Who is an Efficient and Effective Physician? Evidence from Emergence Medicine

Soroush Saghafian
Harvard Kennedy School, Harvard University, Cambridge, MA,

Raha Imanirad
Technology and Operations Management, Harvard Business School, Cambridge, MA,

Stephen J. Traub
Department of Emergency Medicine, Mayo Clinic Arizona, Phoenix, AZ

Improving the performance of the healthcare sector requires an understanding of the effectiveness and efficiency of care delivered by providers. Although this topic is of great interest to policymakers, researchers, and hospital managers, rigorous methods of measuring effectiveness and efficiency of care delivery have proven elusive. Through Data Envelopment Analysis (DEA), we make use of evidence from care delivered by emergency physicians, and develop scores that gauge physicians' performance in terms of effectiveness and efficiency. In order to validate our DEA scores, we independently use various Machine Learning (ML) algorithms, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Random Forest (RF), a Generalized Linear Model (GLM), and Least Absolute Shrinkage and Selection Operator (LASSO). After validating our DEA scores via comparison with predictions made by these algorithms, we make use of them to identify the distinguishing behaviors of highly effective and efficient physicians. We find that highly effective physicians order less tests compared to their peers and maintain their effectiveness when working under high workloads. We also observe that highly efficient physicians order less tests on average and become even more efficient during high-volume shifts. Importantly, our results indicate a statistically significant positive relationship between a physician's effectiveness and efficiency scores suggesting that, contrary to conventional wisdom, effectiveness and efficiency in care delivery should be viewed as compliments not substitutes. In addition, we find that effectiveness is lower among physicians who have higher job tenure or average test order count. Efficiency, however, is lower among physicians with less experience (measured in number of years after graduation from medical school) or high average test order count. Furthermore, our results indicate an increase in a physician's average efficiency and a decrease in his/her average effectiveness when faced with high workloads. Finally, we find evidence of peer influence on a focal physician's effectiveness and efficiency, which suggests an opportunity to improve system performance by taking physician characteristics into account when determining the set of physicians that should be scheduled during the same shifts.

Key words: Data Envelopment Analysis, Physician Performance Evaluation, Physician Effectiveness and Efficiency, Peer Influence

1. Introduction

Motivation. Healthcare spending is projected to rise to 19.9% of the GDP by 2025 (Keehan et al. 2017), spurring interest in finding new ways to increase both the effectiveness and efficiency of care delivery. As most decisions regarding utilization of healthcare services are ultimately made by frontline clinicians (Tsugawa et al. 2017), understanding and evaluating provider performance could help identify sources of waste in the healthcare sector. Although care delivery performance measurement initiatives have proliferated in recent years, there are few rigorous methods to evaluate the effectiveness and efficiency of physicians. A careful method for evaluating the effectiveness and efficiency of physicians is especially needed for understanding what the efficient and effective physicians do differently than their peers. This understanding of best practices can, in turn, result in training more efficient and effective physicians, and thereby, improve the performance of the healthcare sector.

In this study, we focus on care delivery in hospital Emergency Departments (EDs). Specifically, we collect a large dataset of care delivered by ED physicians that includes more than 115,000 patient visits. We employ Data Envelopment Analysis (DEA)—a linear programming (LP) optimization technique that provides a multi-dimensional evaluation tool—to develop and evaluate scores related to physician efficiency and effectiveness. We validate our generated DEA scores by making use of various Machine Learning (ML) algorithms, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Random Forest (RF), a Generalized Linear Model (GLM), and Least Absolute Shrinkage and Selection Operator (LASSO). Our results show that there is a 76% overlap between the results derived from the best ML approach and those obtained from our DEA models, giving us confidence about the validity of our DEA models. Unlike the ML algorithms, however, DEA provides an important advantage in terms of *interpretability*, since it offers a clear input-output view of a physician's performance and avoids any "black-box" operations. Thus, it can be easily communicated to (a) hospital administrators who are interested in improving the efficiency and effectiveness of care delivery at their hospital, and (b) physicians who are interested in boosting their own individual performance.

In order to learn about what the high-performing physicians do differently than other physicians, and thereby generate insights into best practices, we conduct a second-stage analysis in which we use our DEA scores along with a Tobit framework to identify factors (e.g., test order count, experience, etc.) associated with higher levels of performance. Furthermore, we use our framework to study how physicians influence each other's efficiency and effectiveness. In particular, we make use of our DEA scores and consider various peer physician characteristics, including relative effectiveness, efficiency, gender, and type of medical degree (MD vs. DO) to examine how such characteristics affect a focal physician's efficiency and effectiveness.

Data and Setting. Our data consist of detailed care delivery information associated with 115,350 patient visits in a leading U.S. hospital. Our partner ED is equipped with an emergency medicine team comprising 32 board-certified physicians and more than 70 registered nurses. All patients in our partner ED are algorithmically assigned to physicians upon arrival through an automated rotational patient assignment process (Traub et al. 2016). This workflow essentially removes all patient selection biases or preferences of physicians in "cherry-picking" their patients.

We included all patient visits from July 12, 2012, to July 31, 2016 who were identified in the electronic health record system as having been seen by an ED physician in our dataset. Patient-specific data include demographic (age, gender, race, etc.) and insurance information. Encounter-level data include laboratory tests, chief complaint, Emergency Severity Index (ESI) level (a five-level triage scale that categorizes patients according to their acuity level), day of the ED visit, and time of the day, among others. To avoid distortion of the results by outliers, 4 physicians with relatively low patient volumes (fewer than 200 visits over the 4-year period) were excluded from the analysis. Our final dataset comprises 110,325 patient-visit-level observations.

Research Questions. We address four research questions as follows. Research Question 1: Are effectiveness and efficiency of a physician substitutes (negatively correlated) or complements (positively correlated)? Research Question 2: What is the relationship between effectiveness/efficiency of a physician and various characteristics, including those of the physician (e.g., test order count, experience, tenure), patients (e.g., race, gender, age, ESI), and the environment (e.g., ED volume/workload)? Research Question 3: What do highly effective and efficient physicians do differently than their peers? Research Question 4: How do physician peers influence each other's effectiveness and efficiency? Addressing these questions enables us to (a) shed light on factors that affect physicians' effectiveness and efficiency, and (b) provide actionable insights into ways that physicians' effectiveness and efficiency can be improved.

Main Findings. Regarding Research Question 1, our results indicate that a conventional wisdom about the efficiency-effectiveness tradeoff in the healthcare sector might not be true. This conventional wisdom suggests that improving the efficiency of care delivery comes at the price of lowering effectiveness. Contrary to this conventional wisdom, we find that not only there are physicians who obtain high scores on both efficiency and effectiveness dimensions, but that overall there is a statistically significant positive association between physicians' efficiency and effectiveness scores. This implies that physicians who are efficient in care delivery are also more likely to provide effective care (and vice-versa). Our results, hence, suggest that physician effectiveness and efficiency serve as *complements* and not *substitutes*.

With respect to Research Question 2, we find that a physician's efficiency score is negatively associated with his/her average number of test orders per patient visit and positively correlated

with his/her experience (measured in number of years after graduation from medical school). This implies that efficient physicians are those who (a) order less tests, and (b) are more experienced. In addition, we observe a statistically significant negative relationship between a physician's effectiveness and his/her job tenure (measured in number of years the physician has worked in our partner ED). This finding might be related to a selection bias: the ED might have imposed higher hiring standards in recent years or simply has been able to attract physicians with higher effectiveness. However, it might also be due to a difference in motivation level of new hires versus existing physicians. Newly hired employees typically have a higher motivation level of establishing good performance than existing employees (Hackman and Oldham 1980, Kass et al. 2001, Bruursema et al. 2011), and so inherently they might score higher on the effectiveness metric. Our data is insufficient for distinguishing between these potential hiring and motivation differences (which are both difficult to measure and hidden to us). Nevertheless, our finding that job tenure negatively impacts physicians' effectiveness provides an important avenue for future research to shed light on mechanisms that might improve effectiveness of care delivery (e.g., motivational training programs, providing performance-based incentives for physicians with long job tenure, or making use of specific hiring procedures). Furthermore, our results show that patient characteristics have little, if any, effect on a physician's effectiveness and efficiency. We also find that, faced with high workloads (i.e., during high-volume shifts), a physician's efficiency improves while his/her effectiveness declines.

Addressing Research Question 3, our findings suggest that highly effective physicians order less tests on average compared to their peers. This indicates that, compared to other physicians, they are able to order tests more intelligently: they eliminate unnecessary tests while still ordering the necessary ones. Our results also indicate that during high-volume shifts highly effective physicians are able to maintain their effectiveness level more than their peers. Similarly, our results show that highly efficient physicians have a lower average test order count and become even more efficient under high workloads.

Finally, addressing our last research question (Research Question 4), our findings suggest that working alongside more effective and efficient peers is negatively associated with improving a focal physician's effectiveness and efficiency, respectively. This is consistent with the findings in Saghafian et al. (2019), which studies influence of physicians on each other's performance using a different methodology (neither DEA nor Machine Learning), and shows that a "resource spillover" effect caused by the existence of shared resources with limited capacities in the ED is the mechanism driving peer influence.

Implications. Our results have various implications for both hospital administrators and physicians. In particular, our DEA methodology allows hospital administrators to utilize a transparent

and easy-to-understand scoring system to evaluate the efficiency and effectiveness of care delivered in their hospital. Similarly, it allows individual physicians to observe their weaknesses and realize the advantages of following what the highly efficient and effective physicians do in their practice. We expect well-designed training programs to be able to facilitate this learning process. In addition, our results have implications for physician scheduling programs, where hospital administrators need to decide upon the set of physicians who should work during the same shifts. In particular, our analyses of our DEA scores show that effectiveness and efficiency scores of a physician decline while working alongside more effective and efficient peers. This observation can be incorporated in future scheduling programs and utilized as a mechanism for improving the overall performance of physicians. Finally, as noted earlier, our results provide an important avenue for future research to explore and implement mechanisms including designing motivational training programs, providing performance-based incentives for physicians with long job tenure, or making use of specific hiring procedures.

