What Drives Physician Efficiency? Evidence from Emergency Department Operations

Raha Imanirad
Schulich School of Business, York University, Toronto, Ontario

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

Stephen J. Traub
Optum Arizona and New Mexico, Phoenix, AZ

Nicole R. Hodgson
Department of Emergency Medicine, Mayo Clinic Arizona, Phoenix, AZ

Abstract: Emergency Departments (EDs) play a critical role in healthcare, serving as the frontline of acute care delivery. However, increasing demand, constrained resources, and the complexity of patient cases pose significant challenges to maintaining operational efficiency. This study investigates the determinants of ED physician efficiency using a robust multi-stage analytical framework. First, we apply Data Envelopment Analysis (DEA) to quantify individual physician efficiency using data from 110,325 patient visits at a leading U.S. hospital. DEA provides a transparent input-output evaluation method, yielding data-driven efficiency scores that reflect physician practice patterns and offer actionable insights for hospital administrators. Second, we apply a Tobit regression model to analyze how various factors—including physician experience, shift familiarity, and patient severity—influence physician efficiency. Third, we conduct comprehensive robustness checks, including omitted variable bias sensitivity analyses, permutation-based placebo tests, and machine learning-based classification tests, to mitigate concerns regarding endogeneity and potential model misspecification.

Our results demonstrate that physician experience positively correlates with efficiency. Shift familiarity enhances efficiency, especially for early-career physicians, although its marginal benefit diminishes with greater experience. While higher patient severity generally decreases efficiency—underscoring the greater resource intensity of complex patient care—this negative effect is attenuated for more experienced physicians, highlighting their adaptability under challenging clinical conditions. These insights suggest that strategically assigning experienced physicians to higher-severity or less familiar shifts, where their adaptability offsets environmental challenges, can enhance overall ED staffing effectiveness. These findings offer actionable strategies for improving ED operations. For example, targeted scheduling policies could enhance performance by increasing shift consistency for junior physicians while flexibly assigning senior physicians to high-severity or variable shifts. Taken together, our results indicate that integrating severity-responsive resource planning with evidence-based scheduling can improve throughput, reduce resource utilization, and support better patient outcomes.

Key words: Data Envelopment Analysis, Emergency Department Operations, Physician Efficiency, Tobit Regression

1 Introduction

Motivation. Emergency Department (ED) utilization is a critical focus of healthcare research and policy given its implications for patient outcomes, healthcare costs, and overall system efficiency. As frontline providers of acute care, ED physicians frequently operate under high pressure and with limited resources, leading to challenges in managing patient flow and ensuring timely treatment. The increasing demand for emergency services, driven by factors such as population growth, aging demographics, and the rising prevalence of chronic diseases, exacerbates these challenges. Consequently, evaluating and understanding the determinants of ED physician efficiency is crucial for developing strategies to optimize performance, enhance patient care, and reduce waste. In this paper, we examine the multifaceted drivers of ED physician efficiency, offering insights into how EDs can more effectively allocate resources and improve the delivery of emergency care. To this end, we utilize a large dataset capturing the care delivery patterns of ED physicians and apply Data Envelopment Analysis (DEA)—a linear programming optimization technique that provides a multi-dimensional evaluation tool—in the first stage to evaluate their relative performance. DEA offers a key advantage in *interpretability*, yielding efficiency scores from a transparent input–output view of physician performance and avoiding "black-box" operations. This interpretability allows DEA scores to be readily communicated to (a) hospital administrators aiming to improve care delivery efficiency, and (b) physicians seeking to enhance their individual performance. We then use these efficiency scores in a second-stage Tobit regression to examine the impact of key factors—physician experience, shift familiarity, and patient severity—on ED physician efficiency.

Data and Setting. Our data consist of detailed care delivery information associated with 110,325 patient visits at the ED of a leading U.S. hospital. The ED is staffed by 32 board-certified emergency medicine physicians and over 70 registered nurses. During the study period, all care was delivered directly by attending physicians, with no advanced practice providers and fewer than 5% of cases involving resident physicians. All patients in our partner ED are algorithmically assigned to physicians upon arrival through an automated rotational patient assignment process (Traub et al. 2016b,a). This workflow mimics randomized patient assignment, eliminating physician discretion in patient selection and preventing any "cherry-picking" of patients based on preference or case complexity. All patient visits identified in the electronic health record system as having been seen by an ED physician from July 12, 2012, to July 31, 2016, were included in our dataset. Patient-specific data encompass demographic (e.g., age, gender, race) and insurance information. Encounter-level data include laboratory tests, chief complaint, Emergency Severity Index (ESI) level (a five-level triage scale where 1 indicates the most urgent and 5 the least urgent), day and time of the ED visit, among others.

Research Question. This paper investigates a critical question for healthcare operations: How do individual physician attributes, patient characteristics, and environmental factors interact to shape physician efficiency in EDs? To address this, we first apply a DEA model to evaluate ED physician efficiency. We then examine how variation in these efficiency scores relates to three core determinants: physician experience, patient severity (measured by ESI level), and operational conditions such as shift familiarity. These factors were selected as they represent critical elements that influence clinical decision-making, the quality of care, and overall patient outcomes in emergency medicine settings. For instance, physician experience is empirically linked to diagnostic accuracy and treatment efficacy (Leigh and Kravitz 2001), offering direct avenues to inform staffing and resource allocation. Patient severity, by demanding variable time and resources, is crucial for a nuanced understanding of efficiency. Furthermore, system-level factors like shift familiarity are pivotal, impacting physician fatigue, stress, and cognitive function, which are directly tied to performance and patient safety (Linzer et al. 2002, Mason et al. 2002). By adopting an efficiency-based framework, our study evaluates how physician experience, shift familiarity, and patient severity jointly shape the transformation of diagnostic and treatment resources into patient outcomes in the ED, generating actionable insights for improving emergency care delivery and resource allocation.

Main Findings. Our analysis reveals several key insights into the determinants of ED physician efficiency. As expected, more experienced physicians (measured by years in clinical practice) demonstrate higher efficiency. Efficiency also varies with patient severity: it is higher when physicians treat lower-severity patients (i.e., higher proportions of ESI-4) and lower when treating higher-severity patients (i.e., higher proportions of ESI-2), reflecting the greater cognitive and time demands associated with complex cases. More importantly, our findings reveal that physician experience moderates the effects of both patient severity and shift familiarity on efficiency. Specifically, as experience increases, the efficiency penalty associated with high-severity patients diminishes, suggesting that more experienced physicians are better able to manage complex cases without compromising efficiency. In addition, we find that experience and shift familiarity act as substitutes: as physicians accumulate clinical experience, they become less reliant on familiar shift conditions to maintain high performance. Together, these findings challenge assumptions of linear human capital returns and suggest that experience builds adaptability, enabling physicians to maintain consistent performance across varying patient severity levels and shift conditions. This insight provides actionable guidance for designing shift assignments and team structures tailored to physician experience levels.

Managerial Implications. Our study offers insights for improving ED efficiency through targeted scheduling, patient assignment, and resource planning. From a managerial perspective, these findings support consistent scheduling for early-career physicians and more flexible assignment of senior physicians in variable clinical contexts. Complementary strategies, such as mentorship and peer pairing on familiar shifts, may further reinforce learning and efficiency gains among less experienced physicians without requiring additional staffing. Simulations based on our empirical model show that increasing shift familiarity from the 25th to the 75th percentile is associated with a 2.68% improvement in physician efficiency for a physician with 13 years of experience and a 1.24% reduction for a physician with 32 years of experience.

We also find that patient severity is a significant predictor of efficiency. While patient severity cannot be controlled directly, it often follows predictable temporal patterns. EDs can use this information to implement severity-aware scheduling and proactive resourcing strategies. For example, simulating a scenario in which 50% of the efficiency loss associated with a shift toward higher-severity patients—captured by an increase in the proportion of ESI level 2 cases from the 25th to the 75th percentile—is recovered through resourcing interventions yields a 1.31% improvement in efficiency. Our simulations also confirm a strong moderating effect of physician experience. For a less experienced physician (13 years of experience), increasing the proportion of high-severity patients (ESI 2) from the 25th to the 75th percentile is associated with a 4.86% decline in efficiency. In contrast, for a highly experienced physician (32 years of experience), the same increase in the proportion of high-severity patients yields a negligible 0.33% change in efficiency, confirming that experienced physicians absorb severity increases without meaningful efficiency losses. These results suggest that strategically assigning lower-severity patients to early-career physicians can produce greater marginal efficiency gains, while reserving high-severity cases for more experienced physicians helps minimize performance loss in complex conditions. Together, these results support a portfolio of low-cost, high-leverage operational interventions: schedule stability for early-career physicians, flexible staffing for senior physicians, severity-aware shift planning, and anticipatory resourcing during high-severity periods. More broadly, our approach demonstrates how simulation-informed policies grounded in empirical estimates can help EDs optimize performance under real-world constraints.

Robustness Checks. To validate our findings and assess their sensitivity to model specification and potential bias, we conduct a comprehensive series of robustness checks. We examine robustness to alternative DEA specifications and re-estimate the second stage using OLS and Simar and Wilson (2007) double bootstrap truncated regression. To assess omitted variable bias, we implement two complementary approaches: the Altonji et al. (2005) coefficient stability analysis across four nested specifications, and a permutation-based placebo test with 1,000 iterations. In addition, we test for potential model misspecification by adding a nonlinear term for ED volume. Finally, we implement a machine learning (ML)-based validation approach using supervised classifiers to assess whether DEA-identified efficient physicians can be distinguished from others based on observable features. High AUC scores indicate that DEA classifications align with patterns identified using alternative, data-driven methods. Across all checks, our results remain consistent with the main analysis.

2 Related Studies and Hypotheses Development

2.1 Physician Performance Evaluation

Research on physician performance has traditionally focused on condition-specific clinical metrics. For example, Glickman et al. (2008) use indicators such as performing a diagnostic electrocardiogram (ECG) for syncope in patients over 60, while Hess et al. (2011) assess care quality for diabetic patients using measures such as retinal and foot exams and blood pressure tests. To broaden the evaluation of performance, some studies incorporate behavioral dimensions using tools such as physician questionnaires (Smith et al. 2004) and patient chart audits (Goulet et al. 2002), though these methods face challenges related to reliability, scalability, and measurement bias. More broadly, findings from this literature may not generalize to settings like EDs, where physicians treat a diverse and unpredictable mix of patient conditions. In parallel, policymakers have increased their efforts to improve transparency in healthcare outcomes through initiatives such as the public reporting of quality measures (Saghafian and Hopp 2019, 2020), underscoring the growing need for robust, meaningful performance metrics—particularly in high-variability environments such as EDs. Much of the existing literature, however, evaluates physician performance using isolated outcome metrics rather than assessing how physicians transform clinical resources into patient outcomes, providing limited insight into the efficiency of clinical decision-making in complex environments such as EDs.

To assess ED physician performance, prior research has relied on operational metrics such as time from arrival to clinical assessment, LOS, the percentage of patients who left without being seen, 72-hour readmission rates, and mortality or morbidity outcomes (e.g., Fernandes et al. 1997, Spaite et al. 2002). However, while these metrics capture outcomes, they do not account for the resources used in delivering care, making it difficult to distinguish physicians who achieve strong outcomes through efficient resource use from those who rely on more intensive resource utilization. In a resource-constrained setting like an ED—where physicians rely on shared inputs—evaluating performance without considering resource utilization offers an incomplete picture of performance. Hence, a methodology such as DEA, which explicitly evaluates how inputs are transformed into outputs, provides a natural framework for assessing physician efficiency in ED settings.

