Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Forecast-Driven Dynamic Physician Staffing in a Pediatric Emergency Department: A Prospective Quasi-Experimental Pilot Study

Ahmet Ziya Birbilen
Hacettepe University
Izzet Turkalp Akbasli
iakbasli@hacettepe.edu.tr

Hacettepe University
Ozlem Teksam
Hacettepe University

Article

Keywords: Emergency Service, Hospital, Pediatrics, Artificial Intelligence, Machine Learning, Personnel Staffing and Scheduling, Deep Learning, Operations Research

Posted Date: April 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9334490/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Emergency department crowding is a persistent threat to acute care quality, yet predictive models for ED demand have rarely been translated into prospective operational staffing interventions. Here we report a prospective single-center quasi-experimental pilot study evaluating forecast-driven dynamic physician scheduling in the pediatric ED of Hacettepe University Ihsan Dogramaci Children's Hospital (December 2024 to May 2025). Using a deep learning demand forecasting model (TiDE-RIN) combined with linear programming, we determined daily physician counts (range 3 to 6) for evening shifts (16:00 to 24:00) during days 1 to 15 of each month; days 16 to end of month maintained the institution's standard fixed four-physician schedule. Among 13,935 visits (6,957 intervention; 6,978 concurrent controls), propensity score matching (6,949 pairs) estimated a boarding time reduction of 31.9 minutes (21.4%; $p < 0.0001$); interrupted time series analysis confirmed an immediate level change of 22.1 minutes at intervention onset ($p = 0.010$). Diagnostic testing per patient decreased by 8.4% ($p = 0.022$) and physician-level boarding burden declined by 12.0% ($p = 0.013$), with larger effects among lower-acuity patients and during the early-evening demand peak. Spillover analysis confirmed no progressive improvement in the concurrent control period. These findings demonstrate that coupling deep learning demand forecasting with optimization-based physician scheduling can produce measurable and multi-dimensional improvements in pediatric ED operations, directly addressing the translational gap between predictive model development and real-world clinical implementation.

Introduction

Emergency department crowding is one of the most consistently documented threats to acute care quality, independently associated with treatment delays, prolonged length of stay, and adverse clinical outcomes including delays in definitive care[1]. Despite sustained quality improvement efforts, crowding remains prevalent across health systems worldwide and continues to strain emergency care delivery, including pediatric emergency departments during recurring viral illness surges and seasonal demand peaks[2,3].

Health systems have pursued multiple strategies to mitigate crowding, including downstream capacity expansion and front-end process redesign, yet these approaches often struggle to keep pace with rapid pediatric demand surges and persistent access block[4]. Numerous predictive models have been developed to forecast ED length of stay and demand patterns; however, systematic reviews demonstrate substantial methodological heterogeneity and limited external validation among these models[5,6]. Patient arrivals and ED demand exhibit predictable temporal patterns amenable to forecasting[7], but the translation of these predictions into operational staffing interventions remains limited[5,8]. Pediatric EDs face particular operational vulnerability: pronounced seasonal surges driven by respiratory viral illness, a demand profile skewed toward lower-acuity patients, and constrained physician pools create demand-capacity mismatches that fixed staffing schedules are structurally unable to resolve[9].

Prior investigations have demonstrated that ED demand, prolonged waiting times, and hospitalization risk can be forecast using predictive modelling approaches[10,11]. However, most such studies have focused on model performance rather than the prospective translation of predictions into staffing decisions within routine clinical practice. Recent systematic reviews confirm that the overwhelming majority of published AI studies in ED operations remain confined to predictive modelling or simulation environments, with very few demonstrating real-world implementation of AI-driven operational interventions[12]. Broad reviews of AI in healthcare similarly emphasize persistent challenges in translating predictive models into measurable clinical impact[13]. In pediatric emergency settings, where demand variability and physician pool constraints are most pronounced, evidence linking predictive models to dynamic staffing adjustments remains absent. Building on a previously published forecasting-optimization framework for this department, this study progresses from model development to prospective real-world operational evaluation. To our knowledge, this is the first study to implement and prospectively evaluate forecast-driven dynamic physician scheduling in a pediatric emergency department. We therefore aimed to determine whether this demand-responsive staffing strategy produced measurable improvements in boarding time, diagnostic test utilization, physician workload, and overall operational performance.

Methods

Study Design and Setting

This study was designed as a prospective, single-center, quasi-experimental pilot study conducted in the Pediatric Emergency Department (PED) of Hacettepe University Ihsan Dogramaci Children's Hospital. The intervention targeted after-hours shifts (16:00–24:00) and spanned the period from December 2024 to May 2025. The study was conducted under a single-blind operational framework in which clinicians were not informed of the study allocation or the underlying staffing strategy during the study period.

An intra-month crossover allocation design was implemented: during days 1–15 of each month, a machine learning-based optimized shift schedule (intervention arm) was used, while from day 16 to the end of the month, the institution's established fixed four-physician shift schedule (concurrent control arm) was applied. To distinguish seasonality and calendar effects, the same months of the prior year (December 2023–May 2024; fixed four-physician schedule throughout) served as a historical control cohort. The Hacettepe University Non-Interventional Clinical Research Ethics Committee (Approval No: SBA 24/1092) approved the study.

Participants

The target population included all PED visits triaged during after-hours shifts (16:00–24:00). Exclusion criteria were defined as follows: (i) visits solely for consultation or record correction purposes; (ii)

records with timestamp errors or missing triage entry times; (iii) inter-facility transfers without triage documentation; and (iv) patients directly admitted bypassing triage.

Intervention: Demand-Responsive Physician Scheduling

The intervention comprised three sequential phases. In the first phase (demand forecasting), hourly PED visit counts were predicted using the TiDE-RIN (Time-series Dense Encoder with Reversible Instance Normalization) model[14], a deep sequence-to-sequence architecture employing dense encoder-decoder layers rather than self-attention to capture long-range temporal dependencies. The detailed technical description of the forecasting architecture, machine learning operations (MLOps) infrastructure, and shift optimization algorithm has been previously published[2]. The model operated at hourly resolution with visit count as the sole target variable, using calendar-derived auxiliary features (hour of day, day of week, week and month indicators) without external regressors. Reversible Instance Normalization was applied to stabilize distributional scale and shift. Uncertainty was quantified via 100 stochastic forward passes with the mean serving as the point estimate. Four-week-ahead forecasts were generated at least five days before each scheduling cycle. In the second phase (shift optimization), the daily physician count for the 16:00 to 24:00 shift was determined by linear programming (PuLP 2.9.0), with a single integer decision variable constrained to the range 3 to 6 (reflecting institutional minimum staffing requirements, maximum available physician pool, and contractual scheduling constraints) and an objective function minimizing demand-capacity mismatch based on shift-level aggregates derived from the TiDE-RIN hourly forecasts.

In the third phase (implementation), optimised schedules were communicated to the departmental administrative unit at least five days before each month and applied without modification during days 1 to 15; during the remainder of the month, the standard fixed four-physician model was maintained. Physicians were not formally informed of allocation status, although complete blinding could not be guaranteed in an operational staffing intervention. The administrative coordinator verified daily concordance between scheduled and actual physician counts. Throughout the six-month study period, 100% schedule adherence was achieved with no shifts overridden or modified. This complete implementation fidelity distinguishes the present study from prior forecast-driven staffing trials in which partial compliance limited the ability to evaluate the full effect of optimised schedules.

Data Sources and Variables

Data were obtained as time-stamped event records from the hospital's electronic health record (EHR) and operational information systems; data were de-identified in accordance with institutional data governance principles prior to analysis.

