The Impact of Historical Workload on Nurses' Perceived Workload

Carri W. Chan
Decision, Risk, and Operations, Columbia Business School, cwchan@gsb.columbia.edu

Yi Chen
Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, yichen@ust.hk

Jing Dong
Decision, Risk, and Operations, Columbia Business School, jing.dong@gsb.columbia.edu

Sarah C. Rossetti
Department of Biomedical Informatics, Columbia University sac2125@cumc.columbia.edu

Recent and ongoing nursing shortages have highlighted the valuable and skilled work that nurses provide around the clock in hospital inpatient care. Intense and sustained high nursing workload has been linked to nurse burnout and patient safety concerns, necessitating targeted approaches to better managing nursing workload. In this work, we take an empirical approach to understanding the effect of historical workload on nurses' perceived workload. We leverage a unique dataset that records detailed patient-to-nurse assignment information, an order-based workload measure, and a clinically perceived workload measure for each patient during each shift. We also address several identification challenges, including endogeneity, missing values, and measurement errors. Our estimation results show that one level of increase in historical order-based workload can lead to a 0.629 increase in the discrepancy between the clinically perceived workload and the order-based workload. Based on the temporal effect of nursing workload, we design an integer program-based patient-to-nurse assignment policy that achieves a more balanced workload over time while maintaining a high level of continuity of care.

Key words: healthcare operations, nursing workload, workload balancing, continuity of care
History:

1. Introduction

Provider burnout is a major and ongoing challenge in healthcare. Many studies have shown that stress and burnout play an important role in nurses' intention to leave the profession (Jourdain and Chênevert 2010). Health systems are faced with stark workforce challenges which have been further exacerbated by the COVID-19 pandemic due to the increasing nurse turnover rate, the growing aging population, and shifting sites of care for patients (Berlin et al. 2022). With nursing

shortages projected to be nearly 25,000 nurses to meet the annual needs (Stiegler et al. 2021), there is an ever-increasing need to mitigate nursing burnout. Stress and burnout have been shown to affect nurses' overall well-being and the quality of professional care they provide (Aiken et al. 2002, Vahey et al. 2004, Halbesleben et al. 2008). Among various factors, nurses' perceived workload has been shown to be positively associated with nursing burnout (Holland et al. 2019). In this work, we take an empirical approach to understand how nurses' historical workload affects their perceived workload in the setting of a medical intensive care unit. This will help inform the design of patient-to-nurse assignment policies to achieve more balanced workload assignments over time and create a fairer and safer working environment for critical care nurses.

We utilize a novel data set from a nursing workload management software (OptiLink 2012). The data contains detailed patient-to-nurse assignment information and two workload intensity measures: One is an order-based workload intensity measure calculated automatically using the electronic health record; the other is a clinically perceived workload intensity measure reported by the staff nurses and is supposed to reflect the patient's workload in the opinion of the nurses who are caring for the patient. We are interested in understanding the difference in nurses' perceived workload from the objective order-based workload. This gap is likely influenced by various factors including mental stress or physical exhaustion. These factors are further associated with nursing burnout and job dissatisfaction.

Our study setting is the intensive care unit (ICU) whose patients tend to have high acuity and increased risk of deterioration. The skilled work that nurses provide round the clock is critical to patient care. At the same time, there is a lot of heterogeneity in terms of the level of medical attention/supervision patients require. For example, patients on ventilators, antimicrobial desensitization, or hypothermia protocol may require continuous minute-to-minute nursing assessment and monitoring; while other ICU patients may require less intensive, though still hourly, monitoring. Thus, properly measuring nursing workload and using this to determine how to assign patients to nurses can be highly non-trivial. The nursing workload is comprised of numerous factors such as the amount of nursing time required for patient care activities (medications and therapeutic supports, testing, etc); patient and family communication and education needs; and the amount of physical exertion required to cover patient's basic needs (Alghamdi 2016). Even though order-based workload measures are widely used, they are inherently limited and there is a growing consensus in the nursing community that they may not accurately capture many important aspects of nursing workload. To control for various factors that can affect the nurse's perceived workload, we supplement the nursing workload management data with patient hospitalization data, patient flow data, patient hourly laboratory-based acuity-score data, and nurse staffing data.

The main treatment variable we are interested in is nurses' historical workload, which is measured using the order-based workload. While we cannot avoid assigning high-workload-intensity patients to some nurses due to the nature of ICU care, we may be able to avoid continuously assigning high-workload-intensity patients to the same nurse across different shifts that the nurse works. In other words, we can achieve a certain notion of workload balancing over time.

As we utilize retrospective data collected from electronic health records and staffing databases, a number of empirical challenges arise. We address these challenges through a combination of instrumental variable analyses, corrections for sample selection, and sensitivity analyses.

In the workload management software that we utilize, both the order-based and clinically perceived workload intensities are measured in four levels: low (1), median/average (2), high (3), and extremely high (4). Our empirical estimation shows that a higher historical workload is associated with a higher discrepancy between the nurses' perceived workload and the order-based workload. More specifically, a one-unit increase in the nurse's historical workload (measured by the order-based workload measure) can lead to 0.629 units of increase in the discrepancy on average. This implies that for an average patient with an average order-based workload, if the nurse who takes care of the patient has an extremely high historical workload, the nurse will perceive an extremely high workload on average for the focal patient in the focal shift. Note that perceiving an extremely high workload likely will increase stress and exhaustion for the nurse. Thus, it is important to balance the nurses' workload over time to reduce the risk of burnout.

Based on our estimation results, we propose a patient-to-nurse assignment policy that balances nursing workload over time, with the hope to reduce the risk of nurses perceiving an extremely high workload. Our assignment policy also takes continuity of care considerations into account. In particular, we want to assign a "more familiar" nurse to the patient when possible. Continuity of care has been shown to be beneficial for patient outcomes (Haggerty et al. 2003, Ahuja et al. 2020, 2022) and staff productivity (Kajaria-Montag et al. 2022) in various healthcare settings. Maintaining continuity of care may sometimes require assigning the same high workload patient to the same nurse. Thus, there can be a tradeoff between the continuity of care and workload balancing considerations. Our proposed policy carefully balances these two considerations and is able to achieve improvement on both performance metrics compared to the current policy used in our partner hospital. With properly chosen weight on continuity of care and workload balancing, our policy can reduce the amount of time nurses perceive an extremely high workload by 17%, while improving one commonly used continuity of care measure by 3.6%.

Above all, our work has two main contributions. First, we quantify the impact of nurses' historical workload on how their perceived workload deviates from the order-based workload, highlighting

the potential of balancing nursing workload over time to reduce the amount of time nurses perceiving extremely high workloads. To infer causal effects of historical workload, we utilize data from multiple sources and address several identification challenges:

- *Sample Selection Bias.* In our data, 20% of the outcome variables are missing, and they may not be missing at random. Thus, there may be a sample selection bias. To account for sample selection, we use the Heckman selection model with a properly constructed instrumental variable.
- *Endogeneity.* Our estimation may suffer from omitted variable bias. In particular, because nurse assignments are unlikely to be truly random, there may be unobservable factors that are correlated with both nurses' historical workload and their perceived workload. To address the endogeneity issue, we construct appropriate instrumental variables (IVs). We also conduct sensitivity analysis for potential violations of the exclusion restriction for the IVs.
- *Measurement Errors.* In our data, some nurse assignment information is missing when a patient is first admitted to the ICU. Thus, the calculation of the historical workload may incur underestimation due to the missing admission shift information. In addition, because the workload intensities are classified into only four different levels, the outcome variable, i.e., the difference between the nurse's perceived workload and the order-based workload, may be censored. For example, the nurse cannot report a higher perceived workload when the order-based workload is at level 4 (extremely high). We conduct carefully designed sensitivity analysis to gauge the effects of these measurement errors.

We find statistically significant and robust evidence that a higher historical workload increases the likelihood that nurses perceive a higher workload than the order-based workload for their focal patient in the focal shift.

Our second contribution is that we propose a patient-to-nurse assignment policy that strikes a balance between workload balancing and continuity of care. The assignment rule is based on a simple integer linear program and achieves significant improvement in the workload-balancing metrics for our partner hospital while also improving the continuity-of-care metrics.

1.1. Literature Review

Our work contributes to the broader literature on workload in healthcare. There is extensive medical literature on nursing workload management. For example, Miranda et al. (1997, 2003), Muehler et al. (2010), Griffiths et al. (2020) study various nursing workload measures. Aiken et al. (2002), Laschinger and Leiter (2006), Carayon and Gurses (2008), Liu et al. (2018), Shah et al. (2021) study the impact of nursing workload on nurse burnout and patient safety. Our work builds on these works and provides rigorous causal quantification of how nurses' historical workload affects their perceived workload. We also study the implication of our empirical finding on how to match

patients with nurses. In this section, we mainly focus on reviewing works within the operations management space.

Impact of Workload. The impact of workload has been widely studied in the healthcare operations management literature. Most of these works focus on unit-level aggregated workload and measure the impact of workload on patient outcomes or physician behaviors. Kc and Terwiesch (2009) study the association between system-level workload and service times. They find that workers accelerate the service rate as the load increases. However, long periods of increased load can decrease the service rate. In contrast, Berry Jaeker and Tucker (2017) find an inverted U-shape relationship between workload and service time. Patient length of stay (LOS) increases as occupancy increases, until a tipping point, after which patients are discharged early. There could also be a second tipping point, beyond which an additional increase in occupancy leads to a longer LOS. Luo et al. (2022) find a similar inverted U-shape relationship utilizing a task-based measure of workload. Kc and Terwiesch (2012) study the effect of ICU occupancy on patients' length of stay (LOS) and find that a patient is more likely to be discharged early when the ICU occupancy is high. However, early discharge can be associated with a higher chance of readmission, which may end up generating more work for the ICU. Long and Mathews (2018) find that active treatment time in the ICU is unaffected by occupancy levels while boarding time in the ICU is shortened when ICU occupancy is high but prolonged when hospital ward occupancy is high. Soltani et al. (2022) find an increasing concave relationship between emergency department (ED) physician workload and post-ED care use, i.e., the number of post-discharge care events.

Medical staff may also come up with various strategies to navigate high workloads. For example, midwives may ration resource-intensive discretionary services and increase the rate of specialist referral for patients with complex needs to manage their workload (Freeman et al. 2021). Emergency Department (ED) staff may order diagnostic tests during the triage process, i.e., early task initiation, when the ED is crowded (Batt and Terwiesch 2017), or prioritize discharged patients when there are a lot of boarding patients (Li et al. 2021). It has also been found that physicians may prioritize easier tasks or batch similar tasks when working under high load (Ibanez et al. 2018, Kc et al. 2020).

Our work complements these works and studies how historical workload affects nurses' perceived workload. Utilizing a unique data set, we are also able to measure the workload at the patient-shift level. We are also able to measure the difference between the nurse's perceived workload and the order-based workload.

Nursing Workload Management It has been long recognized that it is important and highly nontrivial to properly manage the nursing workload. Most existing literature focuses on nurse staffing policies. The goal is to balance the cost of staffing and the service quality provided. For

example, Green et al. (2013), Wang and Gupta (2014) study nurse staffing in the presence of nurse absenteeism. Véricourt and Jennings (2011) use a closed queueing model to determine efficient nurse staffing policies and study the implication of mandate fixed nurse-to-patient ratios. Kim and Mehrotra (2015) study a two-stage stochastic integer programming for nurse staffing and scheduling. Our work focuses on how to match patients with nurses to balance the nursing workload over time, which has not been well-studied in the literature. Related to patient-to-nurse matching, Niewoehner III et al. (2022) study how familiarity affects the patient selection and multitasking level for ED physicians. Matching customers to agents has also been studied in the broader matching literature, but the focus is on maximizing matching utility or minimizing the amount of communication required, not workload balancing (Arnosti et al. 2021, Shi 2022).