2. Related Studies

Evaluating the performance of physicians has gained attention in research as health policymakers look for ways to drive quality improvement and increase physicians' accountability for achieving quality goals. Most lines of research on this topic have focused on specific patient conditions. For example, Glickman et al. (2008) use clinical measures such as performing a diagnostic electrocardiogram (ECG) for syncope in patients older than 60 years as a performance measurement criterion of physicians. Hess et al. (2011) utilize physician performance measures such as completion of retinal and foot exams and blood pressure test to assess the quality of care provided to diabetic patients. However, the findings generated from such studies may not be generalizable to settings such as EDs where there is heterogeneity in patient population. Other studies have evaluated behavioral aspects of physician performance using questionnaires (Smith et al. 2004) and patient chart audits (Goulet et al. 2002). Qualitative metrics, however, are difficult to measure and may cause bias in performance evaluation.

Various performance-specific measures have been used to assess the performance of ED physicians. A review of the literature highlights ED time intervals such as time between patient arrival to initial clinical assessment, Length of Stay (LOS), as well as patients left without being seen, readmission within 72 hours and mortality/morbidity as most frequently used performance measures (see Fernandes et al. 1997, Spaite et al. 2002). Using pure performance measures in evaluating ED physicians, however, does not account for the amount of resources utilized by physicians. In a setting such as an ED, where resources are shared and constrained, resource utilization plays an important role in assessing physician effectiveness and efficiency. Hence, using a methodology such

as DEA, which incorporates resource utilization into performance evaluation, lends itself well to evaluating physician performance in EDs.

DEA has been applied in a variety of healthcare settings including hospitals (Sherman 1984, Grosskopf and Valdmanis 1987), veterans administration medical centers (Harrison and Ogniewski 2005), and organ procurement organizations (Ozcan et al. 1999) to evaluate the relative performance of healthcare institutions. While performance evaluation of hospitals has been explored in prior literature (Zheng et al. 2018, Castelli et al. 2015, Varabyova and Schreyogg 2013, Hollingsworth 2008), the performance of physicians has proven to be more difficult to assess because of diversity in patient mix and treatments, and differences among specialties, among others (Storfa and Wilson 2015). Hence, macro parameters and proxies such as billing and reimbursement are often used to capture physician performance (Johannessen et al. 2017). For example, Wagner et al. (2003) propose DEA models focused on cost containment by using admission and patient visit payments as input variables. Collier et al. (2006) use the total billable charges attributed to physicians as one of the outputs of their proposed model. The authors, however, assume uniform resource utilization among physicians. Other studies use costs of treating specific patient conditions such as sinusitis (Ozcan et al. 2000) and asthma (Ozcan et al. 1998) in their suggested DEA models.

Our study contributes to this literature by proposing two separate DEA models for evaluating physician efficiency and effectiveness. Our choice of the models' input and output variables reduces the risk of overfitting to our study setting and increases generalizability of the models to any ED setting. Specifically, we do not use parameters specific to patient health conditions or physician practice style in our models. Rather, we investigate the effects of physician-specific factors on physician performance in a second-stage analysis, where we identify characteristics of effective and efficient physicians.

Our work is also related to studies on speed-quality tradeoffs. Anand et al. (2011) use a queueing framework to examine the speed-quality tradeoff in a customer-intensive service setting and study how service providers make the optimal speed-quality tradeoff. Saghafian et al. (2018) study the speed-quality tradeoffs in a telemedical physician triage system in the context of an ED setting. Several other studies have examined the interactions between speed and quality of service in different settings assuming an exogenous customer demand including Hasija et al. (2009) (call centers) and Wang et al. (2010) (medical diagnostic services). Our work contributes to this stream of literature by examining the relationship between physician effectiveness and efficiency using metrics derived from the DEA methodology.

3. DEA Models

DEA, first introduced by Charnes et al. (1978), is a methodology useful in evaluating the relative performance of a set of decision making units (DMUs) in a multiple input, multiple output setting.

A DMU can be viewed as an entity responsible for converting a number of inputs into a set of outputs and whose performance is to be evaluated relative to its peers (Cooper et al. 2007). Contrary to a central tendency approach, which evaluates units relative to an average performer, DEA computes a DMU's relative performance by using the best-performing units as the basis for comparison. One of the key advantages of DEA over other regression-based statistical methods is that it does not require specification of any functional relationship (e.g., a specific linear or non-linear model) between inputs and outputs. As a result, DEA can uncover information that remains hidden from other parametric methodologies, and hence, might capture a more complete picture of a DMU's performance relative to the resources it uses. As a data-driven approach, however, DEA is vulnerable to data errors and outliers.

The conventional input-oriented DEA methodology evaluates each DMU j in the population based upon a set of inputs $\{x_{ij}\}_{i=1}^{I}$ and outputs $\{y_{rj}\}_{r=1}^{R}$ by assuming a proportional reduction in all inputs while maintaining a fixed level of outputs. In an output-oriented setting, this methodology provides for a proportional expansion in outputs rather than a reduction in inputs while keeping inputs constant. For the goals of this study, we make use of both the input- and output-oriented mechanisms.

The original DEA model is based on a Constant Returns to Scale (CRS) methodology. The input-oriented CRS model takes the following form:

$$\begin{aligned} max \quad & \theta = \frac{\sum_r u_r y_{rj_o}}{\sum_i \nu_i x_{ij_o}} \\ s.t. \quad & \frac{\sum_r u_r y_{rj}}{\sum_i \nu_i x_{ij}} \leq 1 \qquad j = 1, ..., n, \\ & u_r, \nu_i \geq 0, \qquad r = 1, ..., R; \quad i = 1, ..., I, \end{aligned} \tag{1}$$

where y_{rj_o} and x_{ij_o} represent the output(s) and input(s) of DMU j_o, respectively, and $\{u_r\}_{r=1}^{R}$ and $\{\nu_i\}_{i=1}^{I}$ are decision variables and represent the set of most favorable weights for the DMU under evaluation in the sense of maximizing the ratio scale. The constraints ensure that, when this set of weights is applied to each DMU in the population, no unit's efficiency exceeds 1. The maximum value obtained for DMU j_o is that unit's DEA score, and a value of 1 signifies a frontier-efficient unit. Contrary to composite scoring methods which apply a single set of weights to each unit in the population, DEA assigns a different set of weights to each DMU under evaluation. Hence, it avoids the subjective nature of weight assignment in multi-objective problems.

In order to evaluate the performance of physicians, we develop two DEA models: (1) an *effectiveness DEA model* (see Section 3.1), and (2) an *efficiency DEA model* (see Section 3.2). To improve the power of our statistical analyses and ensure enough variation across the models' input/output

parameters, we conduct our analyses at the physician-year level. Specifically, we design our DMUs so that they each capture a physician's performance in a particular year. To this end, we construct a panel dataset with 106 physician-year observations. Our DEA models, therefore, evaluate the effectiveness and efficiency of individual physician i who uses hospital resources to deliver care in a given year t relative to his/her peers. To ensure that our findings are not sensitive to this choice, we also conduct our analyses at a physician-quarter level. Results presented in Online Appendix A show that our inferences remain unchanged. Furthermore, since there is no reason to believe that an increase in inputs results in a proportional change in outputs (and vice-versa) in our effectiveness and efficiency DEA models, we have used the Banker- Charnes-Cooper (BCC) model (Banker et al. 1984) which extends the CRS model to allow for variable returns to scale. We tested this assumption by making use of Simar and Wilson's (Simar and Wilson 2002, Simar and Wilson 2011) returns-to-scale tests for both the effectiveness and efficiency DEA models.

The choice of the input and output variables in each model is based on the view of the physician as a "production entity" which utilizes hospital resources (inputs) to generate efficient and effective care (outputs). It is important to note that in DEA there is no objective definition of the right variables to use as inputs and outputs. We have chosen to define the models' input and output variables in terms of parameters (a) that best reflect a physician's performance, (b) for which there is at least face validity and some level of agreement among physicians, and (c) that are discussed in the literature of Emergency Medicine and ED operations as common measures. For example, to define our output variables, we note that efficiency in the ED can be measured in multiple ways. We primarily focus on a physician's average contact-to-disposition time (the time from when the physician initiates the first contact with the patient until a disposition order is issued for the patient), because all else equal a lower contact-to-disposition time means that a higher number of patients can be moved through the ED per unit of time (i.e., a higher ED throughput). Given that ED crowding has reached epidemic proportions in the last several years, improving physicians' contact-to-disposition time has become even more important (Salway et al. 2017).

Similarly, we consider the percentage of discharged patients (i.e., those not admitted to the hospital after their ED visit) who do not return to the ED within 72 hours as our primary output variable for our effectiveness DEA model. Returns to the ED within 72 hours of discharge may result from a sub-optimal (i.e., ineffective) first visit, in which not all medical issues were sufficiently identified or addressed. The 72-hour rate of return has also been proposed as a measure of quality in the Emergency Medicine literature (see, e.g., Abualenain et al. 2013, Pham et al. 2011, Klasco et al. 2015) although using it for measuring quality (which is different than effectiveness) of care is controversial. Nevertheless, to check the robustness of our results, we have repeated our analyses

with different combinations of input/output variables for both our effectiveness and efficiency DEA models, and have observed that our main results hold. (see Section 7 for our robustness checks).

Furthermore, as noted earlier, we validate our DEA results through comparison with results obtained using various ML algorithms (see Section 4) that do not necessarily rely on the same set of variables as those used in our DEA models. In particular, unlike our DEA models, these ML algorithms are given the entire dataset and are able to either use it as a whole or select the important variables using some predetermined regularization techniques. The fact that we observe similar results between our DEA models and these ML algorithms gives us further confidence about the validity of our DEA models.

Finally, we note that due to the nature of the automated rotational patient assignment algorithm implemented in our partner hospital, which randomly assigns arriving patients to physicians, risk-adjustments of outcome measures are likely not essential. Nevertheless, in our statistical analyses aimed at generating insights into best practices (i.e., learning about what effective and efficient physicians do differently than their peers), we control for various patient characteristics that might affect physician performance (see Section 5).

