European Journal of Operational Research 253 (2016) 489–502

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

A DEA based composite measure of quality and its associated data uncertainty interval for health care provider profiling and pay-for-performance[☆]

Michael Shwartz[a,b,1], James F. Burgess Jr.[a,c,2], Joe Zhu[d,*]

[a] Center for Healthcare Organization and Implementation Research, VA Boston Healthcare System (152M),150 South Huntington Avenue, Boston, MA 02130, United States
[b] Boston University Questrom School of Business, Rafik B. Hariri Building, 595 Commonwealth Avenue, Boston, MA 02215, United States
[c] School of Public Health, Boston University, 715 Albany Street, T3W, Boston MA 02118, United States
[d] Foisie School of Business, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, United States

ARTICLE INFO

Article history:
Received 16 December 2013
Accepted 27 February 2016
Available online 5 March 2016

Keywords:
Data Envelopment Analysis (DEA)
Health care quality
Monte Carlo
Bootstrapping
Performance

ABSTRACT

Composite measures calculated from individual performance indicators increasingly are used to profile and reward health care providers. We illustrate an innovative way of using Data Envelopment Analysis (DEA) to create a composite measure of quality for profiling facilities, informing consumers, and pay-for-performance programs. We compare DEA results to several widely used alternative approaches for creating composite measures: opportunity-based-weights (OBW, a form of equal weighting) and a Bayesian latent variable model (BLVM, where weights are driven by variances of the individual measures). Based on point estimates of the composite measures, to a large extent the same facilities appear in the top decile. However, when high performers are identified because the lower limits of their interval estimates are greater than the population average (or, in the case of the BLVM, the upper limits are less), there are substantial differences in the number of facilities identified: OBWs, the BLVM and DEA identify 25, 17 and 5 high-performers, respectively. With DEA, where every facility is given the flexibility to set its own weights, it becomes much harder to distinguish the high performers. In a pay-for-performance program, the different approaches result in very different reward structures: DEA rewards a small group of facilities a larger percentage of the payment pool than the other approaches. Finally, as part of the DEA analyses, we illustrate an approach that uses Monte Carlo resampling with replacement to calculate interval estimates by incorporating uncertainty in the data generating process for facility input and output data. This approach, which can be used when data generating processes are hierarchical, has the potential for wider use than in our particular application.

1. Introduction

Publicly profiling health care provider performance is seen by policy makers in many countries as a way of motivating provider improvements and increasing information available to consumers in order to encourage better decisions. In the United States, the Centers for Medicare and Medicaid Services (CMS) make available to the public, through its Hospital Compare (CMS, 2013a) and Nursing Home Compare (CMS, 2013b) websites, data on the performance of hospitals and nursing homes respectively. Other systems with metrics on hospital outpatient care, physician practices, and accountable care organizations are in the process of being tested and released. In the British National Health Service, NHS Choices (2013) has developed comparison websites at the procedure or treatment level reporting an array of patient experiences and outcome measures. Other European countries are beginning to make available comparable information as well, though much of their efforts have been at the country level working through the European Commission and the Organisation for Economic Co-operation and Development to standardize information for comparisons across countries (European Commission, 2013).

☆ This material is based upon work supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development IIR 06-260. A version of this was presented at the 2013 iHEA World Congress, Sydney, Australia, July 7–10, 2013. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

* Corresponding author. Tel.: +1 508 831 5467.
E-mail addresses: mshwartz@bu.edu (M. Shwartz), James.Burgess@va.gov (J.F. Burgess Jr.), jzhu@wpi.edu (J. Zhu).
[1] Tel.: +1 617 353 2677.
[2] Tel.: +1 857 364 5683.

http://dx.doi.org/10.1016/j.ejor.2016.02.049
0377-2217/

Not only is provider performance on many different metrics increasingly being made available to the public around the world, but performance is being used to adjust payments to providers. In the United States, CMS has implemented the Hospital Value-Based Purchasing program (CMS, 2013c), and other programs mandated by the Affordable Care Act, such as those to adjust payments to new accountable care organizations and physicians, are in the process of being implemented. Similar programs in Europe include more focus on economic evaluation related to measuring cost-effectiveness, since the Europeans are not as market oriented as the systems are in the United States. For example, in the United Kingdom, the National Institute for Health and Clinical Excellence (NICE) is using cost-effectiveness ratings to choose the quality metrics to implement pay-for-performance standards for general practitioners. In Germany, RAND Europe has worked with the German health insurance system to develop a system for incorporating quality indicators into the reimbursement system for their physicians (Nolte et al., 2013).