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 hospital performance has been widely explored using DEA (Zheng et al. 2018, Castelli et al. 2015, Varabyova and Schreyogg 2013, Hollingsworth 2008), applications at the individual physician level remain relatively limited. This challenge arises from factors such as heterogeneity in patient mix and treatments, as well as fundamental differences across medical specialties (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 1998) in their suggested DEA models. Such prior work, by relying on condition-specific or proxy variables often limits the generalizability of the derived efficiency measures, resulting in an incomplete picture of individual physician efficiency across the diverse clinical decisions encountered in ED operations.

To address these limitations, we construct a DEA-based efficiency measure that jointly evaluates multiple clinical outcomes relative to physician resource utilization, enabling a direct assessment of how efficiently physicians transform diagnostic and treatment resources into patient outcomes. Our selection of input and output variables excludes those tied to specific patient health conditions or physician practice styles—reducing the risk of overfitting to our study setting and enhancing the model's transferability across diverse ED environments. This approach evaluates physician efficiency across both discharged and admitted patients while explicitly incorporating physician resource utilization, allowing us to capture clinical decision-making across the full spectrum of ED care. The resulting physician-level efficiency scores provide the foundation for analyzing how physician characteristics, patient severity, and operational conditions shape efficiency. From a practical perspective, our DEA methodology provides hospital administrators with a transparent, easy-to-understand scoring system for assessing the efficiency of care delivery in their hospitals. For individual physicians, the model identifies areas of inefficiency and enables learning from the practices of highly efficient peers. Well-designed training programs tailored to these insights can further support this learning process and promote continuous improvement in care delivery.

2.2 Factors Contributing to Physician Efficiency

2.2.1 Patient Severity

ED physicians routinely manage patients with varying levels of severity, from minor injuries to life-threatening conditions. Understanding how patient severity impacts physician performance is essential for improving clinical outcomes and operational efficiency. Prior research has highlighted the multifaceted influence of patient severity on healthcare delivery. For instance, Kawamoto et al. (2020) find that patient severity in intensive care units (ICUs), measured by the Sequential Organ Failure Assessment (SOFA) score, significantly affects the length of face-to-face interactions among healthcare professionals, particularly during critical periods of treatment.

In ED settings, patient severity has been linked to variations in physician performance. Chilingerian (1995) notes that higher patient severity can reduce physicians' attention spans, negatively affecting their efficiency. Soltani et al. (2022) further demonstrate that ED physicians adapt their test-ordering behaviors based on patient severity, with notable differences in resource utilization between discharged and admitted patients. Other studies such as Jameson et al. (2025a,b) examine ED physicians' decisions to batch order tests and identify associations between batching tendencies, patient severity, and presenting complaints. These findings suggest that as patient severity increases, healthcare professionals adjust their decision-making and resource allocation strategies, which can impact overall efficiency in EDs and other high-pressure environments. Moreover, the ongoing management of high-severity cases poses significant challenges for ED physicians. Panagioti et al. (2018) identify patient severity as a major contributor to physician burnout, which in turn diminishes performance, professionalism, and the quality of care delivered. Together, these studies underscore the complex interplay between patient severity, decision-making, and physician efficiency.

The converse relationship also holds. When physicians treat a higher proportion of low-severity patients, clinical demands are more predictable, resource utilization is more routine, and cognitive load is reduced. Under these conditions, physicians can apply standardized protocols more efficiently, reducing variability in decision-making and resource consumption. This suggests that a higher proportion of low-severity patients should be associated with improved efficiency, as the reduced clinical complexity allows for more consistent and streamlined care delivery.

While prior research has explored the relationship between patient severity and physician performance, it often focuses on isolated metrics, such as test-ordering behaviors or burnout. These studies tend to overlook the disproportionate time, attention, and resources required by high-severity patients compared to low-severity ones. Our study addresses this gap by offering a holistic perspective on physician efficiency that integrates multiple health outcomes across a broad patient population, providing a more complete understanding of how severity influences efficiency and resource utilization in the ED setting. In our study, we operationalize patient severity using ESI, a widely used triage tool in EDs that captures both clinical urgency and anticipated resource needs. Leveraging variation in ESI levels across patients treated in our partner ED, and drawing on DEA-based efficiency scores, we make the following hypotheses:

Hypothesis 1a. A higher proportion of high-severity patients (ESI 2) is associated with lower physician efficiency, ceteris paribus.

Hypothesis 1b. A higher proportion of low-severity patients (ESI 4) is associated with higher physician efficiency, ceteris paribus.

2.2.2 Physician Experience

There is a substantial body of literature exploring the impact of physician experience—commonly defined as the number of years in clinical practice—on patient outcomes. Some studies document a negative relationship between experience and adherence to recommended guidelines, with younger physicians more likely to follow disease management protocols (Jacques et al. 1991, Kenny et al. 1993, Elstad et al. 2010) and perform key components of comprehensive diabetic examinations (Elstad et al. 2010). However, much of this research uses the provision of specific therapies or recommended tests as proxies for quality of care, leaving broader quality metrics, such as in-hospital mortality or post-discharge outcomes, less frequently examined. The few studies addressing these broader metrics are often narrow in scope, focusing on single diagnoses (Norcini et al. 2000), specific patient populations (Reid et al. 2010), or combined data for both surgeons and physicians (Burns and Wholey 1991).

In contrast, other studies report a positive association between physician experience and clinical performance, attributing this relationship to greater knowledge and more refined skills that improve patient care (Sacchetti et al. 1992, Haas et al. 1995, Choudhry et al. 2005). Evidence from surgical and obstetric settings links experience to better outcomes (Venkataraman et al. 2018), while experienced physicians also tend to exhibit stronger doctor–patient communication (Canwell and Ramirez 1997) and lower anxiety levels (Bovier and Perneger 2007), both of which contribute to higher quality of care. Extending this line of research, prior studies suggest that the relationship between physician experience and performance is unlikely to be strictly linear. While early-career experience is often associated with rapid gains in performance due to learning and skill acquisition, mid- to late-career stages may yield diminishing returns—or even performance declines—as clinical knowledge becomes outdated or cognitive flexibility wanes (Choudhry et al. 2005, Dane 2010, Tsugawa et al. 2017). These findings point to an inverted-U shaped relationship between experience and performance, which motivates our expectation that the effect of experience on efficiency is non-linear. Patterns of resource use further underscore this pattern. McDonald et al. (2024) find that emergency physicians' resource utilization patterns vary by years of post-residency experience, with mid-career physicians (3–8 years) ordering fewer advanced imaging tests and admitting fewer patients than both less experienced and senior colleagues. This U-shaped pattern suggests that efficiency may peak at moderate experience levels, reinforcing the idea that accumulated, up-to-date experience enables physicians to balance clinical thoroughness with effective resource use, particularly in complex care environments.

In the context of EDs, the relationship between physician experience and clinical or operational performance is particularly pronounced. Experienced ED physicians are more adept at diagnosing and managing complex medical conditions due to their extensive clinical exposure and well-honed decision-making skills (Chisholm et al. 2000). Additionally, Pines et al. (2012) find that more experienced ED physicians have lower rates of diagnostic errors and shorter patient LOS. Similarly, Kohn et al. (2000) report that physician experience significantly reduces adverse events and improves patient survival rates. Beyond clinical outcomes, experienced ED physicians have demonstrated greater efficiency in patient assessment and management, which reduces wait times and enhances patient throughput (Asplin et al. 2003). Additionally, their ability to multitask and prioritize tasks contributes to improved departmental workflow and overall ED efficiency (Horwitz et al. 2009). While prior research has highlighted specific outcomes such as LOS or patient throughput, less attention has been paid to how physician experience might influence a more comprehensive measure of efficiency within ED settings. Notably, carefully evaluating efficiency requires consideration of both inputs (e.g., resource use) and outputs (e.g., LOS). Drawing on our DEA-based efficiency scores that capture this input-output relationship, we contribute to the literature by testing the following hypothesis:

Hypothesis 2. Physician experience is positively associated with physician efficiency, with the effect diminishing at higher levels of experience, ceteris paribus.

2.2.3 Shift Familiarity

The significance of familiarity within shared work environments extends beyond formal team dynamics, influencing individual efficiency and adaptability. In our study, shift familiarity is measured as the proportion of patients a physician sees during their most historically familiar shift type—day, evening, or night—in a given week, where the most familiar shift type is determined by the physician's cumulative patient volume in each shift type across all prior weeks.

Shift familiarity, as conceptualized in our study, is a broader construct than just familiarity with peers. It encompasses not only interpersonal dynamics but also familiarity with the shift's operational context, workflows, and patient mix. This repeated exposure allows physicians to internalize routines, anticipate workflow demands, and adjust more efficiently to shift-specific conditions. For example, repetition-based learning and cognitive load theory suggest that such familiarity reduces decision-making complexity and enables more streamlined clinical performance in high-variability environments like the ED. Empirical studies in healthcare settings have supported the role of familiarity in improving team dynamics and performance. In the ED context, familiarity among physicians and other staff might reduce communication barriers and support high performance. Extending beyond interpersonal trust, recent work links familiarity to measurable improvements in clinical operations, including increased patient pick-up rate, improved multitasking, and shorter patient wait times (Niewoehner et al. 2023). While much of the existing research focuses on formal team-based environments, our study extends this understanding to non-team-based operational settings. In the ED setting of our study, physicians manage their own patients independently but share the same shift environment, interacting informally and relying on shared resources, workflows, and patient load. By conceptualizing familiarity at the shift level rather than within fixed teams, our study captures how repeated exposure to a shared operational environment shapes individual physician efficiency.

The benefits of shift familiarity, however, may not be uniform. For example, Bray et al. (2026) find that familiarity improves provider decision times in the ED, but these benefits diminish in novel or high-uncertainty contexts. Building on this, we examine whether the effect of familiarity also depends on physician characteristics. In particular, while experience is a known driver of physician efficiency, it may also shape how physicians benefit from shift familiarity. In unfamiliar or less predictable settings, physicians draw more heavily on their experience to maintain performance, leveraging their broader knowledge and adaptive problem-solving skills. In contrast, in settings characterized by routine and familiarity, the added benefit of experience may diminish, as efficiency can be sustained through repetition and procedural knowledge alone. This suggests that physician experience does not exert a uniform, additive effect on performance, but rather plays a dynamic and moderating role, shaping how environmental characteristics influence efficiency.

This hypothesis is supported in part by studies such as Bray (2024), who finds that while both familiarity and expertise independently enhance ED performance, they also interact as substitutes. Specifically, the research shows that the positive effect of familiarity is more pronounced when teams include less experienced members, suggesting that expertise moderates the influence of familiarity. These findings underscore the importance of accounting for interaction effects when designing staffing and team formation strategies in dynamic healthcare settings. Similarly, studies on physician scheduling and performance provide strong support for this moderating effect. Iyasere et al. (2022) demonstrate that team familiarity improves physician-nurse performance, especially for early-career residents. In a randomized trial, residents consistently assigned to the same nursing floor outperformed peers on leadership, communication, and clinical task outcomes after 12 months. The effect was most pronounced among less experienced physicians, suggesting that familiarity and experience may act as substitutes—supporting the idea that experience moderates the efficiency gains from familiarity. Parallel findings have emerged in surgical domains. Maruthappu et al. (2016) find that while individual surgeon experience improved operative efficiency, repeated collaboration with a familiar assistant consistently reduced procedure duration, even for senior clinicians. The effect of such familiarity was especially pronounced among junior surgeons, suggesting it can serve as a partial substitute for experience in supporting performance.