Timestamps and derived time intervals

Three time points were recorded for each visit: triage entry (initial assessment by the triage nurse), examination completion (completion of the physician's initial evaluation), and discharge or exit from the PED. Examination length of stay (Exam LOS) was defined as triage entry to examination completion. Hospital length of stay (Hospital LOS) was measured from triage entry to discharge. Boarding time was operationally defined as the interval from completion of the physician's initial evaluation to departure from the PED (Hospital LOS minus Exam LOS), encompassing post-evaluation delays related to disposition, diagnostic workup, consultation, and administrative discharge processes[4,15].

Patient-level variables

Triage assessment: Patients were dichotomized according to the five-level Pediatric Canadian Triage and Acuity Scale [16] into high risk (T1–T3: resuscitation, emergent, urgent) and low risk (T4–T5: less urgent, non-urgent). Diagnostic test utilization: (a) total number of tests ordered per patient; (b) radiological investigation rate (at least one plain radiograph, ultrasonography, CT, or MRI); (c) biochemical panel rate (at least one complete blood count, liver function test, renal function test, or urinalysis).

Shift-level variables and operational metrics

For each study day, the number of physicians on shift was recorded. Two workload metrics were calculated: patients per physician (daily patient count divided by physician count) and boarding burden per physician (mean boarding time multiplied by patients per physician, expressed as patient-minutes per shift). Hourly census was defined as the number of patients present in the PED at each clock hour[17]. The cumulative waiting burden, expressed in patient-minutes, was computed as the hourly sum of individual patient waiting times and served as a system-level measure of aggregate congestion.

Outcome Measures

Primary outcome

Boarding time (minutes). This metric was selected as the primary outcome because, in our prior work[2], we demonstrated that the principal operational bottleneck in this PED was the post-examination disposition process rather than the time to initial physician evaluation. Boarding time therefore isolates the component of ED length of stay most directly modifiable by physician staffing interventions. Additional capacity primarily accelerates downstream tasks such as test interpretation, consultation, and discharge documentation rather than altering the speed of initial clinical assessment itself.

Secondary outcomes: (i) diagnostic test utilization (total tests per patient, radiological investigation rate, biochemical panel rate); (ii) examination length of stay (Exam LOS); (iii) hospital length of stay (Hospital LOS); (iv) hourly census; (v) hourly cumulative waiting burden; (vi) boarding burden per physician (boarding time × patient count / physician count, patient-minutes/shift).

Statistics and reproducibility

This study was reported in accordance with the TRIPOD + AI guidelines for prediction model studies involving artificial intelligence[18]. The primary analysis employed propensity score matching (PSM) to estimate the average treatment effect on the treated (ATT) while controlling for patient-level confounders[19]. A logistic regression model was fitted using arrival hour, day of week, calendar month, and triage acuity as covariates. One-to-one nearest-neighbor matching without replacement was performed. Covariate balance was assessed using standardized mean differences (SMD), with a threshold of 0.1 indicating adequate balance. As a complementary approach, interrupted time series (ITS) analysis[20]was conducted using segmented linear regression to estimate both the immediate level change and the post-intervention trend change attributable to the intervention; the model coefficients and their interpretation are presented in Table 5. A two-sided α of 0.05 was used as the threshold for statistical significance for all reported p-values.

Table 5
ITS segmented regression coefficients.

Parameter	Coefficient	SE	P value	Interpretation
Intercept (β0)	133.8	5.4	< 0,001	Baseline boarding level
Time (β1)	−0.07	0.06	0,20	Pre-intervention trend (NS)
Intervention (β2)	−22.11	8.56	0,01	Immediate level change
Time × intervention (β3)	+ 0.14	0.08	0,10	Post-intervention trend

Table 5. Interrupted time series segmented regression coefficients for hourly mean boarding time. The model estimates the pre-intervention temporal trend (β1), the immediate level change at intervention onset (β2), and the post-intervention trend change (β3). A non-significant pre-intervention trend confirms the absence of a pre-existing improvement trajectory.

To explore heterogeneous treatment effects, the study population was stratified by triage acuity (high risk vs low risk) and arrival time window (16:00–19:00 vs 20:00–24:00), with boarding time compared within each stratum using the two-tailed Mann–Whitney U test. The internal validity of the intra-month crossover design was assessed through spillover and contamination analysis: the temporal trajectory of boarding time during the control period was examined using Spearman rank correlation to detect any progressive improvement suggestive of learned behaviors or behavioral contamination from the intervention arm. Physician workload was evaluated using two operational metrics: patients per physician and boarding burden per physician (defined as mean boarding time multiplied by patients per physician, expressed as patient-minutes per shift), with between-arm comparisons performed using the Mann–Whitney U test and robustness assessed through census-quartile stratification and day-level PSM.

Sample size

The full study cohort comprised approximately 7,000 patients per arm (6,957 intervention; 6,978 control). With a boarding time standard deviation of 148–160 minutes, this sample provided > 99% power to detect a 30-minute mean difference at α = 0.05. Propensity score matching retained 6,949 pairs (99.9% of the intervention arm), preserving statistical power while adjusting for measured confounders. For the ITS analysis, 1,456 hourly observations provided sufficient power to detect a 20-minute level change. All analysis code, including the TiDE-RIN forecasting model and the linear programming-based shift optimisation framework, is publicly available at https://github.com/turkalpmd/SOSAFED.

Results

Study Cohort

During the six-month study period (December 2024 to May 2025), 13,935 patients presented to the PED during after-hours shifts and met the inclusion criteria (Table 1). The AI-optimised intervention arm (days 1 to 15 of each month) comprised 6,957 visits over 90 study days; the concurrent control arm (day 16 to end of month, fixed four physicians) comprised 6,978 visits over 92 days. The month-matched historical cohort from the prior year (December 2023 to May 2024) included 16,633 visits over 184 days..

Table 1
Baseline characteristics (after-hours visits, 16:00–24:00).

Characteristics	Intervention (AI)	Control (non-AI)	Historic (2024)
Visit, n	6,957	6,978	16,633
Study day (days)	90	92	184
Daily volume, mean (SD)	77.3 (15.1)	75.8 (12.4)	90.4 (22.7)
Sex			
Male, n (%)	3,731 (53.6)	3,721 (53.3)	8,932 (53.7)
Female, n (%)	3,226 (46.4)	3,257 (46.7)	7,701 (46.3)
Age of year, mean (SD)	6.9 (4.9)	6.9 (4.9)	68 (4.9)
Triage levels			
T1, n (%)	6 (0.1)	8 (0.1)	19 (0.1)
T2, n (%)	31 (0.4)	23 (0.3)	108 (0.6)
T3, n (%)	639 (9.2)	646 (9.3)	1,727 (10.4)
T4, n (%)	3,511 (50.5)	3,395 (48.7)	8,314 (50.0)
T5, n (%)	2,747 (39.5)	2,896 (41.5)	6,423 (38.6)
High risk (T1–T3), n (%)	676 (9.7)	677 (9.7)	1,854 (11.1)
Low risk (T4–T5), n (%)	6,258 (90.0)	6,291 (90.2)	14,737 (88.6)
Monthly admission			
December	995	1,290	3,791
January	1,459	1,315	2,789
February	1,181	1,031	2,168
March	1,179	1,137	2,723
April	1,063	1,112	2,689
May	1,080	1,093	2,472

Table 1. Baseline characteristics of after-hours visits (16:00–24:00) across study arms. The intervention arm comprised days 1–15 of each month (AI-optimized scheduling); the concurrent control arm comprised days 16 to end of month (fixed four-physician staffing). The historical cohort includes the same calendar months from the prior year (December 2023–May 2024) under fixed staffing throughout. All volumes and metrics refer to the 16:00–24:00 shift period. Age is reported in years. Triage levels follow the Pediatric Canadian Triage and Acuity Scale.

Table 2. Monthly forecasting performance of the TiDE-RIN demand prediction model.