3.1. Effectiveness DEA Model

Our main effectiveness DEA model uses the following set of variables as inputs and outputs. As noted earlier, in our robustness checks, we test the validity of our main DEA models by repeating our analyses with different combinations of input/output variables. We also validate them using ML algorithms that do not necessarily rely on the same set of variables.

Output:

- *Rate of discharged patients who do not return within 72 hours*: Since a high 72-hour return rate is an undesirable indicator of care delivery effectiveness in the ED (see, e.g., Abualenain et al. 2013, Pham et al. 2011, Klasco et al. 2015), we use the proportion of patients discharged by a physician who did not return to the ED within 72 hours of their original discharge as the model's output variable.

Input:

- *Average patient Length of Stay (LOS):* This variable captures patients' time in the ED from registration to discharge.

For the effectiveness model, we have chosen the output-oriented DEA approach based on which the conceptual goal is to maximize outputs for a given level of inputs. Specifically, we compare physicians' percentage of patients who are discharged home after their ED visit and do not return within 72 hours (output) for a given level of LOS (input), where LOS can be viewed as a surrogate measure for using hospital resources (e.g., using diagnostic test services, ED beds, etc.). Intuitively,

physicians who score higher on the effectiveness metric are those with a lower 72-hour rate of return for a fixed level of overall use of ED resources measured by the surrogate variable, LOS. From a patient perspective, this roughly means that the service is considered to be more effective if the chance of returning to the ED (e.g., due to an unresolved issue) is minimized per hour spent in the ED.[1] Both the LOS and 72-hour rate of return metrics have been used in the literature as valid measures (see, e.g., Chilingerian 1995, Fiallos et al. 2017). We refer to the θ scores (see Eq. (1)) generated by the DEA model with the above input-output parameters as physicians' *effectiveness scores*. Similarly, we refer to the θ scores generated by the DEA model with the input-output parameters described in the next section as physicians' *efficiency scores*.

3.2. Efficiency DEA Model

Our main efficiency DEA model uses the following set of variables as inputs and outputs.

Outputs:

- *Low ESI*: Percentage of patients served by the physician who have ESI levels 1 and 2 (i.e., high-acuity patients);
- *Older than 65*: Percentage of patients served by the physician who are older than 65.

Input:

- *Average contact-to-disposition time:* This variable denotes the time from the physician's initial contact with the patient until a disposition order is issued.

For the efficiency model, we use an input-oriented approach to test whether a DMU (i.e., physician) under evaluation can reduce its input while keeping the outputs at their current levels. Intuitively, physicians with higher efficiency scores in this setting are those who have a lower average contact-to-disposition time for a given mix of patients they serve. Low-ESI patients and those older than 65 are known to be patients that have a relatively higher contact-to-disposition time than other patients (Latham and Ackroyd-Stolarz 2014). Thus, assuming that two physicians serve the same mix of patients (ratio of low-ESI and older patients to other patients), the one who can maintain a lower contact-to-disposition time, will have a better throughput (a widely-used measure of operational efficiency).

Our rationale for the efficiency model's input/output variables described above is mainly based on our discussions with ED physicians.[2] In particular, our discussions indicate that while a physician's ability to serve patients efficiently might be attributable to his/her cognitive abilities, his/her

[1] In EDs, the service is provided by a specific physician who is in charge of the patient, and the ED service is very rarely composed of a teamwork among physicians (see, e.g., Saghafian et al. 2012, Saghafian et al. 2019, and the references therein). Thus, a patient's outcomes are directly related to the physician who serves him/her.

[2] One of the authors of this paper is the chairman of the ED of our partner hospital, which is a leading hospital in the U.S.

average contact-to-disposition time given a fixed mix of low-ESI and older patients s/he sees can serve as a good proxy for measuring such abilities. We also note that while we have chosen patient LOS as the effectiveness model's input variable, our choice for the efficiency model's input variable is a physician's average contact-to-disposition time. The reason is that LOS captures the total time a patient spends in the ED, which is not fully controllable by the physician. In contrast, contact-to-disposition time is at the discretion of physicians. Finally, we note while LOS and contact-to-disposition time are positively correlated, the fact that our effectiveness and efficiency models use different DEA orientations ensures that any potential relationship between physicians' effectiveness and efficiency scores is not merely due to the inherent dependency between these variables.

3.3. Physician-Pair DEA Models

Our DEA models described in the previous sections allow us to capture individual physicians' effectiveness and efficiency scores, and answer our first three research questions (Research Questions 1, 2, and 3). In order to also examine the effects of peers' presence on a focal physician's effectiveness and efficiency scores (Research Question 4), we use a variation of the proposed DEA models in which each DMU comprises physician i who has worked alongside his/her peer physician j in year t. Our physician-pair DEA models, hence, capture a focal physician i's average effectiveness and efficiency scores while working alongside his/her peer physician j in year t. We identify a focal physician's peers as those physicians who have worked alongside the focal physician during the same shifts. We then construct a dataset comprising of every combination of focal-peer physician pairs corresponding to each year of our study period. This leaves us with 2,268 physician-pair observations that we can use for our physician-pair DEA models. Making use of all of our four DEA models (effectiveness and efficiency for both individual and physician-pair performance), in turn, enables us to provide answers to our four research questions (see Section 6).

4. Machine Learning (ML) Algorithms

To test the validity of our generated DEA scores, in addition to re-running our DEA models by making use of different sets of input/output variables (see Section 7), we utilize various ML algorithms including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Random Forest (RF), a Generalized Linear Model (GLM), and Least Absolute Shrinkage and Selection Operator (LASSO). We first compare these algorithms in terms of their performance in predicting the effectiveness and efficiency of physicians. We do so via 5-fold cross-validation, which allows measuring the average out-of-sample performance of these algorithms by creating different training and test datasets. We label the highly effective and efficient physicians in the training sets as those with lower-than-average 72-hour rate of return and

contact-to-disposition time, respectively. The input variables (potential predictors) that are used by the ML algorithms include various patient characteristics (age, gender, race, ESI), physician characteristics (average test order count, job tenure, admission rate, etc.), and ED characteristics (e.g., ED volume). A summary statistics of these variables is presented in Table 1. We omit the 72-hour rate of return and average contact-to-disposition time variables from the set of potential predictors in the effectiveness and efficiency ML models, respectively, since these represent the outcome variables (i.e., what the algorithms are asked to predict).

We compare the predictive power of the ML algorithms using the Area Under the Curve (AUC) measure as well as classification accuracy and the Kappa coefficient (which adjusts for the effect of random chance on accuracy). These measures (calculated using 5-fold cross-validation) are presented in Figures 1-4. As demonstrated in these figures, the RF algorithm results in the highest AUC, accuracy, and Kappa measures compared to the other algorithms. We, therefore, use the RF model to predict the highly effective and efficient physicians in the test sets. We then compare the predictions made by the RF approach to those derived from our DEA models. To this end, we define the following four groups of physicians:

Group 1: Highly effective / Highly efficient;

Group 2: Highly efficient / Lowly effective;

Group 3: Highly effective / Lowly efficient;

Group 4: Lowly efficient/ Lowly effective,

where we use the average effectiveness and efficiency DEA scores to categorize physicians into Groups 1-4. Independently, we use the predictions obtained from the best ML approach — the RF algorithm — to classify physicians into the aforementioned four groups. As noted earlier, we trained all the ML algorithms including RF independent of the DEA scores and after labeling the highly effective and efficient physicians in the training sets as those with lower-than-average 72-hour rate of return and contact-to-disposition time, respectively. We then compare the classifications derived from the DEA and ML approaches as illustrated in Figures 5-8. In these figures, red data points indicate highly effective and efficient physicians that are separated by the average line (depicted in blue color) from all other observations (black data points).

We find an average 76% overlap between the classifications obtained via the DEA and ML approaches. This finding validates the accuracy of our proposed DEA models to a great extent. This is especially the case since the RF algorithm uses a different set of input variables compared to those used in our DEA models. For example, Figures 1 and 2 in Online Appendix B present the variable importance graphs corresponding to the RF effectiveness and efficiency models, respectively.[3] As

[3] To measure variable importance, these figures use the mean decrease in node impurity (the Gini coefficient) such that a higher mean decrease in the Gini coefficient denotes higher variable importance.

Table 1 Summary Statistics - ML Variables

Variable	Mean	SD	Min	Max
Patient Characteristics				
Older than 65 Patients (%)	45	2.88	39	58
Female Patients (%)	53	1.83	48	58
White Patients (%)	91	1.54	87	95
ESI Levels 1 and 2 (%)	15	2.13	7.5	21
Physician Characteristics				
Test Order Count	144.13	24.37	87.55	215.39
Experience (Years)	22.16	7.49	6	39
Job Tenure (Years)	8.38	6.01	0	18
Admission Rate (%)	0.11	0.03	0.05	0.20
Overcalling Rate (%)	0.18	0.05	0.08	0.33
Undercalling Rate (%)	0.04	0.02	0	0.11
LOS (Minutes)	235.02	26.84	180.64	297.81
72-hr Rate of Non-Return (%)	0.97	0.01	0.94	0.99
Contact-to-Disposition Time (Minutes)	144.13	24.37	87.55	215.39
ED Characteristics				
ED Volume (Patients per Physician Shift)	23.77	4.90	12.20	41.85

Note: $N = 106$. Observations are at the physician-year level.