Additional studies underscore the interdependent relationship between experience and familiarity in shaping physician performance, offering further support for a moderating role of experience. For example, Li et al. (2016) find that while experienced ED physicians generally achieve shorter LOS, the added benefits of environmental familiarity are more pronounced among less experienced physicians. Similarly, Mehrotra et al. (2012) show that physician performance is influenced by familiar team members, but the effect appears strongest for those with lower levels of individual experience—suggesting that the efficiency gains from familiarity diminish as experience increases.

Together, these theoretical frameworks and empirical findings suggest that shift familiarity positively influences physician efficiency, particularly for less experienced physicians who benefit most from the cognitive and operational support that familiar shift environments provide. As physician experience increases, this benefit diminishes—experienced physicians maintain high efficiency across varying shift contexts through their accumulated adaptive capacity, reducing their dependence on environmental familiarity. While prior studies have documented the interaction between familiarity and experience in influencing isolated performance outcomes, our study extends this literature by examining this interaction through an efficiency-based framework that evaluates how physicians transform clinical resources into patient outcomes across the full spectrum of ED care. By doing so, we examine how physician experience and shift familiarity jointly shape physician efficiency in a non-team-based ED setting. We therefore test the following hypothesis:

Hypothesis 3. Shift familiarity is positively associated with physician efficiency, with this benefit being more pronounced for less experienced physicians. Specifically, the positive association between shift familiarity and efficiency diminishes as physician experience increases.

While prior research has examined physician experience and patient severity independently, their joint effect on physician efficiency—particularly in high-pressure emergency medicine settings—remains underexplored. Building on the established relevance of patient severity, we investigate whether physician experience moderates the relationship between patient severity and physician efficiency. We hypothesize that the challenging impact of high patient severity on efficiency is substantially mitigated by more experienced physicians. In other words, under more clinically urgent conditions—those involving severely ill or unstable patients—the negative influence of severity on performance is less pronounced for experienced physicians. High-severity cases may demand faster decision-making, greater familiarity with unstable presentations, and more refined clinical judgment—competencies that are more fully developed in more experienced physicians (Croskerry 2002, Klein 1999). By contrast, low-severity cases are typically routine and protocol-driven, allowing relatively inexperienced physicians to deliver care effectively, thereby narrowing the efficiency gap and making experience's moderating role less salient in these settings. These findings suggest that physician experience may help mitigate the efficiency challenges posed by more clinically demanding cases, while offering less marginal benefit in routine, low-severity scenarios.

The broader literature reinforces this view. Tsugawa et al. (2017) observe that patient mortality among elderly inpatients increases with physician age, but this trend is attenuated for high-volume physicians, implying that continuous experience is particularly beneficial in complex cases. Similarly, Peterson et al. (2008) find that surgeon experience significantly moderates the effect of patient risk on outcomes in cardiac surgeries, with a stronger association in high-risk cases than in low-risk ones. In trauma and intensive care settings, physician expertise more strongly mitigates the adverse outcomes associated with high-severity cases, affecting measures such as ICU mortality and failure-to-rescue rates (Ghaferi et al. 2009, Pronovost et al. 2002). From a cognitive perspective, physician experience supports the management of heavier mental loads and the advanced diagnostic and coordination demands of high-severity cases—skills developed through repeated exposure and reflective practice (Dane 2010). Conversely, low-severity encounters are often managed effectively using standard procedures, reducing the marginal benefit of accumulated experience. Although the broader literature consistently suggests that the performance benefits of experience are more pronounced under conditions of high patient severity and clinical complexity, to our knowledge, this moderating effect has not been explicitly tested or directly quantified in the context of physician efficiency within EDs. To address this gap and build on these insights, we test the following hypothesis:

Hypothesis 4. Physician experience moderates the impact of patient severity on efficiency. Specifically, the negative impact of high patient severity on efficiency is mitigated by more experienced physicians.

3 DEA Model

DEA, first introduced by Charnes et al. (1978), is a methodology for evaluating the relative efficiency of decision-making units (DMUs) in settings with multiple inputs and outputs. A DMU is an entity that transforms inputs into outputs, with performance evaluated relative to its peers (Cooper et al. 2007). Unlike central tendency approaches that benchmark against an average performer, DEA evaluates units relative to the best-performing peers. A key advantage of DEA over regression-based methods is that it does not require specification of a functional form between inputs and outputs. Consequently, DEA can uncover patterns not captured by parametric approaches, providing a more complete view of performance. However, as a data-driven method, DEA is sensitive 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, the DEA methodology provides for a proportional expansion in outputs rather than a reduction in inputs while keeping inputs constant. For the purpose of this study, we employ the input-oriented orientation. Since there is no theoretical reason to assume that inputs and outputs scale proportionally in our ED setting, we adopt the Banker-Charnes-Cooper (BCC) model (Banker et al. 1984), which relaxes the constant returns to scale (CRS) assumption and allows for variable returns to scale (VRS). We validate this specification empirically using the returns-to-scale test proposed by Simar and Wilson (2002, 2011). Our results reject the null hypothesis of CRS in favor of VRS, supporting the use of the BCC model for efficiency estimation.

Let DMU_0 be the unit under evaluation, with input and output vectors x_{i0} and y_{r0}. The input-oriented BCC model (Banker et al. 1984) under VRS is formulated as the following linear program: $\min\{\theta : \sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{i0}, \forall i; \sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{r0}, \forall r; \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0, \forall j\}$ The first constraint ensures that the weighted combination of inputs (with λ_j denoting the weight assigned to the jth DMU) across all peer DMUs does not exceed a proportion θ of the inputs used by the DMU under evaluation. This reflects the model's objective of identifying how much input could be proportionally reduced without lowering output levels. The second constraint guarantees that benchmark outputs are at least as great as those produced by the DMU being assessed, thereby ensuring that no reduction in output is allowed when seeking greater input efficiency. The convexity constraint ($\sum \lambda_j = 1$) enforces the VRS assumption by requiring the benchmark to be a convex combination of observed units. Finally, the non-negativity constraint ensures the efficiency benchmark is constructed solely from feasible combinations of actual DMUs.

3.1 Model Variables

The choice of input and output variables in our study is based on viewing the physician as a "production entity" that utilizes hospital resources (inputs) to generate effective care (outputs). In DEA, there is no objective definition of the "right" variables. We therefore defined the model's input and output variables using parameters (a) that best reflect physician performance, (b) for which there is face validity and physician agreement, and (c) that appear as common measures in emergency medicine and ED operations literature. DEA variable selection was developed collaboratively with two board-certified emergency medicine physician co-authors with direct clinical and operational expertise at the partner ED. This clinical input complements the existing literature and supports the validity of the variable definitions.

We define each DMU at the physician-week level to ensure sufficient variation across input and output variables, while maintaining an aggregation level that reflects meaningful differences in physician performance. This level of aggregation provides sufficient patient volume within each unit to compute stable outcome-based quality measures. At finer levels of aggregation (e.g., shifts or individual encounters), quality measures are computed over very few observations and become highly sensitive to random noise and clinical uncertainty, leading to unstable efficiency estimates. Drawing from a detailed dataset of patient-level visits at our partner hospital, we construct a panel of 2,111 physician-week observations. Our DEA model evaluates the efficiency of physician i in week t, relative to peers, based on how efficiently they convert hospital resources into delivered care. We next describe the inputs and outputs used in our DEA model. Inputs proxy ED resource utilization, while outputs reflect standard performance measures of care quality used in the emergency medicine literature.

Outputs:

- *Rate of discharged patients who do not return within 72 hours:* We consider each physician's percentage of discharged patients (i.e., those not admitted to the hospital immediately after their ED visit) who do not return to the ED within 72 hours as one of the model's output variables. Returns to the ED within 72 hours of discharge are widely used as indicators of care quality in the emergency medicine literature (Abualenain et al. 2013, Klasco et al. 2015). While Pham et al. (2011) note that returns can reflect patient-driven factors beyond physician control, the rotational assignment of patients to physicians in our setting reduces systematic case-mix differences, and our second-stage regression controls extensively for patient severity, demographics, and complaint mix. Since return visits following discharge are generally viewed as an undesirable indicator of care delivery in the ED, 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 an output variable in our model.
- *Rate of admitted patients who are not discharged within 24 hours:* The percentage of patients admitted by a physician from the ED to an inpatient unit in the hospital who were subsequently discharged within 24 hours of admission serves as a proxy for how often the physician overcalls their patients' illness severity. Since this overcalling rate is considered an undesirable output, we use the non-overcalling rate (24-hour non-discharge patient admission rate) as an output variable in our model (Traub et al. 2017).

- *Rate of patients admitted to a floor/ward bed who are not upgraded within 24 hours:* Similarly, 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 24 hours of admission can be considered as a proxy for the physician's undercalling rate. Since this undercalling rate is also viewed as an undesirable outcome, we use the non-undercalling rate (24-hour non-upgrade patient admission rate) as an output variable in our model (Traub et al. 2017). To ensure stable efficiency estimates, we require a minimum of 15 patient visits and five general internal medicine admissions per physician-week to ensure reliable measurement of overcalling and undercalling rates, yielding a final sample of 27 physicians observed over 1,679 physician-weeks.

It should be noted that the first output variable pertains only to discharged patients, while the other two output variables are related to admitted patients and directly reflect the quality of the ED physician's admission decision. The choice of threshold values (72 and 24) is consistent with clinical practice and is informed by the literature (Traub et al. 2017, Keith et al. 1989, Gordon et al. 1998) as well as discussions with ED physicians at our partner hospital. As a robustness check, we re-estimate the DEA model using an alternative return window, replacing the typical 72-hour non-return rate with a longer 216-hour non-return rate. Our inferences remain unchanged (Table EC.1 in the E-Companion). As an additional robustness check, we estimate an alternative DEA specification that risk-adjusts outputs for patient demographic and case-mix characteristics (e.g., age distribution, gender composition, insurance status, and presenting condition shares). The resulting efficiency scores yield qualitatively consistent results in the second-stage analysis, reported in Table EC.2 of the E-Companion.

Inputs:

- *Test Order Count*: We use the average number of diagnostic and imaging tests including radiology, ultrasound, MRI, and CT ordered by a physician per patient visit as one of the model's input variables. Diagnostic and imaging test services constitute a significant resource in hospitals and have been used as a valid input variable in DEA in prior literature (Chilingerian 1995).

- *Contact-to-Disposition Time*: This variable denotes the time from a physician's initial contact with the patient to the time that a disposition (admit or discharge) order is issued. A substantial body of literature on hospital utilization has viewed LOS as a surrogate measure of hospital resource utilization (e.g., using diagnostic test services, ED beds, etc.) (e.g., Chilingerian 1995, Fiallos et al. 2017).[1] Since LOS captures the total time a patient spends in the ED, which is not entirely within the physician's control, we use contact-to-disposition time—a metric directly influenced by physicians—as the model's input variable. As a robustness check, we re-run our analysis using LOS as the sole input while keeping the output variables intact and find that our inferences remain unchanged (Table EC.3 in the E-Companion). We note that contact-to-

[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 delivered by a team of physicians (see, e.g., Saghafian et al. 2012, Saghafian and Hopp 2019, and the references therein). Thus, a patient's outcome is directly related to the physician who treats the patient.

disposition time reflects physician case-management effort in an operational environment characterized by multitasking and fluctuating congestion, rather than purely sequential treatment of individual patients. Although it may also capture system-level factors such as ED congestion and concurrent patient load, we account for these effects in the second-stage analysis through controls for patient mix, workload intensity, and operational conditions.