There is also extensive literature on how staffing-related decisions affect staff behavior and well-being in healthcare and beyond. For instance, Bergman et al. (2022) study the effect of schedule volatility on nurses' voluntary turnover in a large home health care agency. Kamalahmadi et al. (2021) show that short-notice schedules do not harm agents' productivity but real-time schedules do in the setting of a restaurant chain. Kesavan et al. (2022) estimate the causal effects of responsible scheduling practices (e.g., consistency, predictability, adequacy, and employee control) on employee well-being and productivity in retail stores. On the modeling side, Armony and Ward (2010), Mandelbaum et al. (2012) study fair routing policies in queuing systems. The goal is to properly balance customer wait time and fairly divide the workload among agents of different skill levels.[1] Wang et al. (2022) consider the workload balancing over multiple periods among couriers in the dispatching of last-mile urban delivery.

1.2. Paper Organization

The remainder of the paper is organized as follows. We provide more details about the empirical setting and datasets used in our analyses in Section 2. We then introduce our model and discuss several econometric challenges, as well as our approach to addressing them in Section 3. Section 4 presents our empirical findings. We conduct extensive sensitivity analysis to demonstrate the robustness of our results in Section 5. In Section 6, we propose a patient-to-nurse assignment policy and test its performance by putting different weights on the reward for continuity of care versus workload balancing. Lastly, we conclude and discuss future research directions in Section 7.

[1] There is another notion of workload balancing that balances the workload among different servers to achieve better resource utilization, i.e., minimizing idleness (see, for example, join the shortest queue Foley and McDonald (2001)). Note that this is quite different from our setting.

2. Setting and Data

Our research setting is the Medical Intensive Care Unit (MICU) at the NewYork-Presbyterian/Columbia University Irving Medical Center, a 738-bed academic medical center. The MICU contains 24 beds and operates under two nursing shifts per day. One shift is from 7:00 am to 7:00 pm. The other is from 7:00 pm to 7:00 am.

We combine data from multiple sources to construct a comprehensive list of variables that may affect nursing workload. First, we obtain hospitalization data, which includes patient-visit level information such as patient demographics and comorbidities. Second, we obtain patient-flow data, which includes detailed time stamps on patient admission and discharge activities to different units within the hospital. Third, we collect hourly Laboratory-based Acute Physiology Score (LAPS) (Escobar et al. 2008) for each patient, which integrates information from 14 laboratory tests and serves as a measure of patients' acuity. Fourth, we obtain data from the workload management software – Kronos OptiLink, which further includes two sets of patient-shift level data: assignment data which contains detailed patient-to-nurse assignment information for each nursing shift, and workload intensity data which contains two workload intensity measures associated with each patient in each nursing shift. Lastly, we obtain data from the human resources system that records the nurses' working schedules and actual punch-in and punch-out times.

There are several advantages of utilizing data from multiple sources. In addition to being able to extract a rich collection of covariates, we utilize the overlapping parts of the data sets to ensure better data accuracy. For example, the patient-flow data and workload management data both contain information on patient admission and discharge times. The human resources data and workload management software data both contain information on the nurses' shift start and end times. Inconsistency in these time stamps would flag a potential error in the data (see Appendix A for more details). Lastly, the detailed patient-to-nurse assignment data allow us to measure the workload at the individual nurse level, rather than the aggregated unit level.

2.1. Data Processing

We collect data from the MICU in 2018. Our analysis is at the patient-shift level. All shifts in January and December are excluded due to some missing data: some information for patients admitted before January 2018 or discharged after December 2018 is missing. Thus, our final cohort for analysis is from Feb 1, 2018 to Nov 30, 2018. We further exclude all patient-shifts that correspond to the admission shifts of the patients (1,079 observations) due to a large amount of missing assignment information. In particular, 28.7% of the patients do not have the nurse assignment information during the shift at which they are admitted to the MICU. However, we do include the admitted patients in the calculation of the unit census. We also control for the number of admissions

in each shift in our regression analysis. Lastly, we conduct sensitivity analysis by including the imputed admission shift workload.

We exclude 94 observations associated with four shifts that have a large amount of missing assignment information: 2018-02-28 day shift, 2018-02-28 night shift, 2018-03-07 day shift, and 2018-04-17 night shift. Next, we exclude 454 observations associated with 32 unique patients whose patient-level information like LAPS or Elixhauser scores are missing. These observations are included in the unit census calculation though. Lastly, we classify a nurse as "primary" versus "non-primary" based on whether the nurse worked for more than 30 shifts during the study period (30 corresponds to the lower 5% percentile). The "non-primary" nurses, which includes all the agency nurses (each of whom works for at most 13 shifts during the study period in our data), may primarily work at other units rather than the MICU, and thus we may not be able to measure their historical workload accurately. Based on this classification rule, our data contains 72 primary nurses and 77 non-primary nurses. The primary nurses worked on average 112 shifts during the study period. We exclude 700 observations that are associated with the 77 non-primary nurses. Again, these observations are included in the unit census calculation. Lastly, we exclude 273 observations in the first week of the study period since the historical workload cannot be accurately evaluated for these observations. See Appendix A for more details on data processing and cleaning. Our final cohort contains 12,625 patient-shift observations, associated with 941 unique patients and 72 unique nurses.

2.2. Main Outcome Measure

The workload intensity data is recorded at the patient-shift level and contains two workload intensity measures: one is a data-driven order-based workload intensity measure – DDCIntensity where DDC stands for "data-driven classification"; the other is a clinically perceived workload intensity measure, which is input by the staff nurse – PJIntensity where PJ stands for "professional judgment". The PJIntensity reflects the clinical or psychosocial situation(s) that best describes the patient's workload in the opinion of the nurses who are caring for the patient.

Table 1 summarizes the distribution of the two intensity scores where NA denotes missing values. Recall that both DDCIntensity and PJIntensity take on four different levels: Low (1), Medium/Average (2), High (3), and Extreme (4). Note that 20.0% of the PJIntensity scores are missing (None of the DDCIntensity scores are missing). Excluding those missing data, among the remaining observations, 52.7% of the patient-shifts have a PJIntensity that is higher than their corresponding DDCIntensity (PJ $>$ DDC), while only 5.3% of the patient-shifts have a PJIntensity that is lower than their DDCIntensity. The remaining shifts have equal PJ and DDC intensities.

For patient p in shift t, our main outcome variable is the difference between the PJ and DDC intensities:

$$Intensity\_Gap_{p,t} = PJIntensity_{p,t} - DDCIntensity_{p,t},$$

which measures how the nurse's perceived workload is different from the order-based workload for the focal patient in the focal shift. Table 2 shows the distribution of $Intensity\_Gap$ conditional on the PJIntensity is not missing.

Table 1 Distribution of PJIntensity and DDCIntensity

Level	1 (Low)	2 (Average)	3 (High)	4 (Extreme)	NA
PJIntensity	166 (1.3%)	3385 (26.8%)	5694 (45.1%)	856 (6.8%)	2524 (20.0%)
DDCIntensity	1171 (9.2%)	8142 (64.5%)	3042 (24.1%)	270 (2.1%)	0 (0.0%)

Table 2 Distribution of $Intensity\_Gap$ conditional on PJIntensity score is nonmissing

Level	-2	-1	0	1	2	3
Frequency	12 (0.12%)	523 (5.2%)	4237 (41.9%)	4693 (46.5%)	629 (6.2%)	7 (0.07%)

2.3. Treatment Variable

We are interested in understanding how nurses' historical workload – measured via the orders-based workload – affects the discrepancy between their perceived workload and the order-based workload. The idea behind this is that prior assignments may impact the nurses' stress and fatigue, which may in turn impact the perception of workload for the current assignment. This has important implications for managing the nursing workload over time. For patient p in shift t, our main treatment variable is $Hist\_Work_{p,t}$, which is defined as the maximum DDCIntensity of the nearest previous assignment(s) over the past week for the nurse(s) taking care of patient p.

Specifically, let $N_t(p)$ be the set of the nurses assigned to patient p in shift t (all nurses in $N_t(p)$ take care of patient p during the whole shift t) and $P_t(n)$ be the set of the patients assigned to nurse n in shift t. We also denote P_t as the set of patients who stay in the MICU in shift t. Let k_n be the time interval between nurse n's most recent prior working shift and the focal shift. That is,

$$k_n = \min \left\{ k > 0; n \in \bigcup_{q \in P_{t-k}} N_{t-k}(q) \right\} \tag{1}$$

where $k_n = \infty$ if nurse n has not worked a shift prior to shift t. Then, the historical workload for the nurses who care for patient p in shift t is defined as

$$Hist\_Work_{p,t} = \frac{1}{|N_t(p)|} \cdot \sum_{n \in N_t(p)} \max_{q \in P_{t-k_n}(n)} \left\{ DDCIntensity_{q,t-k_n} \right\} \cdot 1\big(k_n \leq 14\big), \tag{2}$$

where $DDCIntensity_{q,t-k_n}$ is the DDCIntensity for patient q in shift $t - k_n$. For example, if nurse n is assigned to patient p in shift t, then $n \in N_t(p)$. Suppose the most recent shift for nurse n before shift t is shift $t - k_n$, $1 \leq k_n \leq 14$ (considering shifts within the last 7 days), where the nurse is assigned to patient q. Then, we use the DDCIntensity of patient q in shift $t - k_n$ as the historical workload for nurse n associated with patient p in shift t. If multiple patients were assigned to nurse n in shift $t - k_n$, we take the maximum of their DDCIntensity scores. In our data, 82.3% of the nurse-shifts have more than one patient assigned to a nurse. If the focal nurse had no assignment over the past week, $Hist\_Work_{p,t}$ is set as zero. The idea is that after a long enough rest, the impact of the historical workload on the perceived workload in the focal shift is likely negligible. When calculating the historical workload at the patient-shift level, which is the unit of our analysis, if multiple nurses are assigned to the focal patient in the focal shift, we take the average of these nurses' historical workload. In our data, only 7.3% of the patient-shifts have more than one nurse assigned to a patient.

Figure 1 plots a histogram of the historical workload calculated based on our data. We observe that $Hist\_Work$ mostly takes value 2 (54.7%) or 3 (34.4%). The mean and standard deviation are 2.3 and 0.74 respectively. Note that the non-integer values occur (1.8% of the time) because some patients are assigned to multiple nurses in a shift and in these cases, we take the average of nurses' historical workload as $Hist\_Work$.

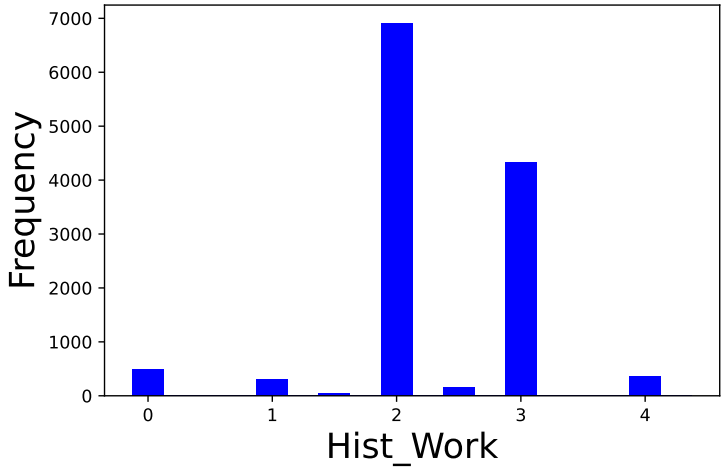


Figure 1 **Distribution of treatment** $Hist\_Work$

We also try other measures of the historical workload (e.g., based on the LAPS, using different length of the look-back time window, etc.) in the sensitivity analysis. See Section 5.3.1 and Appendix C more details.

2.4. Control Variables

We include a number of other variables in our empirical analysis to control for their potential effects on *Intensity_Gap*. Recall that our analysis is at the patient-shift level. We group the control variables into four categories: shift factors, patient factors at the patient record level, patient factors at the patient-shift level, and nursing factors.

Shift factors. For shift t, we control for weekday versus weekend; day versus night; busyness of the shift at the unit level, which includes the total number of admissions in the shift, the total number of discharges in the shift, hourly averaged census, average DDCIntensity scores over all the patients in the unit. We also control for the unit-level nurse-to-patient ratio, i.e., the total number of nurses divided by the total number of patients.

