"Nursevendor Problem": Personnel Staffing in the Presence of Endogenous Absenteeism

Linda V. Green
Graduate School of Business, Columbia University, lvg1@columbia.edu

Sergei Savin
The Wharton School, University of Pennsylvania, savin@wharton.upenn.edu

Nicos Savva
London Business School, nsavva@london.edu

The problem of determining nurse staffing levels in a hospital environment is a complex task due to variable patient census levels and uncertain service capacity caused by nurse absenteeism. In this paper we combine an empirical investigation of the factors affecting nurse absenteeism rates with an analytical treatment of nurse staffing decisions using a novel variant of the newsvendor model. Using data from the emergency department of a large urban hospital, we find that absenteeism rates are consistent with nurses exhibiting an aversion to higher levels of anticipated workload. Using our empirical findings we analyze a single-period nurse staffing problem considering both the case of constant absenteeism rate (exogenous absenteeism) as well as an absenteeism rate which is a function of the number of scheduled nurses (endogenous absenteeism). We provide characterizations of the optimal staffing levels in both situations and show that the failure to incorporate absenteeism as an endogenous effect results in understaffing.

History: February 4, 2012

1. Introduction and Literature Review

In recent years hospitals have been faced with ever-increasing pressure from their major payers - federal and state governments, managed care organizations, and large employers - to cut costs. Since nursing personnel accounts for a very large portion of expenses, the response in many instances has been reductions of the nursing staff. Nurse workloads have been further increased by shorter hospital lengths-of-stay (LOS) and increasing use of outpatient procedures, resulting in sicker hospitalized patients who require more nursing care. The adverse impact of these changes has been documented by a number of studies (Needleman et al. (2002), Aiken et al. (2002), Cho et al. (2003)). These effects include increases in medical errors, delays for patients waiting for beds in emergency rooms, and ambulance diversions. In response a number of state legislatures, e.g. Victoria in Australia (The Victorian Department of Health (2007)) and California in the US (California Department of Health (2004)), have mandated minimum nurse staffing levels.

Establishing the right balance between the quality and cost of patient care is a challenging task. In a hospital environment, nurses are usually scheduled to work 8- or 12-hour shifts, and the choice of appropriate nurse staffing levels for a particular shift is complicated by the need to make staffing decisions well in advance (e.g. 6-8 weeks) of that shift when the patient census is unknown. Even if the patient census could be reliably estimated, nurse staffing decisions are complicated by labor constraints dealing with the number of consecutive and weekend shifts worked per nurse, vacation schedules, personal days, and preferences (see, e.g. Pierskalla et al. (1976)).

It has been increasingly recognized that, in addition to the factors mentioned above, nurse absenteeism is an important consideration in making staffing decisions. According to the US Bureau of Labor Statistics (2008), in 2008 US nurses exhibited 12.5 incidents of illness or occupational injury per 100 Full Time Employees (FTEs), second only to construction workers, as well as the highest number of cases involving days away from work, 7.8 per 100 FTEs. These figures are substantially higher than the national average of 4.2 incidents per 100 FTEs, with only half involving time away from work. Similarly, in Canada nurses have one of the highest absenteeism rates (12.2%) of all workers, and this absenteeism rate has been increasing over the last 10 years (Statistics Canada (2008)).

The goal of this paper is to construct a model for nurse staffing that includes the impact of absenteeism. In order to do this, we must first understand whether and how absenteeism is affected by staffing levels themselves. Since there is no literature that adequately addresses this linkage, we first conduct an empirical hospital-based study to understand how to incorporate absenteeism into an analytical model.

Structural linkages between absenteeism and work environment factors, such as workload, have long been the focus of applied psychology research. In particular, a weak association between workload and absenteeism (see Darr and Johns (2008) for an overview) has been discovered by numerous studies which compare long-run average absenteeism rates across either different industries, different firms, or different employee roles. Most studies find that employees exhibit higher long-run average absenteeism rates when they perform jobs with higher workloads (e.g. Kristensen (1991), Dwyer and Ganster (1991)). This is consistent with theoretical accounts of absenteeism as a manifestation of withdrawal from adverse conditions (Hill and Trist (1955)) or as a coping mechanism that allows employees to replenish their depleted physical and mental resources (Hobfoll (1989)). At the other end of the spectrum there are studies that identify environments where absenteeism is negatively correlated with workload demands (Parkes (1982), Smulders and Nijhuis (1999)). Such behavior is consistent with workload becoming a "pressure-to-attend" (Steers and Rhodes (1978)). The evidence of these opposing effects is also present in a separate literature stream focusing on

the causes of nurse absenteeism. Most studies in this area find that nurse absenteeism is positively related to levels of work-related stress (Shamian et al. (2003)) and nurse workload (Bryant et al. (2000), Tummers et al. (2001), McVicar (2003), Unruh et al. (2007)). However, at least one study examining absenteeism among trainee nurses finds a negative correlation between nurse absenteeism and workload (Parkes (1982)).

Nearly all of the existing studies of nurse absenteeism employ cross-sectional analysis of this phenomenon (see Rauhala et al. (2007) for an exception). In particular, they compare the long-run absenteeism behavior of nurses across different clinical units, rather than tracking the same set of nurses over time. As such, these studies are of limited value to managers responsible for day-to-day staffing decisions. The first part of our paper presents a longitudinal investigation of the link between nurse absenteeism and workload using data from the Emergency Department (ED) of a large New York City hospital. Rather than relying on subjective self-reported workload measures, we use patient census values to calculate nurse-to-patient ratios which are treated as proxies for the workload experienced by nurses working a particular shift. Nurse shortages and other organizational limitations (such as union rules) provide the necessary exogenous variation in staffing decisions that we exploit in order to identify the impact of workload on absenteeism. We hypothesize that fluctuations in nurse workload due to irregularities in scheduling and/or unpredictable demand, have a direct impact on absenteeism. An anticipated increase in workload could result in an increase in nurses' motivation to help their fellow nurses and, therefore, in decreased absenteeism, or, given generally high workloads and adverse working conditions, in increased absenteeism. In either case, the resulting behavior may very well result in optimal nurse staffing levels that are different from those predicated by models that do not consider this workload-induced behavior. Therefore, our empirical investigation aims to provide support for one of these specific hypotheses so that we can incorporate the appropriate assumption into our analytical modeling. Our main finding is that absenteeism increases when there is a higher anticipated workload. Specifically, we find that for our data set with an average absenteeism rate of 7.3%, an extra scheduled nurse is associated with an average reduction in the absenteeism rate of 0.6%. As such, our paper is related to the growing body of literature (KC and Terwiesch (2009), Powell and Schultz (2004), Schultz et al. (1998), and Schultz et al. (1999)) on the effects of workload on system productivity.

The second goal of our paper is to develop a model of optimal staffing in service environments with workload-dependent absenteeism. Extant literature either ignores absenteeism or treats it as an exogenous phenomenon (Bassamboo et al. (2010), Easton and Goodale (2005), Harrison and Zeevi (2005) and Whitt (2006) provide examples of call-center staffing while Fry et al. (2006) provide an example of firefighter staffing). The uncertain supply of service capacity created by nurse absenteeism connects our work with a stream of literature focused on inventory planning

in the presence of unreliable supply/stochastic production yield (Yano and Lee (1995)) with two important distinctions. First, the overwhelming majority of papers which deal with stochastic supply yields model them as being either additive or multiplicative (Ciarallo et al. (1994), Bollapragada and Morton (1999), Gupta and Cooper (2005), Yang et al. (2007)), a justifiable approach in manufacturing settings. The supply uncertainty in our model has a binomial structure, a more appropriate choice in personnel staffing settings. Binomial yield models are a relative rarity in the stochastic yield literature, perhaps due to their limited analytical tractability (Gerchak and Henig (1994), Grosfeld-Nir and Gerchak (2004), Fadiloglu et al. (2008)). Most importantly, our analysis is the first one to introduce and analyze endogenous stochastic yields.

Using our model we characterize the optimal staffing levels under exogenous and endogenous absenteeism. We show that the failure to incorporate absenteeism as an endogenous effect results in understaffing, which leads to a higher-than-optimal absenteeism rate and staffing costs. Specifically, for model parameters that closely match the hospital we study, we find that ignoring the endogenous nature of absenteeism can lead to a staffing cost increase of $2-3\%$. In addition to the cost impact, the understaffing associated with ignoring absenteeism may result in an increase in medical errors, particularly in the pressured and sensitive environment of an ED. Considering that nursing costs is one of the biggest components of overall hospital operating costs, more accurate nurse staffing based on endogenous absenteeism constitutes a substantial opportunity for hospitals to simultaneously reduce costs and improve quality of care.

