevent the tree from splitting further even if outcomes differ between providers. For example, consider a simple case where Provider 1 is better than Provider 2 at treating male patients and equally good at treating young or old patients. The preferred tree should split only on gender. However, if age affects outcome more than gender does, the regression tree will split first on age and may not split further on gender.

The second approach is to fit two separate trees for the two providers using only patient characteristics as predictors. For a patient with given characteristics, we can calculate the outcome difference between providers using average outcomes of the corresponding terminal nodes to which the patient belongs. However, if a provider is too small, the regression tree will not split on any predictors, leading to a single terminal node that consists of all patients treated by that provider. Such an outcome may obscure patient characteristics that matter to outcomes. Furthermore, even when the trees split, the predictors that affect treatment outcomes may be different from those affect outcome differences. Hence, this approach may not yield appropriate patient-centric provider comparisons.

3.3. Causal Tree

Athey and Imben (2016) proposed a causal tree framework to analyze heterogenous treatment effects. Below, we first describe how an analogous approach can be used to identify heterogeneous provider effects when there are two providers, and then extend it to identify heterogenous provider effects when there are multiple providers.

3.3.1. Casual Tree with Two Providers The main difference between a causal tree and a regression tree is the objective function used to define splitting criterion. Recall that the objective of a regression tree is to predict outcomes, and therefore it splits on predictors in a way that minimizes out-of-sample mean squared errors across all groups. In contrast, the objective of a causal tree is to identify heterogeneous treatment effects, and therefore it splits on predictors in a way that maximizes the mean squared treatment effects across all groups. Let $D^{\pi}_{12}(x_l)$ denote the outcome differences between Provider 1 and Provider 2 for a group of patients with characteristics x_l, a causal tree π solves

$$\begin{aligned} &max\ \tfrac{1}{M}[\textstyle\sum_{l=1}^{M} D^{\pi}_{12}(x_l)^2] - \alpha M \\ &s.t.\ L = \{l_1, ..., l_M\}; \\ &l_i \cap l_j = \emptyset, \forall i \neq j. \end{aligned} \tag{1}$$

where α is the tuning parameter that controls the complexity of the tree.

To estimate D^{π}_{12}, we note that each patient can only be treated by one provider, so we cannot observe outcomes of both providers for a specific patient. Let $T_{ij} \in \{0,1\}$ indicate whether patient i was treated by provider $j \in \{1,2\}$. Let Y_{ij} indicate the outcome of patient i at provider j. For patients who are treated by Provider 1, we observe Y_{i1} but not Y_{i2}. Similarly, for patients who are treated by Provider 2, we observe Y_{i2} but not Y_{i1}. Therefore, D^{π}_{12} cannot be calculated by taking the differences of two potential outcomes for each patient. Instead, we estimate it using propensity score matching. Let $P(X_i)$ denote the propensity that patient i with characteristics X_i will be treated at Provider 1 and $1 - P(X_i)$ represent his/her propensity of being treated by Provider 2. Then, we can estimate provider outcome difference $D^{\pi}_{12}(x_l)$ using inverse probability weighting (Horvitz and Thompson, 1952),

$$D_{12}(x_l) = \frac{\sum_{i \in l, T_{i1}=1} Y_{i1}/P(X_i)}{\sum_{i \in l, T_{i1}=1} 1/P(X_i)} - \frac{\sum_{i \in l, T_{i2}=1} Y_{i2}/(1-P(X_i))}{\sum_{i \in l, T_{i2}=1} 1/(1-P(X_i))}$$

Similar to the regression tree method, the parameter α can be chosen through cross validation and the prediction accuracy can be evaluated using a goodness-of-fit measure on a testing set: $\frac{1}{N}\sum_{i=1}^{N}(D^{test}_{12}(X_i) - \hat{D}^{test}_{12}(X_i))^2$, where $D^{test}_{12}(X_i)$ denotes the true difference between Provider 1 and Provider 2 for patient i in the test set, and $\hat{D}^{test}_{12}(X_i)$ denotes the predicted outcome difference between the two providers for patients i in the test set. However, in contrast with a regression tree, where the outcome Y^{test}_i of a patient i in the test set is directly observable, the true outcome difference $D^{test}_{12}(X_i)$ cannot be observed. Therefore, one cannot calculate the mean squared errors in the test set directly.

To address this issue, Su et al. (2009) proposed an “honest” approach to construct unbiased estimates of mean squared errors using one sample to build the tree and an independent sample to estimate treatment effects. Let S^{train}, S^{est} and S^{test} denote training, estimation and testing samples respectively. Given any value of α, we first use the training sample to choose a tree structure that solves the maximization problem in (1). Given the tree structure, we then use the estimation sample to estimate the outcome difference between providers for patient i, i.e., $D^{est}_{12}(x_i)$. We therefore use

$D_{12}^{est}(x_i)$ from the estimation sample as our predicted difference for the training sample. The mean squared error to be minimized can be rewritten as

$$MSE(S^{test}, S^{est}) = \frac{1}{N}\sum_{i\in S^{test}} (D_{12}^{test}(X_i) - D_{12}^{est}(X_i))^2$$

The expected MSE is the expectation of $MSE(S^{test}, S^{est})$ over the test and estimation samples. By exploiting the equality $E(D_{12}^{test}(X_i)) = E(D_{12}^{est}(X_i)) = D_{12}^{\pi}(X_i)$ and observing that $E(D_{12}^{test}(X_i)^2)$ does not depend on the estimator, we have

$$\begin{aligned} EMSE(S^{test}, S^{est}) &= E_{S_{test},S_{est}} MSE(S^{test}, S^{est}) \\ &= -E_{S_{test}}[D_{12}^{\pi}(X_i)^2] + E_{S_{test},S_{est}}[Var(D_{12}^{est}(X_i))] \end{aligned}$$

In the second item, $Var(D_{12}^{est}(X_i))$ is the variance of estimated differences for the corresponding group (see Appendix A). The expected variance $E_{S_{test},S_{est}}[Var(D_{12}^{est}(X_i))]$ can be calculated as a weighted average of the group variances, where the weights are the fractions of observations (of the estimation sample) in the groups.

We can estimate the first term using the square of the estimated means in the training sample, $D_{12}^{train}(X_i)^2$, minus an estimate of its variance

$$\hat{E}_{S_{test}}[D_{12}^{\pi}(X_i)^2] = D_{12}^{train}(X_i)^2 - Var(D_{12}^{train}(X_i))$$

We thus have the expected MSE expressed as (see Appendix B),

$$EMSE(S^{test}, S^{est}) = -D_{12}^{train}(X_i)^2 + Var(D_{12}^{train}(X_i)) + E_{S_{test},S_{est}}[Var(D_{12}^{est}(X_i))]$$

Note that this estimate for EMSE is based on a given α. We can now vary the value of α to minimize expected mean squared error.

3.3.2. Causal Tree with Multiple Providers While it is straightforward to apply the causal tree method to analyze heterogeneous provider effects for two providers, we need to clear several hurdles to extend the method to multiple providers. Recall that the causal tree splits on predictors in a way that maximizes the mean squared treatment/provider effect (i.e., $\frac{1}{M}[\sum_{l=1}^{M} D_{12}^{\pi}(x_l)^2]$). When there are multiple providers, it is unclear which provider or set of providers should be considered as the treatment group and which as the control group. That is, eventually, we must partition providers, as well as patient groups. Note that the partitions of providers can be different for different patient groups and vice versa.

There are several competing alternatives for addressing this issue. Some of these require predefined provider groups, while others involve modifications of the objective function of the causal tree to accommodate differences of all pairs of providers. For instance, the causal tree method can be applied directly if a provider itself is considered as a group and all the other providers are considered as another group. We can build the causal tree using patient characteristics and a provider indicator as predictors. If the tree splits on the provider indicator, it indicates that the provider differs from the other providers as a group. We can estimate outcome difference between the provider and the other providers using the procedures discussed earlier. This approach may work well when the number of providers is relatively small and providers are of similar size. However, when the number of providers under comparison is large, the propensity of a patient to visit a single provider will be very small, while the propensity of patients visiting all the other providers will be very large, which makes the scenario unsuitable for propensity score matching (Crump et al., 2008). Even if there is no issue with matching, the derived outcome information can be confusing, because the baseline group changes as we move to compare another provider with its peers. As a results, a patient can not directly compare the outcomes of two providers when his/her choices of providers are limited.

An alternative is to modify the objective function. For instance, one can partition patients into groups such that, within each group, there is a large outcome variation across all providers. Then the objective function needs to be modified to $\frac{1}{M}[\sum_{l=1}^{M}\sum_{i\neq j} D_{ij}^{\pi}(x_l)^2]$, where $D_{ij}^{\pi}(x_l)$ captures the outcome difference between any pair of providers for patient group x_l. The major problem of this approach is that the groups differentiating one pair of providers may be different from those differentiating another pair of providers. Consider a simple example where Provider 1 is better than Provider 2 only for young patients and Provider 3 is better than Provider 4 only for male patients. The causal tree with above modified objective function is not suitable because it will result in a universal partition that is homogeneous across all provider pairs, and hence is not sensitive to the heterogeneous differences across provider pairs.

A solution to these issues is to apply the causal tree method to each pair of providers. While the approach is methodologically sound, it poses significant interpretation difficulties. For example, a patient considering 10 providers would have to examine 45 pairwise comparisons, which is likely to lead to confusion. To avoid this, we develop a two-stage approach. In the first stage, we analyze pairwise provider differences. In the second stage, we condense the results into a form that enables a patient to make direct comparisons between any provider and the state average. First, we estimate the outcome difference between a provider j and any of the other providers. To do this, we build $N-1$ causal trees using provider j and the other $N-1$ providers one at a time. From these trees, we can estimate the outcome differences between providers j and k for patient i, $D_{jk}(X_i), \forall j \neq k$. Second, we use the estimated results to derive patient-centric outcome information based on the outcome difference between each provider and the state average. To formalize this, we let $D_{j,SA}^{\Pi}(X_i)$

denote the difference between provider j and the state average of H providers from a set of causal trees Π,

$$\begin{aligned} D^{\Pi}_{j,SA}(X_i) &= E[Y_j(X_i) - \tfrac{1}{H}(Y_1(X_i) + Y_2(X_i) + ... + Y_H(X_i))] \\ &= \tfrac{1}{H}\textstyle\sum_{k \neq j} E[Y_j(X_i) - Y_k(X_i))] \\ &= \tfrac{1}{H}\textstyle\sum_{k \neq j} D^{\pi}_{jk}(X_i) \end{aligned}$$

Because we partition patients into groups based on the outcome differences between two providers, the groups we identify by comparing providers j and k may be different from those identified by comparing providers j and l. For example, if provider j is better than provider k at treating male patients but better than provider l at treating white patients, the causal trees will partition patients into {male, female} when comparing providers j and k and {white, non-white} when comparing providers j and l. However, as we will show later, this does not affect our estimation of outcome differences between provider j and the state average.