Patient factors at the patient-visit level. For patient p, we control for age; gender; comorbidity burden as measured by the Elixhauser score; and whether the patient is admitted to the MICU directly from the emergency department.

Patient factors at the patient-shift level. For each patient p in shift t, we control for the severity and workload intensity of the patient, including the DDCIntensity, whether the shift is within 24 hours of the patient's ICU admission, length of stay (LOS) in hours in the ICU prior to shift t, whether the patient is discharged during the shift, and average hourly LAPS during the shift. We also control for the relative severity of patient p, including the relative LAPS, DDCIntensity, and Elixhauser scores, which are defined as the ratios between the patient's average hourly LAPS, DDCIntensity, and Elixhauser score, and the average over all the patients in the unit during shift t respectively.

Nursing factors. For nurse $n \in N_t(p)$ taking care of patient p in shift t, we control for the number of patients assigned to the nurse in shift t, whether nurse n has taken care of patient p before shift t, i.e., a continuity of care flag that is equal to 1 if nurse n has taken care of patient p before, as well as the time interval between nurse n's previous working shift and shift t, which we refer to as the rest time for nurse n. To measure the nursing factors at the patient-shift level, if $N_t(p)$ contains more than one nurse, we take the average over all the nurses in $N_t(p)$. We also control for the number of nurses taking care of patient p in shift t, i.e., the size of $N_t(p)$.

Table 3 provide an overview of the control variables and some summary statistics of these variables.

Table 3 Summary statistics of control variables

Variable category	Variable	Mean/Proportion	Stdev
Shift factors (unit level)	day shift flag	49.8%	
	weekday flag	72.0%	
	# of admissions	1.73	1.10
	# of discharges	1.80	1.15
	census	22.34	1.14
	nurse-to-patient ratio	1.57	0.13
Patient factors (record level)	age	57.11	17.62
	gender-male	52.2%	
	admitted from ED	66.1%	
	Elixhauser	30.90	15.50
Patient factors (shift level)	average hourly LAPS	83.50	32.78
	DDCIntensity	2.19	0.62
	first 24-hour flag	7.6%	
	LOS so far (hours)	185.22	222.90
	relative LAPS	1.008	0.410
	relative DDCIntensity	1.004	0.287
	relative Elixhauser	1.014	0.531
Nursing factors	continuity of care flag	44.1%	
	rest time	4.80	6.15
	# of patients assigned per nurse	1.84	0.43
	# of nurses assigned	1.07	0.27

3. Econometric Model and Estimation

Our goal is to estimate the effect of nurses' historical workload on the discrepancy between the nurses' perceived workload and order-based workload, $Intensity\_Gap_{p,t}$. The hypothesis is that a heavier historical workload is associated with a larger $Intensity\_Gap_{p,t}$, possibly due to stress or fatigue. We model $Intensity\_Gap_{p,t}$ as

$$Intensity\_Gap_{p,t} = \theta \cdot Hist\_Work_{p,t} + X_{p,t}^{\top}\beta + u_{p,t}, \tag{3}$$

where $X_{p,t}$ denotes control variables specified in Section 2.4, $u_{p,t}$ is the unobserved determinant of the workload intensity gap for patient p in shift t. Throughout this section, we use the triple (p,n,t) to denote that nurse n is assigned to patient p in shift t.

We are interested in estimating θ, which measures the impact of nurses' historical workload on how much their perceived workload differs from the order-based workload. A naive estimation of (3) via ordinary least squares (OLS) may suffer from biases due to a number of estimation challenges that arise in our data and setting. In particular, there are four main identification challenges: missing outcome measures due to missing PJIntensity, which may not be missing at random; endogeneity due to omitted variables that are correlated with both $Intensity\_Gap_{p,t}$ and $Hist\_Work_{p,t}$; measurement error of $Hist\_Work_{p,t}$ due to missing admission shift assignment; and censored responses since the workload intensity scores only have four levels. We next describe these estimation challenges in more detail and the econometric approach we take to address them.

3.1. Sample Selection

In our data, 20% of the PJIntensity scores are missing, and thus we are unable to observe the main outcome variable $Intensity\_Gap_{p,t}$ for these observations. If the PJIntensity scores are not missing at random, it may lead to estimation bias. For example, since the PJIntensity scores are entered manually by the nurses, a nurse who has a very high perceived workload may be too busy to report the PJIntensity, resulting in an underestimate of the true treatment effect; alternatively, a nurse who has a low perceived workload may not bother to report the PJIntensity, resulting in an overestimate of the true treatment effect.

We address the potential sample selection bias using the Heckman selection model (Wooldridge 2010). Let $Report\_PJ_{p,t}$ denote the binary variable indicating whether the PJIntensity is reported for patient p in shift t. Note that the response $Intensity\_Gap_{p,t}$ is observable only if $Report\_PJ_{p,t} = 1$. We assume a probit model for sample selection:

$$Report\_PJ_{p,t} = 1(\tilde{X}_{p,t}^{\top}\alpha + v_{p,t} > 0), \tag{4}$$

where $\tilde{X}_{p,t}$ denotes the control variables which is a superset of $X_{p,t}$ and may include more variables; $v_{p,t}$ is assumed to follow a standard normal distribution that is independent of $\tilde{X}_{p,t}$. We estimate (3) and (4) jointly, assuming $u_{p,t}$ in (3) is also independent of $\tilde{X}_{p,t}$, and $u_{p,t}, v_{p,t}$ satisfy the following relation:

$$\mathbb{E}[u_{p,t}|v_{p,t}] = \gamma v_{p,t}.$$

The Heckman selection model introduces a correction term to account for sample selection. To see this, we first rewrite (3) as

$$\begin{aligned} Intensity\_Gap_{p,t} =& \theta \cdot Hist\_Work_{p,t} + X_{p,t}^{\top}\beta + \mathbb{E}[u_{p,t}|\tilde{X}_{p,t}, Report\_PJ_{p,t} = 1] \\ & - \mathbb{E}[u_{p,t}|\tilde{X}_{p,t}, Report\_PJ_{p,t} = 1] + u_{p,t} \\ =& \theta \cdot Hist\_Work_{p,t} + X_{p,t}^{\top}\beta + \mathbb{E}[u_{p,t}|\tilde{X}_{p,t}, Report\_PJ_{p,t} = 1] + e_{p,t}, \end{aligned}$$

where $e_{p,t} = u_{p,t} - \mathbb{E}[u_{p,t}|\tilde{X}_{p,t}, Report\_PJ_{p,t} = 1]$. It is easy to see that $\mathbb{E}[e_{p,t}|\tilde{X}_{p,t}, Report\_PJ_{p,t} = 1] = 0$. We also have

$$\begin{aligned} \mathbb{E}[u_{p,t}|\tilde{X}_{p,t}, Report\_PJ_{p,t} = 1] &= \mathbb{E}[\mathbb{E}[u_{p,t}|\tilde{X}_{p,t}, Report\_PJ_{p,t} = 1, v_{p,t}]|\tilde{X}_{p,t}, Report\_PJ_{p,t} = 1] \\ &= \mathbb{E}[\mathbb{E}[u_{p,t}|v_{p,t}]|\tilde{X}_{p,t}, Report\_PJ_{p,t} = 1] \\ &= \gamma\mathbb{E}[v_{p,t}|\tilde{X}_{p,t}, Report\_PJ_{p,t} = 1] \\ &= \gamma\mathbb{E}[v_{p,t}|v_{p,t} > -\tilde{X}_{p,t}\alpha] = \gamma\lambda(\tilde{X}_{p,t}\alpha). \end{aligned}$$

where the last step follows from the independence between $v_{p,t}$ and $\tilde{X}_{p,t}$ and $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio, ϕ and Φ are the probability density function and cumulative distribution

function of the standard normal distribution respectively. As a result, we obtain the following reduced-form model of $Intensity\_Gap_{p,t}$ when $Report\_PJ_{p,t} = 1$:

$$Intensity\_Gap_{p,t} = \theta \cdot Hist\_Work_{p,t} + X_{p,t}^{\top}\beta + \gamma \cdot \lambda(\tilde{X}_{p,t}^{\top}\alpha) + e_{p,t}. \tag{5}$$

3.2. Endogeneity

Like many observational studies, our estimation can suffer from omitted variable bias. In particular, there can be unobserved variables that are correlated with both nurses' historical workload and their perceived workload. For example, a more experienced nurse is more likely to be assigned to a high-acuity patient and, thus, is more likely to have a higher historical workload. Meanwhile, with extensive experience, this nurse may be less likely to report a higher perceived workload relative to the order-based workload. In this case, the unobservable, captured in $u_{p,t}$, is negatively correlated with $Hist\_Work_{p,t}$ but positively correlated with $Intensity\_Gap_{p,t}$. Direct estimation of (3) is likely to result in an underestimation of θ. Alternatively, a more experienced nurse, who is more likely to be assigned to a high-acuity patient, may be better at characterizing the patient's true needs, which may, in turn, leads to a higher reported (perceived) workload. In this case, $u_{p,t}$ is positively correlated with both $Hist\_Work_{p,t}$ and $Intensity\_Gap_{p,t}$. Direct estimation of (3) may result in an overestimation of θ. Note that it is a priori unclear which of the above scenarios is more likely to occur or has a larger impact.

To address the endogeneity issue, we apply the instrumental variable (IV) strategy (Wooldridge 2010). A valid IV in our case must be correlated with $Hist\_Work_{p,t}$, i.e., the relevance condition; but has no direct effect on $Intensity\_Gap_{p,t}$ other than through its effect on $Hist\_Work_{p,t}$, i.e., the exclusion restriction. The IVs we propose are the average workload, measured by DDCIntensity and LAPS respectively, at the unit level, excluding the focal patient, of one shift prior to the focal nurse's last assignment shift. For example, consider the triple (n, p, t). Suppose the last assignment shift of nurse n prior to shift t is shift $t - k_n$. Recall that P_{t-k_n-1} denotes the set of patients who were in the ICU in shift $t - k_n - 1$. Then, the two IVs can be expressed as:

$$Avg\_Last\_DDC_{p,t} = \frac{1}{|P_{t-k_n-1} \setminus \{p\}|} \sum_{i \in P_{t-k_n-1} \setminus \{p\}} DDCIntensity_{i,t-k_n-1}, \tag{6}$$

$$Avg\_Last\_LAPS_{p,t} = \frac{1}{|P_{t-k_n-1} \setminus \{p\}|} \sum_{i \in P_{t-k_n-1} \setminus \{p\}} LAPS_{i,t-k_n-1}. \tag{7}$$

Note that when the unit is busy with high workload intensity patients in shift $t - k_n - 1$, nurse n is more likely to experience a higher workload in shift $t - k_n$, i.e., the relevance condition is likely to hold. On the other hand, since we exclude the focal patient p when calculating the IVs and nurse n did not work in shift $t - k_n - 1$ (no nurse works in two consecutive shifts in our data), the IVs

are unlikely to affect the workload intensity for patient p in shift t, other than through its effect on $Hist\_Work_{p,t}$. Thus, the exclusion restriction is also likely to be satisfied. With the IVs, we have

$$\begin{aligned} Intensity\_Gap_{p,t} &= \theta \cdot Hist\_Work_{p,t} + X_{p,t}^{\top}\beta + u_{p,t}, \\ Hist\_Work_{p,t} &= X_{p,t}^{\top}\delta + \eta_1 \cdot Avg\_Last\_DDC_{p,t} + \eta_2 \cdot Avg\_Last\_LAPS_{p,t} + \varepsilon_{p,t}, \end{aligned} \tag{8}$$

which can be estimated via two-stage least squares (2SLS).

In practice, we can use data to verify the relevance condition, but cannot assert the exclusion restriction, in general. Specifically, the IVs may be correlated with other unobservables. In Section 5.1, we conduct a sensitivity analysis of the exclusion restriction and show that a mild violation of the exclusion restriction is unlikely to affect the statistical significance of our estimation results.