Month	MAE	RMSE	SMAPE (%)	R^2
Dec-24	2.18	2.79	47.1	0.62
Jan-25	3.38	4.13	51.4	0.43
Feb-25	2.28	2.9	43.7	0.62
Mar-25	2.4	3.18	39.1	0.68
Apr-25	2.51	3.24	47.6	0.53
May-25	2.32	2.87	43.3	0.61
Overall	2.36	3.04	45.4	0.61

Table 2. Monthly forecasting accuracy of the TiDE-RIN model at hourly resolution. All metrics are computed on prospective out-of-sample predictions for the 16:00-24:00 shift window. MAE: mean absolute error (patients/hour). RMSE: root mean square error (patients/hour). SMAPE: symmetric mean absolute percentage error, used because standard MAPE is undefined for hours with zero observed arrivals. R^2: coefficient of determination. Overall values represent the median across months.
The two 2025 arms were well balanced across all pre-specified baseline characteristics. Sex distribution was nearly equal (male 53.6% vs 53.3%), mean age was identical (6.9 +/- 4.9 years in both arms), triage acuity profile was indistinguishable (high-risk 9.7% vs 9.7%), and mean hourly census was comparable (5.73 +/- 4.23 vs 5.49 +/- 3.61; p = 0.979). Mean daily visit volume was similar between arms (77.3 +/- 15.1 vs 75.8 +/- 12.4; p = 0.430). The historical 2024 cohort had higher daily volumes (90.4 +/- 22.7) and a slightly higher high-risk proportion (11.1%), consistent with secular trends in pediatric ED utilization following the 2022 to 2023 viral illness surge.

Forecasting Model Performance

The TiDE-RIN demand forecasting model was evaluated prospectively across each month of the study period; Table 2 summarizes monthly prediction accuracy at hourly resolution. Overall, the model achieved an RMSE of 3.04 patients/hour and an R2 of 0.61. Performance was lowest in January 2025 (R2 = 0.43, RMSE = 4.13), a period coinciding with post-holiday demand variability and the highest monthly patient volume in the intervention arm. The strongest performance was observed in March 2025 (R2 = 0.68, SMAPE = 39.1%). Standard MAPE was undefined for hours with zero observed arrivals. Symmetric MAPE was used as a robust alternative. Despite month-to-month variation, forecasting

accuracy was sufficient to generate operationally actionable scheduling recommendations across all six months of the study period.

Overview of Outcomes by Study Period

Table 3 presents an overview of all outcomes across the three study periods. Within 2025, the AI-optimised arm showed shorter boarding time (117.4 vs 128.1 min), lower test utilization (0.58 vs 0.63 tests/patient), and lower radiological and biochemical ordering rates compared with the concurrent control, despite managing 9.7% more total cumulative waiting burden (105,063 vs 95,794 patient-minutes). Examination length of stay was nearly unchanged between arms (35.4 vs 34.0 min), confirming that the throughput gains were not achieved at the expense of abbreviated clinical evaluations.

Table 3
Comprehensive outcome comparison.

Outcome	2025 Intervention (AI)	2025 Control (Fixed)	Historical 2024
Volume and operational			
Patients (n)	6,957	6,978	16,633
Shift volume, mean (SD)	77.3 (15.1)	75.8 (12.4)	90.4 (22.7)
Total cumulative waiting (pt-min)	105,063	95,794	229,213
Patient-level durations			
Exam LOS (min), mean (SD)	35.4 (28.9)	34.0 (27.9)	33.0 (36.8)
Hospital LOS (min), mean (SD)	152.6 (152.8)	162.0 (158.4)	136.1 (151.1)
Boarding time (min), mean (SD)	117.4 (148.7)	128.1 (154.8)	103.4 (147.7)
Crowding and test utilization			
Census (patients/hour), mean (SD)	9.7 (4.3)	9.5 (4.2)	11.4 (5.3)
Tests/patient, mean (SD)	0.58 (1.16)	0.63 (1.24)	0.64 (1.27)
Radiological investigation (%)	33.5	35.2	36.0
Biochemical panel (%)	19.8	21.7	22.8

Table 3. Comprehensive outcome comparison across study periods. All values pertain to the after-hours shift (16:00–24:00) and are reported as mean (SD) unless otherwise indicated. Boarding time is operationally defined as Hospital LOS minus Exam LOS. Census represents the mean number of patients simultaneously present in the PED per clock hour. Cumulative waiting burden is expressed in patient-minutes. The historical 2024 cohort reflects the same calendar months under fixed four-physician staffing throughout.

Primary Outcome: Boarding Time

Three complementary analytical approaches were applied to the primary endpoint, each controlling for a different source of bias. Results across all three converged on a consistent direction and magnitude of effect.

Unadjusted comparison

At the hourly level, mean boarding time was 121.6 ± 77.1 min in the intervention arm and 131.3 ± 85.1 min in the control arm (−9.7 min,−7.4%; Mann–Whitney p = 0.035). Hospital length of stay showed a directionally consistent trend (156.2 ± 83.7 vs 164.9 ± 90.8 min; p = 0.092); examination length of stay was preserved (34.7 ± 20.8 vs 33.6 ± 18.7 min; p = 0.935) (Fig. 1).

Propensity score matching (PSM)

To control for patient mix differences, 6,949 patient pairs (13,898 individuals) were matched on arrival hour, day of week, calendar month, and triage acuity. Post-matching, all SMD values were below 0.1 (Table 4) and propensity score distributions showed near-complete overlap (Fig. 2). The ATT was 31.9 minutes (intervention 117.4 +/- 148.7 vs matched control 149.3 +/- 160.2 min; p < 0.0001), corresponding to a 21.4% relative reduction. The PSM-adjusted effect was approximately three times larger than the unadjusted estimate (31.9 vs 9.7 minutes), a divergence consistent with confounding attenuation: the intervention arm managed 1.9% more patients and 9.7% more cumulative waiting burden, which biased the unadjusted comparison toward the null. Full covariate balance (all SMD < 0.1) and near-complete propensity score overlap support adequate model specification; however, residual unmeasured confounding cannot be excluded.

Table 4
Covariate balance before and after PSM.

Covariate	Pre-matching SMD	Post-matching SMD	Balanced?
Arrival hour	−0.005	−0.017	Yes
Day of week	0.057	−0.001	Yes
Calendar month	−0.123	−0.001	Yes
Triage acuity	0.000	0.079	Yes

Table 4. Covariate balance before and after propensity score matching. Patients were matched 1:1 on four covariates using nearest-neighbor matching without replacement (6,949 matched pairs). Standardized mean differences (SMD) below 0.1 indicate adequate balance.

Interrupted time series **analysis**

Segmented regression confirmed that the pre-intervention period was stable (β1 =−0.073 min/day, p = 0.203)—a pre-existing improvement trend was excluded. A statistically significant immediate level change of−22.1 minutes was observed at intervention onset (β2 =−22.105, p = 0.010); the post-

intervention trend change was borderline significant (β3 = + 0.136, p = 0.098)—a slight attenuation trend, although the net benefit was sustained.

Heterogeneous Treatment Effects

Temporal stratification by arrival window revealed a 2.5-fold gradient in effect size (Table 6). The early-evening peak window (16:00 to 19:00), coinciding with the post-work visit surge and the period of greatest demand-capacity mismatch, showed a 16.0-minute reduction in boarding time (p < 0.001). The late-evening window (20:00 to 24:00) showed a smaller but still statistically significant reduction of 6.3 minutes (p = 0.012). Triage acuity stratification showed that the intervention effect was concentrated among lower-acuity patients (T4-T5): boarding time decreased by 11.6 minutes (p < 0.001) in this group, which constitutes approximately 90% of the cohort and is most vulnerable to capacity-driven throughput delays. Among high-acuity patients (T1-T3), who are prioritized and fast-tracked regardless of staffing level, the change was a non-significant 3.0 minutes (p = 0.693). This pattern confirms that the intervention acted through capacity-dependent throughput mechanisms rather than any change in clinical prioritization .