Before proceeding with efficiency estimation, we conduct a series of diagnostic tests to ensure that the assumptions underlying our DEA model are satisfied and that the estimation is robust to methodological concerns. To assess the validity of the convexity assumption, we compare DEA scores which impose a convex production frontier with those from a Free Disposal Hull (FDH) model, which does not. Following the guidelines in Kneip et al. (2016), the convexity assumption is supported if DEA scores are consistently lower than or equal to their FDH counterparts. This condition is satisfied across all observations in our sample, with a mean DEA efficiency of 0.555 compared to a mean FDH efficiency of 0.613 and zero violations across all 1,679 observations, confirming the appropriateness of the convex DEA framework and reinforcing the reliability of our efficiency estimates. Moreover, we evaluate other key DEA assumptions to ensure interpretability and reliability. With 2 inputs ($I = 2$) and 3 outputs ($R = 3$), our sample of 1,679 physician-week observations exceeds the minimum discrimination threshold of 15 DMUs ($\max I \times R = 6,;3(I+R) = 15$), consistent with common DEA guidelines (e.g., Bowlin 1998). We also ensure monotonicity by selecting clinically relevant inputs and outputs, validated in consultation with physician collaborators, such that more input use does not reduce output. Recognizing that DEA attributes all inefficiency to deterministic sources, we implement bootstrap DEA (Section 6.1) to account for stochastic noise and obtain bias-corrected efficiency scores. We further assess robustness using supervised ML models that independently classify physician efficiency based on observed characteristics (Section 6.6). The strong predictive performance of these models supports the internal validity of our results. In addition, all DMUs in our sample operate within the same ED and occupy comparable clinical roles, satisfying the homogeneity assumption. Together, these checks provide strong support for the reliability of our DEA-based efficiency estimates.

4 Regression Analysis

Building on the efficiency scores described in the previous section, we use these weekly measures as the dependent variable in a regression framework to test our hypotheses and examine factors influencing physician efficiency. We outline the independent and control variables and describe the estimation approach.

4.1 Independent Variables

We employ three main independent variables. The first independent variable is physician experience, which is quantified as the number of years the physician has been in clinical practice. We expect physician experience to be associated with efficiency in a non-linear manner—improving in early career stages and potentially plateauing or declining thereafter. To capture this potential inverted-U relationship, we include both a linear and a squared term for physician experience. The second independent variable is patient severity, measured using the proportions of ESI levels 2 and 4 patients in the physician's caseload, with ESI level 3 as the reference category. ESI levels 1 and 5 are excluded because they occur very infrequently in our dataset, resulting in insufficient variation for reliable estimation. The proportion of ESI level 2 patients is used to test both H1a and the moderating effect of experience on severity (H4), as accumulated experience is expected to yield the greatest efficiency gains when cases are most complex. The proportion of ESI level 4 patients is used to capture the effect of low-severity case mix on efficiency (H1b). The third independent variable is shift familiarity, measured as the proportion of patients a physician sees during their most historically familiar shift type (day, evening, or night) in a given week. In addition to these direct effects, our model incorporates interaction terms between physician experience and shift familiarity, and between physician experience and the proportion of high-severity patients (ESI 2). These interactions test whether the effects of shift familiarity and high-severity patient mix on efficiency vary with physician experience.

4.2 Control Variables

To account for potential confounders that may influence physician efficiency, we include a comprehensive set of control variables in the regression analysis. First, we control for patient demographics, including the proportion of patients aged 75 and older, the proportion of female patients, and the racial composition of each physician's caseload during the week. We also include the proportion of Medicaid patients as a proxy for patient socioeconomic status. Second, we control for clinical case mix by incorporating the proportions of key diagnostic categories treated by the physician during the week, including cardiac, abdominal, neurological, respiratory, trauma, and infection cases. In addition, we include the standard deviation of ESI scores within the physician's caseload to capture variation in patient severity levels.

Third, we control for operational conditions within the ED. This includes ED volume, defined as the total number of patients treated in the ED in a given week excluding those treated by the focal physician, which captures system-level crowding and resource strain. We also include the average patient load per shift for each physician and the total number of patients treated by the physician during the week. In addition, we include the average door-to-MD time, which measures the time between a patient's arrival and their initial assessment by a physician and serves as an indicator of ED flow and congestion. Fourth, we incorporate shift composition variables, including the proportions of day and weekend shifts worked by each physician during the week. These controls account for heterogeneity in scheduling that may systematically affect physician performance. Fifth, we include physician job tenure, measured as the number of years employed at the same ED, to capture familiarity with the local work environment and to separate site-specific experience from broader clinical experience. Finally, we include month and year fixed effects to account for unobserved temporal factors that may influence efficiency over time.

4.3 Regression Strategy

To gain insights into factors that affect physician efficiency and test our hypotheses discussed earlier, we use the following regression model:

$$\theta_{it} = \beta_1 \mathbf{X}_{it} + \beta_2 Z_{it} + \beta_3 \text{Experience}^2_{it} + \alpha_1(\text{Experience}_{it} \times \text{ShiftFamiliarity}_{it}) + \alpha_2(\text{Experience}_{it} \times \text{ESI}_{it}) + \mu_m + \delta_y + \varepsilon_{it} \tag{2.1}$$

where θ_{it} is the generated DEA score for physician i in week t (defined in Section 3), X_{it} represents a vector of independent variables, and Z_{it} denotes a vector of control variables. μ_m and δ_y denote month and year fixed effects, respectively, and ε_{it} is the error term. Table 1 presents summary statistics of the variables used in our DEA and regression analysis. In order to estimate the coefficients in model (2.1), we require a regression method that accommodates the characteristics of DEA efficiency scores. These scores are bounded between 0 and 1 and often exhibit a mass of observations at the upper limit, which can lead to non-normal and heteroskedastic errors—violating key assumptions of OLS. To address this, we use Tobit regression, which is designed for limited dependent variables and is frequently used in the DEA literature as a second-stage method to examine the relationship between efficiency scores and explanatory variables (Chilingerian 1995). While we use a Tobit specification as our primary empirical model, we re-estimate our model using OLS and double bootstrap truncated regression (Simar and Wilson 2007) as robustness checks (Sections 6.4.1 and 6.4.2). All findings are qualitatively consistent across alternative specifications. Finally, as noted earlier, concerns about endogeneity and related issues are likely mitigated in our setting due to the rotational assignment of patients to providers at our partner ED as well as how our independent variables are defined. Nevertheless, we conduct a sensitivity to omitted variable bias analysis and a permutation-based placebo test in Sections 6.2 and 6.5 respectively, to more rigorously assess the robustness of our results.

5 Results

5.1 Descriptive Analysis of Efficiency Measures

Figure 1 illustrates the distribution of physician efficiency scores obtained from the DEA model. The distribution is unimodal and right-skewed, with a gradual tail extending toward higher values. While most physician–week observations are concentrated between 0.47 and 0.62, this skewness indicates that a small subset achieves substantially higher performance relative to the production frontier. Table 2 provides the corresponding summary statistics. Efficiency scores range from 0.30 to 1.00, with a mean of 0.55 (SD = 0.12) and a median of 0.53. The interquartile range aligns with the visual concentration seen in Figure 1, confirming that while most observations cluster around the center, there remains meaningful dispersion across the sample. Importantly, the DEA specification maintains high discriminatory power. Only eight physician–week observations, representing six individual physicians, attain the maximum score of 1.00 to

Table 1 Summary Statistics of First-Stage (DEA) and Second-Stage (Tobit Regression) Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
Input Variables				
Average Imaging Tests per Patient	1.03	0.26	0.35	3.56
Contact-to-Disposition Time (Minutes)	145.01	30.05	62.41	302.05
Output Variables				
72-hr Rate of Non-Return	0.97	0.04	0.79	1.00
Non-Overcalling Rate	0.82	0.14	0.17	1.00
Non-Undercalling Rate	0.97	0.06	0.60	1.00
Independent Variables				
Experience (Years)	21.19	6.81	6.00	39.00
Shift Familiarity	0.55	0.24	0.02	1.00
Proportion of ESI-2 Patients	0.14	0.06	0.00	0.38
Proportion of ESI-4 Patients	0.12	0.06	0.00	0.39
Control Variables				
Proportion of Age 75+ Patients	0.25	0.07	0.00	0.57
Proportion of Female Patients	0.53	0.08	0.21	0.91
Proportion of White Patients	0.91	0.05	0.65	1.00
Proportion of Medicaid Patients	0.05	0.04	0.00	0.30
Proportion of Cardiac Cases	0.07	0.04	0.00	0.29
Proportion of Abdominal Cases	0.14	0.06	0.00	0.44
Proportion of Neurological Cases	0.03	0.03	0.00	0.22
Proportion of Respiratory Cases	0.04	0.03	0.00	0.29
Proportion of Trauma Cases	0.05	0.04	0.00	0.26
Proportion of Infection Cases	0.06	0.04	0.00	0.31
ED Volume	870.89	472.87	92.00	1663.00
Proportion of Day Shifts	0.47	0.28	0.00	1.00
Proportion of Weekend Shifts	0.27	0.32	0.00	1.00
Average Patients per Physician per Shift	9.81	4.27	3.00	41.00
Average Door-to-MD Time (Minutes)	33.29	17.11	4.07	118.71
ESI Standard Deviation	0.55	0.09	0.22	1.04
Patient Count	54.12	29.51	15.00	177.00
Job Tenure (Years)	9.97	5.90	0.00	18.96

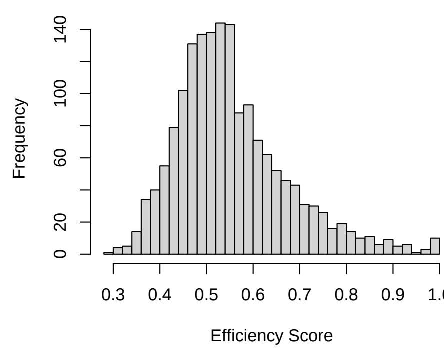
Note: $N = 1,679$. All variables are aggregated at the physician-week level. Inputs and outputs reflect variables used in the DEA model. Independent variables are primary regressors in the second-stage Tobit regression; control variables are included in all regressions.

Table 2 Summary Statistics – DEA Scores

Statistic	Value
Sample Size	1,679
Mean	0.5552
Median	0.5347
Standard Deviation	0.1193
25th Percentile	0.4731
75th Percentile	0.6155
Minimum	0.2994
Maximum	1.0000

Note: Observations are at the physician-week level.