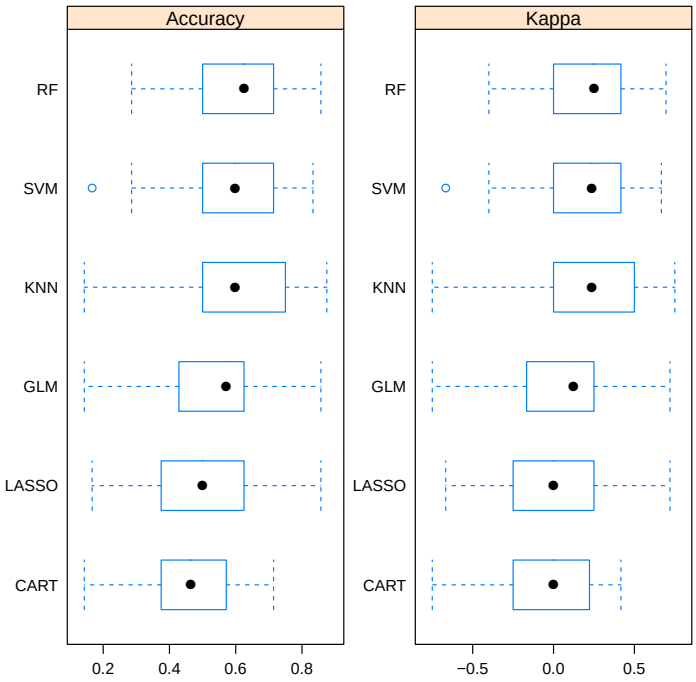


Figure 1 Accuracy - Kappa Measures of Effectiveness ML Models

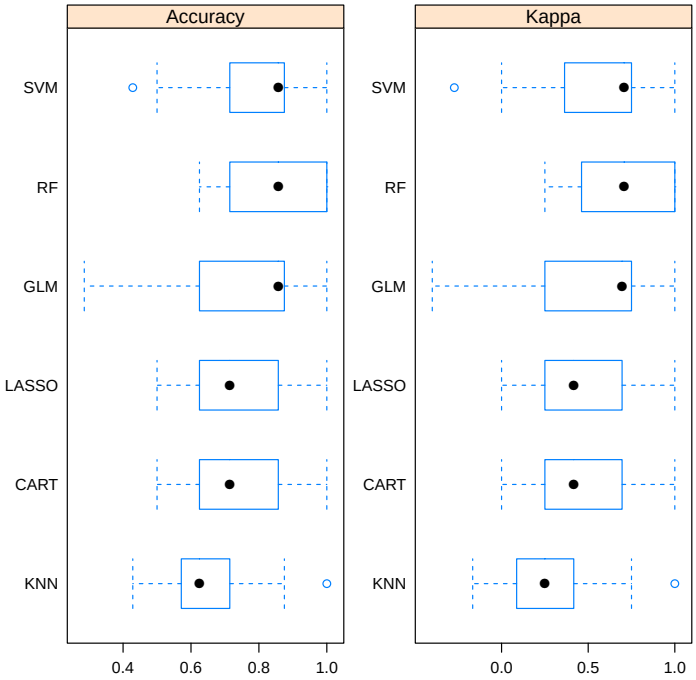


Figure 2 **Accuracy - Kappa Measures of Efficiency ML Models**

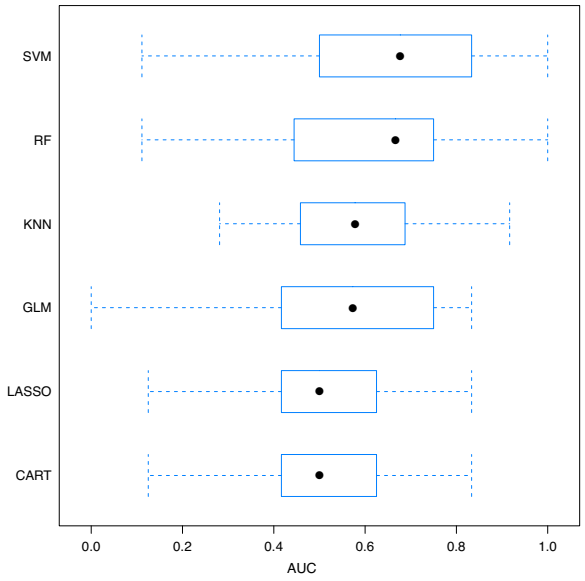


Figure 3 **AUC Measure - Effectiveness ML Models**

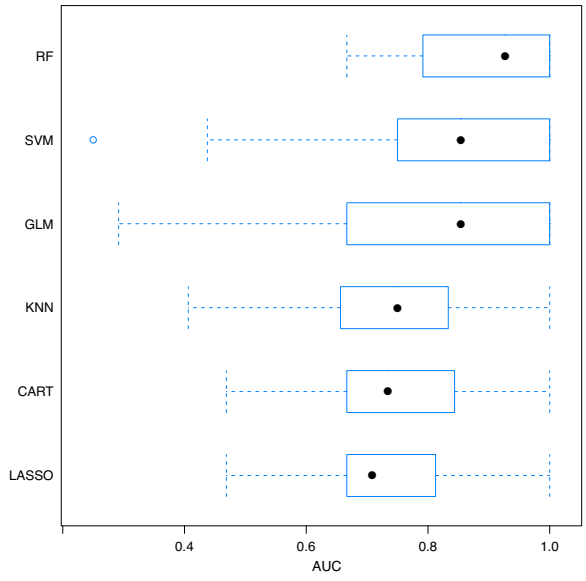
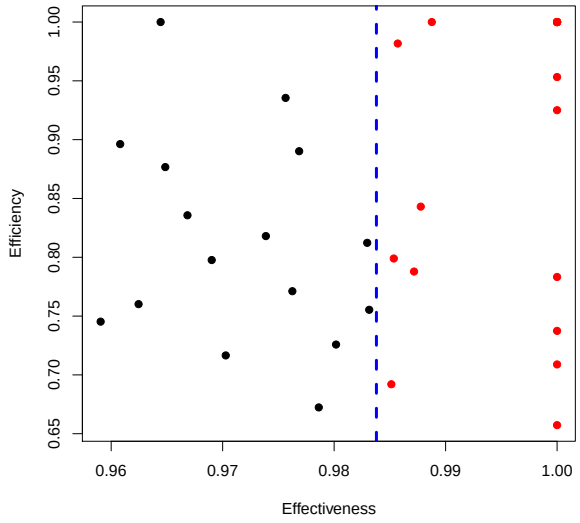
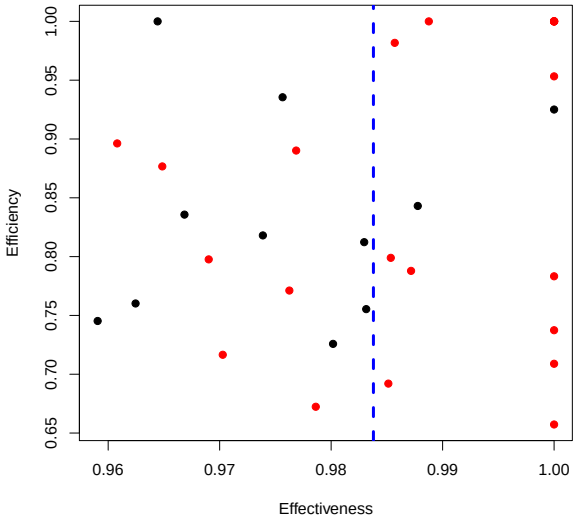


Figure 4 **AUC Measure - Efficiency ML Models**



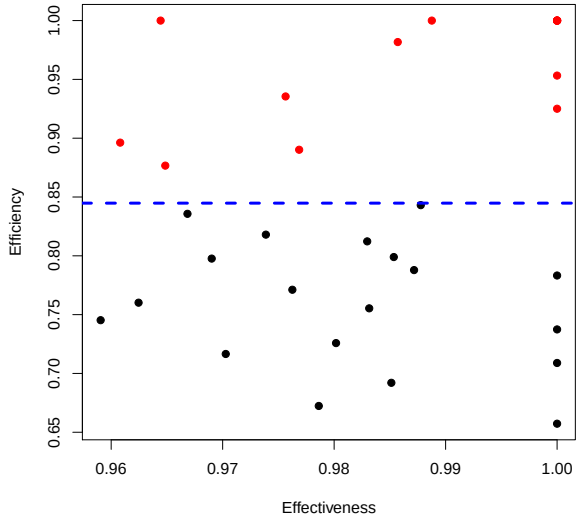
Note: The dotted blue line denotes the average effectiveness score.

Figure 5 DEA Model - Effectiveness Classification



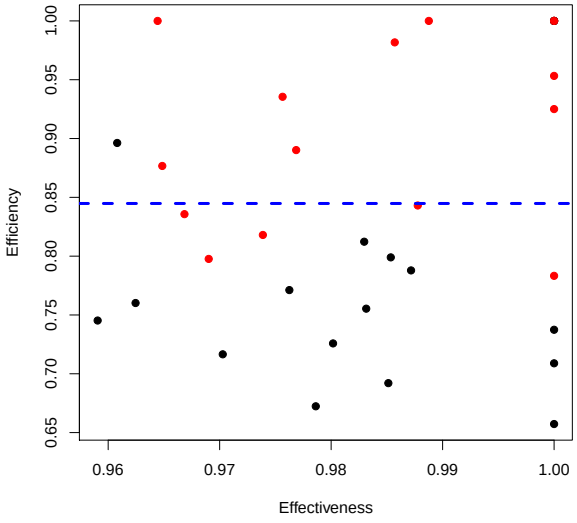
Note: The dotted blue line denotes the average effectiveness score.

Figure 6 ML Model (RF) - Effectiveness Classification



Note: The dotted blue line denotes the average efficiency score.

Figure 7 DEA Model - Efficiency Classification



Note: The dotted blue line denotes the average efficiency score.

Figure 8 ML Model (RF) - Efficiency Classification

demonstrated in these figures, the RF algorithm's selection of important variables is completely different than that of our DEA models. Yet, the results obtained from the RF approach significantly overlaps with those of our DEA models, which gives us confidence about the validity of our DEA models.

In what follows, we make use of our DEA scores to generate insights into our Research Questions 1-4. We make use of our DEA scores as opposed to any of the above-mentioned ML algorithms (e.g., RF) for three main reasons: (1) compared to the black-box nature of the ML algorithms, the DEA methodology provides more transparency to the user, (2) it is easier to interpret the DEA results and communicate the derived insights using an input-output view of a DMU, and (3) the ML algorithms typically require a large set of input variables compared to DEA, which makes them less useful in hospitals in which not all such variables are collected.