Finally, we show that despite understaffing, the exogenous-absenteeism model will appear to be self-consistent in the sense that the assumed exogenous absenteeism rate will be equal to the observed (endogenous) absenteeism rate. This is particularly worrisome for staffing managers as it implies that it is impossible to tell whether the model is well specified just by examining the observed absenteeism rate. In this regard our paper contributes to the literature on model specification errors (e.g. Cachon and Kök (2007) and Cooper et al. (2004)). Cachon and Kök (2007) examine the assumption that the salvage value of the newsvendor model is independent of order quantity, while Cooper et al. (2004) examine the assumption that the the demand estimate for a particular ticket class of a revenue management model is independent of the chosen protection level. In both papers, as in ours, the model specification error cannot be detected by studying data as the misspecified model will produce consistent outcomes. Our paper is the first to study model specification error in the context of staffing, and the first in the model specification literature to start from an empirical observation.

2. Endogeneity in Nurse Absenteeism Rates: An Empirical Study

Our study is based on nurse absenteeism and patient census data from the ED of a large New York City hospital. Nurses employed in this unit are full-time employees, each working on average 3.25

Measure	Mean	Standard Deviation	Minimum	Maximum
Day Shift				
Nurses Scheduled	11.4	1.07	8	16
Absenteeism rate	0.0762	0.0799	0	0.4
Patient visits	141	20.1	77	188
Average Census	116	17.1	56.5	158
Maximum Census	136	20.8	64	182
Night Shift				
Nurses Scheduled	10.5	0.849	9	14
Absenteeism rate	0.0707	0.0829	0	0.4
Patient Visits	66.0	9.45	40	95
Average Census	102	14.2	54.3	142
Maximum Census	127	20.4	57	174
Evening Shift				
Nurses Scheduled	3.63	0.756	2	5
Absenteeism rate	0.0589	0.119	0	0.5
Patient visits	137	16.2	75	196
Average Census	125	18.6	58.2	164
Maximum Census	137	20.7	64	182

Table 1 : Descriptive statistics for nurse and patient data.

shifts per week. The unit uses two primary nursing shifts; the "day" shift starts at 8:00am and ends at 8:00pm, while the "night" shift starts at 8:00pm and ends at 8:00am. Another ("evening") shift is also operated from 12:00pm to 12:00am. The evening shift is fundamentally different from the other two shifts. First, the nurses working on this shift are dedicated to this shift and, unlike the other nurses, do not work on the other two shifts. Thus, it is less likely that they are informed about the nurse staffing schedules for the day/night shifts. Second, this shift consists of fewer nurses who are more experienced, exhibit less absenteeism than the other two shifts, as shown in Table 1.

In our analysis of absenteeism we limit our attention to the nurses on the day and night shifts. However, we do take into account the evening shift when measuring workload since the evening shift overlaps with both the day shift and the night shift. For each shift, for a period of 10 months starting on July 1, 2008 (304 day shifts, 304 evening shifts and 303 night shifts), we collected the following data: the number of nurses scheduled, the number of nurses absent, and the patient census data recorded every two hours.

The nurse scheduling process starts several weeks before the actual work shift when the initial schedule is established. This initial schedule often undergoes a number of changes and corrections due to e.g. family illnesses, medical appointments and jury duty obligations, which may continue until the day before the actual shift. In our study we have used the *final* schedules, i.e. the last schedules in effect before any "last minute" absenteeism is reported for the shift. We record as absenteeism any event where a nurse does not show up for work without giving sufficiently advance notice for the schedule to be revised. In the clinical unit we study, nurses are allowed to use up to ten "personal" days per year which do not require any significant advance notice. The resulting descriptive statistics for three shifts are presented in Table 1.

The average patient census during a shift varies substantially from day to day. Some of this variation (52.2% for day shifts and 32.6% for night shifts) can be explained by day-of-the-week and

week fixed effects. Further, the patient census exhibits significant serial autocorrelation ($\rho = 30.6\%$) with the values recorded during the previous shift. The number of nurses scheduled for a particular type of shift, e.g. day shift on a Wednesday, is highly variable. Approximately 25% of the variation in the number of nurses scheduled can be explained by day-of-the-week and week fixed effects (adjusted $R^2 = 27.5\%$ for the day shift and adjusted $R^2 = 24.1\%$ for the night shift). Also, after controlling for fixed effects the number of nurses scheduled for a shift shows little dependence on either the average patient census during that shift or on the average census values for the 14 previous shifts, which correspond to one calendar week. This indicates that the unit's nurse staffing policy does not seem to be affected in any significant way by actual patient census.

Our discussions with nurse manager indicated that there are two main factors driving the significant variations in the number of scheduled nurses. First, personnel scheduling is subject to numerous constraints (e.g. union rules) that often prevent manager from assigning the number of nurses desired for a particular shift. Second, as mentioned earlier, initial schedules often undergo a series of changes before they are finalized. While such scheduling variations are not desirable from the point of view of managing the match between the demand for nursing services and the supply of nursing capacity, they provide an opportunity to examine how absenteeism rates are related to the numbers of scheduled nurses.

2.1. Nurse Workload and Absenteeism: Empirical Results

We model the phenomenon of nurse absenteeism as follows. We treat all nurses as being identical and independent decision makers and focus on a group of y_t nurses scheduled to work during a particular shift t ($t = 1$ for the first shift in the data set, $t = 2$ for the second shift, etc., up to $t = 607$). For nurse n, $n = 1, ..., y_t$, the binary variable $Y_{n,t}$ denotes her decision to be absent from work ($Y_{n,t} = 1$), or to be present ($Y_{n,t} = 0$). We assume that this absenteeism decision is influenced by a number of factors expressed by the vector $\mathbf{x}_t$ which include parameters related to workload as well as fixed effects such as the day of the week or the shift. Each nurse compares the utility she receives from being absent from work to the utility she receives from going to work. The difference in these utility values is given by $U^*_{n,t} = \mathbf{x}'_t\beta + \epsilon_{n,t}$, where $\epsilon_{n,t}$ are, for each n and t, *i.i.d.* random variables with mean zero. While the utility difference $U^*_{n,t}$ is an unobservable quantity, we can potentially observe each nurse's decision to show up for work. The decision is such that $Y_{n,t} = 1$ if $U^*_{n,t} > 0$, and $Y_{n,t} = 0$ otherwise. Assuming that $\epsilon_{n,t}$ follow the standardized logistic distribution (the standard normal distribution) we obtain the logit (probit) model (Greene (2005)). It is important to keep in mind that our empirical data do not record the attendance decisions of individual nurses. Rather, we measured the aggregate absenteeism behavior of a group of nurses scheduled for a particular shift. Consequently, we treat all nurses scheduled for a given shift as

a homogenous group and build the model for the corresponding group behavior. We examine the impact of relaxing this assumption in Section 2.2. We focus on the maximum-likelihood-based logit estimation of the probability of absenteeism γ_t during shift t. We estimate the model using the maximum likelihood approach for grouped data (see Greene (2005), Chapter 21.4.6). Under this approach the dependent variable is the proportion of nurses that are absent given the number of nurses scheduled for a particular shift y_t.

Since our goal is to study how the nurse absenteeism rate is affected by workload, we need to measure and quantify nurse workload for each shift. We use the nurse-to-patient ratio as a proxy for the workload nurses experience during a particular shift. For shift t, we define the nurse-to-patient ratio variable, denoted as NPR_t, as the ratio of the number of nurses working during a particular shift and the patient census averaged over the duration of that shift. To estimate the number of nurses present we assume that the number of nurses scheduled in a particular shift *is* the number of nurses actually present. This assumption is consistent with the practice in the ED we studied, which uses either an agency nurse or a nurse from the previous shift to work overtime to substitute for an absent nurse. With this assumption, the number of nurses present during each 24-hour period varies as follows: between 8:00am and 12:00pm it is equal to the number of nurses scheduled for the day shift (y_t), between 12:00pm and 8:00pm it is equal to the number of nurses scheduled for the day shift (y_t) plus the number of nurses scheduled for the evening shift (e_t), between 8:00pm and 12:00am it is equal to the number of nurses scheduled for the evening shift (e_t) plus the number of nurses scheduled for the night shift, while between 12:00am and 8:00am it is equal to the number of nurses scheduled for the night shift (y_t). Thus we estimate NPR_t as follows

$$\text{NPR}_t = \frac{y_t + \frac{2}{3}e_t}{C_t} \text{ for the day shift}, \quad \text{NPR}_t = \frac{y + \frac{1}{3}e_t}{C_t} \text{ for the night shift}, \tag{1}$$

where C_t is the patient census averaged over the duration of shift t.