Because propensity score is defined as the probability of a patient being treated by one provider as opposed to another, a patient may have different propensity scores when we compare the same provider with different alternatives. Let $P_j(X_i), P_k(X_i), P_l(X_i)$ denote the unconditional probabilities of patient i going to providers j, k and l respectively. Let $P_{jk}(X_i) = Pr(T_{ij} = 1 | X_i, T_{ij} + T_{ik} = 1)$ denote the probability of patient i being treated by provider j given that he/she is treated at either j or k. Assuming the probability of being treated by a given provider can be modeled using a multinomial logit model, then we have

$$\begin{aligned} P_{jk}(X_i) &= P_j(X_i)/(P_j(X_i) + P_k(X_i)) \\ P_{jl}(X_i) &= P_j(X_i)/(P_j(X_i) + P_l(X_i)) \end{aligned}$$

These equations hold as a result of the Independence of Irrelevant Alternatives (IIA) property. Let l_{jk} denote the terminal node that includes patient i in a causal tree built for providers j and k. For a given matrix of propensity scores, $P(X)$, the proposed estimator of $D^{\prod}_{j,SA}(X_i)$ is

$$\begin{aligned} D^{\prod}_{j,SA}(X_i | P(X)) &= \tfrac{1}{H}\textstyle\sum_{k \neq j} D^{\pi}_{jk}(X_i) \\ &= \tfrac{1}{H}\textstyle\sum_{k \neq j}\left(\frac{\sum_{i \in l_{jk}, T_{ij}=1} Y_{ij}/P_{jk}(X_i)}{\sum_{i \in l_{jk}, T_{ij}=1} 1/P_{jk}(X_i)} - \frac{\sum_{i \in l_{jk}, T_{ik}=1} Y_{ik}/(1-P_{jk}(X_i))}{\sum_{i \in l_{jk}, T_{ik}=1} 1/(1-P_{jk}(X_i))}\right) \end{aligned}$$

It is straightforward to see that $E[D_{j,SA}(X_i)|P(X)] = \frac{1}{H}\sum_{k\neq j} E[Y_j(X_i) - Y_k(X_i)] = D^{\Pi}_{j,SA}(X_i|P(X))$. We can estimate the variance of $D^{S}_{j,SA}(X_i|P(X))$ as follows

$$
\begin{aligned}
Var[D^{\Pi}_{j,SA}(X_i|P(X))] &= Var[\tfrac{1}{H}\textstyle\sum_{k\neq j}(\frac{\sum_{i\in l_{jk},T_{ij}=1} Y_{ij}/P_{jk}(X_i)}{\sum_{i\in l_{jk},T_{ij}=1} 1/P_{jk}(X_i)} - \frac{\sum_{i\in l_{jk},T_{ik}=1} Y_{ik}/(1-P_{jk}(X_i))}{\sum_{i\in l_{jk},T_{ik}=1} 1/(1-P_{jk}(X_i))})] \\
&= \tfrac{1}{H^2}(\textstyle\sum_{k\neq j} Var[D^{\pi}_{jk}(X_i|P(X))] \\
&\quad + \textstyle\sum_{k\neq j}\sum_{l\neq j} Cov[\frac{\sum_{i\in l_{jk},T_{ij}=1} Y_{ij}/P_{jk}(X_i)}{\sum_{i\in l_{jk},T_{ij}=1} 1/P_{jk}(X_i)}, \frac{\sum_{i\in l_{jl},T_{ij}=1} Y_{ij}/P_{jl}(X_i)}{\sum_{i\in l_{jl},T_{ij}=1} 1/P_{jl}(X_i)}) \\
&= \tfrac{1}{H^2}(\textstyle\sum_{k\neq j} Var[D^{\pi}_{jk}(X_i|P(X))] \\
&\quad + \textstyle\sum_{k\neq j}\sum_{l\neq j} \frac{Cov[\sum_{i\in l_{jk},T_{ij}=1} Y_{ij}/P_{jk}(X_i), \sum_{i\in l_{jl},T_{ij}=1} Y_{ij}/P_{jl}(X_i)]}{\sum_{i\in l_{jk},T_{ij}=1} 1/P_{jk}(X_i)\sum_{i\in l_{jl},T_{ij}=1} 1/P_{jl}(X_i)} \\
&= \tfrac{1}{H^2}(\textstyle\sum_{k\neq j} Var[D^{\pi}_{jk}(X_i|P(X))] \\
&\quad + \textstyle\sum_{k\neq j}\sum_{l\neq j} \frac{\sum_{i\in l_{jk},i\in l_{jl}} Var(Y_{ij})}{\sum_{i\in l_{jk},T_{ij}=1} 1/P_{jk}(X_i)\sum_{i\in l_{jl},T_{ij}=1} 1/P_{jl}(X_i)}
\end{aligned}
$$

Similar to what we did for two providers, we can use the law of iterated expectations and the total law of variance to estimate the mean and variance of $D_{j,SA}(X_i)$. A provider is statistically significantly better (or worse) than the state average if the t-statistic of $D_{j,SA}(X_i)$ is smaller (or larger) than the critical values.

Before concluding, we note that the causal tree method is superior to a multivariate regression model for our study, because the regression model assumes that the effects are linearly additive and unobservable errors follow a certain distribution (e.g., normal distribution for linear regression and probit models, and logistic distribution for the logit model). Hence, multivariate regression is not suitable for studying heterogeneous (i.e., nonlinear) effects with multiple patient characteristics, because the number of parameters becomes very large once we include the full set of interaction effects. For example, when there are 10 patient characteristics and 35 providers, a total of $2^{10} \times (35-1) = 34,816$ parameters are required. Methods such as LASSO can reduce the dimensionality, but they still rely on the assumptions of linear additivity, sparsity and distribution of the errors. In contrast, the causal tree method is a non-parametric approach, which makes no assumption on the errors and allows the predictors to interact in a more flexible and sophisticated manner.

4. Empirical Setting and Data

We choose cardiovascular diseases (commonly known as heart diseases) as the empirical setting for personalized health care outcome analysis for several reasons. First, cardiovascular diseases are the leading cause of death worldwide (WHO, 2011). Each year, about 17.5 million people die from cardiovascular diseases, which accounts for 1 in every 4 deaths, and this number is expected to grow to more than 23.6 million by 2030.[5] Second, cardiovascular surgeries are relatively complicated procedures. They require sophisticated skills, advanced technology and intensive post-surgical care, which makes them candidates for sizable variations across providers (hospitals or surgeons). Third, cardiovascular surgeries include several different types of procedures, each requiring a different set of skills and technology. As a result, a hospital may perform well for some procedures but not as well for others.

Cardiovascular diseases refer to (a) conditions when the blood vessels are narrowed or blocked, which can lead to heart attack, (b) chest pain or stroke and (c) conditions that affect the heart's muscles, valves or rhythm. Cardiovascular surgeries are operations performed by surgeons on the heart and blood vessels to repair the damage caused by diseases or disorders of the cardiovascular system. In this study, we focus specifically on three cardiac surgeries — Mitral Valve Replacement (MVR), Aortic Valve Replacement (AVR) and Coronary Artery Bypass Grafting (CABG), and three vascular surgeries — Abdominal Aortic Aneurysm (AAA) repair, Carotid endarterectomy (CE) and Lower Extremity bypass Graft (LEG).

4.1. Data Description and Preparation

Our study makes use of data from New York state that consist of patient-level records of all in- and out-patient discharges from all hospitals in New York from 2008-2012. The data contain detailed clinical and resource use information, including admission status (e.g., elective, emergent and urgent), patient demographics and comorbidities, hospital identifiers, and principal and secondary diagnoses. For each discharge, the data indicate the type of surgery a patient underwent. They also record whether a patient experienced any complications or died during the procedure or post-surgery hospitalization. Finally, we identify readmissions by linking inpatient and outpatient data.

We identify discharges related to the six cardiovascular procedures under this study by using related clinical codes in the International Classification of Disease (9th revision). From 2008-2012, a total of 124,895 patients with cardiovascular diseases were discharged from 144 hospitals. Because some of the hospitals did not perform cardiovascular surgeries every year or have a low volume, we focus on the 41 cardiac hospitals compared by the New York State of Health for Cardiovascular Surgery Quality Report Cards. However, six of these hospitals did not perform vascular surgeries, so we focus on the other 35 hospitals that perform all the six cardiovascular surgeries discussed earlier. This results in a total of 107,252 discharges over the five year period. We focus on isolated surgeries and exclude patients who underwent multiple types of surgeries (6,950 of the sample). This allows us to characterize patient outcomes at each hospital for each surgery type. In addition,

[5] https://www.heart.org/idc/groups/ahamah-public

we exclude patients with missing information such as admission status. The final sample contains a total of 99,378 discharges.

4.2. Outcome Measures and Feature Space

To measure a hospital's outcome quality, we consider the rates of complication, readmission and mortality as potential metrics. We identify complications using the diagnosis codes provided in the data and focus on hospital acquired conditions rather than pre-existing conditions. We are able to separate the two types of complications because the data indicate whether each diagnosis was present at admission. We focus on 23 cardiovascular surgery related complications[6] and use them collectively as an outcome measure (STS, 2016, Tuinen et al., 2005, Williams et al., 1965).

In our sample, 29.58% patients had at least one of the 23 complications, while 10.55% had two or more complications. Because a sizeable number of patients had more than one complications, we cannot simply use a binary variable to indicate whether a patient experienced at least one complication. The 23 complications have different severity levels. For example, complications such as pulmonary embolism or insufficiency are relatively easy to cure, while complications such as coma and multi-organ failure are likely to lead to patient deaths (Glance et al., 2007, Reddy et al., 2013). Therefore, we cannot simply count the number of complications a patient experienced. To capture both the number and the severity of complications associated with a patient during the surgery and hospital stay, we need to translate complications into a numeric score that weights each complication by its severity.

The Elixhauser comorbidity index is a vector of 30 binary variables in which each 1 represents the existence of a comorbidity (Elixhauser et al., 1998). To describe the overall sickness of a patient and to weight the severity of individual comorbidities, van Walraven et al. (2009) modified the Elixhauser comorbidity index into a single numeric score (called "Elixhauser comorbidity score") by using a backward stepwise multivariate logistic regression to determine the correlation between each comorbidity and in-hospital mortality. The parameter estimates of the regression model were modified into a vector of weights based on methods described by Sullivan et al. (2004). The Elixhauser comorbidity score is calculated as the dot product of the index vector and the vector of weights. We follow the same approach to develop a complication score as an outcome measure for the purpose of this study. The complications and their weights are summarized in Appendix C. The average complication score for each procedure in our study ranges from 0.11 (for CE) to 1.65 (for AAA) and the average across all procedures is 0.68 (Table 2).

[6] The complications are stroke, aortic dissection, renal failure, ventilation, multi-organ failure, coma, cardiac arrest, sepsis, gastrointestinal events, tracheal reintubation, surgical complications, tamponade, wound infection, renal dialysis, mediastinum, reoperation for bleeding, pneumonia, pulmonary embolism, heart block, myocardial infarction, pulmonary insufficiency, surgical E codes and other cardiac complications.

To analyze readmission rate, we merge outpatient discharge data with inpatient discharge data using the link provided by the Agency for Healthcare Research and Quality. We focus on 30-day readmission by identifying patients who visited the same or other hospitals within 30 days after discharge. The last month of the data is censored because their re-admissions are not observed. The average readmission rate for each procedure ranges from 10.3% (for CE) to 18.1% (for LBG) and that for all procedures equals 14.7%. Lastly, we observe directly from the data whether a patient died during hospitalization. The average mortality rate for each procedure ranges from 0.3% (for CE) to 4.7% (for MVR) and the average across all procedures is 1.7%.

Table 2 Summary of Outcomes for Different Procedures

Surgical	Procedure	Complication Score		Readmission Rate		Mortality Rate	
name	count	mean	s.d.	mean	s.d.	mean	s.d.
CE	14,539	0.11	0.77	10.3%	30.4%	0.3%	5.8%
CABG	46,098	0.66	1.80	14.0%	34.7%	1.2%	10.9%
LBG	12,227	0.41	1.47	18.1%	38.5%	1.5%	12.2%
AAA	1,356	1.65	2.86	11.6%	32.1%	2.7%	16.3%
AVR	20,061	0.99	2.30	16.4%	37.0%	2.9%	16.7%
MVR	5,097	1.47	2.80	17.0%	37.6%	4.7%	21.3%
Total	99,378	0.68	1.90	14.7%	35.4%	1.7%	12.9%

The features we use to construct the causal trees include six cardiovascular procedures (CE, CABG, LBG, AAA, AVR and MVR), patient genders, races (white, black, hispanic, asian, native and others), admission statuses (emergent, urgent and elective), six age groups (below 50, 50-60, 60-70, 70-80, 80-90 and above 90) and five major comorbidities (chronic heart failure, chronic lung disease, diabetes, hypertension and renal failure) of cardiovascular diseases (STS, 2016). Considering all these features results in a total of 6 procedures × 2 genders × 6 races × 3 admissions × 6 ages × 2^5 comorbidities = 41,472 different combinations of patient characteristics.

5. Results and Discussion

To address the first three key questions we raised in the Introduction, we first provide evidence of heterogeneous outcome differences between hospitals using an exploratory approach in which patient groups are defined a priori (for example, by procedure type, age group or comorbidity). Then we apply the regression and causal tree methods to systematically partition patients and discuss why the causal tree method is better able to detect the outcome differences between hospitals. Finally, we extend the causal tree method to compare multiple hospitals and identify hospitals that are statistically significantly better than the state average for each patient.

5.1. Evidence of Heterogeneous Outcome Differences

To evaluate the extent to which outcome differences between hospitals are indeed heterogenous across groups of patients, we partition patients into groups according to patient characteristics (e.g., procedure, age and comorbidities) and use t-tests to see whether one hospital is significantly different from another for each group. Table 3 summarizes the results for two hospitals of similar size in New York when patients are partitioned by procedure type, age group and/or comorbidities.

The first partition compares the hospital outcomes for the procedures CE, AVR and CABG. This shows that Hospital 1 has a lower complication score for CE (0.05 lower, p-value< 0.1), but a higher complication score for both AVR (0.40 higher, p-value< 0.01) and CABG (0.40 higher, p-value< 0.01). The results suggest that the outcome differences between providers are heterogeneous across procedures.