Most existing provider profiling and related research focuses on individual performance indicators. Individual performance indicators are useful in targeting specific areas for improvement and monitoring improvement progress. However, a multitude of individual performance indicators does not allow an easy assessment of how well a provider organization is performing in the aggregate, the level at which assessments for payment and reporting are focused. To assess overall performance, it is useful to aggregate individual indicators into a composite measure (Institute of Medicine, 2006). Our focus in this paper is on indicators of the quality of care provided by health care facilities, where patients may be affected by different attributes of organizational performance and where we can measure these effects at the patient level. In this context, composite measures of quality are a useful summary for management, consumers and other stakeholders of the extent to which the facility has created a culture of quality and designed structures and processes to ensure quality throughout the organization. Composite measures of quality allow senior leaders to better benchmark the quality performance of their organization against high-performing organizations and to monitor changes over time. They provide information useful to patients when they are selecting where to receive their care. They also can be aligned against other relevant measures of performance, such as costs, to help managers and policy makers understand the value organizations are delivering to their clients. And, perhaps most importantly, they provide a basis for facility profiling that focuses on the "big picture."

In pay-for-performance (P4P) programs, performance on individual indicators is usually mapped into a payment adjustment and then the adjustments are added together to determine the overall impact on provider payment. As Shwartz et al. (2011) have shown in the context of hospitals, facilities that do best on a composite measure are often not in the group of highest performers on many of the individual measures. It seems as though facilities have two strategies: (1) to concentrate on some measures at the expense of others; or (2) to attempt to do pretty well on all of the measures, recognizing that as a result they may not be a top performer on very many of the individual measures. It seems reasonable that policies should be designed to recognize and reward both types of behavior. Our main focus in this paper is the composite measure component of a P4P program.

Our main innovation in methodology is to illustrate the use of Data Envelopment Analysis (DEA) (Charnes, Cooper, & Rhodes, 1978) to calculate a composite measure of quality and an associated uncertainty interval that exploits the hierarchical structure of the data generating process that leads to the inputs at each facility. To understand the implications of using DEA in this way, it is useful to compare the DEA composite measure we construct to composite measures calculated using other approaches. Specifically, we consider two other approaches: (1) Using opportunity-based weights (OBWs) to combine individual performance indicators into a composite measure. This is the approach used by CMS in its early P4P Demonstration Programs (Kahn, Ault, Isenstein, Potetz, & Gelder, 2006; Reeves et al., 2007) and is a commonly used way of calculating a composite measure from individual measures when the individual measures are proportions. We show later in the paper that OBWs are a form of equal weighting; and (2) Using a Bayesian latent variable model (BLVM) to estimate the "underlying" composite measure (Landrum, Bronskill, & Normand, 2000). The main questions we examine are the sensitivity of the resulting facility ranks and identification of high and low performers to the approach used to calculate the composite measure; and, in the context of a specific type of P4P program, the impact of the different approaches on the percentage of the pool of resources available in the program that are allocated to each facility. We identify high and low performers based on whether interval estimates of performance are below overall mean performance (high performers with lower likelihood of adverse events) or above (low performers with higher likelihood of adverse events). Interval estimates are easily determined for a Bayesian hierarchical latent variable model.

The most widely used approach for determining interval estimates in DEA uses bootstrapping with a kernel density estimator applied to the estimated frontier to approximate the DEA frontier data generating process (Simar & Wilson, 1998, 2000a, 2000b, 2011a, Chap. 10; Kneip, Léopold, & Wilson, 2008, 2015). These interval estimates reflect the error in estimating the frontier given the location of a specific set of facilities in multidimensional space. This approach as currently developed does not allow constraints on the DEA weights, something that is important in our context. Also, it takes as fixed each facility's input and output data. However, in our situation, there is uncertainty due to the fact that the inputs are calculated from patients at each facility, i.e., there is a hierarchical structure to our data. The particular set of patients at each facility can be viewed as a random sample from a population of "potential" patients that would use the facility if the health service need arose. Thus, in our situation, there is variation in the estimates of DEA efficiencies due to this source of uncertainty in the inputs. We capture this uncertainty in our interval estimates; they reflect how stable facility performance is likely to be in future periods with different realizations of patient arrivals. Because our approach may be useful in other situations with hierarchical data structures in which DEA is used, we describe the approach in a somewhat more general context than required by our application and provide details in Appendix A.

In order to motivate the different approaches to calculating a composite measure, we provide a little background on composite measures and an example that highlights the underlying conceptual distinctions.

1.1. Background on composite measures

Different approaches have been proposed to create composite measures of health care provider performance (e.g., Caldis 2007; Jacobs, Goddard, & Smith, 2005; Jha, Zhonghe, Orav, & Epstein, 2005; Landrum et al., 2000; Lied, Malsbary, Eisenberg, & Ranck, 2002; O'Brien et al., 2007; Reeves et al., 2007; Staiger, Dimick, Baser, Fan, & Birkmeyer, 2009; Shwartz, Peköz, Christiansen, Burgess, & Berlowitz, 2013; Werner & Bradlow, 2006; Zaslavsky, Shaul, Zaborski, Cioffi, & Clearly, 2002). In this paper, we step back from the specific ways in which these and other such approaches differ and focus on the bigger picture: what are the conceptual and philosophical differences that underlie several different approaches to profiling and what are the implications of these differences for