In making their attendance decisions for shift t nurses may be influenced by the anticipated workload for shift t. The impact of *anticipated* workload arises because nurses are informed in advance of their schedule and they are aware of how many (and which) other nurses are scheduled to work on the same shift as them. Since nurses anticipate a certain patient census $E[C_t]$, consistent with their past experience of working in the ED, nurses form an expectation about the anticipated workload for that shift. Naturally, if fewer (more) nurses are scheduled on that particular shift than the nurses deem appropriate, they will anticipate a higher (lower) workload than normal. The group attendance data do not present a measurement challenge, since the nurses scheduled for the same shift are subjected to the same anticipated workload value.

The anticipated workload can have a dampening effect on absenteeism through the "pressure-to-attend" mechanism (Steers and Rhodes (1978)) or can enhance absenteeism by encouraging "withdrawal behavior" (Hill and Trist (1955), Hobfoll (1989)). To the best of our knowledge, the impact of anticipated workload on absenteeism has not been previously studied. We test this potential impact in our setting by including in the vector of covariates $\mathbf{x}_t$ the anticipated value of nurse-to-patient ratio

$$\text{ENPR}_t^1 = \frac{y_t + \frac{2}{3}e_t}{E[C_t]} \text{ for the day shift}, \quad \text{ENPR}_t^1 = \frac{y_t + \frac{1}{3}e_t}{E[C_t]} \text{ for the night shift}, \tag{2}$$

where e_t is the number of nurses scheduled in the evening shift which overlaps with $\frac{2}{3}$ ($\frac{1}{3}$) of the duration of the day (night) shift in question. While day and night shift nurses are fully informed about the schedule for their shifts, it is not clear that they would be as familiar with the schedule of the evening shift staffed by a different pool of nurses. Motivated by this observation, we estimate two models based on alternative definitions of the expected nurse-to-patient ratio. In the first definition (ENPR_t^1 of equation (2)) we use the exact number of evening nurses scheduled (e_t), while in the second definition (ENPR_t^2) we use the average value of e_t (averaged over all evening shifts in our sample). The latter formulation reflects the situation where day- and night-shift nurses do not know precisely how many evening nurses will be present but form a rational expectation about this value. In other words, in the second model day- and night-shift nurses behave as if they ignore any variation in the number of nurses scheduled for the evening shift that overlaps with their own shift. $E[C_t]$ is set to the expectation of patient census values computed over *all* shifts in our sample. This formulation reflects an assumption that nurses, when making their attendance decisions, use a mental model which captures any potential difference occurring on different days/shifts with a fixed effect and, therefore, focus on expected patient census value. We also assume that the nurses form rational expectations about the patient census which are consistent with empirically observed patient census data. In addition to the models based on (2), we have also estimated several alternative variants which we discuss in 2.2.

In addition to the anticipated nurse-to-patient ratio ($\text{ENPR}_t^i, i = 1, 2$), the vector of covariates $\mathbf{x_t}$ includes a number of controls. In particular, we include a day-of-the-week dummy variable to capture any systematic variation in absenteeism across days, a day/night-shift fixed effect to capture variations between day and night shifts, and a week fixed effect to capture any systematic variations which remain constant over a period of one week and affect absenteeism but are otherwise unobservable. Also, we include a holiday fixed effect which takes the value of 1 on national public holidays and zero on any other day. This last variable is designed to deal with a potential endogeneity problem since nurses may be inherently reluctant to work on some select days, such as

public holidays. These days are known to the management of the clinical unit which tries to accommodate the nurses' aversion by staffing fewer nurses on such days. Nevertheless, the nurses that are scheduled to work on these "undesirable" days are still more likely to be absent, irrespective of the chosen staffing levels. By including the holiday variable we are trying to explicitly account for this effect. It is possible that there exist other correlated variables that we omit, but to the extent that they do not vary drastically over a period of one week, the week fixed effect should be able to capture the influence of those variables. Finally, to account for the possibility that absenteeism might be a delayed response to past workloads, we include the values of 14 lagged nurse-to-patient ratios NPR_{t-j}, $j=1,\ldots,14$, which correspond to one calendar week. As the number of past shifts we use is rather arbitrary, we conducted our statistical analysis for several different values to make sure the results are not sensitive to the number we choose as long as it is sufficiently large.

Specifically, the models we estimate are

$$\begin{aligned}\mathrm{logit}(\gamma_t) = {} & \beta_0^i + \beta_{\mathrm{ENPR}}^i \times \mathrm{ENPR}_t^i + \sum_{j=1}^{14} \beta_{\mathrm{NPR},j}^i \times \mathrm{NPR}_{t-j} + \sum_{d=2}^{7} \beta_{\mathrm{DAY},d}^i \times \mathrm{DAY}_{d,t} \\ & + \sum_{f=2}^{44} \beta_{W,f}^i \times W_{f,t} + \beta_{\mathrm{DAYSHIFT}}^i \times \mathrm{DAYSHIFT}_t + \beta_{\mathrm{HOLIDAY}}^i \times \mathrm{HOLIDAY}_t, \end{aligned} \tag{3}$$

where γ_t is the probability that a nurse is absent in shift t, $i=1,2$ refers to the definition of ENPR^i used, $\mathrm{DAY}_{d,t}$ and $W_{f,t}$ are the day and week fixed effects, $\mathrm{DAYSHIFT}_t$ and $\mathrm{HOLIDAY}_t$ are the shift and holiday fixed effects. The estimation results for equation (3) are presented in Table 1. Model I uses the first definition of the anticipated nurse-to-patient ratio (ENPR_t^1), while Model III uses the second definition (ENPR_t^2). In order to test whether the observed effect of anticipated nurse-to-patient ratio on absenteeism is robust we also estimated the restricted versions of Models I and III (which we denote as Models II and IV), where we omit the 14 lagged nurse-to-patient ratio variables. If the lagged nurse-to-patient ratios are not related to absenteeism (i.e. if $\beta_{\mathrm{NPR},j}^i = 0$ for all $j=1,...,14$) omitting these variables will not introduce any bias even if the lagged nurse-to-patient variables are correlated with the variables included in the model. Model II uses the first definition of the anticipated nurse-to-patient ratio (ENPR_t^1) while Model IV uses the second definition (ENPR_t^2).

As can be seen from Table 1, the anticipated nurse-to-patient ratio has a significant effect (at the 5% or 10% level) on absenteeism rates in Models I, III and IV. In Model II the p-value of the anticipated nurse-to-patient ratio is 10.7%. The more nurses scheduled for a particular shift, the less likely each nurse is to be absent. In particular, according to the first model we estimate the marginal effect of staffing an extra nurse (calculated at the mean values of all remaining independent variables and using the expected patient census value of 109) on the individual absenteeism rate is