Note, however, that the numbers in the first partition in Table 3 are simple averages for different patient groups and are not risk adjusted. To make sure the observed heterogeneity is not an artifact of different patient mixes, we examine some finer partitions based on patient age and comorbidities. The second partition in Table 3 compares the same procedures but focuses on patients in their 70s. This shows that the differences between the two hospitals are still significant for AVR and CABG. However, Hospital 1 has a lower complication score than Hospital 2 for AVR patients in their 70s. The difference between the two hospitals is insignificant for CE. This insignificance can be due in part to the reduction of the sample size. For both AVR and CABG, the magnitudes of the differences between the two hospitals are larger for patients in their 70s than for patients of all ages.

The third partition in Table 3 focuses on CE and patients of different ages (i.e., 60s, 70s and 80s). This shows that Hospital 1 has a higher complication score for patients in their 60s but has a lower complication score for patients in their 80s. The difference between the two hospitals for patients in their 70s is not statistically significant. These results indicate that outcome differences are heterogeneous across patient age groups.

The fourth partition in Table 3 focuses on CE and different comorbidities (i.e., diabetes, lung disease or heart failure). This shows that Hospital 1 has a lower complication score for patients with diabetes but a higher complication score for patients with lung disease. The difference is not significant for patients with hypertension. These results suggest that outcome differences are heterogeneous across patient comorbidities.

To summarize, Table 3 suggests that outcome differences between Hospital 1 and 2 are heterogeneous across procedure types, patient age and comorbidities, which speaks to the first question raised in the Introduction. It also illustrates the tradeoff between precision and power. As we see from the fifth partition in Table 3, fine partitions pose the risk of small sample sizes that are inadequate for statistical testing. Of course, these partitions are for illustration purpose only and there are many other ways to partition patients for outcome comparison. In the subsequent section, we employ tree-based methods to obtain the optimal partition to detect heterogeneous outcome differences between providers.

Table 3 Complication Scores at Two Hospitals for Various Patient Groups

Partition Alternatives	Procedure Type	Age Group/ Comorbidity	Hospital 1 count	Hospital 1 mean	Hospital 1 s.e.	Hospital 2 count	Hospital 2 mean	Hospital 2 s.e.	Difference in Average Complication Score
1	CE	all	489	0.07	0.02	481	0.12	0.03	−0.05∗
	AVR	all	200	0.83	0.16	385	0.43	0.08	0.40 ∗ ∗∗
	CABG	all	891	0.40	0.05	969	0.23	0.03	0.17 ∗ ∗∗
2	CE	70	147	0.05	0.03	158	0.06	0.03	−0.02
	AVR	70	46	0.07	0.07	97	0.53	0.17	−0.46 ∗ ∗∗
	CABG	70	307	0.33	0.08	327	0.15	0.04	0.18 ∗ ∗
3	CE	60	58	0.12	0.08	47	0	0	0.12∗
	CE	70	147	0.05	0.03	158	0.06	0.03	−0.02
	CE	80	209	0.06	0.03	196	0.14	0.05	−0.09∗
4	CE	diabetes	158	0.08	0.04	197	0.17	0.06	−0.10∗
	CE	lung disease	101	0.17	0.06	197	0.06	0.03	0.11∗
	CE	heart failure	19	0.53	0.29	29	0.28	0.24	0.25
5	CE	60/diabetes	49	0.41	0.17	66	0.18	0.12	0.23
	CE	60/lung disease	29	0.17	0.14	30	0.1	0.1	0.07
	CE	60/heart failure	3	1.00	1.00	6	0.50	0.50	0.50

*** p < 0.01, ** p <0.05, * p < 0.1

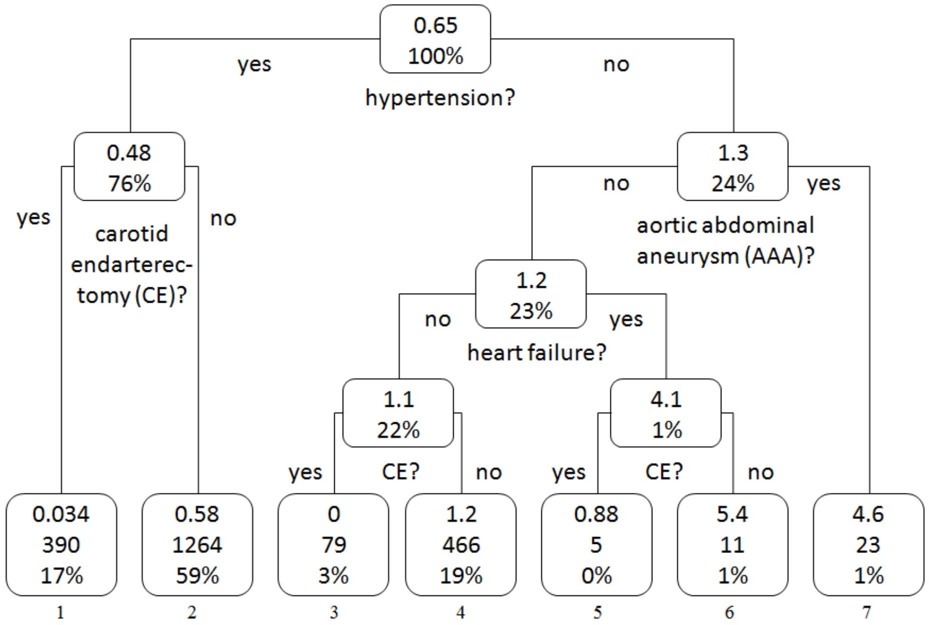
5.2. Comparison of Causal and Regression Trees With Two Providers

To address the second key question of how to identify patient groups that exhibit significant outcome differences, we make use of the statistical methods we presented earlier.

We start with the traditional regression tree method using patient characteristics and a hospital indictor (Hospital 1) as predictors. From Figure 1, we see that the regression tree splits first on hypertension, which indicates that hypertension is the most important factor affecting outcomes. For patients with hypertension, it splits on CE only. But for patients without hypertension, it splits on AAA, chronic heart failure and CE. At the bottom of the tree, there are seven terminal nodes representing seven distinct groups of patients. The numbers in a terminal node indicate the average complication score, the total number and the fraction of patients in the node. Finally, we note that the regression tree does not split on Hospital 1, which indicates that the choice of Hospital 1 over Hospital 2 is not an important determinant of outcomes for any of the patient groups.

Next we apply the causal tree method to the same patients treated at these two hospitals. Our objective is to detect significant differences in complication scores between Hospital 1 and Hospital 2. From Figure 2, we see that the causal tree splits first on CABG, which indicates that CABG is the

Figure 1 Regression Tree Approach



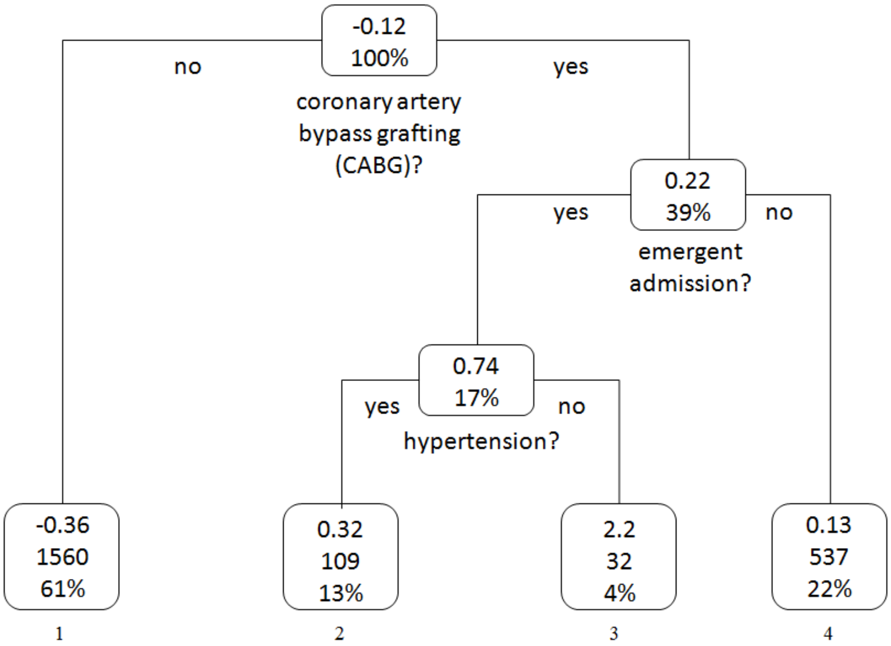
most important factor differentiating outcomes at the two hospitals. For CABG patients, the tree splits on emergent admission and hypertension, but for patients undergoing other types of surgery, the tree does not split at all. This non-splitting is unlikely to be due to a small sample, because the node (non-CABG) includes 61%, or 1560 patients. At the four terminal nodes, a positive value indicates that Hospital 1 has a higher complication score and a negative value indicates otherwise. Compared with the regression tree, the causal tree is smaller and partitions patients in a markedly different way.

Finally, we compare the two trees to analyze which splitting method allows us to detect heterogenous outcomes differences between the two hospitals. From each tree, we first identify patients from the two hospitals at a terminal node and calculate the average complication scores. We then use t-tests to determine wether the two hospitals have statistically significantly different complication scores for the group of patients at the terminal node.

Table 4 shows that the regression tree partitions patients into groups of sizes ranging from 8 (the 5th node) to 2,012 (the 2nd node). Within each terminal node, the average outcome of Hospital 1 is close to that of Hospital 2 and, as a result, the differences for all seven groups are small (from -0.13 to 0.30). More importantly, the differences are not statistically significant for six of the seven groups at conventional significance levels.

Table 4 shows that the causal tree partitions patients into groups of sizes ranging from 153 (the 3rd node) to 2,092 (the 1st node). Within each terminal node of the causal tree, the average outcome of Hospital 1 is very different from that of Hospital 2. The outcome differences of the four groups range from -0.36 to 2.24 and three of the four differences are significant at conventional levels.

Figure 2 Causal Tree Approach



As expected, the causal tree partitions patients in a way that maximizes the outcome differences between hospitals for groups of patients, whereas the regression tree partitions patients in a way that minimizes the outcome differences among patients.

To obtain additional insight into the differences between causal and regression trees, we compare groups of patients at the two hospitals after the trees' initial splitting. The regression tree begins by partitioning patients into those with and without hypertension. The complication scores are 0.48 and 1.3 for patients with and without hypertension, respectively, so the difference is -0.82. For those with hypertension, the complication scores are 0.4 and 0.57 for Hospitals 1 and 2, respectively, so the difference is -0.17. For those without hypertension, the complication scores are 1.33 and 1.21 for Hospitals 1 and 2, respectively, so the difference is 0.12. Comparing magnitudes of these differences, we see that hypertension is important in predicting patient outcomes but not as important in predicting outcome differences between hospitals.

In contrast, the causal tree splits first on the procedure of CABG. The complication scores are 0.75 and 0.59 for CABG and non-CABG patients, respectively, so the difference is 0.16. For CABG patients, the complication scores are 0.86 and 0.63 at Hospitals 1 and 2, respectively, so the difference is 0.23. For those undergoing other types of surgeries, the complication scores are 0.50 and 0.86 at Hospitals 1 and 2, respectively, so the difference is -0.36. Comparing magnitudes of these differences, we see that CABG is important in predicting outcome differences between hospitals, but is not as important in predicting patient outcomes. Since our goal is to identify provider differences that matter to patients, the causal tree is more useful to our purpose.

Table 4 Comparison of Causal Tree And Regression Tree

	Node Index of	Hospital 1			Hospital 2			Difference in Average
	Respective Trees	count	mean	s.e.	count	mean	s.e.	Complication Score
Regression Tree	1	390	1.29	0.15	271	1.11	0.16	0.18
	2	1,264	0.53	0.04	748	0.66	0.07	$-0.13*$
	3	79	0.00	0.00	18	0.00	0.00	0.00
	4	466	0.04	0.02	126	0.02	0.02	0.01
	5	5	0.80	0.80	3	1.00	1.00	-0.20
	6	11	5.55	1.78	8	5.25	1.47	0.30
	7	23	4.65	1.00	5	4.60	3.03	0.05
Causal Tree	1	1,560	0.50	0.05	532	0.86	0.10	$-0.36***$
	2	109	0.88	0.17	327	0.56	0.09	$0.32**$
	3	32	3.16	0.87	121	0.92	0.20	$2.24***$
	4	537	0.72	0.08	199	0.59	0.14	0.13

*** p < 0.01, ** p <0.05, * p < 0.1

5.3. Causal Tree for Multiple Providers

As described in Section 3, to identify hospitals that are statistically significantly different from the state average for certain patient groups, we first construct causal trees for each pair of hospitals, which requires a total of $35 \times 34/2 = 595$ trees. For each patient, we estimate the differences in complication score between a hospital and the state average, and calculate the standard error of the difference using the approach of Section 3. Table 5 summarizes the results for an example of six different patients. The best hospital for each patient is highlighted in bold. We observe that, while some hospitals (e.g., hospitals 3 and 4) are uniformly better than the state average for all six patients, others (e.g., hospital 26) are uniformly worse. However, for hospitals that are uniformly better (or worse) than the state average, the magnitude of the differences varies for individual patients. For example, Hospital 3 is better than the state average by 0.74 for the 2nd patient (AVR, 80s, one comorbidity) and by 0.19 for the 3rd patient (CE, 70s, two comorbidities). There are also hospitals that are better than the state average for some patients but worse for others. For example, Hospital 1 is better for the 3rd (CE, 70s, two comorbidities) and 5th (MVR, 30s, two comorbidities) patients but worse for the 2nd(AVR, 80s, 1 comorbidity), 4th (CABG, 40s, one comorbidity) and 6th (AAA, 60s, two comorbidities) patients. These results indicate that outcome differences between pairs of hospitals are indeed heterogenous across patients, and that different patients have different sets of hospitals that are significantly better that the state average.