5. Statistical Methodology

To gain insights into our Research Questions 1-3, we regress the generated individual DEA scores of physician i in year t (θ_{it}) (defined in Section 3), on a set of explanatory variables related to physician characteristics which we denote by U_{it}. The regression model takes the following general form:

$$\theta_{it} = \beta_1 U_{it} + \beta_2 W_{it} + \beta_3 E_{it} + \gamma_t + \epsilon_{it}, \tag{2}$$

where W_{it} denotes the vector of control variables which include patient characteristics such as average age, gender, race, and ESI. E_{it} indicates the average ED volume of those shifts that physician i is assigned to in year t and γ_t denotes year fixed effects. ϵ_{it} is a statistical noise. In order to examine the potential influence of peers' characteristics on a focal physician's average effectiveness and efficiency (Research Question 4), we make use of the following regression model:

$$\theta_{ijt} = \beta_1 U_{ijt} + \beta_2 W_{ijt} + \beta_3 Z_{ijt} + \beta_4 E_{ijt} + \sigma_i + \gamma_t + \epsilon_{ijt}, \tag{3}$$

where θ_{ijt} (defined in Section 3.3) denotes physician-pair DEA scores corresponding to focal physician i when working alongside peer physician j in year t and Z_{ijt} represents indicator variables coded as 1 if peer physician j has a higher effectiveness score, a higher efficiency score, different medical degree, or opposite gender compared to focal physician i. σ_i denotes physician fixed effects.

In order to estimate the coefficients in (2) and (3), a regression technique other than the standard multivariate regression is needed. This is because the standard regression technique assumes a normal and homoscedastic distribution of the noise. However, since the DEA scores are between 0 and 1, our dependent variable is bounded and error terms may not satisfy these assumptions.

Tobit regression can be used whenever there is truncation, causing a mass of observations at a threshold value such as 0 or 1 (Chilingerian 1995). Although unlike the case of truncation, DEA does not exclude observations greater than 1 (or below 0), it does not allow a DMU to be assigned a value outside the range [0, 1]. Hence, DEA easily fits the requirement of the Tobit model (Chilingerian 1995). Following the normalization approach of Greene (1993), which assumes

a censoring point at zero, we transform the DEA scores to:

$$y_{it} = (1/\theta_{it}) - 1,$$

where θ_{it} is the DEA measure of physician i's performance in year t. The transformed DEA scores then become the dependent variable that takes the form:

$$y_{it} = \begin{cases} B'x_{it} + u_{it}, & \text{if } y_{it} > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where B is a vector of coefficients and x_{it} is a vector of covariates, and u_{it} is the error term that is normally distributed with a mean of zero and a variance of σ^2. To account for unobserved serial correlation in the DEA scores, which might arise as a result of calculating a DMU's DEA score by incorporating all other DMUs in the dataset, we use Simar and Wilson's bootstrap procedure (Simar and Wilson 1998) for bias-correction of the scores.

6. Results

To present our results, we first discuss our findings related to our Research Question 1: are effectiveness and efficiency of a physician substitutes (negatively correlated) or complements (positively correlated)? We then present our results related to our Research Question 2: what is the relationship between effectiveness/efficiency of a physician and various characteristics, including those of the physician (e.g., test order count, experience, tenure), patients (e.g., race, gender, age, ESI), and the environment (e.g., ED volume/workload)? Next, we present our findings in answering our Research Question 3: what do highly effective and efficient physicians do differently than their peers? Finally, we discuss our results with respect to our Research Question 4: How do physician peers influence each other's effectiveness and efficiency?

6.1. Effectiveness and Efficiency: Substitutes or Compliments?

We begin our analysis by generating insights into our Research Question 1. We do so by examining the relationship between physicians' effectiveness and efficiency scores. Importantly, we find that higher scores on the efficiency metric do not lead to lower scores on the effectiveness metric, as conventional wisdom might suggest. Rather, there is a statistically significant positive relationship between the two scores (see Table 2): effective physicians are more likely to be efficient as well. This is an important observation, especially in the view of traditional debates that argue healthcare providers cannot become more effective and more efficient at the same time. Indeed, our finding questions the validity of the conventional wisdom, which postulates that being efficient in providing care might require following less effective treatments, and suggests that physician effectiveness and efficiency should be viewed as complements (not substitutes).

Table 2 Regression Results - Effectiveness Model - Individual Physician

Dependent variable: DEA Effectiveness Score

DEA Efficiency Score	0.0565^{***}
	(0.0110)
Job Tenure	-0.0003^{*}
	(0.0001)
Avg Test Order Count	-0.0043^{**}
	(0.0015)
ED Volume	-0.0022^{**}
	(0.0008)
ED Volume x Avg Test Order Count	0.0002^{**}
	(0.00006)
Contact-to-disposition Time	-0.0001^{***}
	(0.00002)
Overcalling Rate	-0.0186^{*}
	(0.0083)
Undercalling Rate	-0.0094
	(0.0209)

Note: $N = 106$. Observations are at the physician-year level.

Note: *p<0.05; **p<0.01; ***p<0.001

6.2. Physician Characteristics

To provide insights into Research Question 2, we next examine the relationship between physicians' DEA scores and their characteristics. As shown in Table 2, our results indicate a statistically significant negative relationship between a physician's effectiveness score and his/her job tenure. This observation implies that more tenured physicians have on average lower effectiveness scores. A reasonable initial assumption might be that as knowledge and skill increase with greater tenure, effectiveness will also improve (Ng and Feldman 2013). In contrast, our finding is more consistent with the literature on job design and motivation that suggests that, as job tenure increases, employees are likely to become less motivated at work (Hackman and Oldham 1980, Kass et al. 2001, Bruursema et al. 2011). However, our results might also be related to a different reason: the ED might have imposed higher hiring standards in recent years or simply has been able to attract physicians who are more effective. Due to lack of data, we are unable to differentiate between these or other potential reasons behind our finding. Although measuring differences in motivation or

hiring standards is typically a difficult task, we hope future research can use other sources of data to shed light on the reason behind the negative relationship between job tenure and effectiveness.

Our results also indicate a negative relationship between a physician's effectiveness and his/her average number of test order count. This implies that effective physicians are those who order less tests, or more accurately, order tests more intelligently. The fact that physicians with lower number of ordered tests have higher scores on the effectiveness metric supports a theory that not only there exist inherent differences among physicians with respect to effectiveness, but that effectiveness of providers might be improved via training programs that enable providers to decrease their use of unnecessary tests.

Our results regarding physician efficiency are displayed in Table 3. The results indicate a statistically significant positive relationship between a physician's efficiency score and his/her experience level. This is consistent with findings from other studies that indicate higher levels of experience can boost efficiency.[4] In addition, our results show a negative correlation between a physician's efficiency and his/her average number of test order count, implying that a physician's test ordering behavior is a contributing factor to his/her efficiency (similar to his/her effectiveness).

6.3. Patient Characteristics

To provide further answers to Research Question 2, we also examine the relationship between a physician's performance and his/her patient characteristics. Our results presented in Table 4 show no statistically significant relationship between a physician's effectiveness score and his/her average patient characteristics. With regards to physician efficiency scores, the results presented in Table 5 show that physicians' average efficiency scores increase when they encounter older patients, although the size of the coefficient is small (0.04). Overall, our results are consistent with the relevant literature that suggests patient characteristics should ideally have little or no effect on DEA scores (Chilingerian 1995).

6.4. Environment Characteristics

In addition to physician and patient characteristics we discussed in the previous sections, we study the impact of environment characteristics on physicians' effectiveness and efficiency. Specifically, given the large body of literature examining the effects of high workloads on physician performance (KC and Terwiesch 2009, Powell et al. 2012, Berry Jaeker and Tucker 2017, Batt and Terwiesch 2017), we study whether and how physician effectiveness and efficiency are affected by high ED volume. The results presented in Table 3 show that on average physician efficiency improves as ED volume increases. Furthermore, our results regarding physician effectiveness presented in Table 2

[4] For example, Venkataraman et al. (2018) show that more experienced surgeons are more efficient (evidenced by their patients' reduced LOS) in performing surgical procedures.

Table 3 Regression Results - Efficiency Model - Individual Physician

Dependent variable: DEA Efficiency Score	
Experience	0.0028^{*}
	(0.0012)
Avg Test Order Count	-0.0249^{***}
	(0.0059)
ED Volume	0.0254^{***}
	(0.0043)
ED Volume x Avg Test Order Count	-0.00012^{*}
	(0.0005)
Overcalling Rate	−0.1870
	(0.2479)
Undercalling Rate	−0.8911
	(0.5052)

Note: $N = 106$. Observations are at the physician-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

Table 4 Effectiveness Model - Patient Characteristics

Dependent variable: DEA Effectiveness Score	
Age	0.0011
	(0.0001)
ESI	-0.0514
	(0.0330)
Gender	-0.0555
	(0.0719)
Race	-0.0076
	(0.0633)

Note: $N = 106$. Observations are at the physician-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

Table 5 Efficiency Model - Patient Characteristics

Dependent variable: DEA Efficiency Score	
Age	0.0407***
	(0.0099)
ESI	-0.3212
	(0.3818)
Gender	-0.5533
	(0.3693)
Race	1.1730
	(1.0329)

Note: $N = 106$. Observations are at the physician-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

show that high workloads have a negative effect on physicians' average effectiveness scores. Consistent with the extant literature, our findings, thus, highlight the impact of high workloads on physician effectiveness and efficiency. For example, Saghafian et al. (2019) show that the magnitude of the effects of peers on physician speed and quality increases during high-volume shifts. The authors further demonstrate that test ordering and hospital admission are potential channels through which physicians influence each other's speed and quality, respectively. Our dataset, however, is insufficient for establishing statistically significant evidence for the mechanisms driving the effects of ED volume on physician effectiveness and efficiency DEA scores. Exploring such mechanisms provides a potential avenue for future research.