Table 6
Heterogeneous treatment effects by subgroup.

Subgroup	Intervention (min)	Control (min)	Δ (min)	P value
Low risk (T4–T5)	109.9	121.5	−11.6	< 0.001
High risk (T1–T3)	186.8	189.8	−3.0	0.693
16:00–19:00	107.6	123.6	−16.0	< 0.001
20:00–24:00	125.5	131.8	−6.3	0.012

Table 6. Heterogeneous treatment effects by patient subgroup (unadjusted comparisons). Triage risk was dichotomized according to the Pediatric Canadian Triage and Acuity Scale: high risk (T1–T3) and low risk (T4–T5). Temporal stratification reflects the early evening peak demand window (16:00–19:00) and late evening period (20:00–24:00). P values are from Mann–Whitney U tests.

Diagnostic Test Utilization

A consistent reduction in test utilization was observed across all three measured categories (Table 7). Total tests per patient decreased by 8.4% (0.580 vs 0.633; p = 0.022). Radiological investigation rates declined from 35.2% to 33.5% (p = 0.014) and biochemical panel rates decreased from 21.7% to 19.8% (p = 0.005). No clinical protocol changes or test-restriction directives were in effect during the study period, indicating that the reductions reflected changes in physician ordering behaviour rather than externally imposed constraints.

Table 7
Diagnostic test utilization: intervention vs concurrent control.

Metric	Intervention	Control	Relative Δ	P value
Total tests/patient, mean (SD)	0,580 (1,158)	0,633 (1,238)	−8,4%	0,022
Radiological investigation (%)	33,5	35,2	−4,8%	0,014
Biochemical panel (%)	19,8	21,7	−8,8%	0,005

Table 7. Diagnostic test utilization during the intervention and concurrent control periods. Total tests per patient includes all laboratory and imaging orders. Radiological investigation denotes the proportion of patients receiving at least one plain radiograph, ultrasonography, CT, or MRI. Biochemical panel denotes the proportion receiving at least one complete blood count, liver function test, renal function test, or urinalysis. P values are from Mann–Whitney U tests. No clinical protocol changes or test-reduction directives were in effect during the study period.

Physician Workload and Staffing Efficiency

The AI system allocated a mean of 4.31 +/- 0.77 physicians per shift (range: 3 to 6), compared with a fixed 4.00 physicians in the control arm, a 7.8% increase ($p < 0.001$). Despite this modest increase in average staffing, boarding burden per physician decreased by 12.0% (1,787 +/- 729 vs 2,031 +/- 791 patient-minutes; $p = 0.013$) and patients per physician declined from 17.6 to 16.2 (-7.9%; $p = 0.043$) (Table 8, Fig. 3).

Table 8
Physician workload comparison: intervention vs concurrent control.

Metric	Intervention (AI)	Control (Fixed)	Δ	P value
Physicians/shift, mean (SD)	4,31 (0,77)	4,00 (0,00)	+ 7.8%	< 0,001
Patients/physician	16,2 (4,1)	17,6 (3,3)	−7.9%	0,043
Mean boarding time (min)	110,2 (40,8)	115,2 (43,1)	−4.3%	0,159
Boarding burden/physician (pt-min)	1.787 (729)	2.031 (791)	−12.0%	0,013

Table 8. Physician workload comparison between the intervention and concurrent control periods. Boarding burden per physician is defined as mean boarding time multiplied by patients per physician, expressed in patient-minutes per shift. The control arm maintained a fixed four-physician model throughout; the intervention arm allocated 3–6 physicians per shift based on forecasted demand. P values are from Mann–Whitney U tests.

Volume-quartile analysis confirmed that the AI system followed a demand-responsive allocation strategy. On low-volume days (Q1, approximately 60 patients) the system matched the baseline staffing

level (3.96 physicians), avoiding unnecessary resource expenditure. On high-volume days (Q4, approximately 96 patients) it increased to 4.55 physicians, reducing per-physician burden from a projected 24.0 under fixed staffing to 21.6, a 10% reduction.

An exploratory dose-response analysis was conducted at the day level. In a covariate-adjusted regression controlling for daily patient volume, calendar month, and day of week, each additional physician per shift was associated with an approximately 6-minute reduction in per-patient boarding time, providing further mechanistic support for the staffing-throughput relationship.

Internal Validity: Spillover Analysis

A critical assumption of the intra-month crossover design is that the control period is free from spillover effects attributable to the intervention. To test this assumption, the temporal trajectory of boarding time during the control arm was examined using Spearman rank correlation between elapsed time and mean daily boarding time. The correlation was $r = -0.049$ ($p = 0.185$), indicating no progressive improvement over the course of the study period, and ruling out learning-effect contamination or spillover from the intervention arm as an explanation for the observed between-arm differences.

DISCUSSION

This prospective pilot study demonstrates that coupling deep learning demand forecasting with linear programming-based physician scheduling produces clinically meaningful, operationally measurable improvements in boarding time, diagnostic test utilization, and physician workload in a pediatric emergency department, representing to our knowledge the first prospective implementation of forecast-driven physician scheduling in this setting. Three independent analytical approaches converged on a coherent picture: propensity score matching estimated a 31.9-minute reduction in boarding time (21.4%; $p < 0.0001$), interrupted time series analysis confirmed an immediate level change of 22.1 minutes at intervention onset ($p = 0.010$), and the unadjusted comparison yielded a 9.7-minute reduction ($p = 0.035$). The substantially larger PSM-adjusted than unadjusted estimate reflects confounding attenuation, because the intervention arm managed 9.7% greater cumulative waiting burden, which biased the raw comparison toward the null; the matched estimate therefore provides the more appropriate measure of effect. An exploratory dose-response analysis corroborated the causal chain: each additional physician per shift was associated with approximately a 6-minute reduction in per-patient boarding time. Beyond throughput, the intervention was associated with an 8.4% reduction in diagnostic testing per patient ($p = 0.022$) and a 12.0% decrease in boarding burden per physician ($p = 0.013$), indicating that the benefits extended substantially beyond boarding time alone.

The specificity of enhancement lends mechanistic clarity. Boarding time, defined as the post-evaluation interval from completion of initial physician assessment to patient departure, represents the component of ED length of stay most directly sensitive to changes in physician staffing capacity[4,15]. Examination length of stay was virtually unchanged between arms (35.4 vs 34.0 min; $p = 0.935$), while boarding time and total hospital LOS remained highly correlated ($r > 0.95$), confirming that the observed gains reflected

faster post-evaluation throughput rather than abbreviated clinical assessments. This dissociation is consistent with queueing theory. When demand and capacity are better aligned, the nonlinear relationship between server utilization and waiting time predicts disproportionate reductions in queue length, even with only modest increases in average staffing[21]. The AI system's census-adaptive behaviour reinforced this interpretation. On low-volume days staffing closely matched the institutional baseline, while on high-demand days it increased to five or six physicians, yielding disproportionate reductions in per-physician boarding burden consistent with the nonlinear dynamics of congested service systems.