Of course, Table 5 only shows six patients as examples. We have analyzed the outcome differences across hospitals for all of the patients this study. To provide an overall visual illustration of the heterogeneity in outcomes across hospitals for different patients, we group patients by procedure type, age group and comorbidities.[7] For each patient group, we use $Y_{ijk} \in \{-1, 0, 1\}$ to indicate

[7] We tried different ways to group patients and noticed that, when patients are grouped by procedure type, comorbidities and age group, the resulting heat map has obvious patterns. Patients within each group may have different

Table 5 Comparison of Complication Score with the State Average for Different Patients

Hospital	LBG, 70s	AVR, 80s	CE, 70s	CABG, 40s	MVR, 30s	AAA, 60s
Index	1 Comorb	1 Comorb	2 Comorb	1 Comorb	2 Comorb	2 Comorb
	(1)	(2)	(3)	(4)	(5)	(6)
1	**−0.37**+++	−0.59+++	−0.32+++	**−0.35**+++	**−0.62**+++	**−0.39**+++
2	−0.25+++	**−0.74**+++	−0.19+++	−0.24+++	−0.48+++	−0.28+++
3	−0.20+++	−0.44+++	−0.35+++	−0.25+++	−0.23++	−0.21+++
4	−0.17++	−0.43+++	**−0.41**+++	−0.17++	−0.17+	−0.24+++
5	−0.25+	−0.65+++	−0.05	−0.09	−0.05	−0.10
6	−0.04	−0.53+++	−0.16++	−0.03	−0.13	0.02
7	−0.04	−0.23	−0.07	0.02	−0.17	−0.04
8	−0.17++	−0.04	−0.12++	−0.07	0.13	−0.15++
9	0.05	−0.19++	−0.29+++	0.06	0.07	−0.07
10	−0.11	−0.30+++	−0.01	−0.10+	−0.10	−0.12
11	−0.05	−0.17	−0.17	−0.08	−0.10	−0.16++
12	0.00	−0.14	−0.26++	−0.01	−0.06	0.00
13	−0.19+++	−0.16	−0.34++	−0.22+++	−0.28+++	−0.32+++
14	0.01	−0.06	−0.16	−0.16+++	−0.26+++	−0.22+
15	−0.05	0.08	0.05	0.07	0.00	0.04
16	0.00	−0.10	−0.05	−0.05	−0.15+	−0.06
17	0.07	0.25	−0.04	−0.10++	−0.08	0.06
18	−0.15+	0.04	−0.03	0.02	−0.23++	−0.07
19	0.02	0.04	−0.26++	−0.08++	0.03	−0.10
20	0.01	0.13	0.08	−0.10++	0.21−	−0.13+
21	0.13−	0.07	0.07	−0.10+	0.06	0.11
22	0.06	0.19	−0.11	0.06	0.15	0.05
23	0.38	0.84−	0.37	0.18	0.13	0.18
24	0.09	0.03	−0.07	0.03	−0.03	−0.01
25	0.14−	−0.26+	0.30−	0.04	0.12	−0.02
26	0.16−	0.14	0.15	−0.03	0.13	0.27−
27	−0.15++	0.42−	0.34−−	−0.07+	0.11	−0.11
28	0.04	0.12	0.44−−	0.00	0.40−	0.03
29	0.11	0.24−	1.25−	−0.10+	−0.03	0.29−
30	0.28−	0.10	−0.06	0.25−	0.49−	0.28−
31	0.29−	0.34−	0.02	0.15−	0.11	0.13−
32	0.05	0.31−	−0.22+	0.21−	−0.15+	0.30−
33	0.04	0.52−	0.34−−	0.16−	0.11	0.18−
34	0.26−	0.55−	0.17−	0.67−	0.48−	0.46−
35	0.05	0.63−	0.20	0.46−	0.57−	0.37−

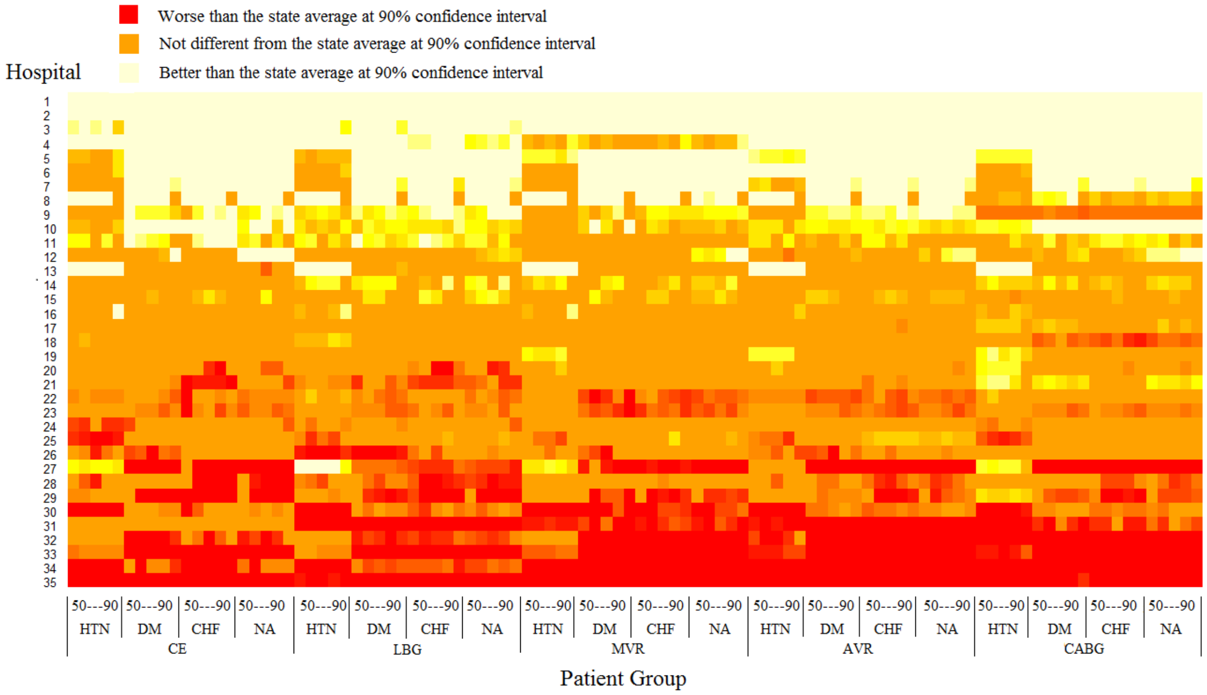
+++, ++, +: better than state average at 99%, 95% and 90% confidence level
−−−, −−, −: worse than state average at 99%, 95% and 90% confidence level

whether hospital j is statistically significantly worse than, the same as, or better than the state average at a 10% significance level for patient i in group k. Then we calculate the overall performance of hospital j for patient group k using $\bar{Y}_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ijk}$ and present the results in a heat map (Figure 3), where the yellow/red colors indicate that a hospital's overall performance is better/worse than the state average, and the intensity of the colors indicates the fraction of patients in a cell for which a hospital is better/worse than the state average.

sets of hospitals that are significantly better than the state average. However, as shown in the heat map, a majority of patients in each group have the same best set of hospitals.

From Figure 3, we observe that many of the cells in the middle (i.e., those associated with hospitals 11-25) are orange, which indicates that these hospitals are not significantly different from the state average for many patient groups. The majority of the cells in rows at the top (e.g., those associated with hospitals 1-3) have the color of yellow, indicating that these hospitals are better than the state average for most patient groups. In contrast, the red color of the cells in rows at the bottom (e.g., those associated with hospitals 34-35) indicates that these hospitals are worse than the state average for most patient groups. Rows near the top having colors of yellow and orange indicate that the corresponding hospitals are better for some patient groups, but are not statistically different from the state average for other patient groups. Likewise, rows near the bottom with a mixture of red and orange cells indicate that these hospitals are worse for some patient groups but are not significantly different from the state average for other groups. Interestingly, there are hospitals (e.g., 13 and 27) that are significantly better than the state average for some patient groups (e.g., patients with hypertension) but are significantly worse than the state average for other patient groups (e.g., patients at 60s with no major comorbidities).

Figure 3 Comparison of Hospital Complication Scores for Patient Groups



Note: Patients are grouped by age group (i.e., 50s to 90s), comorbidity and surgery. Acronyms for comorbidities: HTN - hypertension, DM - diabetes, CHF - chronic heart failure, NA - no comorbidities. Acronyms for surgeries: CE - carotid endarterectomy, LBG - lower extremity bypass graft, MVR - mitral valve repair, AVR - aortic valve repair, CABG - coronary artery bypass grafting.

5.4. Outcome Differences Based on Other Metrics

We have also performed similar analyses of the heterogeneity in outcome differences using in-hospital mortality and 30-day readmission as metrics. The results are displayed in the heat maps of Figures 8 and 9 in Appendices D and E. We see that, because mortality is a rare event for the procedures in this study, most hospitals are not statistically significantly different from the state average, as indicated by the prevalence of orange in Figure 8. However, we observe that Hospitals 2, 3 and 30 are better than the state average for many patient groups and that Hospitals 22, 26 and 33 are worse for some patient groups and not significantly different from the state average for other patient groups.

Readmission is a more common event than mortality, so using readmission rate as the outcome metric allows us to identify more hospitals that are significantly different from the state average. From Figure 9, we see that Hospitals 3, 27, 9 and 10 are better than the state average for most patient groups, whereas Hospitals 6, 23, 31 and 34 are worse than the state average for most patient groups. Most other hospitals (e.g., Hospitals 1, 2, 5 and 8) are either not statistically different from or are better than the state average, depending on the patient group. Similar to the case with complication score as the outcome metric, we see that some hospitals (e.g., Hospitals 17, 19, 26 and 30) are better than the state average for some patient groups but worse than the state average for other patient groups.

Using either complication score or readmission rate allows us to identify a relatively large set of hospitals that are significantly different from the state average. However, for a given patient group, the best hospitals that are significantly better than the state average with respect to complication score may be different than those with respect to readmission rate. For example, for patient group CE50HTN (CE, 50s, with hypertension), Hospitals 1, 2, 4, 8 and 13 are significantly better than the state average when outcomes are measured by complication score, but Hospitals 3, 9, 17 and 27 are significantly better than the state average when outcomes are measured by readmission rate. Likewise, for a given hospital, the set of patient groups for which it produces the best outcomes are different when different outcome metrics are used. For example, Hospital 2 is significantly better than the state average for all patient groups when outcomes are measured by complication score but for only patient group CE90CHF (CE, 90s, with chronic heart failure) when outcomes are measured by readmission rate. The reason is that these metrics measure different capabilities. Readmission rate measures the ability of a hospital to ensure patients are healthy (or at least stable) when released. In contrast, complication rate measures the ability of a hospital to avoid problems such as hospital-acquired infections, while the patient is in the hospital. Prior studies that have found mixed results regarding the correlation between readmission and complication rates (see for example Hospital Compare,[8] 2016, Lawson et al., 2013, Merkov et al., 2015, Robbins, 2013) are consistent with the differences we have observed.

6. Managerial Implications

We now turn to the last of our four key questions, which is what are the benefits of patient-centric information to patients, payers and providers. To evaluate the impacts on patients, we compare the sets of best hospitals and potential outcomes under population-average and patient-centric information. To illustrate the potential benefit to payers, we use the Hospital Acquired Condition Reduction Program as an example of how patient-centric information enables payers to better align payment with hospital performance. To illustrate the benefits to providers, we discuss how patient-centric information can help hospitals better align their strategic focus with their strengthes and focus their process improvement efforts where they will have the greatest impact.