6.5. What Do Highly Effective and Efficient Physicians Do Differently?

Our results presented in the previous sections provide insights into our Research Questions 1 and 2. We now turn our attention into our Research Question 3, and generate insights into the characteristics of highly effective and efficient physicians, defined as those physicians with a higher-than-average effectiveness and efficiency DEA scores, respectively. To this end, we run model (2) on sub-samples of highly effective and efficient physicians. We present our results corresponding to highly effective and efficient physicians in Tables 6 and 7, respectively. As demonstrated in Table 6, we find a negative relationship (though weakly statistically significant) between highly effective physicians and their average test order count. This implies that highly effective physicians order less tests compared to their peers. This indicates that, compared to other physicians, highly effective physicians are able to order tests more intelligently and eliminate only unnecessary tests

Table 6 Highly Effective Physician Model

Dependent variable: DEA Effectiveness Score	
Avg LOS	-0.0003***
	(0.00007)
Avg Test Order Count	-0.0009
	(0.0006)
ED Volume	0.0007
	(0.0006)

Note: $N = 46$. Observations are at the physician-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

Table 7 Highly Efficient Physician Model

Dependent variable: DEA Efficiency Score	
Avg Contact-to-Disposition Time	-0.00514***
	(0.0006)
Avg Test Order Count	-0.0120**
	(0.0045)
ED Volume	0.0117**
	(0.0042)

Note: $N = 40$. Observations are at the physician-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

(i.e., cut tests that do not result in less effective treatments). While we establish a negative relationship between average physician effectiveness and ED volume in Section 6.4, we find no statistically significant evidence that high ED volume impacts the effectiveness of highly effective physicians. This finding, thus, suggests that highly effective physicians maintain their effectiveness under high workloads unlike other physicians.

The results presented in Table 7 provide statistically significant evidence for a positive association between physician efficiency and ED volume among highly efficient physicians, suggesting that a highly efficient physician's efficiency improves during high-volume shifts. In addition, our results show that highly efficient physicians order less tests on average compared to their peers.

6.6. Peer Influence

We now provide insights into our Research Question 4. Our results presented in Table 8 show a statistically significant negative relationship between a physician's effectiveness and the presence

Table 8 Regression Results - Effectiveness Model - Physician-Pair

Dependent variable: DEA Effectiveness Score	
More Efficient Peer	0.0007
	(0.0022)
More Effective Peer	-0.009^{***}
	(0.0021)
Different-Degree Peer	-0.004
	(0.003)
Opposite-Gender Peer	0.0016
	(0.003)

Note: N = 2,268. Observations are at the physician pair-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

of a more effective peer. This finding suggests that, all else equal, scheduling a physician with a less effective peer during the same shift results in an increase in the physician's effectiveness. Similarly, the regression results of physician-pair efficiency analysis displayed in Table 9 show that the presence of a more efficient peer is associated with a decrease in the focal physician's efficiency. These suggest that more effective and efficient providers can have a negative influence on their peers' effectiveness and efficiency, respectively. This is in line with the findings in (Saghafian et al. 2019) in which, using a different statistical methodology, the authors provide evidence for opposite-directional peer influence, and highlight the importance of incorporating peer influence in physician scheduling where a hospital administrator needs to decide on the set of physicians who should work during the same shifts. Finally, our results do not indicate any statistically significant evidence for peer influence with respect to physicians' relative gender and medical degree on a focal physician's average effectiveness and efficiency.

7. Robustness Checks

In this section, we provide various robustness checks. We do so by providing alternative models for measuring physicians' effectiveness and efficiency. We also conduct our ML analysis using an alternative approach to labeling the high-performing physicians in the training sets.

7.1. Alternative Effectiveness Model

In order to ensure that our results with respect to physician effectiveness are not sensitive to the choice of the output variable (72-hour rate of non-return), we repeat our analysis using alternative measures of effectiveness. Specifically, we use a physician's over- and under-calling rates as output variables in addition to the 72-hour rate of non-return patient visits. We define a physician's

Table 9 Regression Results - Efficiency Model - Physician-Pair

Dependent variable: DEA Efficiency Score	
More Efficient Peer	-0.006^{*}
	(0.0023)
More Effective Peer	0.004
	(0.0021)
Different-Degree Peer	0.0002
	(0.0031)
Opposite-Gender Peer	0.0024
	(0.0031)

Note: N = 2,268. Observations are at the physician pair-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

overcalling rate as the percentage of patients admitted by him/her from the ED to the hospital who were subsequently discharged from the hospital within 12 hours of admission. Similarly, we choose the percentage of patients admitted by a physician to the hospital (from the ED) who were upgraded from a floor bed to an intermediate care or ICU bed within 12 hours of admission as a proxy for how often the physician undercalls his/her patients' illness severity. Since the over- and under-calling rates would be considered undesirable outputs, we use the 12-hour non-discharge and 12-hour non-upgrade patient admission rate as output variables. We choose the physician's average number of test order counts as the model's input variables. This effectiveness model's variables, thus, include:

Outputs:

- *Rate of discharged patients who do not return within 72 hours;*
- *Rate of admitted patients who are not discharged within 12 hours;*
- *Rate of patients admitted to a floor/ward bed who are not upgraded within 12 hours.*

It should be noted that the first output variable above is suitable for considering performance among discharged patients, while the two other output variables consider performance among admitted patients. The choice of threshold numbers (72 and 12) is made based on observations made in the literature (see, e.g., Keith et al. 1989, Gordon et al. 1998, and the references therein) as well as conversations with ED physicians. In addition, we perform sensitivity analyses on these thresholds by changing each of them within a range, and observe that our main results still hold.

Inputs:

- *Radiology Order Count*: Average number of the physician's radiology orders per patient visit;
- *Ultrasound Order Count*: Average number of the physician's ultrasound orders per patient visit;

- *MRI Order Count*: Average number of the physician's MRI orders per patient visit.

A positive correlation between the input and the output variables confirms our choice of the model's input variables (P = 0.001). We re-run our second-stage Tobit regression analysis using the scores derived from this model and observe that our main results hold (see Online Appendix C).

7.2. Alternative Efficiency Model

Similar to our robustness test for the effectiveness model, we re-run our analysis using an alternative efficiency model defined as follows:

Output:

- *Throughput:* Average number of patients seen by the physician per shift.

Inputs:

- *High ESI*: Percentage of patients served by the physician who have ESI levels 4 and 5 (i.e., low-acuity patients);
- *Younger than 65*: Percentage of patients served by the physician who are younger than 65.

For our alternative efficiency model, we choose an output-oriented DEA model based on which efficient physicians are identified as those who have a higher throughput rate for a given mix of patients. Based on our discussions with ED physicians, throughput — the average number of patients served by a provider per unit of time — possesses significant face validity for exploratory analysis. Our input variables in this model comprise a low-acuity and younger patient mix which on average requires less time to treat. As such, efficient physicians are identified as those who have a higher throughput rate for a given mix of patients. We re-run our statistical analysis using this alternative efficiency model and observe that our findings are consistent with our main results discussed in Section 6 (see Online Appendix C).

7.3. Alternative ML Algorithm Training Sets

In Section 4, we used the average 72-hour rate of return and contact-to-disposition time to label the highly effective and efficient physicians in our training sets, respectively. In order to examine the robustness of our results to this choice, we repeat our analysis using the median 72-hour rate of return and contact-to-disposition time as an alternative way to identify the high-performing physicians in the training sets. We present the classifications of physicians using this alternative approach and comparison with the DEA classifications in Online Appendix D. Our inferences remain unchanged.

8. Conclusions

Using evidence from emergency medicine, we develop and analyze metrics for physicians' effectiveness and efficiency. We then use our metrics to generate insights into the relationship between physician performance and factors related to patient, physician, environment, and physician peers. Unlike what the conventional wisdom suggests, our findings show that a physician's effectiveness and his/her efficiency are positively associated. In addition, we find that more effective physicians have lower than average test order count and job tenure. We also find that efficient physicians have on average lower test orders per patient visit and more years of experience compared to their peers. In addition, we find that during high-volume shifts, a physician's efficiency improves while his/her effectiveness declines. We also identify some of the characteristics of highly effective and efficient physicians. Our findings indicate that highly effective physicians order less tests compared to their peers. Faced with high workloads, we show that highly effective physicians are able to maintain their effectiveness more so than their peers. In addition, we find that highly efficient physicians utilize less test orders per patient visit and are more efficient during high-volume shifts compared to their peers. Furthermore, our results provide evidence for the existence of peer influence, and suggest that the presence of more effective and efficient peers has negative effects on a focal physician's effectiveness and efficiency, respectively.

We believe that our analysis serves as an early step to explore issues of physician effectiveness and efficiency. Importantly, we do not believe that the scores we develop are the only ways to measure effectiveness or efficiency.[5] That is, our work does not provide a definitive calculus for determining who is (or is not) an effective or efficient physician, but rather uses analytical techniques to explore these issues in an early attempt to better understand them. Nevertheless, our findings shed light on important potential new ways to improve the efficiency and effectiveness of healthcare delivery. For example, they can help individual physicians observe their weaknesses and realize the advantages of following what the highly efficient and effective physicians do differently. Similarly, well-designed training programs can use our findings to facilitate this learning process. Furthermore, our findings can prove useful in the area of physician scheduling as they demonstrate how peer influence can play an important role in effective and efficient care delivery. Thus, our insights on peer influence can be used to understand which physicians should be scheduled during the same shift so as to boost performance without increasing resources.

Finally, we note that our analyses in this paper are purely based on quantitative data. The inclusion of qualitative factors in future studies may improve the strength and applicability of our effectiveness and efficiency models. Future work can also provide a more complete picture of

[5] For example, one may improve our scores by also including aspects of patient satisfaction that correlate with higher provider performance levels.

the channels through which a physician's effectiveness and efficiency can be improved. Given the importance of understanding factors that can improve the efficiency and effectiveness of physicians, we hope to see more future studies in these veins.