These findings build on two intersecting lines of prior work. Hu et al. first demonstrated that machine learning can accurately forecast shift-level nurse staffing requirements in the ED[22] and subsequently embedded those forecasts in a two-stage optimization framework that reduced hourly nursing costs without deteriorating patient flow[13]. Our study advances this physician-scheduling analogue in three key respects. First, physician headcount more directly governs throughput capacity than nurse-patient ratios. Second, the pediatric ED presents a distinct demand profile, characterised by pronounced evening surges, recurring viral illness peaks, and a patient population whose triage and treatment requirements differ systematically from adult settings. Third, the intervention achieved 100% schedule adherence over six months, eliminating the compliance-related dilution of effect that the nurse staffing literature identified as a primary constraint on demonstrable impact[13]. In a second line of relevant work, Poursoltan et al. prospectively compared forecasting architectures for ED boarding volume prediction and found that a VAR-XGBoost hybrid reduced forecast error by up to 41% relative to a moving-average baseline, the highest accuracy reported for this task; the authors explicitly identified the translation of such forecasts into optimization frameworks for staffing and capacity decisions as the logical next step. The present study delivers precisely that translation, coupling hourly deep learning demand prediction with linear programming-based physician scheduling and prospectively measuring the resulting clinical impact[17]. While Hodgson et al. demonstrated that AI-guided individual patient routing can accelerate ED flow by targeting patient-level decisions, our system operates at the level of population-wide capacity scheduling, illustrating that AI interventions targeting different nodes of the ED system represent complementary rather than competing strategies[23].

Subgroup analyses produced an effect pattern that is mechanistically coherent with the operational logic of triage-based emergency care. Lower-acuity patients (T4-T5), who represent approximately 90% of the cohort and whose throughput is most vulnerable to capacity-driven bottlenecks, showed a boarding time reduction of 11.6 minutes ($p < 0.001$). High-acuity patients (T1-T3), who are prioritized and fast-tracked regardless of staffing level, showed a non-significant change of 3.0 minutes ($p = 0.693$)[24,25]. This triage-stratified pattern confirms that the intervention acted through capacity-dependent throughput mechanisms rather than any change in clinical prioritization. Temporal stratification reinforced this interpretation: the early-evening peak window (16:00–19:00) showed the largest effect at 16.0 minutes ($p < 0.001$), while the late-evening period (20:00–24:00) showed a smaller but still significant reduction of 6.3 minutes ($p = 0.012$). This 2.5-fold gradient is consistent with the principle that demand-responsive

staffing generates its greatest benefit precisely during the hours when demand-capacity mismatch is most acute[21].

The clinical significance of these findings extends beyond operational efficiency metrics. Prolonged ED boarding is independently associated with increased adverse events, delayed time-sensitive treatments, and higher rates of patients leaving without being seen[1,2,4,26].

The concurrent reduction in diagnostic test utilization, observed across all three measured modalities and in the absence of any protocol-level restriction policy, warrants specific attention. Operations research evidence indicates that higher physician workload in the ED is associated with increased test ordering, particularly for lower-acuity patients: under time pressure and high cognitive load, risk-aversion drives broader, less selective test panels as a hedge against diagnostic uncertainty, and reducing per-physician workload is expected to reverse this pattern[27]. The observed reductions in radiological investigation rates (35.2% to 33.5%; $p = 0.014$) and biochemical panel rates (21.7% to 19.8%; $p = 0.005$), concentrated among lower-acuity patients who constitute the majority of the cohort, are consistent with this mechanism. Their coexistence with shorter boarding times argues against the interpretation that tests were inappropriately omitted rather than judiciously selected on the basis of more careful pre-test probability assessment.

Systemwide, the overall operational gain is substantial. Applying the conservative unadjusted estimate of 9.7 minutes per patient to a mean daily volume of 77 patients yields approximately 750 patient-minutes saved per day; at a mean boarding time of 120 minutes this is equivalent to roughly one additional patient cleared per 8-hour shift, a capacity gain achieved without physical expansion, capital investment, or any increase in total physician headcount[28]. The demand-adaptive character of the scheduling produced a pattern that fixed staffing cannot replicate. On low-volume days (bottom census quartile, approximately 60 patients) the system matched the institutional baseline, avoiding unnecessary expenditure. On high-volume days (top quartile, approximately 96 patients) it increased to an average of 4.55 physicians, cutting per-physician patient load from a projected 24.0 under fixed staffing to 21.6. This graduated response illustrates the core operational argument for forecast-driven scheduling: the marginal benefit of an additional physician is greatest when the queue is longest, yielding disproportionate throughput gains consistent with the nonlinear properties of congested service systems[21].

Several design features lend credibility to these findings. The study evaluated a prospectively implemented operational intervention in its actual clinical environment rather than a retrospective simulation, directly addressing the well-documented gap between predictive model development and measurable clinical impact in AI healthcare research[5,29]. The intra-month crossover allocation provided concurrent control periods nested within the same institution and season, a feature absent from most PSM-ITS evaluations of ED interventions and one that substantially reduces confounding from temporal trends and institutional changes[19]. Spillover analysis confirmed no progressive improvement in the control period (Spearman $r = -0.049$; $p = 0.185$), ruling out learning-effect contamination as an

explanation for the observed differences. Complete schedule adherence over six months (100%) eliminates the compliance-related attenuation of effect estimates that constrained interpretation in previous forecast-driven staffing trials[13]. The forecasting model maintained stable performance throughout the study period (overall RMSE 3.04 patients/hour; R2 = 0.61). The relatively high SMAPE (45.4%) reflects hourly resolution, which amplifies percentage error during low-volume overnight periods. This did not translate into operationally significant scheduling errors because the optimization algorithm aggregates hourly forecasts to shift-level staffing decisions, smoothing out hour-to-hour variability[6].

One contextual observation warrants note. Patient volumes in 2025 were lower than in the prior-year historical cohort (daily volumes 77.3 and 75.8 versus 90.4 in 2024), most plausibly reflecting secular trends in pediatric ED utilization following the resolution of the 2022–2023 respiratory viral illness surge rather than a study artifact[9]. This volume difference does not affect the primary analysis, which compared the two concurrent 2025 arms under statistically identical census conditions (p = 0.979). The intervention arm achieved shorter boarding times while managing greater cumulative waiting burden, which directly argues against volume-driven confounding as an explanation for the observed differences[21].

This study has important limitations. The single-center quasi-experimental design, restricted to evening shifts over six months, does not support causal inference with the certainty of a randomised trial and does not capture a full annual seasonal cycle. Residual confounding from unmeasured patient-level or contextual factors cannot be excluded despite propensity score adjustment and the concurrent control design[19,20]. The intra-month allocation structure creates a systematic sequence in which the intervention arm always precedes the control arm within each calendar month; although the spillover analysis strongly argued against contamination, the possibility that heightened institutional attention at month start contributed to the observed differences cannot be entirely excluded. The reduction in diagnostic testing should be interpreted as exploratory. The study was not designed or powered to assess diagnostic appropriateness, patient safety events, or adverse outcome rates. Physician satisfaction, burnout risk, and formal cost-effectiveness were not evaluated. Generalizability to adult emergency departments, lower-volume pediatric settings, resource-constrained environments, and health systems with substantially different staffing structures requires prospective multi-center validation before broad implementation can be considered. Future work should integrate formal cost-effectiveness modelling, patient safety outcome assessment, and clinician experience evaluation to build the comprehensive evidence base that widespread adoption demands.

Conclusion

This prospective pilot study demonstrates that integrating deep learning-based demand forecasting with linear programming-based physician scheduling produces clinically meaningful and methodologically robust improvements in boarding time, diagnostic testing intensity, and physician workload in a pediatric emergency department. The convergence of propensity score-matched, interrupted time series, and

unadjusted analyses across three independent frameworks strengthens confidence in the observed association. Subgroup analyses demonstrate that the effect is concentrated among lower-acuity patients and during peak demand periods where staffing-demand mismatch is greatest, confirming a capacity-dependent throughput mechanism. The concurrent reduction in diagnostic test utilization across all three modalities, achieved without any protocol-level restriction, is consistent with operations research evidence linking higher physician workload to broader, less selective test ordering. Complete schedule adherence (100%) and the absence of spillover effects support the internal validity of the intra-month crossover design. These findings position forecast-driven dynamic physician scheduling as a feasible and effective strategy for improving pediatric emergency care operations, warranting multi-center validation with longer follow-up, direct assessment of patient safety outcomes, and formal cost-effectiveness evaluation.