	Model I		Model II		Model III		Model IV	
Variable	Coefficient (Robust Error)	Marginal effect (Robust Error)	Coefficient (Robust Error)	Marginal effect (Robust Error)	Coefficient (Robust Error)	Marginal effect (Robust Error)	Coefficient (Robust Error)	Marginal effect (Robust Error)
ENPR1	-10.01*	-0.623*	-8.534	-0.533				
	(5.669)	(0.352)	(5.291)	(0.330)				
ENPR2					-15.24**	-0.946**	-13.25**	-0.827**
					(6.355)	(0.393)	(5.892)	(0.367)
NPR1	-0.248	-0.0154			-0.474	-0.0294		
	(3.061)	(0.190)			(3.071)	(0.191)		
NPR2	2.463	0.153			3.024	0.188		
	(3.438)	(0.214)			(3.459)	(0.215)		
NPR3	-0.492	-0.0306			-0.712	-0.0442		
	(2.989)	(0.186)			(2.982)	(0.185)		
NPR4	3.314	0.206			3.673	0.228		
	(3.265)	(0.203)			(3.275)	(0.203)		
NPR5	-2.620	-0.163			-2.917	-0.181		
	(3.020)	(0.188)			(3.023)	(0.188)		
NPR6	5.912*	0.368*			5.837*	0.362*		
	(3.249)	(0.201)			(3.251)	(0.201)		
NPR7	0.518	0.0322			0.615	0.0382		
	(3.155)	(0.196)			(3.157)	(0.196)		
NPR8	1.556	0.0968			1.570	0.0974		
	(3.245)	(0.202)			(3.254)	(0.202)		
NPR9	-4.112	-0.256			-4.252	-0.264		
	(2.998)	(0.186)			(3.002)	(0.186)		
NPR10	5.656*	0.352*			5.746*	0.357*		
	(3.205)	(0.199)			(3.207)	(0.198)		
NPR11	-3.320	-0.207			-3.263	-0.203		
	(2.980)	(0.186)			(2.991)	(0.186)		
NPR12	2.462	0.153			2.707	0.168		
	(3.271)	(0.204)			(3.276)	(0.203)		
NPR13	-2.934	-0.183			-2.724	-0.169		
	(3.016)	(0.188)			(3.018)	(0.187)		
NPR14	4.934*	0.307*			4.823	0.299		
	(2.987)	(0.185)			(2.978)	(0.184)		
DayShift	0.112	0.00694	0.233	0.0145	0.211	0.0130	0.327**	0.0203**
	(0.151)	(0.00937)	(0.143)	(0.00893)	(0.163)	(0.0101)	(0.155)	(0.00959)
Holiday	-0.853*	-0.0382***	-0.787**	-0.0364***	-0.848*	-0.0380***	-0.783*	-0.0362***
	(0.440)	(0.0134)	(0.401)	(0.0131)	(0.440)	(0.0135)	(0.401)	(0.0131)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day-of-Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	-2.381*		-2.171***		-1.834		-1.726**	
	(1.426)		(0.721)		(1.449)		(0.752)	
Observations	6481		6631		6481		6631	
Log Likelihood	-1654		-1690		-1653		-1688	
LR-test	102.7		92.35		105.3		95.13	

***, ** * denote statistical significance at 1% 5% and 10% confidence levels, respectively. Standard errors are shown in parentheses. Marginal effects are calculated at the mean values of the independent variables. For the dummy variables the marginal effect shows the differenced caused by changing the variable from 0 to 1.

Figure 1: Estimation results for logit models.

around $0.575\% = 0.626/109$. In other words, the absenteeism rate would decrease from its average value of 7.34% to about 6.78% when an extra nurse is added to the schedule. The coefficient of the anticipated nurse-to-patient ratio is statistically more significant in Models III and IV, where the variation in the number of scheduled evening nurses is ignored. This might suggest that when nurses decide whether to show up for work they place greater emphasis on the number of nurses working in their shift rather than the number of nurses working in the evening shift that overlaps with their own. As with any empirical finding, one might argue the relationship we find between absenteeism and anticipated workload is due to reverse causality (i.e. it is not absenteeism that reacts to anticipated workload but instead staffing, and thus workload, that reacts to absenteeism). However, in the ED study site, we know that absenteeism was not considered by the nurse manager making staffing decisions. More generally, in discussions with managers responsible for nurse staffing in other hospitals, absenteeism patterns were not tracked or used in staffing decisions.

It is interesting to note that the lag 6, lag 10 and lag 14 (lag 6 and lag 10) of the nurse-to-patient ratio variables have positive coefficients in Model I (Model III) which are individually significant at

the 10% level. This seems to imply that the probability of a nurse being absent on any shift would increase if the shift occurring 3, 5 or 7 days ago had a higher nurse-to-patient ratio. Although these lags are individually significant, the Wald test statistic and the Likelihood Ratio test statistic for joint significance of all 14 lagged workload variables in Model I as well as Model III reject the hypothesis (even at the 10% level) that lagged nurse-to-patient ratios have any joint explanatory power. In the absence of any plausible explanation as to why a lighter workload on a similar shift occuring 3, 5 or 7 days ago might increase absenteeism, and in light of the weakness of this statistical relationship, we are inclined to treat this result as spurious.

Interestingly, the holiday variable's effect is significant (at 10% confidence level) and negative, thus suggesting that nurses are about 3.8% less likely to be absent on public holidays. Week fixed effects are jointly significant (at the 1% level) in all four models. One of the effects that week dummies seem to pick up reasonably well is the impact of weather (in particular, heavy snow conditions) on absenteeism. For example, week 35 of our data set includes March 2-4, 2009. During six day and night shifts corresponding to these dates snow on the ground in New York City was recorded to be more than 5 inches,[1] the level identified by the New York Metropolitan Transportation Authority as the one at which the public transportation disruptions are likely to set in.[2] The week-35 fixed effect is positive and significant (at the 5% level) in Models II and IV with a marginal effect equal to 8.75% in Model I and 9.80% in Model IV. Only 3 other days in our dataset had as much snow on the ground (December 3, December 21 and January 20). Turning to the impact of day-of-the week fixed effect, there is some (weak) evidence that nurses are more likely to be absent on weekends. They are also more likely to be absent during a day shift, when conflicting family obligations often cited as an important reason behind nurse absenteeism (Erickson et al. (2000), Nevidjon and Erickson (2001)) are likely to be more prevalent.

2.2. Verification Tests

In order to check the validity of our model estimation procedure we have also conducted several verification tests as described below. (The estimation details are available from the authors).

2.2.1. Alternative Specifications To ensure the robustness of our results we estimated a number of alternative modeling specifications. Namely, we estimated the models of equation (3) under the probit specification. We also estimated variants of our models which use month fixed effect variables instead of week fixed effect variables. The results were almost identical in terms of variable significance, model significance and magnitude of marginal effects. The model of equation (3) assumes that any difference between the night and the day shifts is completely captured by

[1] www.accuweather.com

[2] http://www.mta.info/news/stories/?story=173, last accessed on January 23, 2011

the dummy variable "Dayshift". However, it is possible that the two shifts are inherently different, and so are their best-fit coefficients. To test the hypothesis that the coefficients for two shifts differ we fitted an unrestricted binary choice model to the data for each of the two shifts separately and compared the fit with the restricted model by constructing a Likelihood Ratio test. The test does not reject the hypothesis that the coefficients are the same at the 5% confidence level in all models. Finally, we estimated models with alternative definitions of the expected nurse-to-patient ratio. More specifically, when estimating the expected patient census during the upcoming shift in equation (2), nurses a) distinguish between day and night shifts, but not between days of the week, b) distinguish between days of the week, but not between day shifts or night shifts, and c) distinguish between both the shift type and the day of the week. Our main finding that higher (lower) anticipated nurse-to-patient ratios decrease (increase) nurse absenteeism is robust to these alternative modeling specifications.

2.2.2. Nurse Heterogeneity and Aggregation Bias The aggregate nature of our data does not permit an exact characterization of how workload impacts the absenteeism behavior of each individual nurse. Indeed it would be interesting to measure whether aversion to anticipated increased workloads is a commonly occurring nurse characteristic or limited to a relatively small subset of nurses. However our aggregate results are not invalidated by the lack of such a characterization. Through a simulation study we find that estimating a homogeneous logit model when nurses are in fact heterogeneous does produce biased estimation coefficients. However we find that the bias on the ENPR coefficient is *positive*, i.e. it would bias our (negative) coefficient towards zero and not away from it. Therefore, our estimate should be treated as a lower bound on how workload affects nurse absenteeism. Furthermore, we also find that for most reasonable assumptions about nurse heterogeneity the magnitude of the bias is small compared to the estimated value of the coefficient. This result is similar to Allenby and Rossi (1991) whose analysis of marketing data lead the authors to conclude that in most realistic settings there exists no significant aggregation bias in logit models.

2.2.3. Within-Shift Correlation One of the assumptions we have made when estimating our models is that conditional on the vector of covariates $\mathbf{x_t}$, the nurses are independent decision makers. In this section we relax this assumption and assume that nurses can exhibit within-shift correlation. We estimate a model within the Generalized Estimating Equations (GEE) framework which allows the individual nurse decisions $Y_{n,t}$ to exhibit within shift correlation. The correlation structure estimated is of the "exchangeable form", i.e., $\text{Corr}(Y_{n,t}, Y_{m,l}) = \rho$ for $t = l$, and is 0 otherwise. The least-square estimation of these models yields very similar results to the ones presented in Table 1, and the estimated correlation ρ is around negative 1%.