References

Abualenain J, Frohna WJ, Smith M, Pipkin M, Webb C, Milzman D, Pines JM (2013) The prevalence of quality issues and adverse outcomes among 72-hour return admissions in the emergency department. *The Journal of Emergency Medicine* 45(2):281-8.

Anand K, Paç MF, Veeraraghavan S (2011) Quality-speed conundrum: Trade-offs in customer-intensive services. *Management Science* 57(1):40-56.

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* 30(9):1078-1092.

Batt R, Terwiesch C (2017) Early Task Initiation and Other Load-adaptive Mechanisms in the Emergency Department. *Management Science* 63(11):3531-3551.

Berry Jaeker J, Tucker A (2017) Past the Point of Speeding Up: The Negative Effects of Workload Saturation on Efficiency and Patient Severity. *Management Science* 63(4):1042-1062.

Bruursema K, Kessler SR, Spector PE (2011) Bored employees' misbehaviour: The relationship between boredom and counterproductive work behavior. *Work and Stress* 25:93-107.

Castelli A, Street A, Verzulli R, Ward P (2015) Examining variations in hospital productivity in the English NHS. *European Journal of Health Economics* 16(3):243-254.

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. *European Journal of Operational Research* 2(6):429-444.

Chilingerian JA (1995) Evaluating physician efficiency in hospitals: A multivariate analysis of best practices. *European Journal of Operational Research* 80(3):548-574.

Collier DA, Collier CE, Kelly TM (2006) Benchmarking physician performance, part 1. *Journal of Medical Practice Management* 21(4):185-189.

Cooper WW, Seiford LM, Tone K (2007) *Data Envelopment Analysis A Comprehensive Text with Models, Applications, References and DEA-Solver Software.*, 2nd ed. (Springer, New York).

Fernandes CMB, Price A, Christenson JM (1997) Does reduced length of stay decrease the number of emergency depart-ment patients who leave without seeing a physician? *J Emerg Med* 15(3):397-9.

Fiallos J, Patrick J, Michalowski W, Farion K (2017) Using data envelopment analysis for assessing the performance of pediatric emergency department physicians. *Health Care Management Science* 20(1):129-140.

Glickman SW, Schulman KA, Peterson ED, Hocker M B, Cairns CB (2008) Evidence-based perspectives on pay for performance and quality of patient care and outcomes in emergency medicine. *Annals of Emergency Medicine* 51(5):622-631.

Gordon JA, An LC, Hayward RA, Williams BC (1998) Initial emergency department diagnosis and return visits: risk versus perception. *Annals of Emergency Medicine* 32(5):569-573.

Goulet F, Jacques A, Gagnon R, Bourbeau D, Laberge D, Melanson J, Mnard C, Racette P, Rivest R (2002) Performance assessment. Family physicians in Montreal meet the mark! *Canadian family physician* 48(8):1337-1344.

Greene WH (1993) *Econometric Analysis* 2nd ed. (Macmillan, New York).

Grosskopf S, Valdmanis V (1987) Measuring Hospital Performance. A Non-Parametric Approach. *Journal of Health Economics* 6(2):89-107.

Hackman JR, Oldham GR (1980) *Work redesign.* Reading, MA: Addison-Wesley.

Harrison JP, Ogniewski RJ (2005) An Efficiency Analysis of Veterans Health Administration Hospitals. *Military Medicine* 170(7):607-11.

Hasija S, Shumsky RA, Pinker E (2008) Call Center Outsourcing Contracts Under Information Asymmetry. *Management Science* 54(4):793-807.

Hess BJ, Weng W, Lynn LA, Holmboe ES, Lipner RS (2011) Setting a fair performance standard for physicians' quality of patient care. *Journal of General Internal Medicine* 6(5):467-473.

Hollingsworth B (2008) The measurement of efficiency and productivity of health care delivery. *Health Economics* 17(10):1107-1128.

Johannessen KA, Kittelsen SAC, Hagen TP (2017) Assessing physician productivity following Norwegian hospital reform: A panel and data envelopment analysis. *Social Science & Medicine* 175:117-126.

Kass SJ, Vodanovich SJ, Callender A (2001) State-trait boredom: Relationship to absenteeism, tenure, and job satisfaction. *Journal of Business and Psychology* 16(2):317-327.

KC DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486-1498.

Keehan SP, Stone DA, Poisal JA, Cuckler GA, Sisko AM, Smith SD, Madison AJ, Wolfe CJ, Lizonitz JM (2017) National health expenditure projections, 2016-25: price increases, aging push sector to 20 percent of economy. *Health Affairs (Millwood)* 36(3):553-563.

Keith KD, Bocka JJ, Kobernick MS, Krome RL, Ross MA (1989) Emergency department revisits. *Annals of Emergency Medicine* 18(9):964-8.

Klasco RS, Wolfe RE, Wong M (2015) Assessing the rates of error and adverse events in the ED. *The American Journal of Emergency Medicine* 33:1786-9.

Latham LP, Ackroyd-Stolarz, S (2014) Emergency department utilization by older adults: a descriptive study. *Canadian Geriatrics Journal* 17(4):118-125.

Ng TWH, Feldman DC (2013) Does longer job tenure help or hinder job performance? *Journal of Vocational Behavior* 83:305-14.

Ozcan YA (1998) Physician benchmarking: measuring variation in practice behavior in treatment of otitis media. *Health Care Management Science* 1(1):5-17.

Ozcan YA, Begun JW, McKinney MM (1999) Benchmarking Organ Procurement Organizations: A National Study. *Health Services Research* 34(4):855-74.

Ozcan YA, CW Pai, HJ Jiang (2000) Regional variation in physician practice pattern: an examination of technical and cost efficiency for treating sinusitis. *Journal of Medical Systems* 24(2):103-117.

Pham JC, Kirsch TD, Hill PM, DeRuggerio K, Hoffmann B (2011) Seventy two-hour returns may not be a good indicator of safety in the emergency department: a national study. *Academic Emergency Medicine* 18:390-7.

Powell A, Savin S, Savva N (2012) Physician Workload and Hospital Reimbursement. *Manufacturing and Service Operations Management* 14(4):512-528.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient Streaming as a Mechanism for Improving Responsiveness in Emergency Departments. *Operations Research* 60(5):1080-1097.

Saghafian S, Hopp WJ, Iravani SMR, Cheng Y, Diermeier D (2018) Workload Management in Telemedical Physician Triage and Other Knowledge-Based Service Systems. *Management Science* 64(11):5180-5.

Saghafian S, Imanirad R, Traub SJ (2019) Do Physicians Influence Each Other's Performance? Evidence from the Emergency Department. *Working Paper,* Harvard University.

Salway RJ, Valenzuela R, Shoenberger JM, Mallon WK, Viccellio A (2017) Emergency Department (ED) Overcrowding: Evidence-Based Answers To Frequently Asked Questions. *Revista Medica Clinica Las Condes* 28(2):213-219.

Sherman HD (1984) Hospital Efficiency Measurement and Evaluation. Empirical Test of a New Technique. *Medical Care* 22(10):922-938.

Simar L, Wilson P (1998) Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science* 44(1):49-61.

Simar L, Wilson P (2002) Non-parametric tests of returns to scale. *European Journal of Operational Research* 139(1):115-132.

Simar L, Wilson P (2011) Inference by the m out of n bootstrap in nonparametric frontier models. *Journal of Productivity Analysis* 36:33-53.

Smith CA, Varkey AB, Evans AT, Reilly BM (2004) Evaluating the performance of inpatient attending physicians: A new instrument for today's teaching hospitals. *Journal of General Internal Medicine* 19(7):766-771.

Spaite D, Bartholomeaux F, Guisto J, Lindberg E, Hull B, Eyherabide A, Lanyon S, Criss EA, Valenzuela TD, Conroy C (2002) Rapid process redesign in a university-based emergency department:decreased waiting time intervals and improving patient satisfaction. *Ann Emerg Med* 39(2):168-77.

Storfa AH, Wilson, ML (2015) Physician productivity: issues and controversies. *American Journal of Clinical Pathology* 143(1):6-9.

Traub SJ, Stewart CF, Didehban R, Bartley AC, Saghafian S, Smith VD, Silvers SM, LeCheminant R, Lipinski CA (2016) Emergency department rotational patient assignment. *Annals of Emergency Medicine* 67(2):206-15.

Tsugawa Y, Ashish KJ, Newhouse J (2017) Variation in physician spending and association with patient outcomes. *JAMA Internal Medicine* 177(5):675-682.

Varabyova Y, Schreyogg J (2013) International comparisons of the technical efficiency of the hospital sector: panel data analysis of OECD countries using parametric and non-parametric approaches. *Health Policy* 112(1):70-79.

Venkataraman S, Fredendall LD, Taaffe KM, Huynh N, Ritchie G (2018) An Empirical Examination of Surgeon Experience, Surgeon Rating, and Costs in Perioperative Services. *Journal of Operations Management* 61(1):68-81.

Wagner JM, Shimshak DG, Novak MA (2003) Advances in physician profiling: the use of DEA. *Socio-Economic Planning Sciences* 37(2):141-163.

Wang X, Debo LG, Scheller-Wolf A, Smith SF (2010) Design and Analysis of Diagnostic Service Centers. *Management Science* 56(11):1873-1890.

Zheng W, Sun H, Zhang P, Zhou G, Jin Q, Lu X (2018) A four-stage DEA-based efficiency evaluation of public hospitals in China after the implementation of new medical reforms. *PLoS ONE* 13(10):e0203780.

Online Appendix A - Regression Results - Physician-Quarter Level Analysis

Table 1 Regression Results - Effectiveness Model - Individual Physician

Dependent variable: DEA Effectiveness Score	
DEA Efficiency Score	0.2142^{***}
	(0.1014)
Patient Characteristics	
ESI Level 1 & 2 Patients	1.3746
	(2.0121)
Patients Over 65 Years of Age (%)	1.3216
	(1.8152)
Female Patients (%)	−0.2214
	(1.8214)
White Patients (%)	−1.4982
	(2.9821)
Physician Characteristics	
Physician Tenure	-0.0048^{*}
	(0.0021)
Physician Contact-to-Disposition	-0.0011^{*}
	(0.0005)
Avg MRI Count per Patient Visit	3.6921
	(2.6821)
Avg IV Count per Patient Visit	−0.0019
	(0.1031)
Avg CT Scan Count per Patient Visit	-0.1982^{*}
	(0.2991)

Note: $N = 415$. Observations are at the physician-quarter level.