Declarations

Ethics approval and consent to participate

This study was approved by the Hacettepe University Non-Interventional Clinical Research Ethics Committee (Approval No: SBA 24/1092). The requirement for individual informed consent was waived because the analysis used de-identified routine operational and electronic health record data collected as part of standard clinical care, in accordance with institutional data governance principles. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Data availability

All data produced in the present study are available upon reasonable request to the authors.

Code availability

The underlying code supporting the TiDE-RIN demand forecasting and linear programming-based physician scheduling framework used in this study is publicly available at https://github.com/turkalpmd/SOSAFED.

Author contributions

A.Z.B. contributed to the conceptualization, validation, resources, data curation, and drafting the original manuscript. I.T.A. conceptualized the study, developed the methodology, handled the software, validated the results, performed the formal analysis, and contributed to data curation, writing the original draft, and visualization. O.T. was responsible for the investigation, provided resources, reviewed and edited the manuscript, supervised the project, managed project administration, and acquired funding. All authors have read and approved the final manuscript.

Use of Generative Artificial Intelligence

Generative artificial intelligence (Claude Opus 4.6) was used solely for language editing and clarity. No artificial intelligence was used in the generation of scientific content, data analysis, or interpretation. All authors reviewed and approved the final manuscript and take responsibility for its content.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests

All authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Forero, R., McCarthy, S. & Hillman, K. Access block and emergency department overcrowding. *Crit. Care* 15, 216 (2011).
2. Akbasli, I. T., Birbilen, A. Z. & Teksam, O. Artificial intelligence-driven forecasting and shift optimization for pediatric emergency department crowding. *JAMIA Open* 8, ooae138 (2025).
3. Sangal, R. B., Rothenberg, C., Taylor, R. A. & Venkatesh, A. K. Emergency Care Access Based on a Proposed CMS National Quality Measure. *JAMA Health Forum* 6, e250417 (2025).
4. Morley, C., Unwin, M., Peterson, G. M., Stankovich, J. & Kinsman, L. Emergency department crowding: A systematic review of causes, consequences and solutions. *PLOS ONE* 13, e0203316 (2018).
5. Chan, S. L. *et al.* Implementation of Prediction Models in the Emergency Department from an Implementation Science Perspective—Determinants, Outcomes, and Real-World Impact: A Scoping Review. *Ann. Emerg. Med.* 82, 22–36 (2023).
6. Porto, B. M. & Fogliatto, F. S. Enhanced forecasting of emergency department patient arrivals using feature engineering approach and machine learning. *BMC Med. Inform. Decis. Mak.* 24, 377 (2024).
7. Jones, S. S. *et al.* A multivariate time series approach to modeling and forecasting demand in the emergency department. *J. Biomed. Inform.* 42, 123–139 (2009).
8. Blanco, J., Ferreras, M. & Cosido, O. Predictive modeling of hospital emergency department demand using artificial intelligence: A systematic review. *Int. J. Med. Inf.* 207, 106215 (2026).
9. Janke, A. T. *et al.* Emergency Department Care for Children During the 2022 Viral Respiratory Illness Surge. *JAMA Netw. Open* 6, e2346769 (2023).
10. Wang, H., Sambamoorthi, N., Hoot, N., Bryant, D. & Sambamoorthi, U. Evaluating fairness of machine learning prediction of prolonged wait times in Emergency Department with Interpretable eXtreme gradient boosting. *PLOS Digit. Health* 4, e0000751 (2025).

11. Tyler, S. *et al.* Use of Artificial Intelligence in Triage in Hospital Emergency Departments: A Scoping Review. *Cureus* 16, (2024).
12. Ahmadzadeh, B. *et al.* Artificial Intelligence Solutions to Improve Emergency Department Wait Times: Living Systematic Review. *J. Emerg. Med.* 75, 174–187 (2025).
13. Hu, Y. *et al.* Implementing a prediction driven framework for emergency department nurse staffing to optimize real time decisions. *Npj Health Syst.* 2, 16 (2025).
14. Das, A. *et al.* Long-term Forecasting with TiDE: Time-series Dense Encoder. Preprint at https://doi.org/10.48550/arXiv.2304.08424 (2023).
15. McCarthy, M. L. *et al.* The Emergency Department Occupancy Rate: A Simple Measure of Emergency Department Crowding? *Ann. Emerg. Med.* 51, 15–24.e2 (2008).
16. J Murray, M. The Canadian Triage and Acuity Scale: A Canadian perspective on emergency department triage. *Emerg. Med.* 15, 6–10 (2003).
17. Poursoltan, L. *et al.* Prospective comparison of econometric, machine learning, and foundation models for forecasting emergency department boarding patients. *Npj Health Syst.* 2, 49 (2025).
18. Collins, G. S. *et al.* TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385, e078378 (2024).
19. Yun, C.-C., Huang, S.-J., Kuo, T., Li, Y.-C. & Juang, W.-C. Impact of New Bed Assignment Information System on Emergency Department Length of Stay: An Effect Evaluation for Lean Intervention by Using Interrupted Time Series and Propensity Score Matching Analysis. *Int. J. Environ. Res. Public. Health* 19, (2022).
20. Bernal, J. L., Cummins, S. & Gasparrini, A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int. J. Epidemiol.* 46, 348–355 (2017).
21. Green, L. V., Soares, J., Giglio, J. F. & Green, R. A. Using Queueing Theory to Increase the Effectiveness of Emergency Department Provider Staffing. *Acad. Emerg. Med.* 13, 61–68 (2006).
22. Hu, Y. *et al.* Use of Real-Time Information to Predict Future Arrivals in the Emergency Department. *Ann. Emerg. Med.* 81, 728–737 (2023).
23. Hodgson, N. R., Saghafian, S., Martini, W. A., Feizi, A. & Orfanoudaki, A. Artificial Intelligence-Assisted Emergency Department Vertical Patient Flow Optimization. *J. Pers. Med.* 15, (2025).
24. Asplin, B. R. *et al.* A conceptual model of emergency department crowding. *Ann. Emerg. Med.* 42, 173–180 (2003).
25. Hoot, N. R. & Aronsky, D. Systematic Review of Emergency Department Crowding: Causes, Effects, and Solutions. *Ann. Emerg. Med.* 52, 126–136.e1 (2008).
26. Bernstein, S. L. *et al.* The Effect of Emergency Department Crowding on Clinically Oriented Outcomes. *Acad. Emerg. Med.* 16, 1–10 (2009).
27. Soltani, M., Batt, R. J., Bavafa, H. & Patterson, B. W. Does What Happens in the ED Stay in the ED? The Effects of Emergency Department Physician Workload on Post-ED Care Use. *Manuf. Serv. Oper. Manag.* 24, 3079–3098 (2022).

28. Saghafian, S., Austin, G. & Traub, S. J. Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Trans. Healthc. Syst. Eng.* 5, 101–123 (2015).
29. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 17, 195 (2019).