Combining the IVs with the Heckman selection model, we have

$$\begin{aligned} Intensity\_Gap_{p,t} &= \theta \cdot Hist\_Work + X_{p,t}^{\top}\beta + \gamma \cdot \lambda(\tilde{X}_{p,t}^{\top}\alpha) + e_{p,t}, \\ Hist\_Work_{p,t} &= X_{p,t}^{\top}\delta + \eta_1 \cdot Avg\_Last\_DDC_{p,t} + \eta_2 \cdot Avg\_Last\_LAPS_{p,t} + \varepsilon_{p,t}, \\ Report\_PJ_{p,t} &= 1(\tilde{X}_{p,t}^{\top}\alpha + v_{p,t} > 0), \end{aligned} \tag{9}$$

where the covariate $\tilde{X}_{p,t}$ contains $X_{p,t}$, the two IV's: $Avg\_Last\_DDC_{p,t}$, $Avg\_Last\_LAPS_{p,t}$, and $Careful\_Nurse_t$, which is defined as the number of 'careful' nurses working in shift t. A nurse is referred to as being careful if less than 20% of his/her/their assignment PJIntensity is missing over the study period. We hypothesize that if there are more careful nurses working in a shift, they may positively influence other nurses, making them less likely to forget to report their PJIntensity scores as well. This can lead to an overall lower missing rate for PJIntensity in that shift. On the other hand, the number of careful nurses is unlikely to be correlated with factors that may influence the nurse's perceived workload, and thus we can exclude it from the outcome equation.

Model (9) is our main model. A two-stage least-square estimation will be used to estimate θ. We make a few remarks about the estimation of (9). First, all exogenous variables appear in the selection equation, i.e., $\tilde{X}_{p,t}$. We believe that these patient and shift level factors are likely to impact both the selection and our main outcome, $Intesity\_Gap$. An added benefit is that by making the covariates in $\tilde{X}_{p,t}$ a superset of the covariates in $X_{p,t}$, we do not need to make additional assumptions on the excluded variables to ensure the exclusion restriction on the reduced form equations is satisfied. Second, in theory, $\tilde{X}_{p,t}$ can be the same as $(X_{p,t}, Avg\_Last\_DDC_{p,t}, Avg\_Last\_LAPS_{p,t})$. The identifiability holds due to the nonlinearity of the inverse Mills ratio. However, $\lambda(\tilde{X}_{p,t}^{\top}\hat{\alpha})$ can sometimes be approximated well by a linear combination of $\tilde{X}_{p,t}$. In this case, setting $\tilde{X}_{p,t} = (X_{p,t}, Avg\_Last\_DDC_{p,t}, Avg\_Last\_LAPS_{p,t})$ can lead to a multi-collinearity issue, which is known as "variance inflation" (Chapter 19.6 of Wooldridge (2010)). Thus, we introduce the variable $Careful\_Nurse_t$, which is likely to be correlated with $Report\_PJ_{p,t}$, but uncorrelated with $PJIntensity_{p,t}$ to ensure numerical stability.

3.3. Measurement Error

The treatment variable $Hist\_Work_{p,t}$, which is defined as the maximum DDCIntensity of patients assigned to the focal nurse in his/her previous working shift, may incur a measurement error due to the missing workload information for patients in admission shift. In particular, consider the triple (p,n,t), when calculating $Hist\_Work_{p,t}$, which corresponds to nurse n's workload in shift $t-k_n$, we exclude patients who are admitted in shift $t-k_n$ due to a large amount of missing assignment information for these patients. As such, the observed historical workload may be an underestimate of the nurse's true historical workload. To gauge the effect of this measurement error, we conduct extensive sensitivity analyses using different imputation schemes for the admission shift assignments.

We start by discussing the potential impacts of the measurement error. Let $Hist\_Work^*_{p,t}$ denote the true historical workload of the nurses taking care of patient p in shift t, which may contain the workload associated with some assigned patient(s) who is/are admitted in the nurse's last assignment shift, i.e., shift $t-k_n$. Recall that $Hist\_Work_{p,t}$ is the historical workload we observe, which does not contain the workload associated with the newly admitted patients. We define

$$\Delta Hist\_Work_{p,t} = Hist\_Work^*_{p,t} - Hist\_Work_{p,t}.$$

Then, by (8), we have

$$\begin{aligned} & Hist\_Work^*_{p,t} \\ =& Hist\_Work_{p,t} + \Delta Hist\_Work_{p,t} \\ =& X_{p,t}^\top \delta + \eta_1 \cdot Avg\_Last\_DDC_{p,t} + \eta_2 \cdot Avg\_Last\_LAPS_{p,t} + \varepsilon_{p,t} + \Delta Hist\_Work_{p,t}, \end{aligned} \tag{10}$$

and the outcome equation takes the form:

$$Intensity\_Gap_{p,t} = \theta \cdot Hist\_Work^*_{p,t} + X_{p,t}^\top \beta + \gamma \lambda(\tilde{X}_{p,t}^\top \alpha) + e_{p,t},$$

where the noise $e_{p,t}$ has mean zero conditional on $X_{p,t}$ and $Report\_PJ_{p,t} = 1$. By plugging (10) in the outcome equation, we obtain

$$\begin{aligned} Intensity\_Gap_{p,t} =& X_{p,t}^\top(\beta + \theta\delta) + \gamma \cdot \lambda(\tilde{X}_{p,t}^\top \alpha) + \theta\eta_1 \cdot Avg\_Last\_DDC_{p,t} + \theta\eta_2 \cdot Avg\_Last\_LAPS_{p,t} \\ & + (e_{p,t} + \theta\varepsilon_{p,t}) + \theta \cdot \Delta Hist\_Work_{p,t}. \end{aligned}$$

To simplify the discussion, we assume that there is only a single IV, which is denoted as $IV_{p,t}$, and positively correlated with $Hist\_Work_{p,t}$, i.e, the IV has coefficient $\theta\eta$ with $\eta > 0$ in outcome equation. We further assume that the effect of covariates $Z_{p,t} := (X_{p,t}, \lambda(\tilde{X}_{p,t}^\top \alpha))$ is negligible, i.e.,

$$Intensity\_Gap_{p,t} = \beta_0 + \theta\eta \cdot IV_{p,t} + \theta \cdot \Delta Hist\_Work_{p,t} + (e_{p,t} + \theta\varepsilon_{p,t}).$$

In this case, when using 2SLS to estimate the treatment effect, the estimator $\hat{\theta}$ converges to

$$\theta + \frac{Cov(IV_{p,t}, e_{p,t} + \theta\varepsilon_{p,t} | Report\_PJ_{p,t} = 1)}{\eta Var(IV_{p,t} | Report\_PJ_{p,t} = 1)} + \theta \cdot \frac{Cov(IV_{p,t}, \Delta Hist\_Work_{p,t} | Report\_PJ_{p,t} = 1)}{\eta Var(IV_{p,t} | Report\_PJ_{p,t} = 1)}. \tag{11}$$

We next show that the second term in (11) is equal to zero. First, $Cov(IV_{p,t}, \varepsilon_{p,t} | Report\_PJ_{p,t} = 1) = 0$ because by assumption, $IV_{p,t}$ is independent with $\varepsilon_{p,t}$. In addition,

$$\begin{aligned}
&Cov(IV_{p,t}, e_{p,t} | Report_PJ_{p,t} = 1) \\
=&E[IV_{p,t} \cdot e_{p,t} | Report_PJ_{p,t} = 1] - E[IV_{p,t} | Report_PJ_{p,t} = 1] \cdot E[e_{p,t} | Report_PJ_{p,t} = 1] \\
=&E[IV_{p,t} \cdot (u_{p,t} - E[u_{p,t} | \tilde{X}_{p,t}, Report_PJ_{p,t} = 1]) | Report_PJ_{p,t} = 1] \\
=&E[IV_{p,t} \cdot u_{p,t} | Report_PJ_{p,t} = 1] - E[E[IV_{p,t} \cdot u_{p,t} | \tilde{X}_{p,t}, Report_PJ_{p,t} = 1] | Report_PJ_{p,t} = 1] \\
=&0,
\end{aligned}$$

where the last equation holds because $IV_{p,t}$ is contained in $\tilde{X}_{p,t}$. Then, we incur an estimation bias of

$$\theta \cdot \frac{Cov(IV_{p,t}, \Delta Hist\_Work_{p,t} | Report\_PJ_{p,t} = 1)}{\eta Var(IV_{p,t} | Report\_PJ_{p,t} = 1)},$$

which is the third term in (11)

It is easy to see that when $IV_{p,t}$ and $\Delta Hist\_Work_{p,t}$ are negatively correlated, we underestimate the treatment effect. Otherwise, we overestimate the treatment effect. When considering the effect of other covariates in $X_{p,t}$, the above rationale should still hold by considering the residuals of $IV_{p,t}$ and $Admin\_Work_{p,t}$ regressed on $Z_{p,t}$ respectively. There, the estimation bias takes the form

$$\theta \cdot \frac{Cov(Res_{Z_{p,t}}(IV_{p,t}), Res_{Z_{p,t}}(\Delta Hist\_Work_{p,t}) | Report\_PJ_{p,t} = 1)}{\eta Var(Res_{Z_{p,t}}(IV_{p,t}) | Report\_PJ_{p,t} = 1)},$$

where the residual operator $Res_Z(\cdot)$ is defined as $Res_Z(Y) = Y - Z^\top (E[ZZ^\top]^{-1} E[ZY])$.

We next introduce the three schemes to impute the missing admission-shift assignments. The idea is to create different levels of correlation between $IV_{p,t}$ and $\Delta Hist\_Work_{p,t}$. Consider the triple (p, n, t) where nurse n's previous assignment shift is shift $t - k_n$. Let $t' = t - k_n$. We first impute the missing admission assignments in shift t' and then calculate $Hist\_Work^*_{p,t}$ based on the full assignment information for nurse n in shift t'. Recall that $P_{t'}(n)$ is the set of non-admission shift patients that nurse n is taking care of in shift t'. Let

$$Non\_Admin\_Work_{n,t'} = \max_{p' \in P_{t'}(n)} \{DDCIntensity_{p',t'}\}$$

denotes nurse n's non-admission assignment workload in shift t'. Note that $Non\_Admin\_Work_{n,t'} = Hist\_Work_{p,t}$, if nurse n is the only nurse assigned to patient p in shift t. When imputing the missing

admission assignments in shift t', we assume a nurse can be assigned at most one admission-shift patient. In the first scheme, we try to match an admission-shift patient with a higher DDCIntensity with a nurse with a higher non-admission assignment workload. This creates a positive correlation between $IV_{p,t}$ and $\Delta Hist\_Work_{p,t}$. In the second scheme, we try to match an admission-shift patient with a higher DDCIntensity with a nurse with a lower non-admission assignment workload. In the third scheme, we randomly match the admission-shift patients with nurses (see Algorithm 1 in Appendix B for more details).

Let $AP_{t'}(n)$ denote the set of admission patients assigned to nurse n in shift t'. After imputing the missing admission assignments, we consider two measures of the true historical workload. In the first one, we define $Hist\_Work^*_{p,t}$ as the maximum measure:

$$Hist\_Work^*_{p,t} = \frac{1}{|N_t(p)|} \cdot \sum_{n \in N_t(p)} \max_{q \in (P_{t-k_n}(n) \cup AP_{t-k_n}(n))} \{DDCIntensity_{q,t-k_n}\} \cdot 1(k_n \leq 14), \quad (12)$$

In the second one, we define $Hist\_Work^*_{p,t}$ as:

$$\begin{aligned} Hist\_Work^*_{p,t} =& \frac{1}{|N_t(p)|} \cdot \sum_{n \in N_t(p)} \max_{q \in P_{t-k_n}(n)} \{DDCIntensity_{q,t-k_n}\} \cdot 1(k_n \leq 14) \\ &+ \frac{1}{|N_t(p)|} \cdot \sum_{n \in N_t(p)} \max_{q \in AP_{t-k_n}(n)} \{DDCIntensity_{q,t-k_n}\} \cdot 1(k_n \leq 14), \end{aligned} \quad (13)$$

which we refer to this as the additive version. The estimation results for the sensitivity analysis can be found in Section 5.2.