3. Endogenous Nurse Absenteeism: Implications for Nurse Staffing

To study the implications of absenteeism and, in particular, of its endogenous nature on staffing level decisions, we construct a stylized model of nurse staffing. The aim of such a model is not to produce a decision support tool for the clinical unit in question, but rather to generate managerial insights regarding the impact of nurse absenteeism, in general, and the endogenous nature of absenteeism, in particular, on a decision of how many nurses to staff. While our model is parsimonious, we believe that an appropriately calibrated version of it can be used by nursing management in making tactical staffing decisions. In particular, by periodically running the model for all types of shifts (which differ in the distribution of patient census and in nurse absenteeism propensity), the nurse manager can decide how many nurses the unit will need for each shift. Thus, coupled with rostering considerations, the model can help the manager decide what the appropriate aggregate staffing level for the unit is.

We assume that a clinical unit uses the primary nursing care (PNC) mode of nursing care delivery (Seago (2001)) which was employed in the ED we studied. Under the PNC mode, the nursing staff includes only registered nurses (as opposed to licensed practical nurses or unlicensed nursing personnel) who provide all direct patient care throughout the patient's stay in the clinical unit. The nurse staffing process starts several weeks in advance of the actual shift for which planning is performed, and it is then that a hospital staff planner needs to decide how many nurses (y) to schedule for that particular shift. Due to the phenomenon of absenteeism the actual number of nurses who show up for work on that shift, N, is uncertain. We model N as a binomial random variable $B(y, 1-\gamma(y))$, where $\gamma(y)$ is the probability that any scheduled nurse will be absent from work:

$$\text{Prob}(N=k|y,\gamma(y)) = p(k;y,\gamma(y)) = \begin{cases} \frac{y!}{k!(y-k)!}(\gamma(y))^{(y-k)}(1-\gamma(y))^k, & \text{for } 0 \le k \le y, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

We assume that the clinical unit follows a policy of specifying, for each value of the average patient census during a shift, C, a target integer number of nurses $T = R(C)$ required to provide adequate patient care during a particular shift. We assume that C takes on discrete values and that $R(C)$ is a monotone increasing function with $R(0) = 0$. A simple example of $R(C)$ is provided by a "ratio" approach, under which $R(C) = \lceil \alpha C \rceil$, with $\alpha \in [0,1]$ representing a mandated nurse-to-patient ratio. Alternately, if a clinical unit is modeled as a queueing system in which patients generate service requests and nurses play the role of servers, as was done in Yankovic and Green (2011), $R(C)$ can take a more complex form to ensure that certain patient service performance measures, such as the expected time patients wait to be served, conform to pre-specified constraints. At the time of the nurse staffing decision, we assume that the decision maker uses a known probability

density function of the average patient census C during the shift for which the personnel planning is conducted:

$$\text{Prob}\,(C=n) = p_C(n), n \in \mathcal{N}^+, \sum_{n=0}^{\infty} p_C(n) = 1. \tag{5}$$

We treat the uncertain factors in our model (the demand uncertainty expressed by the patient census C and the supply uncertainty expressed by N) as being independent and assume that the realized values of C and N become known shortly before the beginning of the shift. Any nursing shortage $(R(C)-N)^+$ is covered by either hiring agency nurses or asking nurses who have just completed their shift to stay overtime. We further assume that nurses that do show up are paid $\$w_r$ per shift, while nurses that do not show up are paid $\$w_n$ where $w_r \geq w_n$. Setting $w_r = w_n$ represents a clinical unit where nurses are paid the full wage whether they actually show up for work or not. This setting is consistent with the PNC mode of nursing care delivery where nurses are salaried employees who can only be scheduled to work on a fixed number of shifts per week. When a scheduled nurse does not show up for work she can not be rescheduled in-lieu of the shift she missed. Thus, in effect, nurses are paid for each shift for which they are scheduled and are not penalized for being absent, as long as their absenteeism does not exceed the annual limit of ten "personal" days. In contrast, setting $w_n = 0$ represents a setting where nurses are hourly employees that receive no pay when absent.

In addition, we assume that if more nurses show up for work than required given the number of patients present ($N > R(C)$) they all have to be paid and cannot be "sent home". The per-shift cost of extra/overtime nurses is $\$w_e$, which, we assume, is greater than $\$w_r$. The goal of the decision maker is to choose a nurse staffing level y which minimizes the expected cost $W(y)$ of meeting the target $R(C)$:

$$W(y) = w_n y + (w_r - w_n)\,\mathbf{E}_N[N|y] + w_e \mathbf{E}_{C,N}\left[(R(C)-N)^+\,|y\right], \tag{6}$$

where $\mathbf{E}_N$ denotes expectation taken with respect to the number of of nurses who show up for work and $\mathbf{E}_{C,N}$ denotes expectation taken with respect to both the number of patients and the number of nurses who show up for work. Note that since there is a one-to-one correspondence in our model between the patient demand C and the number of required nurses T, we can re-cast the calculation of the expectation with respect to the demand value in terms of an equivalent calculation over the distribution of T using the corresponding probability distribution function. In particular, let $\mathcal{S}_n$ be the set of average patient census values, all corresponding to the same number of required nurses n:

$$\mathcal{S}_n = \left(C \in \mathcal{N}^+ | R(C) = n\right). \tag{7}$$

Then the probability distribution for T is given by

$$\text{Prob}\,(T=n)=p_T(n), n\in\mathcal{N}^+, p_T(n)=\sum_{l\in\mathcal{S}_n}p_C(l), \sum_{n=0}^{\infty}p_T(n)=1. \tag{8}$$

In turn, the cost minimization based on (6) becomes

$$\min_{y\in\mathcal{N}^+}\left(w_n y+(w_r-w_n)\,(1-\gamma(y))\,y+w_e\mathbf{E}_{T,N}\left[(T-N)^+\,|y\right]\right). \tag{9}$$

Note that with no absenteeism ($\gamma(y)=0$) the number of nurses showing up for work N is equal to the number of scheduled nurses y, and the nurse staffing problem reduces to a standard newsvendor model with the optimal staffing level given by

$$y_0^*=\min\left(y\in\mathcal{N}^+|F_T(y)\geq 1-\frac{w_r}{w_e}\right), \tag{10}$$

with

$$F_T(y)=\sum_{n=0}^{y}p_T(n) \tag{11}$$

being the cumulative density function of the demand function evaluated at y, and the value $1-\frac{w_r}{w_e}$ playing the role of the critical newsvendor fractile. Below we present an analysis of the staffing decision (9) starting with the case of exogenous absenteeism which we will use as a benchmark.

3.1. Optimal Nurse Staffing Under Exogenous Absenteeism Rate

Consider a clinical unit which experiences an endogenous nurses' absenteeism rate $\gamma(y)$, but treats it as exogenous. For example, the schedule planner uses the average value of all previously observed daily absenteeism rates, γ_{ave}. The cost function to be minimized under this approach is given by

$$W_{\text{ave}}(y)=y\,(w_r(1-\gamma_{\text{ave}})+w_n\gamma_{\text{ave}})+w_e\sum_{k=0}^{y}\sum_{n=0}^{\infty}(n-k)^+p_T(n)p(k;y,\gamma_{\text{ave}})=wy+w_e\sum_{k=0}^{y}q(k)p(k;y,\gamma_{\text{ave}}), \tag{12}$$

where $w=w_r(1-\gamma_{\text{ave}})+w_n\gamma_{\text{ave}}$ is the effective cost per scheduled nurse, and

$$q(k)=\sum_{n=0}^{\infty}(n-k)^+p_T(n), \tag{13}$$

represents an expected nursing shortage given that k regular nurses show up for work. The optimal staffing level in this case is expressed by the following result.