Figure 1 Distribution of Efficiency Scores



define the best-practice frontier. Under the input-oriented DEA framework, these frontier observations correspond to physician–week observations that achieve the highest relative efficiency by minimizing resource utilization—imaging use and contact-to-disposition time—given the observed patient outcomes included in the model. The limited mass at this threshold ensures the model effectively differentiates performance rather than clustering a large share of observations at the frontier. To characterize the empirical source of efficiency differences across physicians, we compare the inputs and outputs of physician-week observations in the top efficiency quartile (n = 420) with the remaining observations (n = 1,259). While an input-oriented DEA model defines efficiency through input reduction, output variation could, in principle, contribute to score differences. However, our results confirm that efficiency is predominantly driven by input variation. Efficient physician-weeks are associated with 26.9% fewer imaging tests per patient (0.81 vs. 1.11, $p < 0.001$) and 26.6% shorter contact-to-disposition times (114.0 vs. 155.3 minutes, $p < 0.001$), while output differences are modest ($r \leq 0.258$ between efficiency and any output measure). These findings confirm that more efficient physicians achieve comparable care quality while consuming substantially fewer resources. To explore the potential drivers of physician efficiency, we now turn to the hypotheses outlined in Section 2 and test their validity.

5.2 Tobit Regression Results

Table 3 reports the results of the Tobit regression model used to test our hypotheses. Complete regression results are provided in Table EC.4 in the E-Companion. The analysis reveals that patient severity significantly affects physician efficiency in both directions, supporting Hypotheses 1a and 1b. A higher proportion of high-severity patients (ESI 2) is associated with lower efficiency ($B = -0.2013$, $p < 0.01$), while a higher proportion of low-severity patients (ESI 4) is associated with higher efficiency ($B = 0.2228$, $p < 0.001$), consistent with the increased cognitive load and resource demands associated with complex cases. This result carries important implications for ED operations. It highlights the importance of targeted training and staffing strategies to mitigate efficiency losses when managing higher-severity cases, where clinical complexity and resource demands are greatest. Additionally, optimizing workflows and resource allocation for lower-severity patients may improve overall efficiency and help preserve capacity for more critical cases.

Our results support Hypothesis 2, revealing a statistically significant positive relationship between physician experience and efficiency ($B = 0.0049$, $p < 0.01$), consistent with recent empirical evidence showing that senior physicians use fewer imaging resources while maintaining lower 72-hour revisit rates compared with younger physicians (Querin et al. 2025). The negative and significant squared term ($B_{exp^2} = -0.0004$, $p < 0.05$) further indicates that efficiency gains plateau and may decline at higher experience levels, consistent with an inverted U-shaped relationship. Our results support Hypothesis 3, with the interaction term between physician experience and shift familiarity being negative and statistically significant ($B = -0.0031$, $p < 0.05$), confirming that experience and shift familiarity act as substitutes. Less experienced physicians benefit more from working familiar shifts, whereas more experienced physicians can maintain efficiency even in less familiar environments. This insight presents an opportunity to optimize staffing schedules by strategically assigning experienced physicians to less familiar shifts, where their experience compensates for the lack of familiarity, while prioritizing familiar shifts for less experienced physicians where efficiency gains are maximized. Such an approach could ensure more balanced efficiency across shifts, ultimately improving overall operational performance in the ED.

Table 3 Tobit Regression Results (Dependent Variable: DEA Efficiency Score)

Variable	**Estimate**
Proportion of ESI-2 Patients	−0.2013** (0.0717)
Proportion of ESI-4 Patients	0.2228*** (0.0628)
Shift Familiarity	0.0148 (0.0228)
Experience	0.0049** (0.0019)
Experience^2	−0.0004* (0.0002)
Experience × Shift Familiarity	−0.0031* (0.0016)
Experience × Proportion of ESI-2	0.0210*** (0.0055)
Experience × Proportion of ESI-4	0.0050 (0.0075)
Controls	Yes
Month and Year Fixed Effects	Yes
Observations	1,679
Wald Statistic	353.23***
Log-Likelihood	1,344.85
AIC	$-2,603.69$
McKelvey-Zavoina R^2	0.1739

Notes: Standard errors in parentheses. Model includes controls for patient mix, demographics, shift composition, and ED operational factors. Standard errors are clustered at the physician level.
$^\dagger p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Finally, we find a positive and statistically significant interaction between physician experience and the proportion of ESI-2 patients ($B = 0.0210$, $p < 0.001$). This result suggests that physician experience moderates the impact of patient severity on efficiency. In other words, the typically challenging impact of higher patient severity on efficiency is mitigated or better managed by more experienced physicians. This finding supports Hypothesis 4 and aligns with the notion that under more clinically complex and urgent conditions, the benefits of accumulated experience become more pronounced. High-severity cases often require faster decision-making, greater familiarity with unstable presentations, and more refined clinical judgment—competencies that tend to develop with greater clinical experience.

6 Robustness Tests

In this section, we present a series of robustness checks conducted to assess the validity of our findings and the methods used to generate them.

6.1 Bootstrap DEA

Bootstrap DEA leverages the resampling methodology of bootstrapping to enhance the robustness of efficiency scores obtained from traditional DEA. The method involves repeatedly drawing pseudo-samples with replacement from the original dataset, applying DEA to each sample to create a distribution of efficiency scores. Traditional DEA can be sensitive to sample size and noise, often yielding biased scores, particularly in small or noisy datasets. By generating multiple pseudo-samples, bootstrap DEA mitigates these issues, stabilizing the obtained scores and reducing the impact of outliers. This correction improves the accuracy and consistency of the efficiency estimates, resulting in more reliable conclusions. We apply a bootstrapping approach with 5,000 repetitions to enhance the reliability of the DEA results. As shown in Table EC.5 in the E-Companion, re-estimating the Tobit regression using bootstrapped efficiency scores yields qualitatively consistent results, supporting the robustness of our main findings.

6.2 Sensitivity to Omitted Variable Bias

While rotational patient assignment and our variable construction mitigate several sources of endogeneity, omitted variable bias remains a potential concern. Despite controlling for time-varying factors—including patient mix, ED volume, and shift composition—and incorporating time fixed effects, residual confounding may persist due to unobserved determinants of physician efficiency. In particular, unobserved physician characteristics—such as innate ability, decision-making style, learning speed, or adaptability—may be correlated with physician experience, shift familiarity, and observed efficiency. In addition, systematic differences in patient exposure within shifts may further bias our estimates.

Our empirical setting and variable construction substantially limit the scope for such bias. First, shift familiarity is a predetermined measure based on historical exposure prior to the focal week, rather than an assignment based on current performance, which limits concerns about reverse causality. Second, the rotational assignment of patients minimizes the potential for systematic sorting based on physician discretion, as patients are assigned independently of physician characteristics. Third, physician experience evolves mechanically over time and is therefore unlikely to be jointly determined with contemporaneous efficiency. While these features reduce concerns about reverse causality, they do not eliminate the possibility that these variables remain correlated with unobserved physician characteristics.

To further assess sensitivity to omitted variable bias, we examine how key coefficients evolve across nested specifications following Altonji et al. (2005). Results are reported in Table 4. The coefficient on physician experience remains positive and statistically significant across all four specifications, decreasing slightly from 0.0050 in the restricted model to 0.0049 in the fully specified model, a negligible change of 2.0%, with significance maintained at $p < 0.01$. The quadratic experience term remains negative and statistically significant across all specifications, consistent with the inverted U-shaped relationship between experience and efficiency. The interaction term between shift familiarity and experience remains negative and statistically significant across all four specifications, with a magnitude change of 16.2% from the restricted to the fully specified model—confirming the robustness of the substitution effect between experience and familiarity. The ESI-2 main effect grows in magnitude from -0.1068 to -0.2013 as controls are added, with significance strengthening from $p < 0.05$ to $p < 0.01$. The ESI-4 coefficient remains positive and statistically significant ($p < 0.001$) across all specifications with a decrease in magnitude from 0.2886 to

Table 4 Coefficient Stability Across Nested Specifications

	(M1)	(M2)	(M3)	(M4)
Experience	0.0050**	0.0050**	0.0049**	0.0049**
	(0.0019)	(0.0018)	(0.0018)	(0.0019)
Experience^2	−0.0004*	−0.0005*	−0.0004*	−0.0004*
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Experience × Familiarity	−0.0037*	−0.0035*	−0.0039**	−0.0031*
	(0.0015)	(0.0015)	(0.0015)	(0.0016)
Proportion of ESI-2 Patients	−0.1068*	−0.1136†	−0.1163*	−0.2013**
	(0.0535)	(0.0585)	(0.0533)	(0.0717)
Proportion of ESI-4 Patients	0.2886***	0.2824***	0.2898***	0.2228***
	(0.0547)	(0.0549)	(0.0553)	(0.0628)
Proportion of ESI-2 × Experience	0.0197***	0.0214***	0.0200***	0.0210***
	(0.0059)	(0.0060)	(0.0058)	(0.0055)
McKelvey-Zavoina R^2	0.1206	0.1371	0.1556	0.1739
Observations	1,679	1,679	1,679	1,679

Notes: Each column reports Tobit estimates from progressively richer specifications on the same sample of observations as the main Tobit model. All specifications include month and year fixed effects. M1 includes key hypothesis variables only. M2 adds patient demographic controls. M3 adds complaint mix controls. M4 includes the full set of controls and operational variables. Standard errors clustered at the physician level are reported in parentheses. $^\dagger p < 0.10$, $^* p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$.

0.2228. The interaction between the proportion of high-severity patients (ESI 2) and physician experience is stable, remaining positive and statistically significant across all four specifications with a magnitude change of only 6.6%. Collectively, the pattern of coefficient stability as controls are added is consistent with Altonji et al. (2005), suggesting that omitted variable bias, if present, is more likely to attenuate than inflate the estimated relationships. Nonetheless, we caution that our results should mainly be interpreted as associations rather than causal effects, in line with the scope and contribution of our study.

6.3 Alternative Regression Specification

To further test the robustness of our findings, we re-estimate model (2.1) using an alternative specification. Drawing on prior work that highlights the influence of ED congestion on performance (e.g., KC and Terwiesch 2009, Kuntz et al. 2015), we incorporate both linear and quadratic terms for ED volume in the model. Specifically, we make use of the following model, estimated via Tobit regression:

$$\begin{aligned}\theta_{it} = \beta_1 \mathbf{X}_{it} + \beta_2 Z_{it} + \beta_3 \text{Experience}_{it}^2 + \beta_4 \text{EDVolume}_{it}^2 + \alpha_1(\text{Experience}_{it} \times \text{ShiftFamiliarity}_{it})\\ + \alpha_2(\text{Experience}_{it} \times \text{ESI}_{it}) + \mu_m + \delta_y + \varepsilon_{it} \qquad (2.2)\end{aligned}$$

The regression results, presented in Table EC.6 in the E-Companion, are consistent with our main findings.

6.4 Alternative Second-Stage Estimation

6.4.1 Ordinary Least Squares (OLS) Regression

To strengthen the credibility of our findings, we re-estimate our primary model using OLS with standard errors clustered at the physician level. This analysis serves to assess the sensitivity of our results to alternative modeling and inference assumptions. As demonstrated in Table EC.7 in the E-Companion, across all key variables—ESI proportions, physician experience, shift familiarity, and their interactions—the OLS model yields estimates that are consistent in both magnitude and directional sign with our primary Tobit results. This consistency supports the validity of our conclusions even when the parametric assumptions of the Tobit model are relaxed, and demonstrates that the observed patterns are not artifacts of model choice or estimation method.

6.4.2 Simar–Wilson Double Bootstrap Truncated Regression

A potential concern in two-stage DEA models is that efficiency scores are constructed relative to a common frontier. Consequently, the scores are not independent observations, and standard second-stage estimators may produce biased inference. To address this concern, we implement the double bootstrap truncated regression proposed by Simar and Wilson (2007), which accounts for the dependence structure of DEA efficiency scores by bootstrapping both the DEA scores and the truncated regression to construct bias-corrected estimates and confidence intervals. We re-estimate the second-stage model using $B = 2{,}000$ double bootstrap replications, with results reported in Table EC.8 of the E-Companion. All findings are qualitatively consistent with our baseline Tobit estimates, confirming that our conclusions are not driven by the distributional assumptions of the Tobit specification or by the relative nature of DEA efficiency scores.