3.4. Censoring

Our fourth empirical challenge is that the outcome variable $Intensity\_Gap$ may be censored since the workload intensity scores can only take four different values/levels. For example, when a patient's DDCIntensity is 4, the nurse cannot report a higher PJIntensity, which leads to a censored outcome measure.

Suppose we can remove the cap at 4 and the floor at 1. Let $PJIntensity^*_{p,t}$ be the true PJIntensity the nurse wants to report and define $Intensity\_Gap^*_{p,t} = PJIntensity^*_{p,t} - DDCIntensity_{p,t}$. We then have

$$\begin{aligned} Intensity\_Gap^*_{p,t} &= \theta \cdot Hist\_Work_{p,t} + X_{p,t}^\top \beta + \gamma \lambda(\tilde{X}_{p,t}^\top \alpha) + e_{p,t}, \\ Hist\_Work_{p,t} &= X_{p,t}^\top \delta + \eta_1 \cdot Avg\_Last\_DDC_{p,t} + \eta_2 \cdot Avg\_Last\_LAPS_{p,t} + \varepsilon_{p,t}, \\ Intensity\_Gap_{p,t} &= \mathcal{C}(PJIntensity^*_{p,t}) - DDCIntensity_{p,t} \end{aligned}$$

where the censoring operator $\mathcal{C}$ is defined as $\mathcal{C}(x) = 1 \cdot (x < 1) + x \cdot 1(1 \leq x \leq 4) + 4 \cdot (x > 4)$. Let $\Delta Gap_{p,t} = Intensity\_Gap_{p,t} - Intensity\_Gap^*_{p,t}$, which is the gap between the censored response and true response. Then,

$$Intensity\_Gap_{p,t} = \theta \cdot Hist\_Work_{p,t} + X_{p,t}^\top \beta + \gamma \lambda(\tilde{X}_{p,t}^\top \alpha) + e_{p,t} + \Delta Gap_{p,t}$$

$$
\begin{aligned}
=&X_{p,t}^{\top}(\beta+\theta\delta)+\gamma\cdot\lambda(\tilde{X}_{p,t}^{\top}\alpha)+\theta\eta_1\cdot Avg\_Last\_DDC_{p,t}+\theta\eta_2\cdot Avg\_Last\_LAPS_{p,t}\\
&+(e_{p,t}+\theta\varepsilon_{p,t})+\Delta Gap_{p,t}.
\end{aligned}
$$

where the noise terms satisfy

$$
\mathbb{E}[e_{p,t}+\theta\varepsilon_{p,t}|Z_{p,t},Avg\_Last\_DDC_{p,t},Avg\_Last\_LAPS_{p,t},Report\_PJ_{p,t}=1]=0,
$$

and $Z_{p,t}=(X_{p,t},\lambda(\tilde{X}_{p,t}^{\top}\alpha))$. Again, to simplify the discussion, we assume there is only one IV, which is denoted as $IV_{p,t}$ with corresponding coefficient $\theta\eta$ for $\eta>0$ (i.e., the IV is positively correlated with $Hist_Work+p,t$), and the effect of the covariates $Z_{p,t}$ is negligible. Similar to the analysis in Section 3.3, when using 2SLS to estimate the treatment effect, the estimator $\hat{\theta}$ converges to

$$
\theta+\theta\cdot\frac{Cov(IV_{p,t},\Delta Gap_{p,t}|Report\_PJ_{p,t}=1)}{\eta Var(IV_{p,t}|Report\_PJ_{p,t}=1)}.
$$

Note that $\mathcal{C}(x)-x$ is a monotonically decreasing function of x. Since $IV_{p,t}$ is positively correlated with $Hist\_Work_{p,t}$, which is hypothesized to have a positive impact on $PJIntensity^{*}_{p,t}$, we have $Cov(IV_{p,t},\Delta Gap_{p,t}|Report\_PJ_{p,t}=1)<0$. This implies that the censored response leads to an underestimation of the treatment effect. As such, we do not explicitly correct for this potential bias, but note that it may result in *conservative* estimates of the true treatment effect. In particular, if we find a positive and significant effect of $Hist\_Work_{p,t}$ on $Intensity\_Gap_{p,t}$, this implies the true treatment effect is positive and significant, possibly with an even larger effect size. We do some additional analysis to account for the potential censoring bias in Section 5.3.2 when we conduct stratified analysis for heterogeneous treatment effects.

4. Estimation Results

Table 4 summarizes the estimation results based on our main model (9). We also provide the estimation results without the Heckman selection part, i.e., using IVs only (**IVs-only**) based on (8), and direct estimation of (3) using OLS (**OLS**).

We observe from the main model estimation that a higher historical workload is associated with a higher deviation of the nurse's perceived workload from the order-based workload. In particular, one level of increase in historical workload is associated with a 0.629 level higher *Intensity_Gap*. To provide the context of our estimated effects, consider an average patient (i.e., a patient whose features take the average value of the covariates) with an average order-based workload, i.e., with DDCIntensity equal to 2. If the nurse who takes care of the patient has a historical workload of 3 rather than 2, the nurse will perceive a high workload, i.e., a PJIntensity equal to 3. If the nurse who takes care of the patient has a historical workload of 4 rather than 2, the nurse will perceive extremely high workload, i.e., a PJIntensity equal to 4. Alternatively, if we increase all nurses'

historical workload from 2 to 3, this will increase the proportion of patient-shifts with an extremely high perceived workload from 0.9% to 11.5% based on the patient characteristics in our data.

We see in both the main and IVs-only models that the coefficients of the IVs are positive and significant, which verifies the relevance condition. In the main model, the coefficient for the inverse Mills ratio, which captures the correlation between sample selection and outcome, is not significantly different from zero. This suggests that sample selection is unlikely to cause significant estimation bias in our setting. Indeed, we note that the estimated treatment effects are quite similar in magnitude and significance for the main and IVs-only models. When comparing the main and IVs-Only models to OLS, we note that direct estimation of (3) leads to an underestimation of the treatment effect due to the endogeneity of $Hist\_Work$. This supports one of the potential endogeneity mechanisms discussed in Section 3.2.

Table 4 Estimation Results: A Comparison under Different Models

	Estimation Model		
	Main Model	IVs-Only	OLS
$Hist\_Work$	0.629^{***}	0.678^{***}	0.011
	(0.010)	(0.120)	(0.008)
$Last\_Avg\_DDC$	0.289^{***}	0.252^{***}	
	(0.041)	(0.047)	
$Last\_Avg\_LAPS$	0.405^{***}	0.448^{***}	
	(0.081)	(0.090)	
Inverse Mills Ratio	0.193		
	(0.171)		
Observations	12625	10101	10101

standard error in parentheses
$^{*}\,p < 0.05, \quad ^{**}\,p < 0.01, \quad ^{***}\,p < 0.001$

5. Sensitivity Analysis

We conduct several sensitivity analyses to verify the robustness of our estimation results, focusing particularly on endogeneity and measurement error. We focus on these two estimation challenges because 1) we did not find evidence that selection bias was significantly impacting our results and 2) our discussion in Section 3.4 suggests the estimates in Table 4 are conservative when the outcome variable is censored.

5.1. Sensitivity Analysis on the Validity of the IV's

As discussed in Section 3.2, it is in general hard to verify the exclusion restriction for the IV's. The main concern is that there can be unobserved patient or nurse characteristics that are correlated with both the instruments and the outcome. For example, when the MICU is busy with a lot of high-acuity patients, the nursing manager may tend to schedule more experienced nurses; these nurses may also be less likely to report a higher clinically perceived workload than the order-based workload because their experience has provided them the skills to navigate these high-stress shifts. Alternatively, the more experienced nurses may be more likely to report a higher clinically perceived as their experience provide them with a better assessment of the workload. In addition, when the MICU is busy, the hospital may be more selective in terms of who to admit to the ICU, i.e., they may tend to reserve the ICU beds for the more severe patients. Severe patients are also more likely to be associated with a higher perceived workload. In this section, we conduct a sensitivity analysis of the IVs regarding the exclusion restriction.

Following Baiocchi et al. (2014), suppose there exists an unobserved confounder $\zeta_{p,t}$ with mean zero and variance one that is correlated with the outcome $Intensity\_Gap_{p,t}$ and the IV's, i.e., $Avg\_Last\_DDC_{p,t}$ and $Avg\_Last\_LAPS_{p,t}$, but is uncorrelated with the other measurable covariates $X_{p,t}$. We assume a linear model for this relation:

$$\begin{aligned} Intensity\_Gap_{p,t} &= \theta \cdot Hist\_Work_{p,t} + X_{p,t}^{\top}\beta + \gamma \cdot \lambda(\tilde{X}_{p,t}^{\top}\alpha) + \psi\zeta_{p,t} + e_{p,t}, \\ \zeta_{p,t} &= \xi_1 \cdot Avg\_Last\_DDC_{p,t} + \xi_2 \cdot Avg\_Last\_LAPS_{p,t} + \omega_{p,t}, \end{aligned} \tag{14}$$

where the noise terms $e_{p,t}$ and $\omega_{p,t}$ satisfy

$$\begin{aligned} &\mathbb{E}[e_{p,t}|X_{p,t}, Report\_PJ_{p,t} = 1, Avg\_Last\_DDC_{p,t}, Avg\_Last\_LAPS_{p,t}] = 0, \\ &\mathbb{E}[\omega_{p,t}|X_{p,t}, Report\_PJ_{p,t} = 1, Avg\_Last\_DDC_{p,t}, Avg\_Last\_LAPS_{p,t}] = 0. \end{aligned}$$

Note that ψ, ξ_1, and ξ_2 are sensitivity parameters, where ψ measures the effect of a one standard deviation increase in the unobserved confounder on $Intensity\_Gap_{p,t}$, and ξ_1, ξ_2 measure the effect of one unit of increase in the IVs on the unobserved confounder respectively. Under this model, if we can control for ζ, Avg_Last_DDC and Avg_Last_LAPS would be valid IVs. Thus, in the sensitivity analysis, we treat

$$Intensity\_Gap_{p,t} - (\psi\xi_1 \cdot Avg\_Last\_DDC_{p,t} + \psi\xi_2 \cdot Avg\_Last\_LAPS_{p,t})$$

as the modified outcome variable and apply 2SLS to estimate θ. In particular, the outcome equation takes the form