PROPOSITION 1. *a) The minimizer of (12) is given by*

$$y_{\text{ave}}^*=\min\left(y\in\mathcal{N}^+|\sum_{k=0}^{y}F_T(k)p(k;y,\gamma_{\text{ave}})\geq 1-\frac{w_r(1-\gamma_{\text{ave}})+w_n\gamma_{\text{ave}}}{w_e(1-\gamma_{\text{ave}})}\right), \tag{14}$$

and is a non-increasing function of $\frac{w_r}{w_e}$ and $\frac{w_n}{w_e}$.

b) Consider two cumulative distribution functions for the required number of nurses T, $F_T^1(k)$ and $F_T^2(k)$ such that $F_T^1(k)\geq F_T^2(k)$ for all $k\in\mathcal{N}^+$, and let $y_{\text{ave}}^{*,i}$ be the optimal staffing levels corresponding to $F_T^i(k)$, $i=1,2$. Then, $y_{\text{ave}}^{*,1}\leq y_{\text{ave}}^{*,2}$.

We relegate all the proofs to the Appendix. Note that (14) represents a generalization of the expression for the optimal staffing levels without absenteeism (10). As in the no-absenteeism setting, it is never optimal to decrease staffing levels when the target nursing level increases or when the cost advantage associated with earlier staffing becomes more pronounced. While this behavior of the optimal policy is intuitive, the dependence of the optimal staffing levels on the value of the absenteeism rate is not as straightforward. In particular, depending on the interplay between the ratios of the cost parameters $\frac{w_r}{w_e}$ and $\frac{w_n}{w_e}$, the characteristics of the target nursing level distribution and the absenteeism rate, the increase in the absenteeism rate can increase or decrease the optimal staffing level. The following result describes the properties of the optimal staffing levels in general settings.

PROPOSITION 2. *a) There exists $\gamma^{\mathrm{u}}_{\mathrm{ave}}$ such that the optimal staffing level y^*_{ave} is a non-increasing function of the absenteeism rate γ_{ave} for all $\gamma_{\mathrm{ave}} \in [\gamma^{\mathrm{u}}_{\mathrm{ave}}, 1]$.*

b) For

$$w_n \leq w_e \left(\left\lceil F_T^{-1}\left(1-\frac{w_r}{w_e}\right) \right\rceil \right) p_T \left(\left\lceil F_T^{-1}\left(1-\frac{w_r}{w_e}\right) \right\rceil \right) \tag{15}$$

there exists $\gamma^{\mathrm{l}}_{\mathrm{ave}}$ such that the optimal staffing level y^*_{ave} is a non-decreasing function of the absenteeism rate γ_{ave} for all $\gamma_{\mathrm{ave}} \in [0, \gamma^{\mathrm{l}}_{\mathrm{ave}}]$.

A more detailed characterization of the optimal staffing levels can be obtained for some target nursing level distributions, for example, for a discrete uniform distribution.

COROLLARY 1. *Let*

$$F_T(k) = \begin{cases} \frac{k+1}{T_{\max}+1}, & \text{for } 0 \leq k \leq T_{\max}, \\ 1, & k \geq T_{\max}. \end{cases} \tag{16}$$

Then, for $\frac{w}{w_e} \geq \frac{1}{4}$, the optimal nurse staffing level is given by

$$y^*_{\mathrm{ave}} = \left\lceil \left(\frac{T_{\max}}{1-\gamma_{\mathrm{ave}}} - \frac{T_{\max}+1}{(1-\gamma_{\mathrm{ave}})^2} \frac{w}{w_e} \right) \right\rceil, \tag{17}$$

and is a non-decreasing (non-increasing) function of γ_{ave} for $\gamma_{\mathrm{ave}} \leq \gamma^{\mathrm{u}}_{\mathrm{ave}}$ ($\gamma_{\mathrm{ave}} > \gamma^{\mathrm{u}}_{\mathrm{ave}}$), where

$$\gamma^{\mathrm{u}}_{\mathrm{ave}} = \max\left(0, 1 - \max\left(0, \frac{2\,(T_{\max}+1)\,w_r}{T_{\max} w_e - (T_{\max}+1)\,(w_r - w_n)}\right)\right). \tag{18}$$

In order to illustrate the monotonicity properties of the optimal staffing levels formalized in Proposition 2, we use the distribution for the number of required nurses obtained from our empirical data for the average patient census using 1-to-10 nurse-to-patient ratio. For this distribution, Figure 2 shows the dependence of the optimal staffing level on the absenteeism rate for $w = w_r = w_n$. For a given value of the cost ratio $\frac{w}{w_e}$ there exists a critical value of the absenteeism rate $\gamma^{\mathrm{u}}_{\mathrm{ave}}$ for which the optimal response to an increase in absenteeism switches from staffing more nurses to staffing fewer. Note that irrespective of the distribution for targeted nursing level, for high value of the

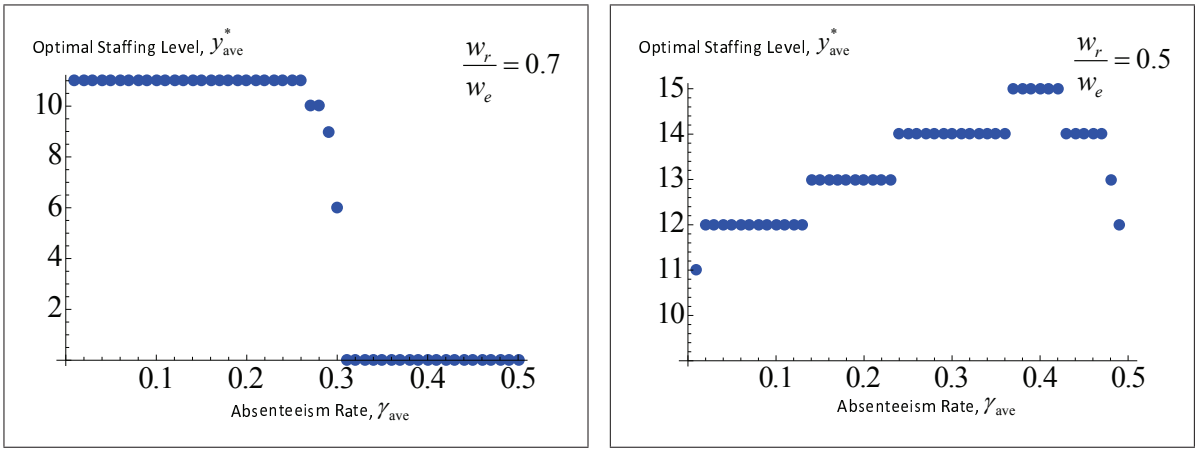

Figure 2: Optimal staffing level as a function of the absenteeism rate for different values of the cost ratio $\frac{w}{w_e}$ for $w = w_r = w_n$ and the empirical targeted nursing level distribution.

absenteeism rate or high value of the cost ratio $\frac{w}{w_e}$ (to be precise, for $\gamma_{\text{ave}} \geq 1 - \frac{w}{w_e}$), it is more cost-effective not to staff any nurses in advance and to rely exclusively on the extra/overtime mechanisms of supplying the nursing capacity. For low values of the absenteeism rate and low values of the cost ratio $\frac{w}{w_e}$, higher absenteeism can induce an increase in staffing levels, as it is cheaper to counter the increased absenteeism by staffing more nurses. However, as the cost ratio $\frac{w}{w_e}$ increases it becomes more cost-effective to staff fewer nurses, relying increasingly on the extra/overtime supply mechanism.

3.2. Endogenous Absenteeism: Optimal Staffing

In the endogenous absenteeism setting the expected staffing cost (9) becomes

$$W(y) = yw_r - (w_r - w_n)a(y) + w_e L(y, \gamma(y)), \tag{19}$$

where

$$a(y) = y\gamma(y), \tag{20}$$

is the expected number of absent nurses, and

$$L(y, \gamma(y)) = \sum_{k=0}^{y} q(k)p(k; y, \gamma(y)), \tag{21}$$

with $q(k)$ defined by (13), $p(k; y, \gamma(y))$ is the probability mass function of the binomial distribution where k nurses show up for work when y are scheduled and $\gamma(y)$ is the (endogenous) probability of a nurse being absent. Note that for general absenteeism rate function $\gamma(y)$ the increasing marginal property of the "exogenous" staffing cost function (12) with respect to the number of scheduled nurses may not hold. Below we formulate a sufficient condition for this property to be preserved

under endogenous absenteeism. First, for a given distribution of the targeted nursing level $p_T(k), k \geq 0$ we introduce the following quantity:

$$\gamma_T(y) = 1 - \min\left(1, \left(\frac{\sum_{k=y-2}^{\infty} p_T(k)}{y p_T(y-1) + p_T(y-2)}\right)^{\frac{1}{y-1}}\right), \; y \in \mathcal{N}^+, y \geq 2. \tag{22}$$

As shown below, (22) represents one of the bounds on the absenteeism rate function which ensures the optimality of the greedy-search approach to finding the optimal nurse staffing level.