6.5 Permutation-Based Placebo Test

To assess whether our findings reflect true relationships between the covariates and physician efficiency—rather than artifacts of the DEA construction or regression specification—we implement a permutation-based placebo test. Specifically, we randomly permute the DEA efficiency scores across physician-week observations, breaking any systematic relationship between efficiency and the covariates, and re-estimate the Tobit specification on each permuted dataset. This procedure is repeated 1,000 times with `set.seed(42)`, generating an empirical null distribution for each coefficient of interest.

The results are reported in Table 5. The observed coefficients for physician experience (0.0049) and its interaction with high-severity cases (0.0210) fall in the extreme tails of their respective permutation distributions (two-sided permutation $p < 0.001$ and $p = 0.005$). Similarly, the effects of patient severity (ESI-2: -0.2013; ESI-4: 0.2228) are statistically significant (two-sided permutation $p = 0.002$ for both coefficients). In contrast, the permutation distributions are centered near zero; all observed coefficients, except the familiarity–experience interaction, lie outside the 95% permutation confidence intervals, while the familiarity-experience interaction remains marginally significant at the 10% level ($p = 0.097$). Together, these results provide strong evidence that the estimated associations are unlikely to arise by chance or spurious correlation.

Table 5 Permutation-Based Placebo Test

	Observed Coef.	Placebo Mean	Placebo SD	Permutation p-value	95% Permutation CI Lower	Upper
Experience	0.0049	0.0000	0.0006	<0.001	−0.0010	0.0011
Familiarity × Experience	−0.0031	0.0000	0.0018	0.097	−0.0036	0.0037
Proportion of ESI-2 Patients	−0.2013	0.0012	0.0667	0.002	−0.1263	0.1323
Proportion of ESI-4 Patients	0.2228	−0.0023	0.0708	0.002	−0.1407	0.1370
Proportion of ESI-2 × Experience	0.0210	−0.0001	0.0075	0.005	−0.0155	0.0137
Permutations				1,000		

Notes: Permutations randomly reassign DEA efficiency scores across physician-week observations and re-estimate the Tobit model. p-values are two-sided and report the share of permutations yielding coefficients with absolute value at least as large as the observed coefficient. CIs denote the 2.5th and 97.5th percentiles of the permutation distribution.

6.6 Machine Learning-Based Validation Approach

Given that DEA relies on deterministic input-output relationships, concerns may arise regarding its sensitivity to model specifications, variable selection, and the flexibility in weight assignment. To address these concerns, we implement a set of supervised ML models as independent validation tools. The objective is to test whether the efficiency classifications identified by DEA—particularly those corresponding to the most efficient physicians—can be independently identified using supervised learning models that do not rely on the production assumptions or weight constraints inherent in DEA. Specifically, to avoid the circularity of validating DEA scores using the same inputs and outputs from which they are derived, we exclude DEA input and output variables from the ML feature set. This ensures that the validation provides an independent assessment of what the efficiency scores capture. Prior research has demonstrated the value of combining DEA with ML techniques, particularly in enhancing robustness and interpretability (Zhu et al. 2021, Liu and Zhang 2019, Misiunas et al. 2016). While much of this prior work aims to predict DEA scores directly, we use ML classifiers to examine whether physicians identified as efficient by DEA can be distinguished from their peers based on observable characteristics. Specifically, we label the top 25% of physician-week observations based on their efficiency scores as efficient (n = 420) and train four supervised ML classifiers—Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Logistic Regression (LR)—to predict this binary label using only contextual variables including physician experience, shift familiarity, patient severity composition, clinical case mix, patient demographics, and operational controls. DEA input and output variables—specifically, average imaging tests, contact-to-disposition time, 72-hour non-return rate, non-overcalling rate, and non-undercalling rate—are deliberately excluded from the feature set (presented in Table 1). We chose the 25% threshold to create a statistically viable and reasonably balanced class of efficient physician-week observations, which is critical for robust training and reliable generalization of the ML classifiers. Stratified 10-fold cross-validation with three repeats is used to ensure that model performance reflects out-of-sample generalization. Figure 2 presents the cross-validated

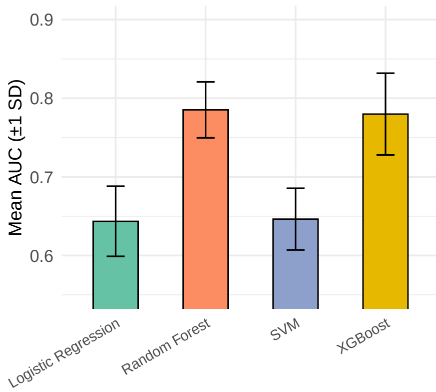


Figure 2 Cross-Validated AUC by ML Model

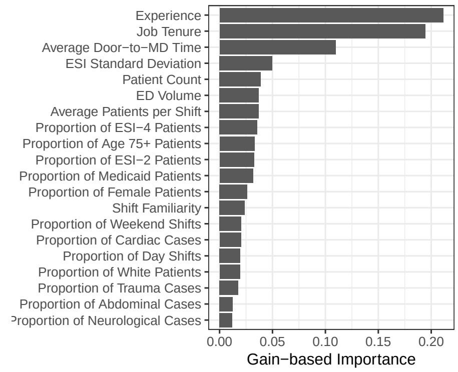


Figure 3 Feature Importance from XGBoost Classifier

Area Under the Curve (AUC) values for each ML model. AUC measures the model's ability to distinguish between efficient and non-efficient physician-week observations across various classification thresholds. The tree-based classifiers—Random Forest and XGBoost—achieve AUC values of 0.784 and 0.781, respectively, indicating meaningful discriminatory performance using contextual variables alone. Logistic Regression and SVM achieve AUC values of 0.647 and 0.646, respectively, suggesting that the relationship between contextual variables and efficiency is nonlinear, consistent with the flexible production frontier approach of DEA. The meaningful predictive performance of the tree-based models using only contextual variables suggests that the DEA efficiency classifications align with independently identifiable patterns in physician, patient, and operational characteristics, rather than being solely driven by DEA-specific modeling assumptions.

To gain additional insights, we also examine the features most important for the classification task. We focus on the XGBoost model, which achieved the second highest AUC among the classifiers and provides interpretable gain-based feature importance scores. Figure 3 reports the gain-based feature importance ranking, based exclusively on contextual variables. Among these, physician experience and job tenure emerge as the two most important predictors of efficiency classification, consistent with their central role in our theoretical framework. Operational factors including average door-to-MD time and ESI standard deviation rank third and fourth in importance, followed by patient severity composition—particularly the proportions of ESI level 2 and ESI level 4 patients. Shift familiarity contributes a smaller but meaningful signal, consistent with its role as a moderating variable in our second-stage analysis. These results indicate that the DEA efficiency classifications capture meaningful variation in physician performance that is independently detectable from observable contextual characteristics, lending robustness to our interpretation.

7 Managerial Implications

Our findings provide actionable guidance for ED managers to improve physician efficiency through scheduling, staffing, and resource allocation. We translate these insights into operational changes and use simulation-based analysis to estimate expected efficiency gains. Simulations are based on Tobit coefficient estimates, with other covariates held at their sample means, and quantify the impact of each change on predicted ED efficiency scores.

7.1 Operational Change 1: Targeting Shift Familiarity for Early-Career Physicians

Our findings reveal that the impact of shift familiarity on efficiency varies with physician experience. To explore the managerial implications of this heterogeneity, we simulate the effect of increasing shift familiarity from the 25th to the 75th percentile for physicians at different points in the experience distribution. We use the 10th (13 years) and 90th (32 years) percentiles of the experience distribution to represent less and more experienced physicians, respectively, providing a data-driven contrast across career stages.

The results, summarized in Table 6, show that for a physician with 13 years of experience, increasing familiarity from the 25th to the 75th percentile yields an efficiency gain of 0.0149, corresponding to approximately a 2.68% reduction in required inputs. In operational terms, this translates to approximately 3.88 minutes of physician time (as measured by contact-to-disposition time) saved per patient and 0.03 fewer imaging tests per patient. In contrast, for a physician with 32 years of experience, the same increase in familiarity yields a marginal efficiency reduction of 0.0069, corresponding to a 1.24% increase in required inputs. This suggests that at high levels of experience, familiar shift assignments may offer no efficiency benefit and could slightly constrain performance.

While the implied changes are small at the individual patient level, aggregating across a physician's weekly caseload yields more meaningful operational estimates. For a physician with 13 years of experience, increasing shift familiarity from the 25th to the 75th percentile translates to approximately 3.5 physician hours saved and 1.5 fewer imaging tests per week. For a physician with 32 years of experience, the same change corresponds to 1.6 additional hours and 0.7 additional imaging tests per week. These results suggest that familiarity improves efficiency primarily through reductions in physician time and improved coordination for less experienced physicians, while highly experienced physicians maintain high performance regardless of shift familiarity and may be better utilized in more variable or unfamiliar shift conditions where their adaptability can be fully leveraged. Together, these findings support a differentiated staffing strategy: ED managers should prioritize scheduling consistency for less experienced physicians, where the efficiency gains from familiarity are greatest, while allowing greater flexibility in scheduling more experienced physicians. This includes allocating more experienced physicians to roles or shifts that require adaptability—such as periods of high demand or less predictable scheduling environments—without compromising efficiency. Although we do not simulate specific deployment patterns, the observed interaction between familiarity and experience provides an empirical basis for these recommendations. Pairing less experienced physicians with more experienced colleagues may offer additional benefits by supporting learning and leveraging experience-based adaptability in dynamic clinical settings.

Table 6 Summary of Simulated Interventions and Input Savings

Intervention Scenario	Efficiency Gain ($\Delta\theta$)	Post-Intervention Efficiency	Input Reduction
Targeted Scheduling by Experience			
Increase shift familiarity from 25th to 75th percentile (13 years experience)	0.0149	0.5701	2.68%
Increase shift familiarity from 25th to 75th percentile (32 years experience)	−0.0069	0.5483	−1.24%
Severity-Based Physician Deployment			
Increase proportion of ESI-2 patients from 25th to 75th percentile (13 years experience)	−0.0270	0.5282	−4.86%
Increase proportion of ESI-2 patients from 25th to 75th percentile (32 years experience)	0.0019	0.5571	0.33%
Proactive Resource Mitigation			
Recover 50% of efficiency loss from increased severity	0.0073	0.5625	1.31%

Note: Simulations assume an input-oriented DEA model with a baseline efficiency score of 0.5552. Input changes reflect the percentage change in physician resource use (e.g., time and diagnostics) required to deliver the same output under the simulated conditions relative to the baseline.