6.1. Implications for Patients

Existing hospital rating systems, such as those of US News and the LeapFrog Group, and quality report cards, such as the New York Cardiac Surgery Quality Report Cards, compare hospitals using O/E ratios of observed to expected metrics (e.g., mortality rate). The expected rates are population averages estimated from a multivariate logit/probit model that includes patient demographics and comorbdities to control for patient severity of illness and hospital dummies to capture the fixed effects of individual hospitals. US News aggregates ratings into broad categories such as heart surgery and cancer, rather than reporting them for individual procedures such as mitral valve or aortic valve surgeries. As a result, it captures only the average effect of a hospital for all discharged patients. The LeapFrog Group and NY quality report cards report ratings for individual procedures such as CABG, mitral valve, aortic valve surgeries, so they capture the average effect of a hospital for a procedure. But they still make use of population-average O/E ratios that do not capture the heterogeneity of outcome differences across groups of patients undergoing the same procedure.

Because population-average based rankings, including those making use of O/E ratios, assume away heterogeneity in provider performance across patient groups, they suggest that the same hospitals (or surgeons or physicians) are best for all patients. This leads to two problems. First, as we discussed in the previous section, some hospitals that are high performers on average have average or below average outcomes for some patient groups. So, O/E ratios will guide some patients to suboptimal choices of providers. Second, because they suggest a "one size fits all" picture of hospital quality, population-average based rankings encourage patients to concentrate unnecessarily in a small subset of hospitals. The resulting capacity overloads will lead to longer patient wait times that could negatively impact patient outcomes.

[8] https://www.medicare.gov/hospitalcompare/compare.html

6.1.1. Comparison of Best Hospitals To illustrate the difference between patient-centric and population-average information in terms of their ability to guide patients to the best hospitals, we focus on complication score as the outcome metric.[9] We use each type of information to identify the best hospital(s) (i.e., those that achieve the minimum complication score) for each patient group. Finally, we compute the weighted average complication score across all patients. The difference between the average complication score under patient-centric and population-average information is a measure of the expected incremental value of patient-centric information to a randomly selected patient who chooses the best hospital for him/her based on the available information.

Because the dependent variable (complication score) is left truncated at zero, we use a tobit model instead of a logit/probit model to identify the best hospital under population-average information. For all models, we have robust standard errors clustered by hospital to allow for differences in the variance/standard errors due to arbitrary intra-group correlation (KC and Terwiesch, 2011, Jaeker and Tucker, 2015). The hospital with the smallest O/E ratio is designated as the best hospital for all patient groups. To rank hospitals using patient-centric information, we use the causal tree method discussed earlier. As we noted earlier, this method can identify different hospitals as best for different patient groups. Furthermore, if the outcome differences between hospitals are not significant, the tree may not differentiate between them. As a result, multiple hospitals may be identified as best for a given patient group.

Applying these methods to data for NY patients discharged in 2012 after one of the six cardiovascular surgeries listed earlier generates the results in Table 6. These identify the set of best hospitals and the number of patients for whom each hospital is best under population-average and patient-centric information. The difference in hospital rankings, and the patient complication scores they produce, that occur when we switch from population-average information to patient-centric information, characterize the value of patient-centric information to an individual patient who seeks out the best hospital for him/her using the available information. In addition to guiding patients to hospitals that will reduce their expected complication score, patient-centric information guides patients to a wider range of hospitals, which will be more feasible from a capacity standpoint to provide patients with the best available treatment.

[9] We use complication score because it captures a wide range of negative patient outcomes and shows substantial variation across hospitals. But the difference between patient-centric and population average information can be evaluated in terms of any of the outcome metrics we introduced earlier, or a composite score that combines them, without changing the overall conclusions about the value of patient-centric information.

6.1.2. Comparison of Patient Outcomes There are two main insights from Table 6. The first is that the hospital that is best on average across the entire population is not best for most patients. Patient-centric information reveals that different hospitals are best for different patients. For most of the surgical procedures, the top-ranked hospital under population-average information is the top hospital only for a minority of patients. For CE, the top-ranked hospital under population-average information is only best for 36 out of 2681 patients. For CABG, it is optimal for 30 out of 7953 patients. For AAA, it is optimal for 4 out of 185 patients. For AVR, it is optimal for 9 out of 4025 patients. For MVR, it is optimal for 13 out of 1054 patients. And for LBG, the top-ranked hospital under population-average information is not the best hospital for any group of patients.

The second insight from Table 6 is that choosing the best hospital on the basis of patient-centric, rather than population-average, information results in a significant reduction in average complication score. This reduction ranges from 0.11 to 0.40, which is equivalent to a 4.5% to 16% reduction in mortality, across the six cardiac specialties. The average reduction across all patients is 0.21, which is equivalent to a 8.8% reduction in mortality.

To get a better sense of which patient groups benefit most from patient-centric information, we group patients by procedure type, age group and major comorbidities (as what we did for the earlier heat maps). The average reduction of complication score for each patient group is summarized in Figure 4. Generally speaking, patients with diabetes or chronic heart failure benefit more than those with other (or no) comorbidities from patient-centric information. Among all the patient groups, LBG patients with diabetes or chronic heart failure benefit the most and MVR patients with chronic heart failure or no comorbidities benefit the least.

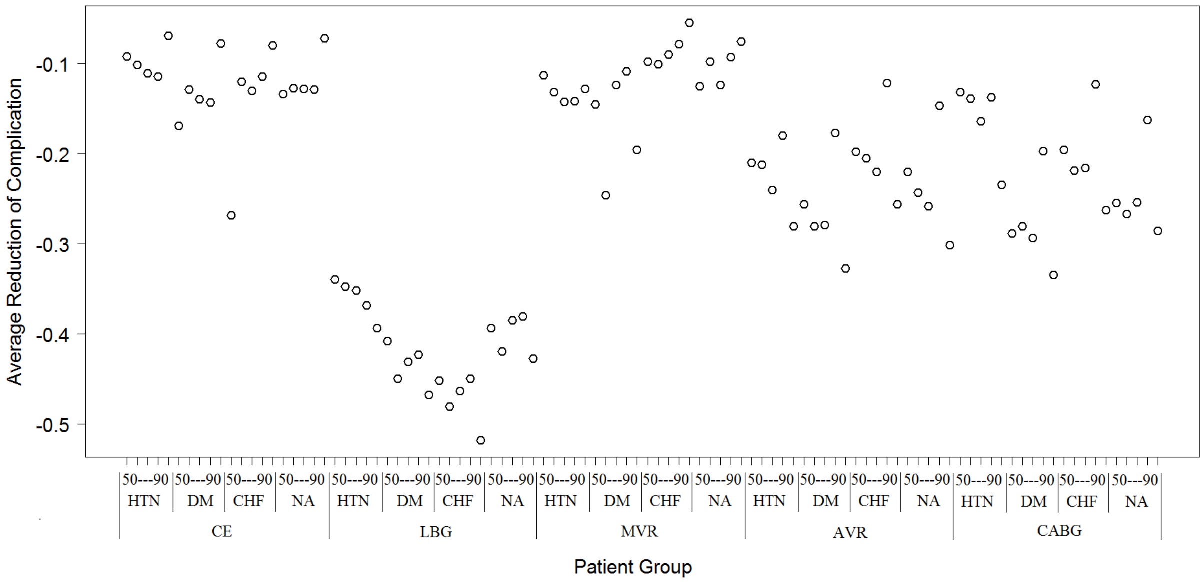
Figure 5 displays the distribution of percentage reduction in complication score. From this histogram, we see that around 97.5% of patients achieve a positive reduction in their complication score under patient-centric information. Only a small fraction of patients are equally well off under population-average and patient-centric information.

6.1.3. Other Outcome Metrics As noted earlier, we can make use of patient-centric information to rank hospitals according to other outcome metrics besides complication score. To illustrate how hospitals perform differently on other metrics, Table 7 compares hospitals based on complication score, readmission and mortality rates for two patients. As we did in the heat maps earlier, we identify hospitals that are statistically significantly better than the state average. Table 7 shows that the set of above average hospitals changes when different outcome metrics are used. Hospitals 3 and 4 are significantly better than the state average for Patient 1 with respect to complication score but are not different from the state average with respect to readmission rate. In this example, there are hospitals that are significantly better than the state average with respect

Table 6 Impact on Average Patient Complication Score From Using Patient-Centric Instead of Population-Average Information in Hospital Selection

		Population-Average Information	Patient-Centric Information					Avg. Reduction of Complication Score
CE	hospital index	2	1	2	4	5	11	
	number of patients	2681	2049	36	97	472	27	
	change in complication score		-0.10	0	-0.20	-0.12	-0.17	-0.11
CABG	hospital index	3	1	2	4	5	3	
	number of patients	7953	5573	4	12	2334	30	
	change in complication score		-0.20	-0.13	-0.03	-0.20	0	-0.19
LBG	hospital index	12	1	2	4	5	3	
	number of patients	2366	1810	494	12	42	8	
	change in complication score		-0.40	-0.41	-0.18	-0.48	-0.20	-0.40
AAA	hospital index	2	1	2	4	5	12	
	number of patients	185	138	4	1	39	3	
	change in complication score		-0.12	0	-0.05	-0.12	-0.06	-0.12
AVR	hospital index	3	1	2	4	5	3	
	number of patients	4025	2499	1139	29	349	9	
	change in complication score		-0.25	-0.23	-0.06	-0.25	0	-0.24
MVR	hospital index	3	1	2	3	5		
	number of patients	1054	658	14	13	359		
	change in complication score		-0.11	-0.13	0	-0.15		-0.12

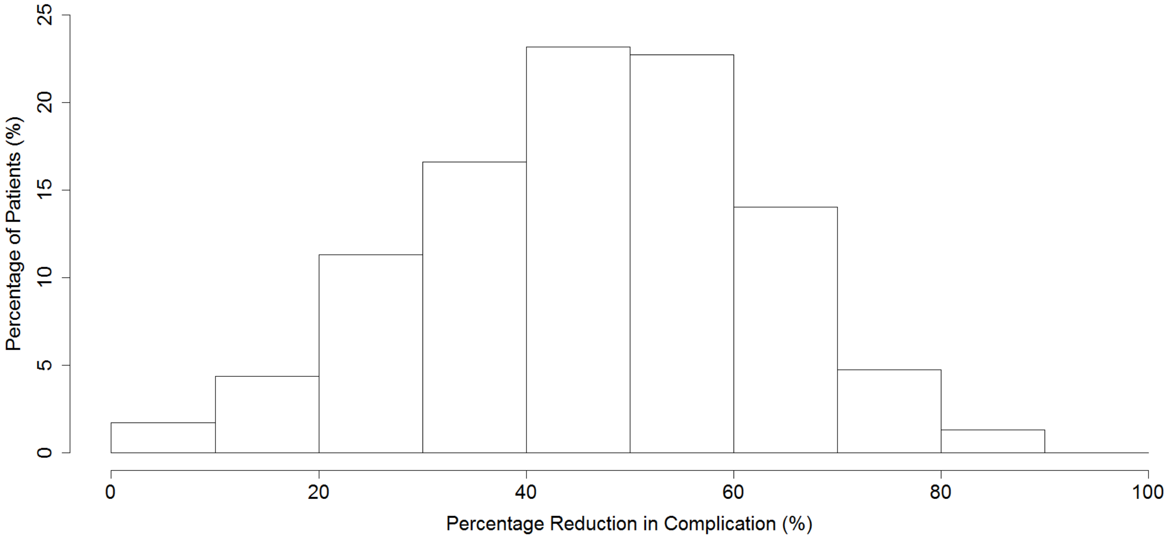
Figure 4 Complication Reduction by Patient Groups



Note: Patients are grouped by age group (i.e., 50s to 90s), comorbidity and surgery. Acronyms for comorbidities: HTN - hypertension, DM - diabetes, CHF - chronic heart failure, NA - no comorbidities. Acronyms for surgeries: CE - carotid endarterectomy, LBG - lower extremity bypass graft, MVR - mitral valve repair, AVR - aortic valve repair, CABG - coronary artery bypass grafting.

to all three outcome metrics used (e.g., Hospital 33 for Patient 1 and Hospital 15 for Patient 2). However, it may be the case that no hospital is above average for all metrics for a given patient. If this is the case, then a patient with his/her primary care physician must either evaluate the multi-

Figure 5 Complication Reduction Under Patient-Centric Information



dimensional outcomes subjectively or place weights on the various outcome metrics and perform a quantitative ranking based on the composite metric.

A sensible approach for a ranking service like LeapFrog or US News would be to create a website that allows a patient to enter his/her characteristics and medical condition, along with weights on outcome metric, and then use the methodology of this paper to generate a personalized ranking of providers (Huckman and Kelly, 2013).