PROPOSITION 3. *Let $\gamma(x) \in C^2(0,\infty), 0 \leq \gamma(x) \leq 1$ be a non-increasing, convex function defined on $x \geq 0$. Consider an endogenous absenteeism setting characterized by the absenteeism rate $\gamma(y)$ for $y \in \mathcal{N}^+$ scheduled nurses. Then, the optimal staffing level is given by*

$$y^* = \min\left(y \in \mathcal{N}^+ | L(y+1, \gamma(y+1)) - L(y, \gamma(y)) \geq -\frac{w_r}{w_e}(1-\gamma(y)) - \frac{w_n}{w_e}\gamma(y) + \frac{w_r - w_n}{w_e} y(\gamma(y+1) - \gamma(y))\right), \tag{23}$$

and is a non-increasing function of $\frac{w_r}{w_e}$, provided that

$$\gamma(y) \leq \min\left(\frac{2}{y}, \gamma_T(y)\right) \tag{24}$$

and

$$\frac{d^2 a(y)}{dy^2} \leq 0 \tag{25}$$

for any $y \geq y^*$. In addition, consider two cumulative distribution functions for the required number of nurses T, $F_T^1(k)$ and $F_T^2(k)$ such that $F_T^1(k) \geq F_T^2(k)$ for all $k \in \mathcal{N}^+$, and let $y^{*,i}$ be the optimal staffing levels corresponding to $F_T^i(k)$, $i = 1, 2$. Then, $y^{*,1} \leq y^{*,2}$, provided that (24) holds for any $y \geq y^{*,2}$.

The sufficient condition (24) states, intuitively, that the increasing marginal shape of the staffing cost function with respect to the number of scheduled nurses is preserved under endogenous absenteeism if the absenteeism rate is not too high, so that (19) is not too different from the cost function in (9). More specifically, this sufficient condition requires that the absenteeism rate function is limited from above by two separate bounds. The first bound implies that the expected number of absent nurses does not exceed 2 irrespective of the number of nurses actually scheduled for work. The sufficient condition (25) requires that the expected number of absent nurses exhibits non-increasing returns to scale.

To study the endogenous absenteeism case further we use a parametric specification consistent with our empirical findings, in particular with the logit model specification. We specify

$$\gamma(y) = \frac{1}{1 + e^{\alpha + \beta y}}, \tag{26}$$

where both α and β are positive constants. The assumption about positive values for these absenteeism rate parameters is plausible in a wide range of settings, $\beta > 0$ implies that the absenteeism rate declines with the number of scheduled nurses, while $\alpha > 0$ ensures that the absenteeism rate is not too high even when the number of scheduled nurses is low and the anticipated workload is high. In particular, evaluating the best-fit logit model in (3) using the estimates reported in Table 1 we obtain $\beta = -\frac{\beta^1_{\text{ENPR}}}{E[C_t]} = 0.092$, with $\beta^1_{\text{ENPR}} = -10.01$, $E[C_t] = 109.0$. In order for the average absenteeism rate to match our sample average of 7.34% we set $\alpha = 1.533$. Note that the endogenous absenteeism rate $\gamma(y)$ characterized by the logistic function given by (26) with $\alpha \geq 0$ and $\beta \geq 0$ is a monotone decreasing convex function. Thus, the result of Proposition 3 is ensured by the following restrictions on the values of α and β:

LEMMA 1. *For $\gamma(y) = \frac{1}{1+e^{\alpha+\beta y}}$, with $\alpha, \beta \geq 0$, $\beta e^{1+\alpha+2\beta} \geq \frac{1}{2}$ implies (24), and $\beta \leq 2$ implies (25).*

In the ED we studied the estimated values of $\alpha = 1.533$ and $\beta = 0.092$ satisfy Lemma 1. In particular, the maximum value of the product of number of scheduled nurses y and the estimated absenteeism rate calculated using these values is equal to 0.804, well below 2. The second bound on the right-hand side of (24) takes the form of an effective absenteeism rate function which depends exclusively on the distribution of the targeted nursing level. Note that $\gamma_T(y) \geq 0$ if and only if

$$\frac{p_T(y-1)}{\sum_{k=y-1}^{\infty} p_T(k)} \geq \frac{1}{y}. \tag{27}$$

The expression of the left-hand side of (27) is the hazard rate function for the distribution of the targeted nursing level. Thus (27) stipulates that the bound described by (24) is meaningful only in settings where such a hazard rate evaluated at y exceeds $\frac{1}{y+1}$. For the absenteeism rate function given by (26), the constraint $\gamma(y) \leq \gamma_T(y)$ implies, in the same spirit as Lemma 1the lower-bound restriction on the values of α and β:

$$\gamma(y) \leq \gamma_T(y) \Leftrightarrow \alpha + \beta y \geq \log\left(\frac{1-\gamma_T(y)}{\gamma_T(y)}\right). \tag{28}$$

Figure 3 compares the absenteeism rate (26) computed for $\alpha = 1.533$ and $\beta = 0.092$ with the effective absenteeism rate $\gamma_T(y)$ from (22) computed using the empirical targeted nursing level distribution. Note that the sufficient condition of Proposition 3 ($\gamma(y) \leq \gamma_T(y)$) is satisfied for virtually any staffing level above 10 nurses which is approximately equal to the expected number of required nurses in our setting.

3.3. Endogenous Absenteeism: Implications of Model Misspecification on Staffing Decisions

In this section we compare the optimal nurse staffing levels with those made by a clinical unit which incorrectly treats the absenteeism rate as being exogenous and employs a trial-and-error procedure

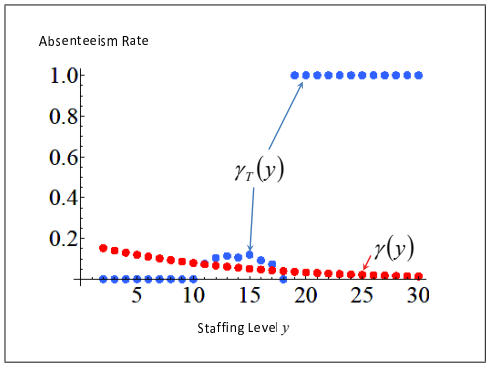

Figure 3: Absenteeism rate $\gamma(y)$ computed for $\alpha = 1.533$ and $\beta = 0.092$ and the effective absenteeism rate $\gamma_T(y)$ computed using the empirical targeted nursing level distribution.

under which the assumed exogenous absenteeism rate is updated every time a new staffing decision in made. In this latter case which we label "misspecified-with-learning", the clinical unit selects staffing level y^{ML} such that

$$y^{\mathrm{ML}} = \min\left(y \in \mathcal{N}^+ | \sum_{k=0}^{y} F_T(k)p(k; y, \gamma(y^{\mathrm{ML}})) \geq 1 - \frac{w_r(1-\gamma(y^{\mathrm{ML}})) + w_n\gamma(y^{\mathrm{ML}})}{w_e(1-\gamma(y^{\mathrm{ML}}))}\right), \tag{29}$$

with $\gamma(y)$ denoting the endogenous absenteeism rate. Note that (29) reflects a self-consistent way of selecting the staffing level; y^{ML} is the best staffing decision in the setting where the absenteeism rate is exogenous and determined by $\gamma(y^{\mathrm{ML}})$. In other words, a clinical unit assuming that the absenteeism rate is given by constant value $\gamma(y^{\mathrm{ML}})$ will respond by scheduling y^{ML} nurses and, as a result of this decision, will observe exactly the same value of the absenteeism rate, even if the true absenteeism process is endogenous and described by $\gamma(y)$. An intuitive way of rationalizing the choice of y^{ML} is to consider a sequence of "exogenous" staffing levels $y_n, n \in \mathcal{N}^+$, such that

$$y_{n+1} = \min\left(y \in \mathcal{N}^+ | \sum_{k=0}^{y} F_T(k)p(k; y, \gamma(y_n)) \geq 1 - \frac{w_r(1-\gamma(y_n)) + w_n\gamma(y_n)}{w_e(1-\gamma(y_n))}\right), n \in \mathcal{N}^+. \tag{30}$$

Equation (30) reflects a sequence of repeated adjustments of staffing levels, starting with some y_0, each based on the value of the absenteeism rate observed after the previously chosen staffing level is implemented. In this updating scheme, y^{ML} can be thought of as the limit $\lim_{n\to\infty} y_n$, if such a limit exists. It is important to note that for a general demand distribution $F_T(k)$ and a general absenteeism rate function $\gamma(y)$, the set of staffing levels E satisfying (29) may be empty or may contain multiple elements. The analysis of existence and uniqueness of y^{ML} is further complicated by the discrete nature of staffing levels. In the following discussion we by-pass this analysis and assume that there exists at least one staffing level satisfying (29). As the following result shows, even if E contains multiple elements, each of them is bounded from above by the optimal "endogenous" staffing level in settings where the expected number of absentees decreases with the number of scheduled nurses.