Figures

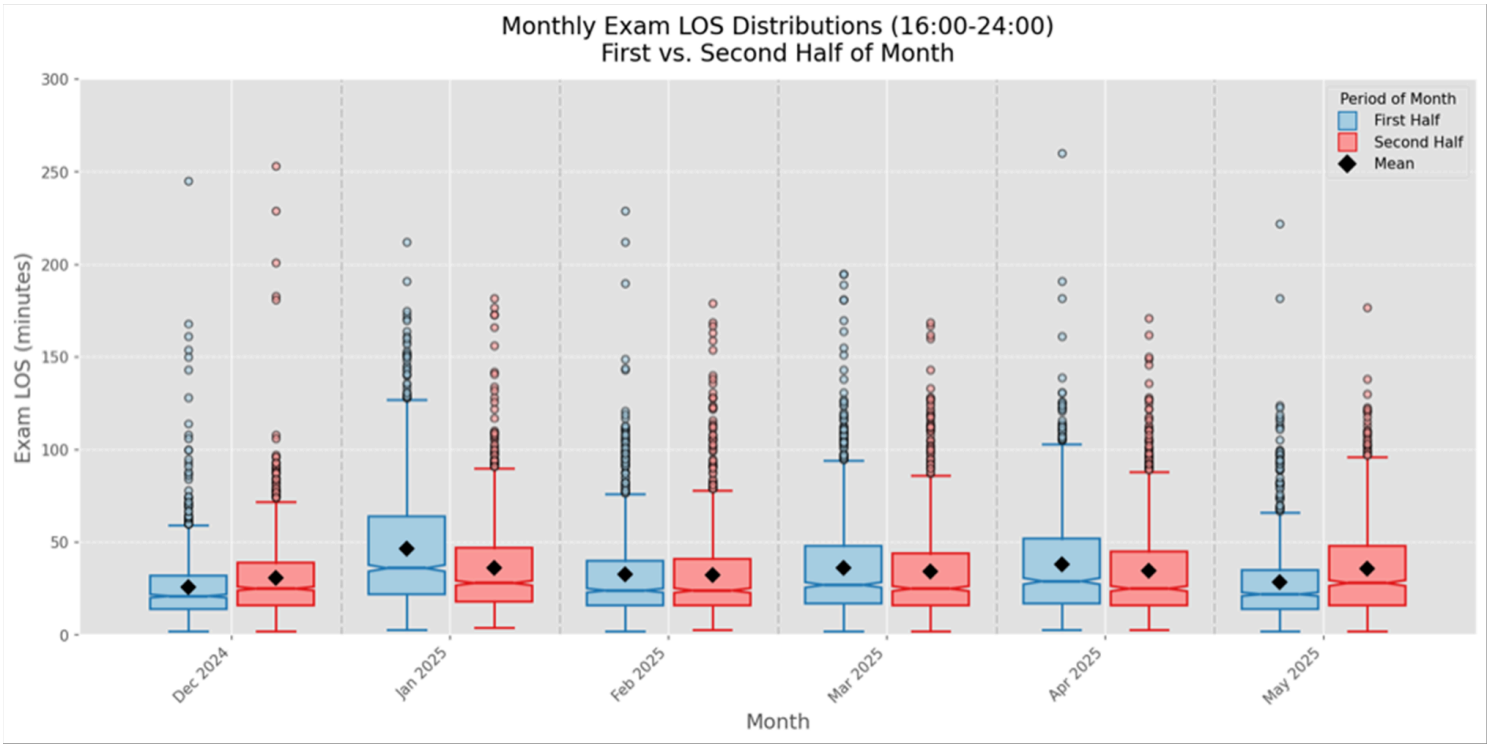


Figure 1

Examination length of stay (Exam LOS) by study period. Monthly distribution of Exam LOS (minutes) for the AI-optimized intervention arm and the concurrent control arm across the six-month study period (December 2024–May 2025). Exam LOS remained stable and comparable between arms (intervention 35.4 ± 28.9 vs control 34.0 ± 27.9 min; p = 0.935), confirming that the observed reductions in boarding time were not achieved at the expense of abbreviated clinical evaluations.

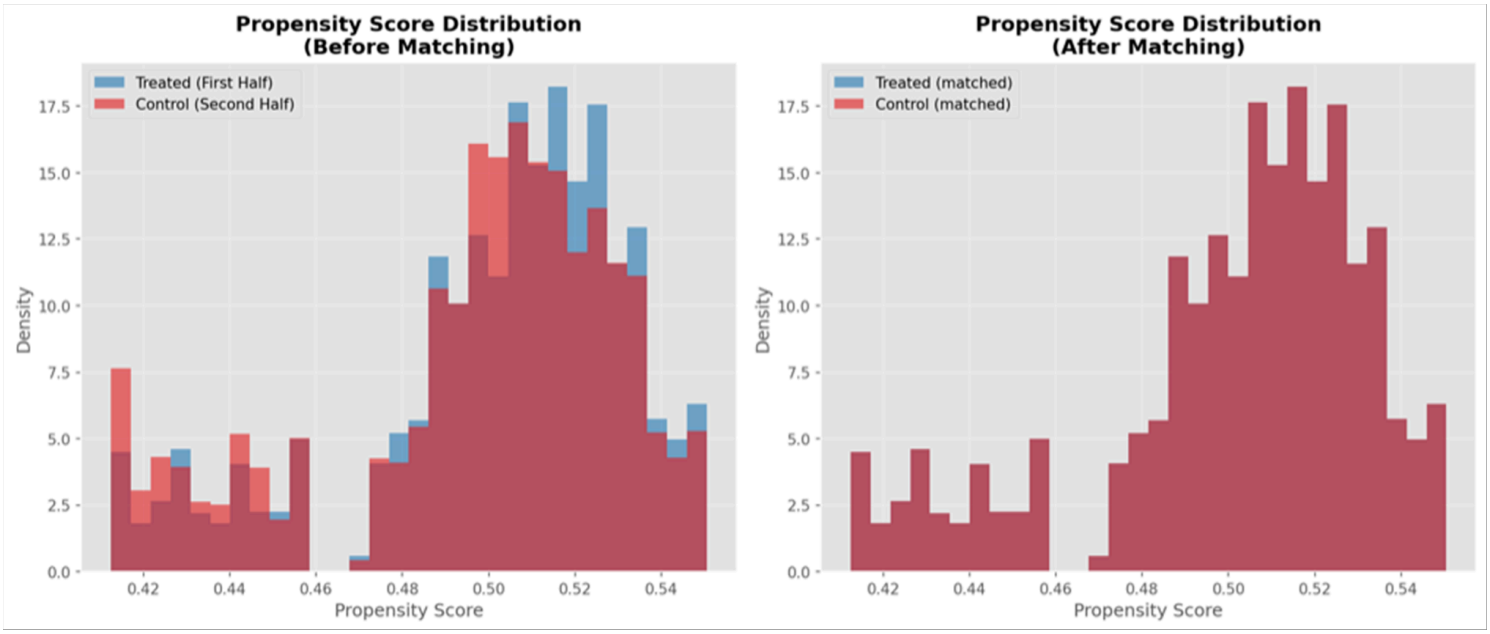


Figure 2

Propensity score distributions before and after matching. Left panel: pre-matching propensity score distributions for the AI-optimized intervention arm (blue) and concurrent control arm (red). Right panel: post-matching distributions demonstrating near-complete overlap, confirming adequate covariate balance across all four matching variables (arrival hour, day of week, calendar month, triage acuity). SMD < 0.1 for all covariates after matching.