6.2. Implications for Hospitals and Payers

Payers are increasingly seeking ways to tie hospital reimbursement to performance. For example, the Hospital Acquired Condition Reduction Program (HACRP) was established in 2013 as a response to increasing costs of complications. This program penalizes low-performing hospitals with regard to the Patient Safety Indicator (PSI) 90 Composite Index Value (Domain 1) and five infection measures (Domain 2).[10] For each measure, CMS uses two years of historical data to calculate risk-adjusted infection rates and then ranks hospitals accordingly. Each hospital is assigned a score between 1 and 10 for each measure based on its relative rank in deciles for that measure. There is only one score for Domain 1. A hospital's Domain 2 score is calculated as the average of the domain's individual measures. The total score is calculated as the weighted average of Domain 1 and Domain 2 scores, where the weights are 15% and 85% for the two domains. In 2015, CMS

[10] The PSI measures include rates of pressure ulcer, iatrogenic pneumothorax, central venous catheter-related bloodstream infection, postoperative hip fracture, perioperative pulmonary embolism or deep vein thrombosis, postoperative sesis, postoperative wound dehiscence and accidental puncture or laceration. The five infection measures are rates of central line-associated bloodstream infection, catheter-associated urinary tract infection, colon and hysterectomy surgical site infection, methicillin-resistant staphlococcus aureus bacteremia, and clostrium dfficile infection.

Table 7 Outcome Metrics for Individual Patients

Hospital	Patient 1 (LBG, 70s, 1 Comorb)			Patient 2 (AVR, 80s, 1 Comorb)		
Index	Complication	Readmission	Mortality	Complication	Readmission	Mortality
1	−0.37+++	−0.03-	0.00-	−0.59+++	−0.01-	−0.02+++
2	−0.25+++	0.04-	−0.01+++	−0.74+++	0.01-	−0.02+
3	−0.20+++	−0.09+++	−0.01+++	−0.44+++	−0.11+++	−0.02+
4	−0.17++	−0.05+++	0.00-	−0.43+++	−0.02-	0.00-
5	−0.25+	−0.01-	0.01-	−0.65+++	−0.01-	−0.02+++
6	−0.04-	0.13--	−0.01-	−0.53+++	0.29--	−0.02+++
7	−0.04-	−0.03+	0.00-	−0.23-	−0.03-	0.01-
8	−0.17++	0.02-	0.00-	−0.04-	0.04-	0.00-
9	0.05-	−0.06+++	0.00-	−0.19++	−0.10+++	−0.01-
10	−0.11-	−0.05++	0.00-	−0.30+++	−0.04+	−0.01-
11	−0.05-	0.00-	−0.01+	−0.17-	−0.02-	0.01-
12	0.00-	0.07--	0.01-	−0.14-	0.12--	0.02-
13	−0.19+++	−0.03++	0.00-	−0.16-	−0.04-	−0.01-
14	0.01-	−0.04++	−0.01+	−0.06-	−0.07-	−0.01-
15	−0.05-	−0.06++	0.00-	0.08-	−0.02-	0.00-
16	0.00-	−0.01-	0.02-	−0.10-	0.01-	0.00-
17	0.07-	0.05-	0.00-	0.25-	0.11--	0.01-
18	−0.15+	−0.02-	−0.01-	0.04-	−0.05++	0.00-
19	0.02-	−0.03+	0.00-	0.04-	−0.03++	0.00-
20	0.01-	−0.04++	0.01-	0.13-	−0.07+++	0.00-
21	0.13-	−0.02-	0.01-	0.07-	−0.05+++	0.00-
22	0.06-	−0.02-	0.00-	0.19-	−0.03-	0.02-
23	0.38-	0.18--	0.01-	0.84-	0.12--	0.03-
24	0.09-	0.00-	0.02-	0.03-	0.00-	0.02-
25	0.14-	−0.03+	0.00-	−0.26+	−0.04++	−0.03+++
26	0.16-	0.01-	0.01-	0.14-	−0.06+++	0.02-
27	−0.15++	−0.08+++	−0.01-	0.42--	−0.10+++	0.01-
28	0.04-	0.02-	0.01-	0.12-	−0.09+++	0.01-
29	0.11-	−0.03-	0.00-	0.24-	−0.07+++	0.00-
30	0.28--	−0.07+++	−0.01-	0.10-	−0.08+++	−0.03++
31	0.29--	0.14--	0.00-	0.34--	0.33--	−0.01-
32	0.05-	0.02-	0.00-	0.31--	−0.01-	0.01-
33	0.04-	0.01-	0.00-	0.52--	0.02-	0.00-
34	0.26--	0.11--	0.00-	0.55--	0.13--	0.02-
35	0.05-	−0.01-	−0.01-	0.63--	−0.01-	0.00-

+++, ++, +: better than state average at 99%, 95% and 90% confidence level
---, --, -: worse than state average at 99%, 95% and 90% confidence level

reduced total payments (i.e., across all patients) by 1% for hospitals that ranked among the worst quartile with regard to hospital acquired infections.

6.2.1. Impact of Patient-Centric Information on Hospital Payments The Hospital Acquired Condition Reduction Program is based on population-average outcome information and so does not recognize heterogenous outcome differences across patient groups. Consequently, applying a uniform penalty to these hospitals does not recognize their acceptable or even high performance for some patient groups. Similarly, hospitals that are not penalized under the HACRP may perform poorly for some patient groups. In addition to misaligning penalties with performance, an incentive system based on population-average information can hide areas of poor performance and discourage hospitals from addressing them. In contrast, patient-centric information allows payers to assess hospital performance by patient group and better align payments with quality to provide shaper incentives for quality improvement.

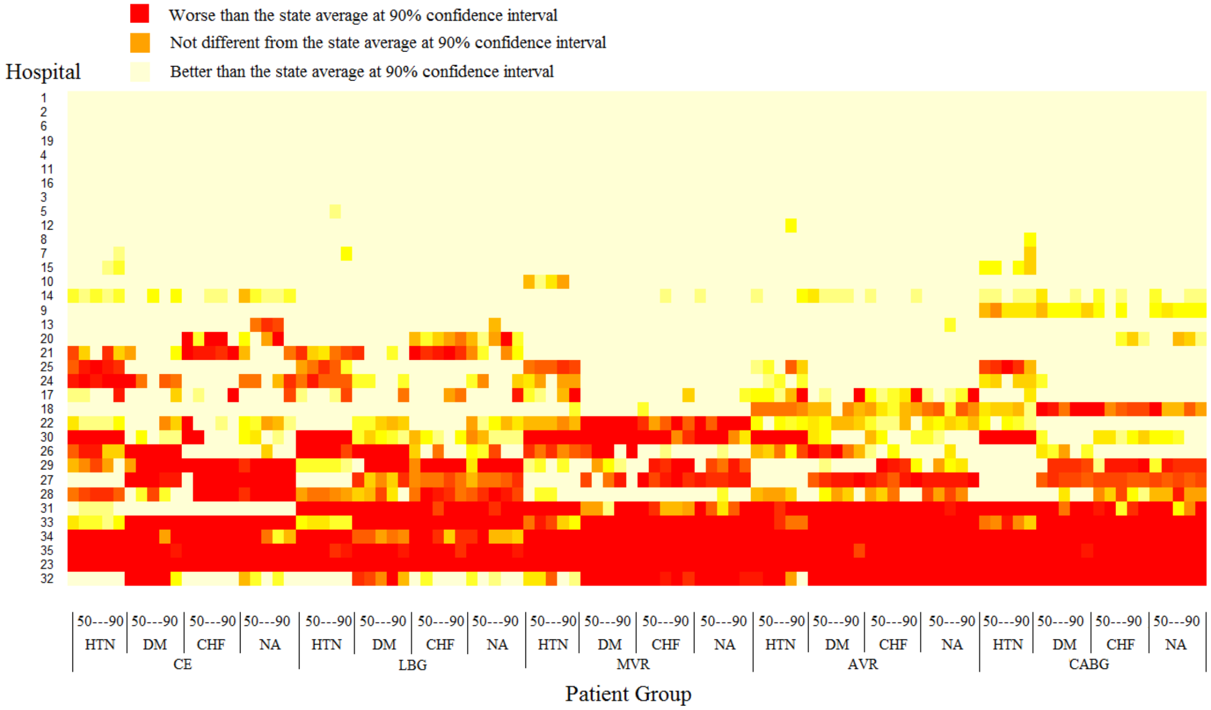
To illustrate a HACRP-type program under patient-centric information, we group patients by procedure type, age group and comorbidities. For each patient group, we use $Y_{ijk} \in \{0, 1\}$ to indicate whether hospital j is among the worst quartile for patient i in group k. We then calculate the overall performance of hospital j for patient group k using $\bar{Y}_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ijk}$ and display the results in the heat map of Figure 6. We see that only Hospitals 23 and 35 are among the worst quartile across all patient groups. Hospitals 31, 33 and 34 are among the worst quartile for a majority of patient groups, but they have areas (e.g., procedure CE for Hospital 31) that are not among the worst quartile. Likewise, Hospitals 20 and 21 are not among the worst quartile for the majority of patient groups, but they have areas (e.g., old CE patients with chronic heart failure for Hospital 20) that are among the worst quartile.

Payments would be better aligned with performance if hospitals were penalized for only their low-performing areas. To see how, in Figure 7, we compare scenarios in which hospitals are penalized based on population-average and patient-centric information. Under population-average information, there are eight hospitals with average performance among the worst quartile, each of which would be penalized by 1% on all payments. The other hospitals are not penalized at all. In contrast, under patient-centric information, only two hospitals are not penalized at all. The rest are penalized on some portion of their payments. Hence, more hospitals would have a financial incentive to improve under patient-centric than under population-average information.

6.2.2. Impact on Hospital Strategy and Improvement Efforts Payments based on patient-centric information provide more focused incentives for hospitals to improve quality, because they reward hospitals for incremental improvements. For example, consider a hospital that discharges 1,000 patients a year, of which 100 are CABG patients. The infection rate across all patients is 1%, but is 5% for CABG patients. If, under the current HACRP, the hospital is not penalized, then it has no economic incentive to improve. Even if it is being penalized, it may be the case that reducing infections among CABG patients will not have a large enough effect on the overall infection rate to eliminate the penalty. However, if HACRP penalties were based on patient-centric information, and therefore individually penalized payments for CABG patients, then the hospital would have economic incentives to reduce the CABG patient infection rate, regardless of whether payments for other types of patients were being penalized or not.

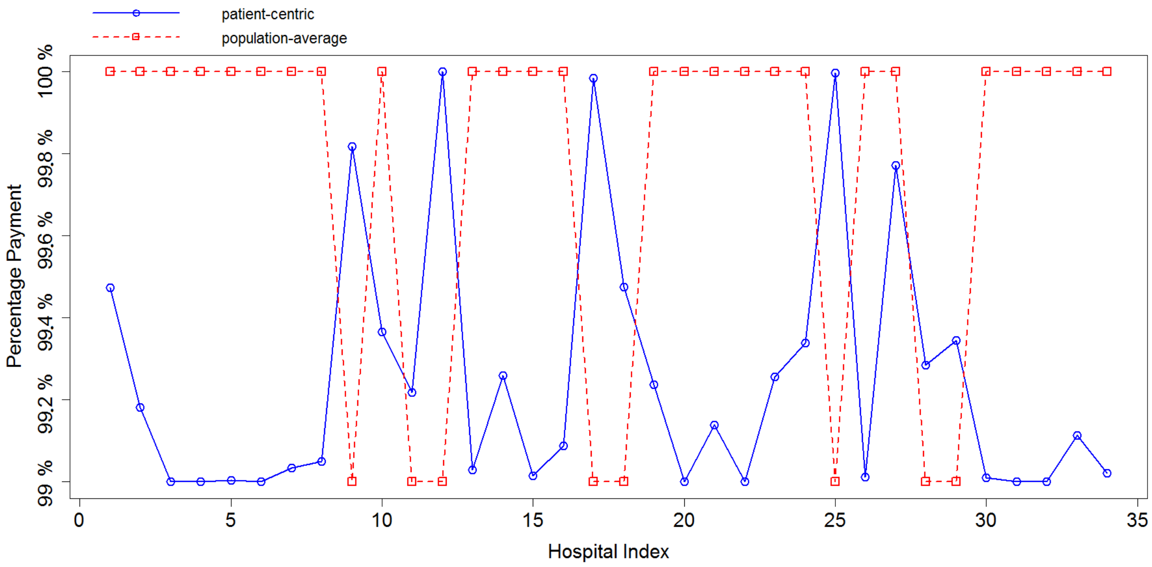
Beyond its use in targeted incentives, transparent patient-centric outcome information can help hospitals learn from one another. For example, the heat map in Figure 3 shows that Hospital 13 has

Figure 6 Comparison of Hospitals' Performance for Patient Groups



Note: Patients are grouped by age group (i.e., 50s to 90s), comorbidity and surgery. Acronyms for comorbidities: HTN - hypertension, DM - diabetes, CHF - chronic heart failure, NA - no comorbidities. Acronyms for surgeries: CE - carotid endarterectomy, LBG - lower extremity bypass graft, MVR - mitral valve repair, AVR - aortic valve repair, CABG - coronary artery bypass grafting.