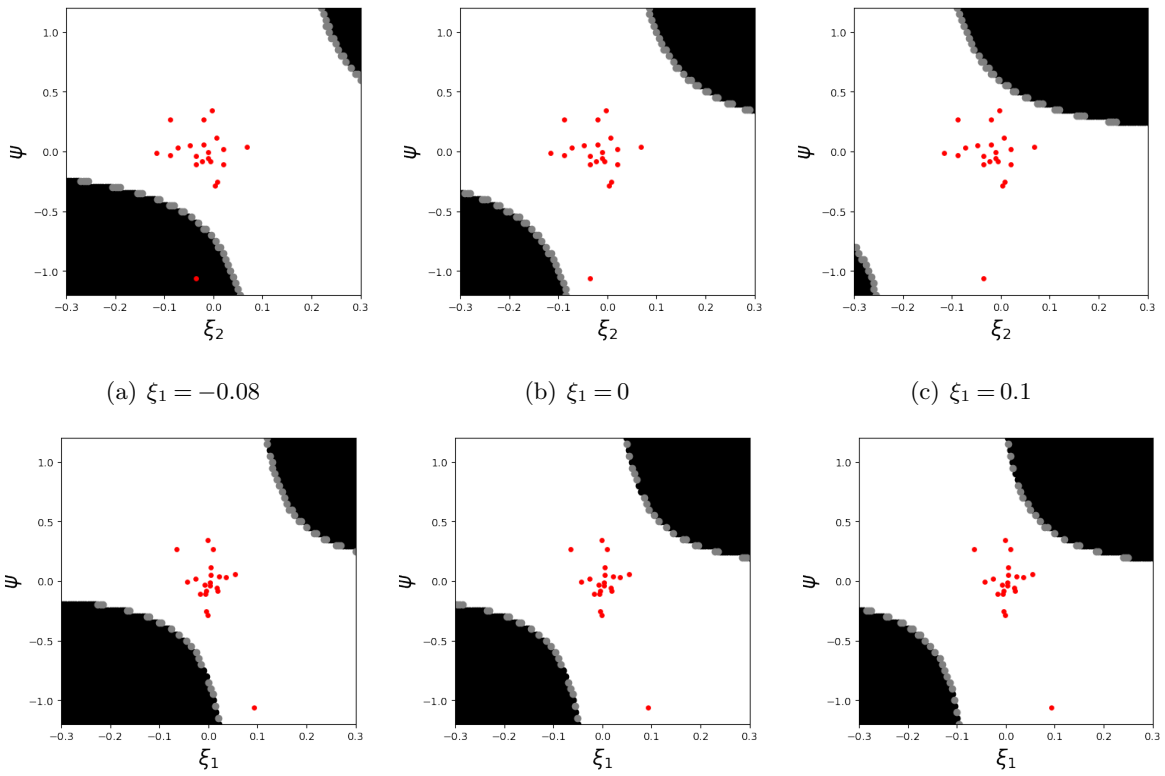
$$\begin{aligned} &Intensity\_Gap_{p,t} - (\psi\xi_1 \cdot Avg\_Last\_DDC_{p,t} + \psi\xi_2 \cdot Avg\_Last\_LAPS_{p,t}) \\ &= \theta \cdot Hist\_Work_{p,t} + X_{p,t}^{\top}\beta + \gamma \cdot \lambda(\tilde{X}_{p,t}^{\top}\alpha) + (e_{p,t} + \psi\omega_{p,t}), \end{aligned}$$

where $(e_{p,t} + \psi\omega_{p,t})$ has mean zero conditional on the controls $X_{p,t}$, $\lambda(\tilde{X}_{p,t}^{\top}\alpha)$, $Avg\_Last\_DDC_{p,t}$, $Avg\_Last\_LAPS_{p,t}$, and $Report\_PJ_{p,t} = 1$. We estimate the model with different values of ψ, ξ_1, and ξ_2 to determine the parameter regimes where the estimate of θ is a) positive and statistically significant at the 5% level, b) negative and statistically significant at the 5% level, or c) statistically not significantly different than 0 at the 5% level. If the regime where a) holds is large, we are more confident that our estimate for θ is reasonably robust against the potential violation of exclusion restriction for the IVs.

Figure 2 summarizes the results of our sensitivity analysis. For this analysis, we standardize all the control variables (i.e., subtract the mean and normalized by the standard deviation accordingly) to ensure that the coefficients are comparable. In the first row of Figure 2, we set $\xi_1 = -0.08, 0, 0.1$, where -0.08 and 0.1 correspond to the minimal and maximal coefficients when regressing the control variables X against Avg_Last_DDC, and vary the values of ξ_2 and ψ. In the second row of Figure 2, we set $\xi_2 = -0.12, 0, 0.08$, where -0.12 and 0.08 correspond to the minimal and maximal coefficients when regressing the control variables X against Avg_Last_LAPS respectively, and vary the values of ξ_1 and ψ. The white region depicts the area where the estimate for θ is positive and statistically significant at the 5% level. The grey region is where the estimate for θ is statistically not different than 0, and the black region is where the estimate for θ is negative and statistically significant. We observe that the white region is quite large. The dots in the plot correspond to the (ξ_1, ψ)-values (in the first row) or (ξ_2, ψ)-values (in the second row) for all observed covariates. Note that most of the dots are in the white region, suggesting that in order for there to be an unobserved confounder that would explain away our result, the effect size of this unobserved confounder would have to be much larger than those of most of the observed covariates. The only dot that falls into the black region is the dot corresponding to DDCIntensity in Figure 2 (a). DDCIntensity is known to be highly correlated with the outcome, and it is quite unlikely that the unobserved confounder has as large an effect on the outcome as DDCIntensity.

5.2. Sensitivity Analysis for Measurement Errors

In this section, we conduct the sensitivity analysis with respect to the measurement error. We consider the three imputations schemes introduced in Section 3.3: The first one aims to create a positive correlation between the IVs and $\Delta Hist\_Work_{p,t}$ (Positive Cor); The second aims to create a negative correlation between the two (Negative Cor); The third use random assignment (No Cor). Tables 5 and 6 summarize the estimation results under the three imputation schemes for two slightly different measures of $Hist\_Work^{*}_{p,t}$, i.e., the maximum measure (12) and additive measure (13) respectively. We observe when the IVs and $\Delta Hist\_Work_{p,t}$ are positively correlated, we see a smaller treatment effect, i.e., if not taking the measurement effort into account, we overestimate



(a) $\xi_1 = -0.08$ (b) $\xi_1 = 0$ (c) $\xi_1 = 0.1$

(d) $\xi_2 = -0.12$ (e) $\xi_2 = 0$ (f) $\xi_2 = 0.08$

Figure 2 Sensitivity analysis for the validity of IVs: ψ is the impact of unobserved confounder ζ on the outcome; ξ_1 and ξ_2 are the sensitivity parameters of IVs Avg_Last_DDC and Avg_Last_LAPS on ζ; and the dots denote corresponding coefficients of the observed covariates.

the treatment effect. When the IVs and $\Delta Hist\_Work_{p,t}$ are negatively correlated, we see a larger treatment effect. This is consistent with our analysis in Section 3.3. Meanwhile, among all the scenarios studied, the estimated treatment effect is quite similar to the coefficient of $Hist\_Work$ in our main model estimation, i.e., 0.629, suggesting the robustness of the result.

Table 5 Effect of Measurement Error: Estimation with Imputed $Hist\_Work^*_{p,t}$ according to the maximum measure in (12)

	Treatment Correction Strategy		
	Positive Cor	No Cor	Negative Cor
$Hist\_Work^*$	0.632^{***}	0.656^{***}	0.658^{***}
	(0.010)	(0.010)	(0.010)
Observations	12625	12625	12625

standard error in parentheses
$^{*}\,p<0.05$, $^{**}\,p<0.01$, $^{***}\,p<0.001$

Table 6 Effect of Measurement Error: Estimation with Imputed $Hist\_Work^*_{p,t}$ according to the additive measure in (13)

	Treatment Correction Strategy		
	Positive Cor	No Cor	Negative Cor
$Hist\_Work^*$	0.561^{***}	0.585^{***}	0.649^{***}
	(0.007)	(0.008)	(0.011)
Observations	12625	12625	12625

standard error in parentheses
$^{*}\,p<0.05$, $^{**}\,p<0.01$, $^{***}\,p<0.001$

5.3. Alternative Model Specifications

We also consider alternative model specifications. In particular, a) we test different measures of historical workload, i.e., the treatment variable; and b) we check for heterogeneous treatment effects. We next provide more details about these analyses.

5.3.1. Historical Workload Measures We propose several alternative measures of nurses' historical workload. Recall that in our main model, $Hist\_Work_{p,t}$ is defined as the maximum of the DDCIntensity scores among all the patients assigned to the focal nurse in his/her last assignment shift. In this section, we consider four alternative measures of the historical workload: i) We take the sum, instead of the maximum, of the DDCIntensity scores among all the assigned patients in the focal nurse's last assignment shift. ii) We add an additional point to the max DDCIntensity score in the event the focal nurse is assigned more than one patient. For example, if nurse n is assigned two patients in her previous assignment shift with DDCItensity 2 and 3. The nurse's adjusted score would be $\max\{2,3\}+1=4$. iii) We take the maximum of the average LAPS (instead of the DDCIntensity) among all the patients assigned to the focal nurse in the nurse's last assignment shift. iv) We take the sum of the average LAPS among all the assigned patients. Table 7 summarizes the estimation results using the four alternative treatment variables. We observe that in all cases, the historical workload has a statistically significant positive effect on $Intensity\_Gap_{p,t}$. The coefficients of LAPS-based treatment variables are smaller than the DDCIntensity-based treatment variables. This is because the LAPS (mean 83.5 and standard deviation 32.7) are in general much larger than the DDCIntensity scores (mean 2.2 and standard deviation 0.62). When considering the effect of a one-standard-deviation increase of the treatment variable, we observe very similar magnitude in the treatment effects as demonstrated in the "standardized effect" row in Table 7.

We also use different look-back time windows to measure the historical workload and observe similar estimated treatment effect. See Table 10 in Appendix C for more details.

Table 7 Robustness Check: Alternative Specification of Treatment

	Treatment Specification				
	Max DDC	Sum DDC	Adjusted Max DDC	Max LAPS	Sum LAPS
$Hist\_Work$	0.629^{***}	0.298^{***}	0.566^{***}	0.014^{***}	0.008^{***}
	(0.010)	(0.005)	(0.009)	(0.0002)	(0.0001)
Standardized Effect	0.458	0.399	0.543	0.499	0.479
Observations	12625	12625	12625	12625	12625

standard error in parentheses
$^{*}\,p<0.05$, $^{**}\,p<0.01$, $^{***}\,p<0.001$
Standardized effect denotes the change of outcome when the treatment increases by one unit of standard deviation.

5.3.2. Heterogeneous Treatment Effect To examine potential heterogeneity in the treatment effect, we divide the observations into two subsets: high versus low-intensity patients. We define a focal patient-shift as high-intensity if its DDCIntensity is larger than or equal to 3, which according to the Optilink manual indicates a 'High' or 'Extreme' workload; otherwise, the patient-shifts are in the low-intensity group. Then, we estimate the impact of historical workload on the gap between the PJIntensity and DDCIntensity scores for the high and low-intensity groups separately.

For the high-intensity group, the estimated treatment effect is 0.532 (standard error 0.018); for the low-intensity group, the estimated treatment effect is 0.671 (standard error 0.013). One may have expected that the burden of historical workload is higher for high-intensity patients as multiple shifts of high workload likely would translate into more fatigue, which in turn increases the intensity gap. However, our results find the counter-intuitive relationship that the treatment effect is larger for the low-intensity group than for the high-intensity group. This may be due to the censoring bias discussed in Section 3.4. Recall that when the DDCIntensity is 4, the nurses cannot report a higher PJIntensity than the DDCIntensity.

To gauge the effect of the censoring bias, we conduct a sensitivity analysis. Since less than 5% of patient-shifts have a lower PJIntensity than DDCIntensity, we would expect the censoring bias to mostly impact the high-intensity group due to censoring from above, where PJIntensity is limited to 4. We assume the PJIntensity can be at most two levels above the DDCIntensity because we rarely see a gap larger than two (only 0.05% of the data). This implies that to impute the uncensored PJIntensity, we only need to focus on two scenarios: (i) The DDCIntensity is 3 and PJIntensity is 4, but the PJIntensity could have been equal to 5 if there were no censoring. (ii) Both the DDCIntensity and PJIntensity are 4, but the PJIntensity could have been equal to 5 (or even 6) if there were no censoring. In our data, there are 484 patient-shifts with DDCIntensity

equal to 3 and PJIntensity equal to 4; 98 patient-shifts with DDCIntensity and PJIntensity both equal to 4. We consider the following imputation schemes. For the 484 patient-shifts in scenario (i), we select $k\%$, $k = 0, 4, 8, 12, 16$, of them and increase their PJIntensity from 3 to 5. For the 98 patient-shifts in scenario (ii), we select $\ell\% \times 11.3\%$, $\ell = 0, 25, 50, 75, 100$, of them and increase their PJIntensity from 4 to 6. We also select $\ell\% \times (1 - 11.3\%)$ of them and increase their PJIntensity from 4 to 5. (11.3% is the empirical proportion of patient-shifts with an intensity gap of 2.) Since a higher historical workload is associated with a higher perceived workload, when "selecting" the patients, we prioritize observations with a larger historical workload (breaking ties randomly when the historical workload is the same).