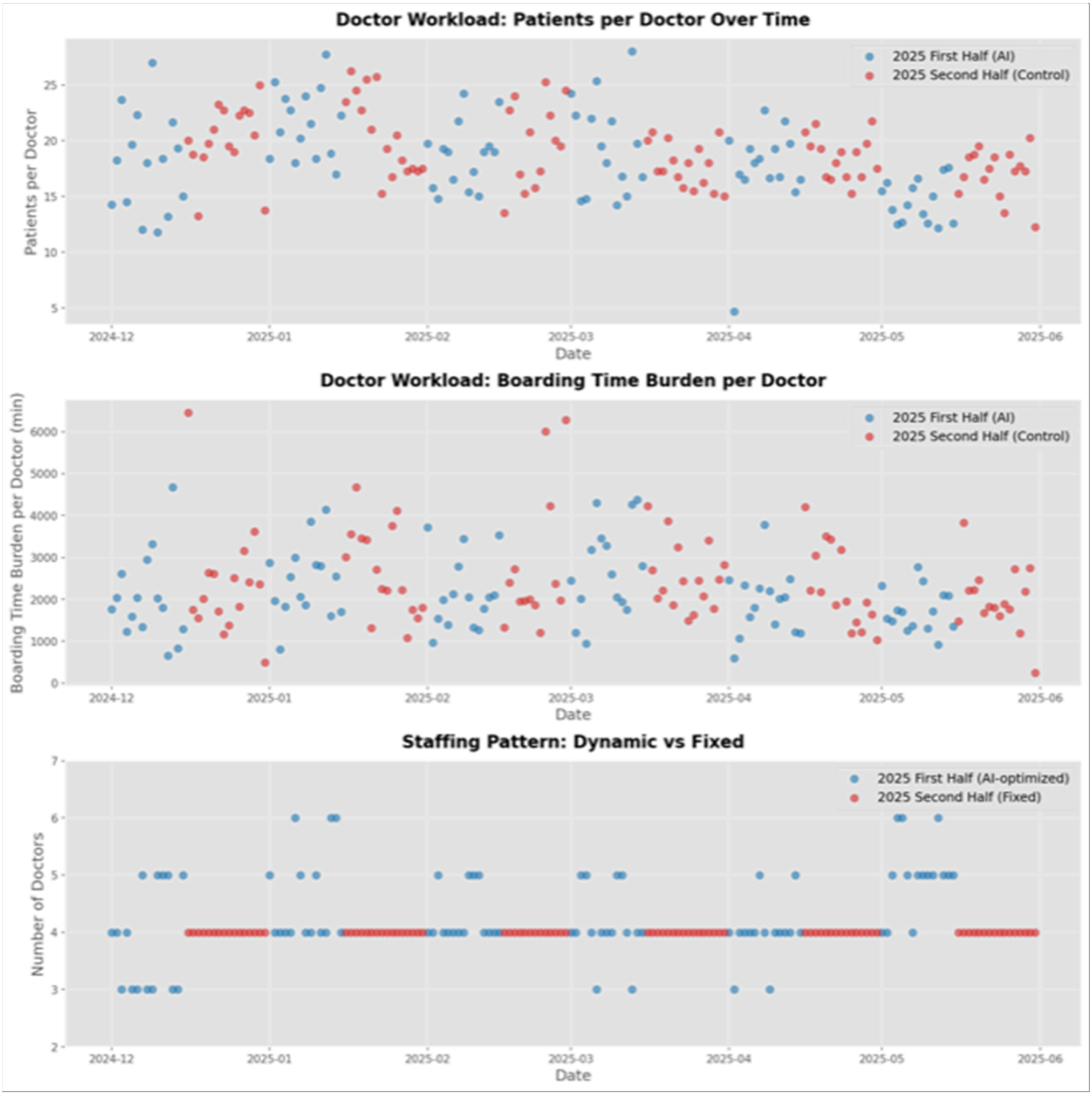


Figure 3

Physician workload analysis: AI-optimized vs fixed staffing. Upper panel: scatter plot of patients per physician by daily patient volume, comparing AI-optimized staffing (blue) with fixed four-physician staffing (red). Lower panel: boarding burden per physician (patient-minutes per shift). The AI system allocated 3–6 physicians per shift based on forecasted demand, resulting in a 12.0% reduction in per-physician boarding burden ($p = 0.013$) despite managing comparable patient volumes.

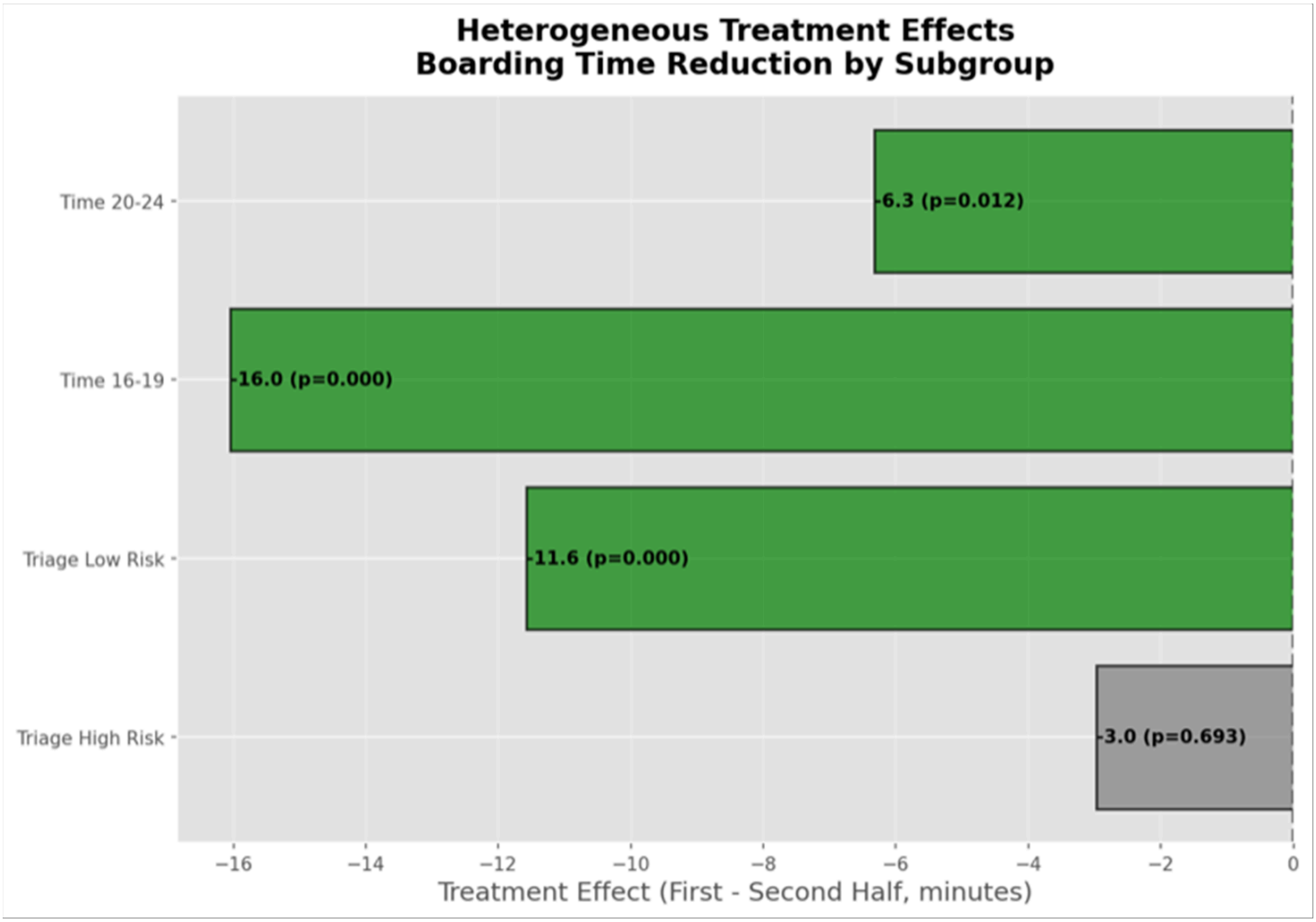


Figure 4

Heterogeneous treatment effects by patient subgroups. Forest plot displaying the difference in mean boarding time (minutes) between the AI-optimized intervention arm and the concurrent control arm across pre-specified subgroups. Low-risk patients (T4–T5) showed a statistically significant reduction of 11.6 minutes (P < 0.0001), whereas the effect among high-risk patients (T1–T3) was not significant (−3.0 min; p = 0.693). Temporal stratification revealed a 2.5-fold gradient: early evening hours (16:00–19:00) showed the largest effect (−16.0 min; p = 0.0001), followed by late evening hours (20:00–24:00; −6.3 min; p = 0.012). Error bars represent 95% confidence intervals.