Figure 7 Percentage Payment under Patient-Centric and Population-Average Measure



very low complication rates for hypertension patients, despite having average performance for other patients. This may indicate that Hospital 13 has made some kind of innovation that enables them to better protect these patients. Hence, patient-centric information in Figure 3 can help hospitals spot best practices that might be shared to elevate performance across the industry.[11]

Finally, in addition to supporting incentives for hospitals to improve outcomes for specific patient groups, patient-centric information may also incent hospitals to focus on the patients they are able to treat most successfully. For example, suppose a hospital has exceptionally good outcomes (e.g., low complication scores), relative to the state average, for elderly patients, but poor outcomes for younger patients. The penalties from an HACRP-type program would make the younger patients less economically attractive to the hospital. And, if patient-centric information were transparently available to patients, demand from younger patients would presumably be weaker as well. Both factors would encourage the hospital to focus on elderly patients, in its process design and marketing efforts. Other hospitals might be incented to focus on particular medical procedures or patient groups (e.g., patients with hypertension, diabetes or cancer). Over time, this would encourage a network of providers that leverage their individual strengths to produce better patient outcomes.

7. Conclusion

In recent years, there have been many wide-ranging efforts to improve the delivery of health care in the United States. Perhaps the most straightforward of these has been the push for better and more transparent outcome information to help patients find the best available care for them. Unfortunately, as we have shown, the standard approach of computing risk-adjusted outcomes produces population averages that do not accurately represent the likely outcomes for all patients. In this paper, we have shown that the relative performance of hospitals is heterogeneous across patient groups. Consequently, patient-centric rankings of hospitals are significantly different than rankings based on population-average information.

In this study, we have addressed the challenges of generating patient-centric outcome information and hospital ranking. Using six cardiovascular surgeries as the clinical setting, we studied the outcomes of thirty-five hospitals in NY based on different metrics. We extended the causal tree method for multiple hospitals to recursively partition patients into groups that exhibit significant outcome differences between hospitals. We quantified the outcome differences for groups of patients using propensity score matching and derived patient-centric estimates of outcome differences between hospitals for individual patients. Our analysis shows that outcome differences between hospitals are heterogeneous not only across procedure types, but also along other dimensions such as patient age and comorbidities.

[11] Competition may hinder sharing of best practices across hospitals. But there are platforms for such sharing. For example, the Quality Collaborative of the Michigan Society of Thoracic Surgeons http://mstcvs.org/qc.html has been set up precisely to encourage the open heart programs in the state of Michigan to share data and practices.

We compared the best hospitals based on population-average and patient-centric information. We found that, for the majority of patients (around 97.5%), the best hospitals are different than those indicated as best by a population-average rating. Furthermore, we found that patient-centric information results in a larger set of best hospitals, which suggests more opportunities for distributing patient workload across hospitals to reduce patient waiting time. Most importantly, we compared the potential outcomes when patients are treated at the best hospitals based on the two types of information, and estimated that the complication score could be reduced by 46% (equivalent to a 8.8% reduction in mortality) by using patient-centric information instead of population-average information.

In addition to the manifest benefits to patients, patient-centric information offers potential benefits to hospitals and payers as well. Using the Hospital Acquired Infection Reduction Program as an example, we showed that patient-centric information allows the CMS to better align payments (and penalties) with patient outcomes. This in turn provides sharper incentives for hospitals to improve quality. Finally, the more detailed patient-centric information can help hospitals to understand their strengths and weaknesses, as well as those of their peers. This can help them better align their strategies with their strengths, and also to learn from one another.

Lastly, providers may select patients they are most skilled at treating, and patients may select providers from whom they are likely to receive the best outcome. This will create an attenuation bias and will make it more difficult to detect differences among providers. In other words, our approach tend to generate a conservative estimate of outcome differences, which means that the impact of using patient-centric information may be even larger than our analysis indicates. It may be possible to combine the tree method with causal inference methods, and we leave this for future research.

References

Ang, E., Kwasnick, S., Bayati, M., Plambeck, E. L., & Aratow, M. (2015). Accurate emergency department wait time prediction. Manufacturing & Service Operations Management, 18(1), 141-156.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360.

Ban, G. Y., El Karoui, N., & Lim, A. E. (2016). Machine Learning and Portfolio Optimization.

Barro, J. R., Huckman, R. S., & Kessler, D. P. (2006). The effects of cardiac specialty hospitals on the cost and quality of medical care. Journal of health economics, 25(4), 702-721.

Bastani, H., & Bayati, M. (2016). Online decision-making with high-dimensional covariates. Available at SSRN 2661896.

Bavafa, H., Hitt, L. M., Terwiesch, C. (2013). Patient portals in primary care: Impacts on patient health and physician productivity. Working Paper.

Bertsimas, D., OHair, A., Relyea, S., & Silberholz, J. (2016). An Analytics Approach to Designing Combination Chemotherapy Regimens for Cancer. Management Science, 62(5), 1511-1531.

Birkmeyer, J. D., Siewers, A. E., Finlayson, E. V., Stukel, T. A., Lucas, F. L., Batista, I., ..., Wennberg, D. E. (2002). Hospital volume and surgical mortality in the United States. New England Journal of Medicine, 346(15), 1128-1137. Clark, J. R., & Huckman, R. S. (2012). Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. Management Science, 58(4), 708-722.

Clark, J. R., Huckman, R. S., & Staats, B. R. (2013). Learning from customers: Individual and organizational effects in outsourced radiological services. Organization Science, 24(5), 1539-1557.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. The Review of Economics and Statistics, 90(3), 389-405.

Cui, G., Wong, M. L., & Lui, H. K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. Management Science, 52(4), 597-612.

Elixhauser, A., Steiner, C., Harris, D. R., & Coffey, R. M. (1998). Comorbidity measures for use with administrative data. Medical care, 36(1), 8-27.

Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2015). Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management, 18(1), 69-88.

Finlayson, S., Birkmeyer, J., Tosteson, A., Nease, R. (1999). Patient preferences for location of care, implications for regionalization. Medical Care, 37:204-209.

Freeman, M., Savva, N., Scholtes, S. (2015). Gatekeepers at work: An empirical analysis of a maternity unit. Working Paper.

Gammie, J. S., Sheng, S., Griffith, B. P., Peterson, E. D., Rankin, J. S., O'Brien, S. M., Brown, J. M. (2009). Trends in mitral valve surgery in the United States: Results from the Society of Thoracic Surgeons Adult Cardiac Surgery Database. The Annals of Thoracic Surgery, 87(5):1431-7.

Gerteis, M. (1993). Through the patient's eyes: Understanding and promoting patient-centered care.

Glance, L. G., Osler, T. M., Mukamel, D. B., & Dick, A. W. (2007). Effect of complications on mortality after coronary artery bypass grafting surgery: evidence from New York State. The Journal of thoracic and cardiovascular surgery, 134(1), 53-58.

Groux, P., Anchisi, S., & Szucs, T. (2014). Are Cancer Patients Willing to Travel More or Further Away for a Slightly More Efficient Therapy?. Cancer and Clinical Oncology, 3(1), 36.

Guajardo, J. A., Cohen, M. A., & Netessine, S. (2015). Service competition and product quality in the US automobile industry. Management Science.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association, 47(260), 663-685.

Huckman, R. S., & Kelley, M. A. (2013). Public reporting, consumerism, and patient empowerment. New England Journal of Medicine, 369(20), 1875-1877.

Huckman, R. S., & Pisano, G. P. (2006). The firm specificity of individual performance: Evidence from cardiac surgery. Management Science, 52(4), 473-488.

Huckman, R. S., & Zinner, D. E. (2008). Does focus improve operational performance? Lessons from the management of clinical trials. Strategic Management Journal, 29(2), 173-193.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics, 7(1), 443-470.

Institute of Medicine (US). Committee on Quality of Health Care in America. (2001). Crossing the quality chasm: A new health system for the 21st century. National Academy Press.

Jaeker, J., Tucker, A. (2016). Past the point of speeding up: The negative effects of workload saturation on efficiency and quality. Management Science.

Kattan, M. W., Vickers, A. J. (2004, August). Incorporating predictions of individual patient risk in clinical trials. In Urologic Oncology: Seminars and Original Investigations (Vol. 22, No. 4, pp. 348-352). Elsevier.

KC, D., Staats, B. R. (2012). Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. Manufacturing & Service Operations Management, 14(4), 618-633.

KC, D., Staats, B. R., Gino, F. (2013). Learning from my success and from others' failure: Evidence from minimally invasive cardiac surgery. Management Science, 59(11), 2435-2449.

KC, D., Terwiesch, C. (2011). The effects of focus on performance: Evidence from California hospitals. Management Science, 57(11), 1897-1912.

Keeler, E. B., Rubenstein, L. V., Kahn, K. L., Draper, D., Harrison, E. R., McGinty, M. J., ... , Brook, R. H. (1992). Hospital characteristics and quality of care. Jama, 268(13):1709-1714.

Kent, D. M., Hayward, R. A. (2007). Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. Jama, 298(10), 1209-1212.

Kim, S. H., Chan, C. W., Olivares, M., Escobar, G. (2014). ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. Management Science, 61(1), 19-38.

Kolstad, J. T. (2013). Information and quality when motivation is intrinsic: Evidence from surgeon report cards. The American Economic Review, 103(7), 2875-2910.

Kravitz, R. L., Duan, N., Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Quarterly, 82(4), 661-687.

Lawson, E. H., Hall, B. L., Louie, R., Ettner, S. L., Zingmond, D. S., Han, L., ... & Ko, C. Y. (2013). Association between occurrence of a postoperative complication and readmission: implications for quality improvement and cost savings. Annals of surgery, 258(1), 10-18.

Lu, S. F., & Lu, L. X. (2016). Do Mandatory Overtime Laws Improve Quality? Staffing Decisions and Operational Flexibility of Nursing Homes. Management Science, Forthcoming.

Lu, Y., Musalem, A., Olivares, M., & Schilkrut, A. (2013). Measuring the effect of queues on customer purchases. Management Science, 59(8), 1743-1763.

Merkow, R. P., Ju, M. H., Chung, J. W., Hall, B. L., Cohen, M. E., Williams, M. V., ... & Bilimoria, K. Y. (2015). Underlying reasons associated with hospital readmission following surgery in the United States. Jama, 313(5), 483-495.

Ramdas, K., Saleh, K., Stern, S., Liu, H. (2014). Variety and experience: Learning and forgetting in the use of surgical devices. Working Paper.

Reddy, H. G., Shih, T., Englesbe, M. J., Shannon, F. L., Theurer, P. F., Herbert, M. A., ... & Prager, R. L. (2013). Analyzing failure to rescue: is this an opportunity for outcome improvement in cardiac surgery?. The Annals of thoracic surgery, 95(6), 1976-1981.

Robbins, R. A., & Gerkin, R. D. (2013). Comparisons between Medicare mortality, morbidity, readmission and complications. Southwest J Pulm Crit Care, 6(6), 278-86.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Schneider, C. R., Cobb, W., Patel, S., Cull, D., Anna, C., & Roettger, R. (2009). Elective surgery in patients with end stage renal disease: what's the risk?. The American surgeon, 75(9), 790-793.

Signorovitch, J. E. (2007). Identifying informative biological markers in high-dimensional genomic data and clinical trials.

Society of Thoracic Surgeons (2016). Online STS risk calculator.

Song, H., Tucker, A. L., Murrell, K. L. (2015). The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. Management Science.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. Journal of Machine Learning Research, 10(Feb), 141-158.

Sullivan, L. M., Massaro, J. M., & D'Agostino, R. B. (2004). Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Statistics in medicine, 23(10), 1631-1660.

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. Journal of the American Statistical Association, 109(508), 1517-1532.

Tsai, T. C., Jha, A. K., Gawande, A. A., Huckman, R. S., Bloom, N., & Sadun, R. (2015). Hospital board and management practices are strongly related to hospital performance on clinical quality metrics. Health Affairs, 34(8), 1304-1311.

Van Tuinen, M., Elder, S., Link, C., Li, S., Song, J. H., & Pritchett, T. (2005). Surveillance of surgery-related adverse events in Missouri using ICD-9-CM codes.

US Food and Drug Administration (2013). Paving the way for personlized medicine: FDA's role in the new era of medical product development.