7.2 Operational Change 2: Staging Clinical Experience Through Severity-Based Learning Pathways

In addition to optimizing immediate physician efficiency, our findings provide actionable guidance for structuring both staffing and physician development. The statistically significant positive interaction between physician experience and proportion of ESI-2 patients indicates that experienced physicians experience smaller efficiency losses—and slight gains—when treating high-severity patients, while the marginal efficiency benefits of their experience diminish as patient severity improves. This suggests a dual benefit: high-severity patients should be prioritized for more experienced physicians, where their expertise yields the greatest returns, while lower-severity patients are well-suited for early-career physicians, enabling them to perform efficiently and gain experience in lower-complexity scenarios. To assess the magnitude of this moderating effect, we simulate a shift toward higher-severity patients by increasing the proportion of ESI-2 cases from the 25th percentile (0.104) to the 75th percentile (0.177), based on observed variation in the data. To illustrate the full range of the moderating effect of experience on patient severity, we use the 10th percentile (13 years) and 90th percentile (32 years) of the experience distribution, representing less and more experienced physicians at the boundaries of the observed range. While a physician at the sample mean (21.19 years of experience) faces a moderate efficiency loss from such a shift, our results show significant heterogeneity at the extremes. Based on our Tobit estimates, this change is associated with an efficiency loss of 0.0270 for a physician with 13 years of experience, corresponding to a 4.86% increase in required inputs. In operational terms, this translates to approximately 7.05 additional minutes of physician time per patient, or 6.4 additional hours per physician per week. In contrast, for a physician with 32 years of experience, the same increase in patient severity yields a negligible efficiency change of 0.0019, corresponding to a 0.33% reduction in inputs—indicating that highly experienced physicians absorb increases in patient severity without meaningful efficiency losses. These results reinforce the value of assigning lower-severity patients to less experienced physicians, for whom higher-severity cases are associated with efficiency losses.

Conversely, assigning high-severity cases to more experienced physicians allows these patients to be managed more efficiently, reflecting the greater ability of experienced physicians to handle clinical complexity. This finding supports a staged, severity-based learning pathway. By initially assigning junior physicians to lower-severity patients, EDs can create a stable environment that supports learning and builds efficiency. As physicians gain experience and familiarity with clinical operations, they can be gradually transitioned to higher-severity assignments—not only to support clinical development but also to maintain operational efficiency in more complex care settings. To support this transition, ED managers may consider severity-based mentorship pairings, where early-career physicians shadow or co-manage patients with more experienced colleagues. This approach facilitates skill development in progressively complex clinical scenarios while maintaining operational efficiency. Over time, such an approach can promote sustainable performance growth and strengthen workforce resilience.

7.3 Operational Change 3: Proactive Resourcing in High-Severity Periods

Beyond its interaction with physician experience, our findings highlight the operational importance of patient severity as a direct driver of efficiency. During high-severity periods, physicians face greater clinical complexity, which can reduce efficiency. By anticipating these periods and adjusting staffing, diagnostic capacity, and ancillary support, ED managers can sustain performance and optimize resource use. Although EDs cannot control incoming patient severity, many experience predictable patterns by time of day, week, or season, making severity-responsive resourcing both necessary and feasible.

To explore the potential value of proactive resourcing strategies, we simulate a severity surge scenario based on observed variation in the data. Specifically, we simulate a shift toward higher-severity patients by increasing the proportion of ESI-2 cases from the 25th percentile (0.104) to the 75th percentile (0.177). Recovering 50% of the associated efficiency loss through proactive resource adjustments yields an estimated efficiency gain of 0.0073 points. This corresponds to a 1.31% reduction in input use under an input-oriented DEA framework, equivalent to approximately 1.90 minutes of physician time and 0.01 imaging tests per patient, or approximately 1.7 hours and 0.7 fewer imaging tests per physician per week. While these per-physician gains may appear modest, they compound over time and across shifts. In high-volume environments where staffing and capacity are already stretched, such gains can meaningfully reduce pressure on frontline teams. Looking ahead, EDs may benefit from developing or adopting severity forecasting tools that use historical ESI patterns to predict high-severity periods. These tools can be linked to operational triggers, such as pre-specified staffing adjustments, flexible support pool deployment, or triage workflow modifications, allowing departments to act in advance of anticipated demand spikes. Beyond improving efficiency, such anticipatory resourcing also strengthens system resilience by enabling EDs to absorb short-term severity surges without compromising care quality or patient throughput. These findings underscore the value of incorporating patient severity into real-time and forward-looking ED planning.

7.4 Operational Change 4: Aligning Resource Use with Efficiency Benchmarks

Frontier physicians according to our DEA model—those operating at maximum observed efficiency—demonstrated consistently lower diagnostic test usage and shorter contact-to-disposition times. This indicates that resource efficiency and output quality are not necessarily in conflict. Sharing anonymized efficiency benchmarks, such as test-order rates and throughput times, may help foster peer learning and reduce practice variation. Monthly performance dashboards can serve as a low-cost mechanism to disseminate these insights and encourage standardization of best practices. In parallel, EDs can adopt data-driven scheduling systems that incorporate familiarity and performance metrics, allowing managers to continuously monitor, adjust, and optimize physician assignments. These tools also support scenario testing, enabling the simulation of projected input savings under alternative scheduling structures.

8 Conclusions

Using evidence from emergency medicine, we develop and analyze an efficiency metric for ED physicians and use it to identify factors shaping physician efficiency. Our findings show that physician experience mitigates the efficiency losses associated with high-severity settings where tacit knowledge is critical, but its impact is less pronounced in low-severity cases governed by routine protocols. The benefits of experience also attenuate with repeated shift familiarity, indicating diminishing marginal returns to familiarity for more experienced physicians. Together, these results highlight the importance of strategic staffing and resource allocation in optimizing ED operations.

Our analysis contributes to the ongoing discourse on improving the efficiency of EDs in meeting the needs of diverse patient populations. While the metric we propose is not the only way to evaluate physician efficiency,[2] it provides a structured, data-driven approach to examining performance in high-pressure care settings. We do not view our efficiency scores as a definitive measure or ranking of physicians, but rather as a tool to explore how efficiency varies across individuals and contexts. Our analysis yields actionable insights for clinical administration: implementing stable scheduling to increase familiarity for early-career physicians, and assigning senior physicians to high-severity or high-variability shifts. Simulation results suggest these strategies generate substantial efficiency gains and input savings without additional staffing. By translating these improvements into measurable resource reductions, we provide guidance for optimizing ED performance, reducing provider strain, and enhancing care delivery under real-world constraints. This research also serves as a valuable resource for future studies aiming to further explore and address the multifaceted challenges faced by EDs.

It is important to note the limitations of our study. First, although we control for observed determinants of physician performance and conduct robustness checks, unobserved factors may still influence efficiency.

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

Second, our analysis relies solely on quantitative data; incorporating qualitative insights in future work could enrich interpretation and applicability. Third, our study is based on a single ED, which may limit generalizability; extending the analysis across multiple EDs would strengthen external validity. We also acknowledge implementation challenges. Increasing scheduling stability for early-career physicians may reduce flexibility and affect work–life balance, while limiting their exposure to high-severity cases could hinder long-term skill development. Increasing schedule variability for senior physicians may be difficult in practice, as many EDs allow reduced or no overnight shifts due to the challenges of day–night transitions with age. In addition, desirable shifts (e.g., 7:00 AM–3:30 PM) may raise fairness concerns under stable scheduling policies. Addressing these trade-offs requires careful scheduling design and physician input to balance efficiency, fairness, sustainability, and development. Similarly, assigning more experienced physicians to higher-severity or less predictable shifts should be approached as an optimization problem, potentially requiring incentives (e.g., reduced hours or increased compensation) to ensure feasibility.

Finally, while this study identifies key determinants of ED physician efficiency, the underlying mechanisms and causal pathways remain an important area for continued investigation. Future research could build on these findings by designing and evaluating interventions that leverage these drivers while accounting for their implications for physician learning, supporting more effective long-term staffing and resource allocation in acute care settings.

References

Abualenain, J., W. J. Frohna, M. Smith, M. Pipkin, C. Webb, D. Milzman, J. M. Pines. 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), 281-288.

Altonji, Joseph G., Todd E. Elder, Christopher R. Taber. 2005. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113 (1), 151-184.

Asplin, B. R., D. J. Magid, K. V. Rhodes, L. I. Solberg, N. Lurie, C. A. Camargo. 2003. A conceptual model of emergency department crowding. *Annals of Emergency Medicine*, 42 (2), 173-180.

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

Bovier, P. A., T. V. Perneger. 2007. Stress from uncertainty from graduation to retirement: a population-based study of swiss physicians. *Journal of General Internal Medicine*, 22 632-638.

Bowlin, W. F. 1998. Measuring performance: An introduction to data envelopment analysis (dea). *J Cost Anal*, 15 (2), 3-27.

Bray, A. 2024. Expertise or expert teams: Team familiarity, expertise, and performance in the emergency department. *Academy of Management Proceedings*, 1.

Bray, A., R. B. Sangal, M. D. King. 2026. Teams in crisis: The effect of team familiarity on performance under conditions of crisis and uncertainty. *Organization Science*, 37 (2), 664-684.

Burns, Lawton R., Douglas R. Wholey. 1991. The effects of patient, hospital, and physician characteristics on length of stay and mortality. *Medical Care*, 29 251-271.

Canwell, B. M., A. J. Ramirez. 1997. Doctor-patient communication: a study of junior house officers. *Medical Education*, 31 17-21.

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

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

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

Chisholm, C. D., E. K. Collison, D. R. Nelson, W. H. Cordell. 2000. Emergency department workplace interruptions: are emergency physicians "interrupt-driven" and "multitasking"? *Academic Emergency Medicine*, 7 (11), 1239-1243.

Choudhry, Niteesh K., Robert H. Fletcher, Stephen B. Soumerai. 2005. Systemic review: the relationship between clinical experience and quality of health care. *Annals of Internal Medicine*, 142 260-273.

Collier, D. A., C. E. Collier, T. M. Kelly. 2006. Benchmarking physician performance, part 1. *Journal of Medical Practice Management*, 21 (4), 185-189.

Cooper, W. W., L. M. Seiford, K. Tone. 2007. *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*. 2nd ed. Springer, New York.

Croskerry, P. 2002. Achieving quality in clinical decision making: Cognitive strategies and detection of bias. *Academic Emergency Medicine*, 9 (11), 1184-1204.

Dane, E. 2010. Reconsidering the trade-off between expertise and flexibility: A cognitive entrenchment perspective. *Academy of Management Review*, 35 (4), 579-603.

Elstad, E. A., K. E. Lutfey, L. D. Marceau, S. M. Campbell, O. von dem Knesebeck, J. B. McKinlay. 2010. What do physicians gain (and lose) with experience? qualitative results from a cross-national study of diabetes. *Social Science & Medicine*, 70 (11), 1728-1736.

Fernandes, C. M. B., A. Price, J. M. Christenson. 1997. Does reduced length of stay decrease the number of emergency department patients who leave without seeing a physician? *Journal of Emergency Medicine*, 15 (3), 397-399.

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

Ghaferi, A. A., J. D. Birkmeyer, J. B. Dimick. 2009. Variation in hospital mortality associated with inpatient surgery. *New England Journal of Medicine*, 361 (14), 1368-1375.

Glickman, S. W., K. A. Schulman, E. D. Peterson, M. B. Hocker, C. B. Cairns. 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, J. A., L. C. An, R. A. Hayward, B. C. Williams. 1998. Initial emergency department diagnosis and return visits: risk versus perception. *Annals of Emergency Medicine*, 32 (5), 569-573.

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

Grosskopf, S., V. Valdmanis. 1987. Measuring hospital performance. a non-parametric approach. *Journal of Health Economics*, 6 (2), 89-107.