The estimation results with the adjusted PJIntensity are summarized in Table 8. k and l capture the amount of censoring in the data. We observe that – as expected – as the adjusted proportion of observations goes up, i.e., as k or ℓ increases, we tend to see a larger treatment effect. When $k \geq 12\%$ and $\ell \geq 75\%$, the treatment effect for the high-intensity group is of a similar magnitude as the low-intensity group. Thus, it is possible that the treatment effects are similar for both the low- and high-intensity groups.

5.3.3. Heterogeneous Nurse-Specific Fixed Effect Lastly, to control for potential heterogeneity in nurse behaviors, we consider an alternative model that includes nurse-specific fixed effects. When multiple nurses are assigned to a single patient in a shift, we randomly pick one nurse to control for. Such a fixed effect captures how likely a specific nurse caring for the focal patient would upward adjust the DDCIntensity intrinsically. For this model, the estimated treatment effect is 0.601 with a standard error of 0.010, which is quite close to the result of the main model. In addition, the estimated fixed effect is quite small. Among the 72 unique nurses, 45 of them has a fixed effect that is not significantly different from zero at a 0.01 level. None of the nurse-specific fixed effect is significantly different from zero at 0.01 level. This suggests that the heterogeneity in nurse behaviors likely has limited impact on our results.

6. Counterfactual Analysis

Our empirical analysis shows that the historical workload has a positive impact on nurses' perceived workload relative to the order-based workload, i.e., a higher historical workload leads to a larger discrepancy between the PJIntensity and DDCIntensity. Since a high perceived workload can lead to increased mental strain for nurses and possibly increase patient safety concerns, it is important to balance the nursing workload temporally to avoid nursing burnout and achieve better quality of care.

Matching patients with nurses appropriately is important in the daily operations of the ICU. To better balance the nursing workload temporally, we want to avoid constantly "overloading" a

Table 8 Robustness Check: Impact of Censored Response in High Severity Group

% of adjustment when DDC=3, PJ=4	% of Adjustment When DDC=4, PJ=4				
	0%	25%	50%	75%	100%
0%	0.532^{***} (0.018)	0.557^{***} (0.019)	0.585^{***} (0.020)	0.611^{***} (0.020)	0.636^{***} (0.020)
4%	0.544^{***} (0.019)	0.565^{***} (0.020)	0.589^{***} (0.020)	0.614^{***} (0.019)	0.650^{***} (0.020)
8%	0.561^{***} (0.018)	0.587^{***} (0.019)	0.602^{***} (0.018)	0.637^{***} (0.021)	0.658^{***} (0.020)
12%	0.567^{***} (0.019)	0.610^{***} (0.019)	0.625^{***} (0.021)	0.657^{***} (0.021)	0.682^{***} (0.022)
16%	0.590^{***} (0.018)	0.605^{***} (0.019)	0.638^{***} (0.022)	0.660^{***} (0.022)	0.695^{***} (0.021)

standard error in parentheses
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

nurse with high-intensity patients. However, this might not always be achievable, since in addition to workload balancing, we also have to take the continuity of care into consideration. In particular, when matching patients and nurses, it is also desirable to maintain certain continuity of care, since continuity of care has been shown to be associated with better quality of care (Haggerty et al. 2003, Ahuja et al. 2020, 2022). In order to maintain continuity of care, it may be necessary to assign the same high-intensity patients to the same nurse in the nurse's consecutive working shifts. In this section, we propose a patient-to-nurse assignment policy that tries to balance the nursing workload and continuity of care.

Our assignment policy solves an integer program that maximizes the continuity of care reward minus the penalty of overloading the nurses. Specifically, for each nurse-patient pair (n, p) in each shift t, we define a reward for continuity of care, $COC_{n,p,t}$, which counts the number of shifts where patient p was assigned to nurse n prior to shift t. We also define a penalty for consecutive high workload shifts, $CHW_{n,p,t}$, which is equal to 1 if both patient p's DDCIntensity score and nurse n's historical workload are larger than or equal to 3, and is equal to 0 otherwise. Next, we define

$$\omega_{npt} = COC_{n,p,t} - \rho \cdot CHW_{n,p,t}, \tag{15}$$

where the parameter $0 \leq \rho \leq \infty$ measures the importance of workload balancing relative to continuity of care. Recall that N_t denotes the set of nurses working in shift t and P_t denotes the set of patients who are in the ICU in shift t. We also define $Patient\_Num_{n,t}$ as the number of patients

assigned to nurse n in shift t in the data, and $Nurse\_Num_{p,t}$ denote the number of nurses assigned to patient p in shift t in the data. Then, the integer program takes the form

$$\begin{aligned}
&\max \sum_{n\in N_t}\sum_{p\in P_t} \omega_{n,p,t}\cdot x_{n,p,t}\\
&\text{s.t.} \sum_{p\in P_t} x_{n,p,t}\cdot DDCIntensity_{p,t}\le 5, \forall n\in N_t\\
&\quad \sum_{n\in N_t} x_{npt} = Nurse_Num_{p,t}, \forall p\in P_t, \quad \sum_{p\in P_t} x_{npt} = Patient_Num_{nt}, \forall n\in N_t\\
&\quad x_{npt}\in\{0,1\}, \forall n\in N_t, p\in P_t,
\end{aligned} \tag{16}$$

where $x_{n,p,t}$ is the decision variable, with $x_{n,p,t}=1$ indicating patient p is assigned to nurse n in shift t. The first constraint requires that the sum of DDCIntensity among all the patients assigned to a nurse can not exceed 5. As a result, we cannot assign two patients with DDCIntensity≥ 3 to the same nurse in a shift. We impose this constraint in order to avoid assigning too heavy a workload to a nurse in a shift. The second set of constraints requires that we maintain the same patient-to-nurse ratio as in the data for each patient and each nurse. This set of constraints is imposed to facilitate a more fair comparison between our proposed assignment policy and the status quo in the hospital. It can be relaxed or replaced with other constraints in practice.

To evaluate the effectiveness of the patient-to-nurse assignment policy, we consider the following performance metrics. For continuity of care, which is measured at the patient level, we adopt two classic measures in the medical literature. The first one is the Herfindahl-Hirschman index HHI, which is also used as a measure of market concentration in economics (Eriksson and Mattsson 1983). It is defined as the sum of the squared proportion of time each nurse cares for the patient during his/her stay. For example, suppose patient p stays in ICU for 10 shifts and is assigned to nurse n_1, n_2, and n_3 for 3, 3, and 4 shifts respectively. Then $HHI_p = 0.3^2 + 0.3^3 + 0.4^2 = 0.34$. The other continuity of care measure is one minus the ratio between the number of unique nurses assigned to the patient during his/her stay and his/her length of stay (Jee and Cabana 2006). We refer to this index as CI. In the previous example, $CI_p = 1 - 3/10 = 0.7$. Note that both HHI_p and CI_p take value in the interval $[0,1]$ and a larger value implies higher continuity of care. We take the average of HHI_p and CI_p among all patients as two metrics of continuity of care. For workload balancing, which is measured at the nurse level, we define the workload balancing index WBI as the proportion of time a nurse experiences a high workload (DDCIntensity larger than or equal to 3) in both the focal shift and the prior working shift. For example, suppose nurse n works in 5 shifts in total and the corresponding maximal DDCIntensity scores are $2,3,3,2,3$, respectively. Then $WBI_n = 1/5$. We take the average of WBI over all nurses as a performance metric of workload balancing. Note that WBI takes value in the interval $[0,1]$ and a smaller value implies a more balanced workload. We also calculate the predicted PJIntensity using our empirical estimation

under different patient-to-nurse assignment policies and count the number of patient-shifts that has a predicted PJIntensity equal to 4, i.e., extremely high perceived workload.

We test the performances of our integer-programming-based patient-to-nurse assignment policy for different values of ρ, i.e., $\rho = 10, 5, 1.75$, corresponding to a high, medium, and low weight on the workload balancing penalty relative to the continuity of care reward in the objective in (15), respectively. The results are summarized in Table 9. We also list the performance metrics under the hospital's current patient-to-nurse assignment policy. First, we observe that for all weights, the workload balance measures improve; this is also true for the continuity of care measures, except for the high WB weight under the average CI measure. This suggests there are real opportunities for operational improvement through careful balancing of the nursing workload and continuity of care when making patient-to-nurse assignment decisions. We also observe that as we put more weight on workload balancing, the nursing workload metrics improve, i.e., the average WBI decreases, and the number of shifts with PJIntensity=4 decreases, but the continuity of care metrics decrease slightly. When we put a high weight on workload balancing, we are able to reduce the average WBI by 53% and the number of patient-shifts with PJIntensity equal to 4 by 22%, while maintaining the same level of continuity of care as the current hospital policy. When we put a low weight on workload balancing, we are able to improve the average HHI by 6% and the average CI by 7%, while reducing the average WBI by 41% and maintaining a similar (slightly smaller) number of patient-shifts with PJIntensity equal to 4.

Table 9 Counterfactual Analysis: A Comparison of Different Policies

	continuity of care metrics		workload balancing metrics	
	average HHI	average CI	average WBI	# $PJIntensity = 4$
High WB weight	0.307	0.320	5.35%	617
Medium WB weight	0.313	0.326	5.66%	654
Low WB weight	0.319	0.346	6.74%	774
Current hospital policy	0.302	0.323	11.40%	788

HHI is the sum of the squared proportion of time each nurse cares for the patient. CI is one minus the ratio between the number of unique nurses assigned to the patient during his/her stay and his/her length of stay. WBI is the proportion of time that a nurse experiences a high workload in both the focal shift and the prior working shift.

7. Concluding Remarks

In this work, we quantify the causal effect of nurses' historical workload on the discrepancy between their perceived workload and the order-based workload. In particular, we show that a higher

historical workload leads to a higher perceived workload relative to the order-based workload. This highlights the importance of properly balancing the nursing workload over time. We also propose a patient-to-nurse assignment policy that aims to temporally balance the nursing workload while maintaining a good level of continuity of care. We show that our policy can improve both workload balancing and continuity of care metrics compared to the current practice of the hospital.

There are several interesting future research directions. First, it would be interesting to directly measure the effect of nursing workload on burnout. This would require new survey data on nurses' job satisfaction and burnout scores. This could further enable quantifying the cost of high nursing workload on nurse turnover. Second, in this work, we demonstrate how to leverage optimization to properly balance nursing workload and continuity of care. Our proposed method is a heuristic one. It remains an important topic to optimize patient-to-nurse assignment policies that achieve a more balanced workload and create a fairer and safer working environment. Lastly, even though balancing the nursing workload through better patient-to-nurse assignment policies can help reduce nursing burnout, in many situations, we need to increase the nurse staffing level to meet the high patient volumes and acuity levels. Thus, it would be of great value to jointly optimize the nurse staffing and assignment decisions.

References

Ahuja V, Alvarez CA, Staats BR (2020) Maintaining continuity in service: An empirical examination of primary care physicians. *Manufacturing & Service Operations Management* 22(5):1088–1106.

Ahuja V, Alvarez CA, Staats BR (2022) An operations approach for reducing glycemic variability: Evidence from a primary care setting. *Manufacturing & Service Operations Management* 24(3):1474–1493.

Aiken LH, Clarke SP, Sloane DM, Sochalski J, Silber JH (2002) Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of the American Medical Association* 288(16):1987–1993.

Alghamdi MG (2016) Nursing workload: A concept analysis. *Journal of Nursing Management* 24(4):449–457.

Armony M, Ward AR (2010) Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research* 58(3):624–637.

Arnosti N, Johari R, Kanoria Y (2021) Managing congestion in matching markets. *Manufacturing & Service Operations Management* 23(3):620–636.

Baiocchi M, Cheng J, Small DS (2014) Instrumental variable methods for causal inference. *Statistics in Medicine* 33(13):2297–2340.

Batt RJ, Terwiesch C (2017) Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science* 63(11):3531–3551.

Bergman A, David G, Song H (2022) "I quit": Schedule volatility as a driver of voluntary employee turnover. *Available at SSRN 3910077* .

Berlin G, Lapointe M, Murphy M, Wexler J (2022) Assessing the lingering impact of COVID-19 on the nursing workforce. Mckinsey & Company. May 11, 2022.