Van Tuinen, M., Elder, S., Link, C., Li, S., Song, J. H., & Pritchett, T. (2005). Surveillance of surgery-related adverse events in Missouri using ICD-9-CM codes.

van Walraven, C., Austin, P. C., Jennings, A., Quan, H., & Forster, A. J. (2009). A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Medical care, 626-633.

Vassileva, C. M., Shabosky, J., Boley, T., Markwell, S., Hazelrigg, S. (2012). Cost analysis of isolated mitral valve surgery in the United States. The Annals of Thoracic Surgery, 94:1429-1436.

Wang, G., Li, J., Fazzalari, F. L., Hopp, W. J., & Bolling, S. F. (2016). Using Patient-Centric Quality Information to Unlock Hidden Health Care Capabilities.

Williams, J. F., MoRRow, A. G., & Braunwald, E. (1965). The incidence and management of "medical" complications following cardiac operations. Circulation, 32(4), 608-619.

World Health Organization. (2011). Global Atlas on Cardiovascular Disease Prevention and Control. World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization.

Xu, Y., Armony, M., & Ghose, A. (2016). The Effect of Online Reviews on Physician Demand: A Structural Model of Patient Choice. Available at SSRN 2778664.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. Journal of Computational and Graphical Statistics, 17(2), 492-514.

Appendix A: Estimation of the Variance of Outcome Difference Estimator

We assume that outcomes of patients treated by the same provider are i.i.d. and outcomes of patients treated by Providers 1 and 2 are independent. From sample S, we have

$$\begin{aligned} Var[D^S_{12}(X_i)|P(X)] = & \frac{\sum_{i\in l,T_{i1}=1} 1/P^2(X_i)}{(\sum_{i\in l,T_{i1}=1} 1/P(X_i))^2} Var(Y_{i1}) \\ & + \frac{\sum_{i\in l,T_{i2}=1} 1/P^2(X_i)}{(\sum_{i\in l,T_{i2}=1} 1/P(X_i))^2} Var(Y_{i2}) \end{aligned}$$

By the low of total variance, we have

$$Var[D^S_{12}(X_i)] = E_{P(X)}[Var(D^S_{12}(X_i)|P(X))] + Var[E_{i\in l}(D^S_{jk}(X_i)|P(X))]$$

Appendix B: Estimation of Mean Squared Errors

The expected MSE is the expectation of $MSE(S^{test}, S^{est})$ over test and estimation samples

$$\begin{aligned} EMSE &= E_{S_{test},S_{est}} MSE(S^{test}, S^{est}) \\ &= E_{S_{test},S_{est}}[(D^{test}_{12}(X_i) - D^{est}_{12}(X_i))^2] \\ &= E_{S_{test},S_{est}}[(D^{test}_{12}(X_i) - D^{\pi}_{12}(X_i))^2 + D^{est}_{12}(X_i)^2 - D^{\pi}_{12}(X_i)^2 + 2D^{test}_{12}(X_i)(D^{\pi}_{12}(X_i) - D^{est}_{12}(X_i))] \\ &= E_{S_{test},S_{est}}[(D^{test}_{12}(X_i) - D^{\pi}_{12}(X_i))^2 + D^{est}_{12}(X_i)^2 - D^{\pi}_{12}(X_i)^2 + 2D^{\pi}_{12}(X_i)(D^{\pi}_{12}(X_i) - D^{est}_{12}(X_i))] \\ &= E_{S_{test},S_{est}}[(D^{test}_{12}(X_i) - D^{\pi}_{12}(X_i))^2 + (D^{est}_{12}(X_i) - D^{\pi}_{12}(X_i))^2] \\ &= E_{S_{test}}[(D^{test}_{12}(X_i)^2 - 2D^{test}_{12}(X_i)D^{\pi}_{12}(X_i) + D^{\pi}_{12}(X_i)^2]^{\pi} + E_{S_{test},S_{est}}[(D^{est}_{12}(X_i) - D^{\pi}_{12}(X_i))^2] \\ &= E_{S_{test}}[D^{test}_{12}(X_i)^2 - D^{\pi}_{12}(X_i)^2] + E_{S_{test},S_{est}}[Var(D^{est}_{12}(X_i))] \end{aligned}$$

where we exploit the equality $E(D^{test}_{12}(X_i)) = E(D^{est}_{12}(X_i)) = D^{\pi}_{12}(X_i)$. Because $E(D^{test}_{12}(X_i)^2)$ does not depend on the estimator,[12] minimizing above EMSE is equivalent to minimizing

$$EMSE(S^{test}, S^{est}) = -E_{S_{test}}[D^{\pi}_{12}(X_i)^2] + E_{S_{test},S_{est}}[Var(D^{est}_{12}(X_i))]$$

[12] For each observation i in the test sample, there is a true outcome difference between Providers 1 and 2, $D^{test}_{12}(X_i)$, which we do not observe.

Appendix C: Complications and Weights

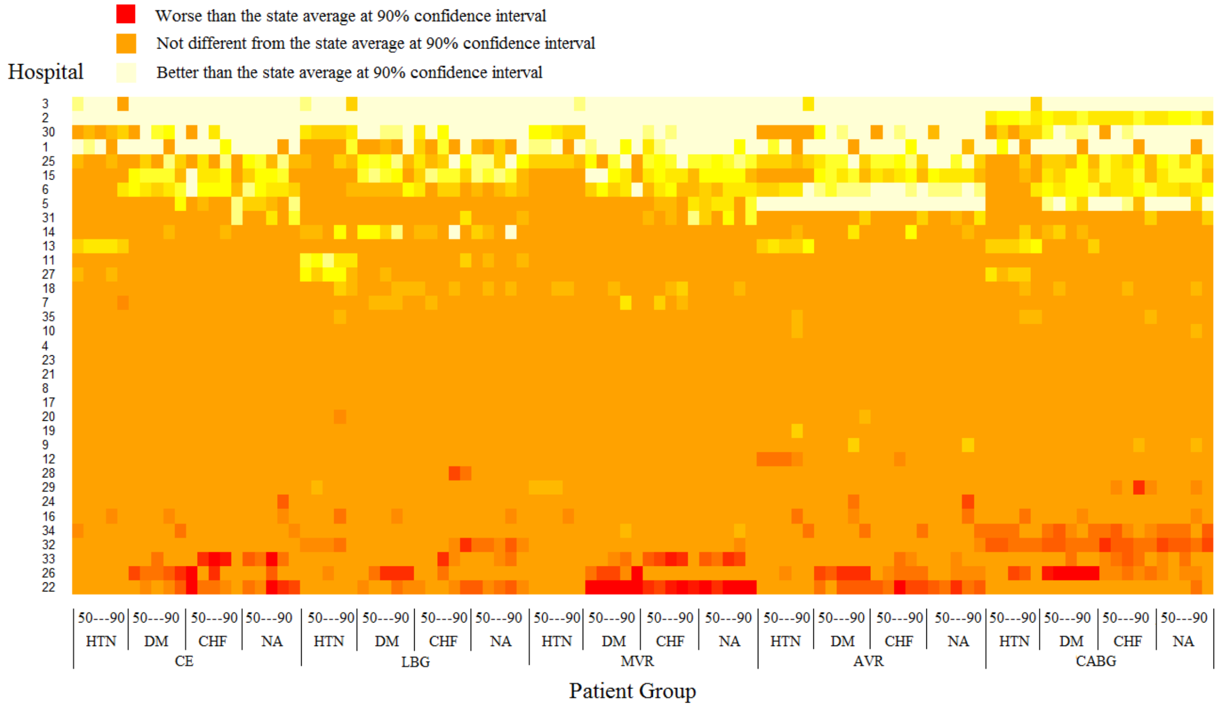
Table 8 The Weights of Complications Used to Calculate Complication Score

Complication	Coefficient	*Std.Err.*	Weight
Stroke	1.03 ∗ ∗∗	0.12	2
AorticDissection	3.16 ∗ ∗∗	0.33	7
RenalFailure	1.46 ∗ ∗∗	0.05	3
Ventilation	0.85 ∗ ∗∗	0.07	2
MultiOrganFailure	2.16 ∗ ∗∗	0.07	5
Coma	2.76 ∗ ∗∗	0.25	6
CardiacArrest	1.79 ∗ ∗∗	0.09	4
Sepsis	1.03 ∗ ∗∗	0.14	2
GIEvent	0.44 ∗ ∗∗	0.10	1
TrachealReintubation	1.22 ∗ ∗∗	0.06	3
SurgComp	1.11 ∗ ∗∗	0.15	2
Tamponade	1.02 ∗ ∗∗	0.14	2
PulmonaryInsuff	0.46 ∗ ∗∗	0.06	1
Constant	−4.93 ∗ ∗∗	0.04	

Note: Robust standard errors are clustered by hospital.
The outcome variable is death during hospitalization. Complications dropped from backward stepwise multivariate logistic regression include wound infection, renal dialysis, mediastinum, reoperation for bleeding, pneumonia, pulmonary embolism, heart block, myocardial infarction, surgical E codes and other cardiac complications.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

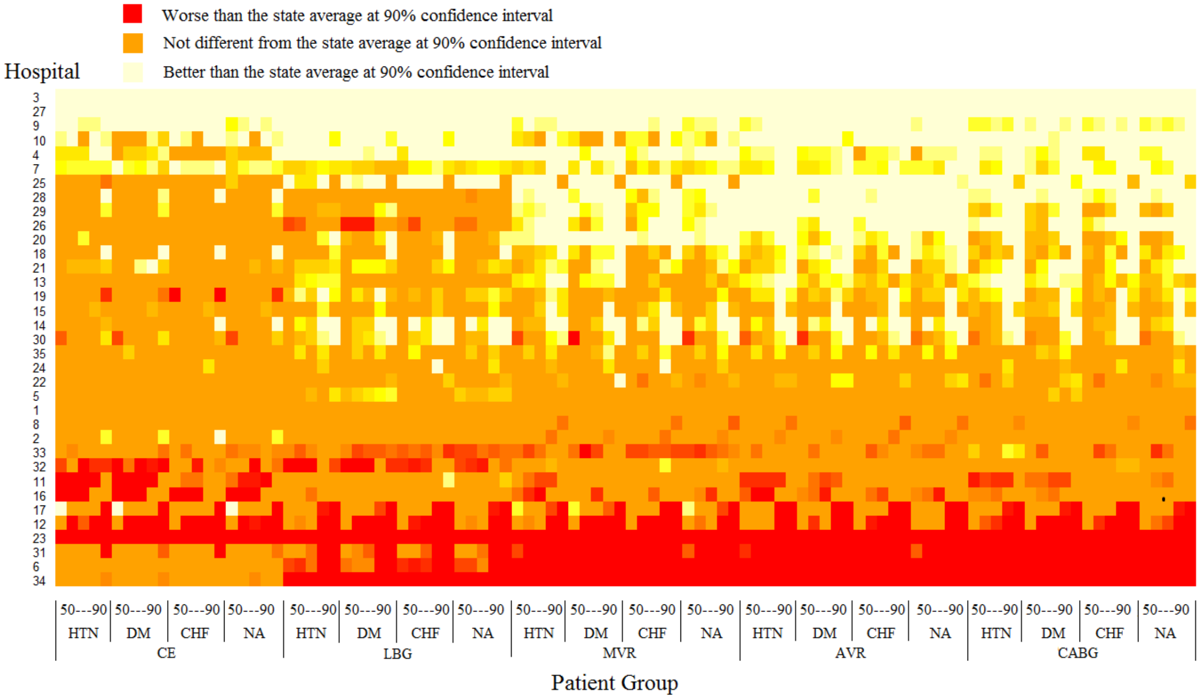
Appendix D: Mortality and Readmission for Groups of Patients

Figure 8 Comparison of Hospitals' Mortality for Groups of Patients



Note: Patients are grouped by age group (i.e., 50s to 90s), comorbidity and surgery. Acronyms for comorbidities: HTN - hypertension, DM - diabetes, CHF - chronic heart failure, NA - no comorbidities. Acronyms for surgeries: CE - carotid endarterectomy, LBG - lower extremity bypass graft, MVR - mitral valve repair, AVR - aortic valve repair, CABG - coronary artery bypass grafting.

Figure 9 Comparison of Hospitals' Readmission for Groups of Patients



Note: Patients are grouped by age group (i.e., 50s to 90s), comorbidity and surgery. Acronyms for comorbidities: HTN - hypertension, DM - diabetes, CHF - chronic heart failure, NA - no comorbidities. Acronyms for surgeries: CE - carotid endarterectomy, LBG - lower extremity bypass graft, MVR - mitral valve repair, AVR - aortic valve repair, CABG - coronary artery bypass grafting.