Haas, J., E. J. Orav, L. Goldman. 1995. The relationship between physicians' qualifications and experience and the adequacy of prenatal care and low birthweight. *American Journal of Public Health*, 85 1087-1091.

Harrison, J. P., R. J. Ogniewski. 2005. An efficiency analysis of veterans health administration hospitals. *Military Medicine*, 170 (7), 607-611.

Hess, B. J., W. Weng, L. A. Lynn, E. S. Holmboe, R. S. Lipner. 2011. Setting a fair performance standard for physicians' quality of patient care. *Journal of General Internal Medicine*, 26 (5), 467-473.

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

Horwitz, L. I., M. L. Green, E. H. Bradley. 2009. Us emergency department performance on wait time and length of visit. *Annals of Emergency Medicine*, 55 (2), 133-141.

Iyasere, C. A., J. Wing, J. N. Martel, M. G. Healy, Y. S. Park, K. M. Finn. 2022. Effect of increased interprofessional familiarity on team performance, communication, and psychological safety on inpatient medical teams: A randomized clinical trial. *JAMA Internal Medicine*, 182 (11), 1190-1198.

Jacques, C. H., R. L. Jones, P. Houts, L. C. Bauer, K. M. Dwyer, J. C. Lynch. 1991. Reported practice behaviors for medical care of patients with diabetes mellitus by primary-care physicians in pennsylvania. *Diabetes Care*, 14 (8), 712-717.

Jameson, J., S. Saghafian, R. S. Huckman, N. Hodgson, J. Baugh. 2025a. The impact of batching advanced imaging tests in emergency departments, Working paper, Harvard University. Available at SSRN: `https://ssrn.com/abstract=5193207`.

Jameson, J. C., S. Saghafian, R. S. Huckman, N. Hodgson. 2025b. Variation in batch ordering of imaging tests in the emergency department and the impact on care delivery. *Health Services Research*, 60 (1), e14406.

Johannessen, K. A., S. A. C. Kittelsen, T. P. Hagen. 2017. Assessing physician productivity following norwegian hospital reform: A panel and data envelopment analysis. *Social Science & Medicine*, 175 117-126.

Kawamoto, E., A. Ito-Masui, R. Esumi, H. Imai, M. Shimaoka. 2020. How icu patient severity affects communicative interactions between healthcare professionals: A study utilizing wearable sociometric badges. *Frontiers in Medicine*, 7 606987-606987.

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

Keith, K. D., J. J. Bocka, M. S. Kobernick, R. L. Krome, M. A. Ross. 1989. Emergency department revisits. *Annals of Emergency Medicine*, 18 (9), 964-968.

Kenny, S. J., P. J. Smith, M. G. Goldschmid, J. M. Newman, W. H. Herman. 1993. Survey of physician practice behaviors related to diabetes mellitus in the u.s. physician adherence to consensus recommendations. *Diabetes Care*, 16 (11), 1507-1510.

Klasco, R. S., R. E. Wolfe, M. Wong. 2015. Assessing the rates of error and adverse events in the ed. *The American Journal of Emergency Medicine*, 33 1786-1789.

Klein, G. A. 1999. *Sources of Power: How People Make Decisions*. MIT Press, Cambridge, MA.

Kneip, A., L. Simar, P. W. Wilson. 2016. Testing hypotheses in nonparametric models of production. *J Bus Econ Stat*, 34 (3), 435-456.

Kohn, L. T., J. M. Corrigan, M. S. Donaldson. 2000. *To Err Is Human: Building a Safer Health System*. National Academy Press.

Kuntz, L., R. Mennicken, S. Scholtes. 2015. Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science*, 61 (4), 754-771.

Leigh, J. P., R. L. Kravitz. 2001. Physician career satisfaction across specialties. *Archives of Internal Medicine*, 161 (13), 1577-1584.

Li, C. J., Y. J. Syue, T. C. Tsai, K. H. Wu, C. H. Lee, Y. R. Lin. 2016. The impact of emergency physician seniority on clinical efficiency, emergency department resource use, patient outcomes, and disposition accuracy. *Medicine*, 95 (6), e2706.

Linzer, M., M. Gerrity, J. A. Douglas, J. E. McMurray, E. S. Williams, T. R. Konrad. 2002. Physician stress: results from the physician worklife study. *Stress and Health: Journal of the International Society for the Investigation of Stress*, 18 (1), 37-42.

Liu, L., X. Zhang. 2019. Analysis of financing efficiency of chinese agricultural listed companies based on machine learning. *Complexity*, 66 (6), 11-17.

Maruthappu, M., A. Duclos, C. D. Zhou, S. R. Lipsitz, J. Wright, D. Orgill, M. J. Carty. 2016. The impact of team familiarity and surgical experience on operative efficiency: A retrospective analysis. *Journal of the Royal Society of Medicine*, 109 (4), 147-153.

Mason, S., E. Knowles, A. Boyle. 2002. Exit block in emergency departments: a rapid evidence review. *Emergency Medicine Journal*, 29 (12), 863-866.

McDonald, N., P. S. Antkowiak, R. Burke, D. T. Chiu, B. A. Stenson, L. D. Sanchez. 2024. Emergency physician resource utilization varies by years of experience. *Journal of the American College of Emergency Physicians Open*, 5 (2), e13162.

Mehrotra, A., R. O. Reid, J. L. Adams, M. W. Friedberg, E. A. McGlynn, P. S. Hussey. 2012. Physicians with the least experience have higher cost profiles than do physicians with the most experience. *Health Affairs (Millwood)*, 31 (11), 2453-2463.

Misiunas, N., A. Oztekin, Y. Chen, K. Chandra. 2016. Deann: A healthcare analytic methodology of data envelopment analysis and artificial neural networks for the prediction of organ recipient functional status. *Omega*, 58 46-54.

Niewoehner, R. J., D. KC, B. Staats. 2023. Physician discretion and patient pick-up: How familiarity encourages multitasking in the emergency department. *Operations Research*, 71 (3), 958-978.

Norcini, John J., Harry R. Kimball, Rebecca S. Lipner. 2000. Certification and specialization: do they matter in the outcome of acute myocardial infarction? *Academic Medicine*, 75 1193-1198.

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

Ozcan, Y. A., J. W. Begun, M. M. McKinney. 1999. Benchmarking organ procurement organizations: A national study. *Health Services Research*, 34 (4), 855-874.

Ozcan, Y. A., C. W. Pai, H. J. 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.

Panagioti, M., K. Geraghty, J. Johnson, A. Zhou, E. Panagopoulou, C. Chew-Graham, D. Peters, A. Hodkinson, R. Riley, A. Esmail. 2018. Association between physician burnout and patient safety, professionalism, and patient satisfaction: A systematic review and meta-analysis. *JAMA Internal Medicine*, 178 (10), 1317-1331.

Peterson, E. D., L. P. Coombs, E. R. DeLong, et al. 2008. Procedural volume as a marker of quality for cabg surgery. *Journal of the American Medical Association*, 295 (21), 2511-2517.

Pham, J. C., T. D. Kirsch, P. M. Hill, K. DeRuggerio, B. Hoffmann. 2011. Seventy two-hour returns may not be a good indicator of safety in the emergency department: a national study. *Academic Emergency Medicine*, 18 390-397.

Pines, J. M., C. V. Pollack, D. B. Diercks, A. M. Chang, F. S. Shofer, J. E. Hollander. 2012. The association between emergency department crowding and adverse cardiovascular outcomes in patients with chest pain. *Academic Emergency Medicine*, 16 (7), 617-625.

Pronovost, P. J., D. C. Angus, T. Dorman, K. A. Robinson, T. T. Dremsizov, T. L. Young. 2002. Physician staffing patterns and clinical outcomes in critically ill patients. *JAMA*, 288 (17), 2151-2162.

Querin, L. B., A. McGary, N. R. Hodgson, R. Allen, W. A. Martini. 2025. Association between physician age and imaging utilization in emergency department abdominal pain evaluation. *Journal of Emergency Medicine*, 72 9-16.

Reid, Rachel O., Mark W. Friedberg, John L. Adams, Elizabeth A. McGlynn, Ateev Mehrotra. 2010. Associations between physician characteristics and quality of care. *Archives of Internal Medicine*, 170 (16), 1442-1449.

Sacchetti, A., C. Carraccio, R. H. Harris. 1992. Resident management of emergency department patients: Is closer attending supervision needed. *Annals of Emergency Medicine*, 21 749-752.

Saghafian, S., W. J. Hopp. 2019. The role of quality transparency in health care: Challenges and potential solutions. *National Academy of Medicine (NAM) Perspectives*, 10-31478.

Saghafian, S., W. J. Hopp. 2020. Can public reporting cure healthcare? the role of quality transparency in improving patient-provider alignment. *Operations Research*, 68 (1), 71-92.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research*, 60 (5), 1080-1097.

Sherman, H. D. 1984. Hospital efficiency measurement and evaluation. empirical test of a new technique. *Medical Care*, 22 (10), 922-938.

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

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

Simar, L., P. W. Wilson. 2007. Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, 136 (1), 31-64.

Smith, C. A., A. B. Varkey, A. T. Evans, B. M. Reilly. 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.

Soltani, M., R. J. Batt, H. Bavafa, B. W. Patterson. 2022. Does what happens in the ed stay in the ed? the effects of emergency department physician workload on post-ed care use. *Manufacturing & Service Operations Management*, 24 (6), 2797-3306.

Spaite, D., F. Bartholomeaux, J. Guisto, E. Lindberg, B. Hull, A. Eyherabide, S. Lanyon, E. A. Criss, T. D. Valenzuela, C. Conroy. 2002. Rapid process redesign in a university-based emergency department: decreased waiting time intervals and improving patient satisfaction. *Annals of Emergency Medicine*, 39 (2), 168-177.

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

Traub, S. J., A. C. Bartley, V. D. Smith, R. Didehban, C. A. Lipinski, S. Saghafian. 2016a. Physician in triage versus rotational patient assignment. *The Journal of Emergency Medicine*, 50 (5), 784-790.

Traub, S. J., C. F. Stewart, R. Didehban, A. C. Bartley, S. Saghafian, V. D. Smith, S. M. Silvers, R. LeCheminant, C. A. Lipinski. 2016b. Emergency department rotational patient assignment. *Annals of Emergency Medicine*, 67 (2), 206-215.

Traub, S. J., M. Temkit, S. Saghafian. 2017. Emergency department holding orders. *Journal of Emergency Medicine*, 52 (6), 885-893.

Tsugawa, Y., J. P. Newhouse, A. M. Zaslavsky, D. M. Blumenthal, A. B. Jena. 2017. Physician age and outcomes in elderly patients in hospital in the us: Observational study. *BMJ*, 357 j1797.

Varabyova, Y., J. Schreyogg. 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., L. D. Fredendall, K. M. Taaffe, N. Huynh, G. Ritchie. 2018. An empirical examination of surgeon experience, surgeon rating, and costs in perioperative services. *Journal of Operations Management*, 61 (1), 68-81.

Wagner, J. M., D. G. Shimshak, M. A. Novak. 2003. Advances in physician profiling: the use of dea. *Socio-Economic Planning Sciences*, 37 (2), 141-163.

Zheng, W., H. Sun, P. Zhang, G. Zhou, Q. Jin, X. Lu. 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.

Zhu, N., C. Zhu, A. Emrouznejad. 2021. A combined machine learning algorithms and dea method for measuring and predicting the efficiency of chinese manufacturing listed companies. *Journal of Management Science and Engineering*, 6 (4), 435-448.