Berry Jaeker JA, Tucker AL (2017) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* 63(4):1042–1062.

Carayon P, Gurses AP (2008) Nursing workload and patient safety—a human factors engineering perspective. *Patient Safety and Quality: An Evidence-Based Handbook for Nurses* .

Eriksson EA, Mattsson LG (1983) Quantitative measurement of continuity of care: Measures in use and an alternative approach. *Medical care* 858–875.

Escobar GJ, Greene JD, Scheirer P, Gardner MN, Draper D, Kipnis P (2008) Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Medical Care* 232–239.

Foley RD, McDonald DR (2001) Join the shortest queue: Stability and exact asymptotics. *Annals of Applied Probability* 569–607.

Freeman M, Robinson S, Scholtes S (2021) Gatekeeping, fast and slow: An empirical study of referral errors in the emergency department. *Management Science* 67(7):4209–4232.

Green LV, Savin S, Savva N (2013) "Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science* 59(10):2237–2256.

Griffiths P, Saville C, Ball J, Jones J, Pattison N, Monks T, Group SNCS, et al. (2020) Nursing workload, nurse staffing methodologies and tools: A systematic scoping review and discussion. *International Journal of Nursing Studies* 103:103487.

Haggerty JL, Reid RJ, Freeman GK, Starfield BH, Adair CE, McKendry R (2003) Continuity of care: A multidisciplinary review. *BMJ* 327(7425):1219–1221.

Halbesleben JR, Wakefield BJ, Wakefield DS, Cooper LB (2008) Nurse burnout and patient safety outcomes: Nurse safety perception versus reporting behavior. *Western Journal of Nursing Research* 30(5):560–577.

Holland P, Tham TL, Sheehan C, Cooper B (2019) The impact of perceived workload on nurse satisfaction with work-life balance and intention to leave the occupation. *Applied Nursing Research* 49:70–76.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389–4407.

Jee SH, Cabana MD (2006) Indices for continuity of care: a systematic review of the literature. *Medical Care Research and Review* 63(2):158–188.

Jourdain G, Chênevert D (2010) Job demands–resources, burnout and intention to leave the nursing profession: A questionnaire survey. *International Journal of Nursing Studies* 47(6):709–722.

Kajaria-Montag H, Freeman M, Scholtes S (2022) *Continuity of care increases physician productivity in primary care* (INSEAD).

Kamalahmadi M, Yu Q, Zhou YP (2021) Call to duty: Just-in-time scheduling in a restaurant chain. *Management Science* 67(11):6751–6781.

Kc DS, Staats BR, Kouchaki M, Gino F (2020) Task selection and workload: A focus on completing easy tasks hurts performance. *Management Science* 66(10):4397–4416.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: an econometric analysis of hospital operations. *Management science* 55(9):1486–1498.

Kc DS, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* 14(1):50–65.

Kesavan S, Lambert SJ, Williams JC, Pendem PK (2022) Doing well by doing good: Improving retail store performance with responsible scheduling practices at the gap, inc. *Management Science* 68(11):7818–7836.

Kim K, Mehrotra S (2015) A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. *Operations Research* 63(6):1431–1451.

Laschinger HKS, Leiter MP (2006) The impact of nursing work environments on patient safety outcomes: The mediating role of burnout engagement. *Journal of Nursing Administration* 36(5):259–267.

Li W, Sun Z, Hong LJ (2021) Who is next: Patient prioritization under emergency department blocking. *Operations Research* .

Liu X, Zheng J, Liu K, Baggs JG, Liu J, Wu Y, You L (2018) Hospital nursing organizational factors, nursing care left undone, and nurse burnout as predictors of patient safety: A structural equation modeling analysis. *International Journal of Nursing Studies* 86:82–89.

Long EF, Mathews KS (2018) The boarding patient: effects of icu and hospital occupancy surges on patient flow. *Production and Operations Management* 27(12):2122–2143.

Luo Q, Wang R, Chen H, Xie X, Tang CS (2022) Going beyond occupancy rate: The impact of doctors' and nurses' task workload on health outcomes. *Available at SSRN* .

Mandelbaum A, Momčilović P, Tseytlin Y (2012) On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Science* 58(7):1273–1291.

Miranda DR, Moreno R, Iapichino G (1997) Nine equivalents of nursing manpower use score (nems). *Intensive Care Medicine* 23(7):760–765.

Miranda DR, Nap R, de Rijk A, Schaufeli W, Iapichino G, of the TISS Working Group M, et al. (2003) Nursing activities score. *Critical Care Medicine* 31(2):374–382.

Muehler N, Oishi J, Specht M, Rissner F, Reinhart K, Sakr Y (2010) Serial measurement of therapeutic intervention scoring system-28 (TISS-28) in a surgical intensive care unit. *Journal of Critical Care* 25(4):620–627.

Niewoehner III RJ, Diwas K, Staats B (2022) Physician discretion and patient pick-up: How familiarity encourages multitasking in the emergency department. *Operations Research* .

OptiLink K (2012) Acuity-based scheduling supports quality-conscious and cost-effective care.

Shah MK, Gandrakota N, Cimiotti JP, Ghose N, Moore M, Ali MK (2021) Prevalence of and factors associated with nurse burnout in the US. *JAMA network open* 4(2):e2036469–e2036469.

Shi P (2022) Optimal matchmaking strategy in two-sided marketplaces. *Management Science* .

Soltani M, Batt RJ, Bavafa H, Patterson BW (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):3079–3098.

Stiegler K, Martiniano R, Forte G (2021) Health care employment projections, 2019–2029: An analysis of bureau of labor statistics projections by setting and by occupation. *Rensselaer, NY: Center for Health Workforce Studies, School of Public Health, SUNY Albany* .

Vahey DC, Aiken LH, Sloane DM, Clarke SP, Vargas D (2004) Nurse burnout and patient satisfaction. *Medical Care* 42(2 Suppl):II57.

Véricourt Fd, Jennings OB (2011) Nurse staffing in medical units: A queueing perspective. *Operations research* 59(6):1320–1331.

Wang WY, Gupta D (2014) Nurse absenteeism and staffing strategies for hospital inpatient units. *Manufacturing & Service Operations Management* 16(3):439–454.

Wang Y, Zhao L, Savelsbergh M, Wu S (2022) Multi-period workload balancing in last-mile urban delivery. *Transportation Science* 56(5):1348–1368.

Wooldridge JM (2010) *Econometric analysis of cross section and panel data* (MIT press).

Appendix A: Details of Data Cleaning Procedure

In this study, we combine data from multiple sources: patient-flow data, hospitalization data, LAPS data, OptLink assignment data, OptLink intensity data, and nurse staffing data. To process the data, we first correct erroneous records by cross-checking different data sources. When inconsistency arises, we use the majority-vote rule to correct the records. Then we impute some missing non-admission shift assignment data leveraging the nurse staffing data. For this part, there are 83 patient-shifts belonging to 62 shifts that do not have the nurse assignment information. By checking the staffing data, we find that in 33 of the 62 shifts, there is only a single patient missing the nurse assignment information and a single "idle" nurse who is working in that shift according to the staffing data but does not have an assignment recorded in the assignment data. For these 33 shifts, we match the idle nurse and the patient who does not have an assignment. In 16 of the 62 shifts, there is a single patient missing nurse assignment information but more than one idle nurse. For these 16 shifts, we match all the idle nurses with the patient that does not have an assigned nurse. Lastly, in 13 of the 62 shifts, there is no idle nurse. For these shifts, we randomly pick a nurse among those who have only one patient assigned and match the nurse with the patient that does not have an assigned nurse. Finally, we remove some observations with missing information. See Figure 3 for a summary of the final cohort construction.

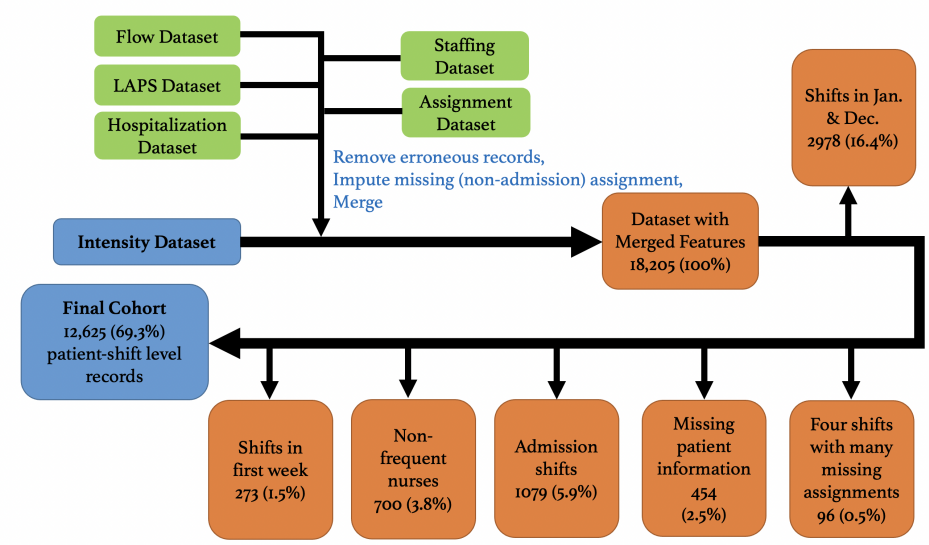


Figure 3 Flow chart of data cleaning procedure

Appendix B: Details of Sensitivity Analysis

In this section, we provide the pseudo-code to impute the missing admission assignment in Section 3.3. The details are summarized in Algorithm 1.

Algorithm 1 Pseudo-code of missing admission assignment imputation strategy

for shift $t' = 1, 2, \cdots T$ **do**

Find $\mathcal{AMP}_{t'}$, the set of admission patients in shift t who do not have the nurse assignment information. Sort $\mathcal{AMP}_{t'}$ in the decreasing order of $DDCIntensity_{p,t'}$.

Let $\mathcal{NAN}_{t'}$, denote the set of nurses working in shift t' who do not have an admission assignment. Calculate their current non-admission assignment workload, i.e., $Non\_Admin\_Work_{n,t'}$ for all $n \in \mathcal{NAN}_{t'}$. Sort $\mathcal{NAN}_{t'}$ in the decreasing order of $Non\_Admin\_Work_{n,t'}$.

while $\mathcal{AMP}_{t'}$ is not empty **do**

if Want to achieve a positive correlation **then**

Match the first patient in $\mathcal{AMP}_{t'}$ with the first nurse in $\mathcal{NAN}_{t'}$.

else if Want to achieve a negative correlation **then**

Match first patient in $\mathcal{AMP}_{t'}$ with the last nurse in $\mathcal{NAN}_{t'}$.

else

Match the first patient in $\mathcal{AMP}_{t'}$ with a randomly selected nurse from $\mathcal{NAN}_{t'}$.

end if

Remove the matched patient from $\mathcal{AMP}_{t'}$ and remove the matched nurse from $\mathcal{NAN}_{t'}$.

end while

end for

Appendix C: Robustness Check: Varying Look-Back Window Size

Recall that when defining the historical workload, we look back for one week and set the historical workload as zero if the focal nurse had no assignment over the past week. The idea is that after a long enough rest, the impact of historical workload on the perceived workload in the focal shift would be negligible. In this section, we test alternative look-back time windows: two weeks and three weeks. The estimation results are summarized in Table 10. Note that the sample size in the three columns is slightly different because we have to remove the first one/two/three weeks of data respectively to ensure that $Hist_Work$ is well-defined. We observe that the estimated treatment effects are similar in the three specifications of $Hist_Work$.

Table 10 Robustness Check: Estimation with Different Look-Back Window Size in the definition of $Hist\_Work_{p,t}$

	Truncation window size		
	One week	Two week	Three week
$Hist\_Work^*$	0.629^{***}	0.602^{***}	0.577^{***}
	(0.010)	(0.010)	(0.011)
Observations	12625	12319	12023

standard error in parentheses
$^{*}\, p < 0.05, \quad ^{**}\, p < 0.01, \quad ^{***}\, p < 0.001$