Note: *p<0.05; **p<0.01; ***p<0.001

Table 2 Regression Results - Efficiency Model - Individual Physician

Dependent variable: DEA Efficiency Score	
Patient Characteristics	
ESI Level 1 & 2 Patients	−2.1022
	(1.9281)
Patients Over 65 Years of Age (%)	1.5021
	(1.8921)
Female Patients (%)	−1.4162
	(1.5102)
White Patients (%)	2.1982
	(3.4921)
Physician Characteristics	
Experience	0.0022*
	(0.0019)
Avg IV order Count	−0.0293
	(0.1023)
Avg Radiology order Count	0.6102
	(0.3321)
Avg MRI Order Count	−5.5930*
	(2.4201)

Note: $N = 415$. Observations are at the physician-quarter level.
Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table 3 Regression Results - Effectiveness Model - Peer Physician

Dependent variable: DEA Effectiveness Score	
Peer Characteristics	
More Efficient Peer	0.0049
	(0.1982)
More Effective Peer	-0.0119^{*}
	(0.0318)
Different-Degree Peer	0.0022
	(0.3392)
Opposite-Gender Peer	-0.0075
	(0.1762)

Note: N = 8,915. Observations are at the physician-quarter level.
Note: *p<0.05; **p<0.01; ***p<0.001

Table 4 Regression Results - Efficiency Model - Peer Physician

Dependent variable: DEA Efficiency Score	
Peer Characteristics	
More Efficient Peer	-0.0048^{*}
	(0.0068)
More Effective Peer	-0.0281
	(0.0075)
Different-Degree Peer	0.01922
	(0.0079)
Opposite-Gender Peer	-0.0102
	(0.0110)

Note: N = 8,915. Observations are at the physician-quarter level.
Note: *p<0.05; **p<0.01; ***p<0.001

Online Appendix B - Variable Importance Graphs

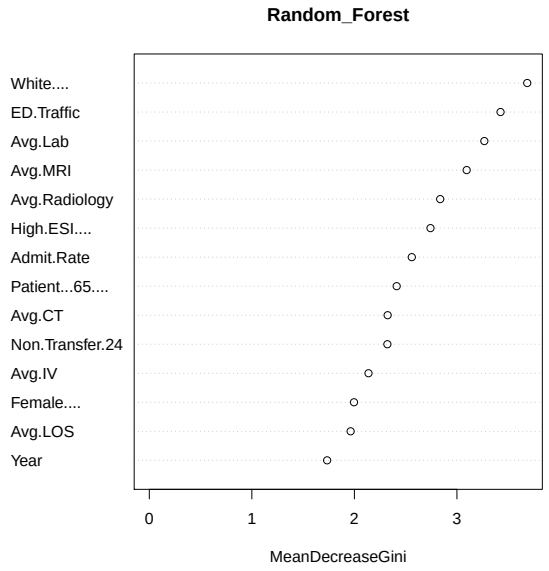


Figure 1 Variable Importance Graph - RF Effectiveness Model

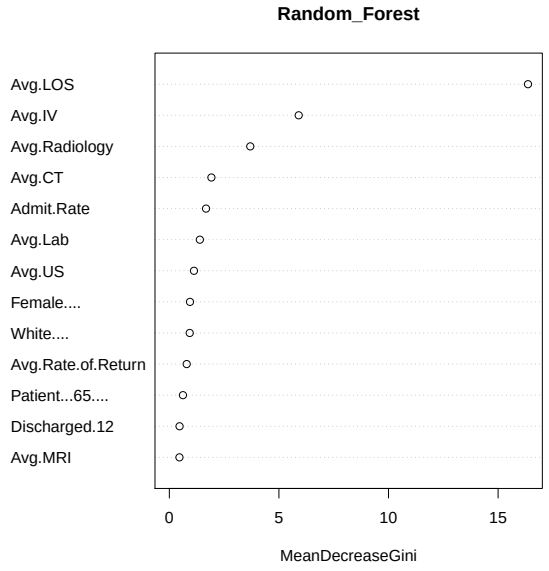


Figure 2 Variable Importance Graph - RF Efficiency Model

Online Appendix C - Alternative DEA Models - Regression Results

Table 1 Regression Results - Effectiveness Model - Individual Physician

Dependent variable: DEA Effectiveness Score	
DEA Efficiency Score	0.3618^{***}
	(0.1567)
Patient Characteristics	
ESI Level 1 & 2 Patients	2.3798
	(2.1063)
Patients Over 65 Years of Age (%)	1.6429
	(1.9013)
Female Patients (%)	−0.4475
	(2.3294)
White Patients (%)	−1.5158
	(3.5544)
Physician Characteristics	
Physician Tenure	-0.0057^{*}
	(0.0025)
Physician Contact-to-Disposition	-0.0016^{*}
	(0.0007)
Avg MRI Count per Patient Visit	4.0121
	(3.1387)
Avg IV Count per Patient Visit	−0.0026
	(0.1163)
Avg CT Scan Count per Patient Visit	-0.2445^{*}
	(0.3874)
Avg Total ED Patients per Shift	−0.0024
	(0.0018)

Note: $N = 106$. Observations are at the physician-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

Table 2 Regression Results - Efficiency Model - Individual Physician

Dependent variable: DEA Efficiency Score	
Patient Characteristics	
ESI Level 1 & 2 Patients	−2.5637
	(2.3786)
Patients Over 65 Years of Age (%)	1.8061
	(2.1200)
Female Patients (%)	−1.7110
	(2.1270)
White Patients (%)	3.2178
	(4.6979)
Physician Characteristics	
Experience	0.0035*
	(0.0033)
Avg IV order Count	−0.0343
	(0.1185)
Avg Radiology order Count	0.7178
	(0.4133)
Avg MRI Order Count	−6.9205*
	(3.5250)
Avg Total ED Patients per Shift	−0.0020
	(0.0032)

Note: $N = 106$. Observations are at the physician-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

Table 3 Regression Results - Effectiveness Model - Peer Physician

Dependent variable: DEA Effectiveness Score	
Peer Characteristics	
More Efficient Peer	0.0069
	(0.2535)
More Effective Peer	-0.0126^{*}
	(0.0411)
Different-Degree Peer	0.0041
	(0.5593)
Opposite-Gender Peer	-0.0098
	(0.2118)

Note: N = 2,268. Observations are at the physician-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

Table 4 Regression Results - Efficiency Model - Peer Physician

Dependent variable: DEA Efficiency Score	
Peer Characteristics	
More Efficient Peer	-0.0059^{*}
	(0.0081)
More Effective Peer	-0.0234
	(0.0082)
Different-Degree Peer	0.0232
	(0.0093)
Opposite-Gender Peer	-0.0105
	(0.0105)

Note: N = 2,268. Observations are at the physician-year level.
Note: *p<0.05; **p<0.01; ***p<0.001

Online Appendix D - Alternative ML Models

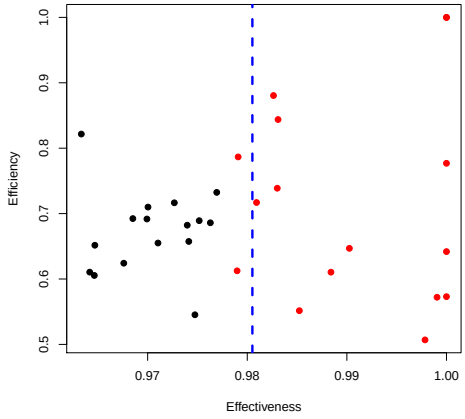


Figure 1 DEA Model - Effectiveness Classification

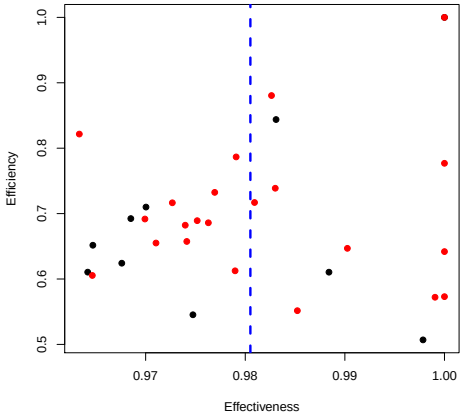


Figure 2 ML Model (RF) - Effectiveness Classification

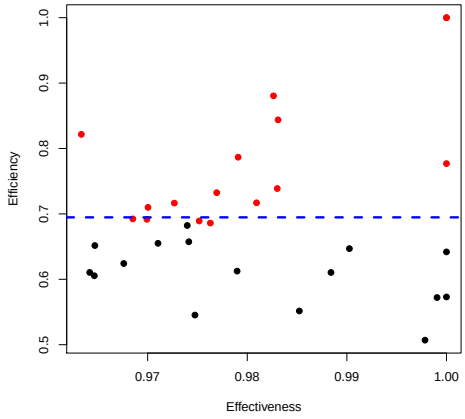


Figure 3 DEA Model - Efficiency Classification

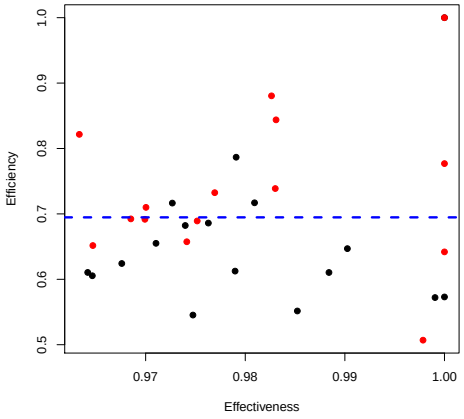


Figure 4 ML Model (RF) - Efficiency Classification