PROPOSITION 4. *Suppose that the conditions of Proposition 3 hold and that the set of staffing levels satisfying (29), E, is non-empty. Then,* $y^{\mathrm{ML}} \leq y^*$*, for any* $y^{\mathrm{ML}} \in E$*, provided that, at the optimal staffing level* y^**, the expected number of absent nurses decreases with the number of nurses scheduled,* $a(y^*+1) < a(y^*)$*.*

Proposition 4 implies that ignoring the endogenous nature of absenteeism can lead to understaffing in settings where both the endogenous absenteeism rate and the expected number of absent nurses decline with the number of scheduled nurses. Figure 4 illustrates the results of the numerical experiment designed to quantify a potential cost impact of using heuristic staffing policies for realistic values of problem parameters. In our study, we have varied the cost ratio $\frac{w_r}{w_e}$ from $\frac{w_r}{w_e} = 0.5$ to $\frac{w_r}{w_e} = 0.7$. The lower limit of this interval, $\frac{w_r}{w_e} = 0.5$ corresponds to the setting in which use of agency nurses carries a 100% cost premium, a realistic upper bound on the values encountered in practice. The upper limit, $\frac{w_r}{w_e} = 0.7$, reflects the use of overtime to compensate for absenteeism, with w_e at a 50% premium with respect to w_r. In the ED we studied, absent nurses were paid at the same rate as the nurses who showed up for work (so that $w_n = w_r$). In order to investigate the effect of lower compensation levels for absent nurses, we have included in our study the cost ratios $w_n = 0.5w_r$ and $w_n = 0$. As the absenteeism rate function, we have used $\gamma(y) = \frac{1}{1+e^{\beta y}}$ with the value of β varied to explore different average absenteeism rates calculated as

$$\gamma_{\mathrm{ave}} = \sum_{y=y_{\min}}^{y_{\max}} \frac{p_T(y)}{1+e^{\beta y}}, \tag{31}$$

where $y_{\min} = 6$ and $y_{\max} = 16$ reflect the smallest and largest possible targeted nursing level realizations, and $p_T(y)$ reflecting the empirical distribution for the required number of nurses. In our study, we have compared the performance of three staffing policies: the optimal policy described in Proposition 3, the ML policy described by (29), and the "exogenous" policy which assumes that the absenteeism rate does not depend on the staffing level and is given by (31). Figure 4 shows the worst-case performance gaps (calculated over the range of cost ratios $\frac{w_r}{w_e} \in [0.5, 0.7]$ of the ML and the "exogenous" policies as functions of γ_{ave} for three ratio levels $\frac{w_r}{w_e} = 1, 0.5$, and 0. Our numerical results indicate that in the settings where the average absenteeism rate γ_{ave} is small, ML and "exogenous" policies represent good approximations for the optimal staffing policies. For example, in the ED we studied the average absenteeism rate was 7.3%, and the corresponding worst-case performances for these two policies were between 2% and 3% for $w_n = w_r$. However, as γ_{ave} increases, so does the worst-case performance gap for both policies: in particular, in the same setting, the worst case performance gaps approximately double to 4% (6%) for the "exogenous" (ML) policy when the average absenteeism rate reaches 15%. As it turns out, a reduction in the amount of hourly compensation paid to absent nurses significantly closes these performance gaps for both policies: for $w_n = 0$ the maximum performance gaps drop below 1%.

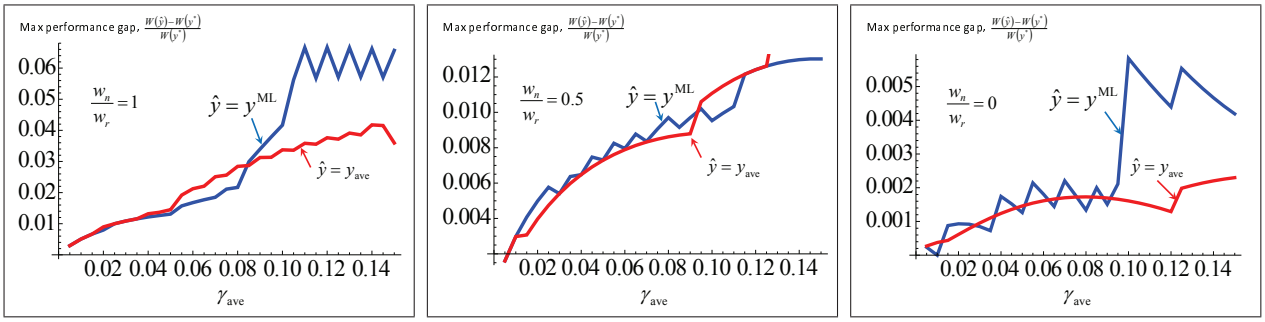

Figure 4: Worst-case performance gaps between the optimal staffing policy and the "misspecified-with-learning" (ML) and "exogenous" policies as functions of the average absenteeism rate γ_{ave} under the empirical targeted nursing level distribution for $\frac{w_n}{w_r}=1, 0.5, 0$.

4. Discussion

In our empirical study we use observations from a large urban hospital ED to study nurse absenteeism behavior at the shift level. We find that nurse absenteeism is exacerbated when fewer nurses are scheduled for a particular shift. This is consistent with nurses exhibiting an aversion to higher levels of anticipated workload. Our study relies on aggregate data from a single department and thus does not permit a detailed investigation of the impact of workload at the individual nurse level. We leave such an extension to further research. It is nevertheless the first study to demonstrate that staffing decisions have an impact on shift-level worker absenteeism. This is a fact that seems not to have been examined by extant staffing literature.

On the analytical front we examine the implications of absenteeism, both exogenous and endogenous, on optimal staffing policies. To do so we develop an extension to the single-period newsvendor model which explicitly accounts for uncertainty in patient census and in the number of nurses that show up for work. We use the model to derive structural properties and to demonstrate that the failure to properly account for the endogenous nature of nurse absenteeism can lead to deviations from optimal staffing decisions. In particular, in settings where higher numbers of scheduled nurses result in lower absenteeism rates and lower expected absentee numbers, a clinical unit which treats absenteeism as an exogenous phenomenon will often under-supply nursing staff capacity even if allowed to repeatedly adjust its staffing decisions in response to observed absenteeism rates. As such, our paper can be viewed as a contribution to the emerging literature on model misspecification error and as the first paper to show the potential impact of model specification error in the context of staffing.

We believe that the presence of endogenous absenteeism gives rise to systematic understaffing, which, in turn, has important practical consequences for hospitals. First, as our model demonstrates, endogenous absenteeism gives rise to noticeable cost increases even in settings with low absenteeism rates, as long as absenteeism exhibits a substantial degree of endogeneity and as long

as monetary compensation for absent nurses is comparable to that of nurses who show up for work. For model parameters that are representative of the hospital we study, we find that the cost of ignoring the endogenous nature of absenteeism can be about $2-3\%$. Second, such chronic understaffing harms patients, especially in the life-and-death setting of an ED. Third, it is likely to be a contributing factor to the widely reported nurse job dissatisfaction (Aiken et al. (2002)). Our research points to an important opportunity for the cash-constrained hospitals to improve quality of patient care as well as nurse working conditions, while reducing operating costs.

Turning to the specific context of our analysis, it is important to note that our assumption about the unlimited availability of extra/overtime nursing capacity may not be valid in some clinical environments. In such environments it may be impossible to replace absent nurses at a reasonable cost or in reasonable time, and the endogeneity of absenteeism can lead to significant understaffing with the possibility of serious deterioration of service quality and longer ED delays. In other clinical units the use of agency nurses who may be less familiar with the unit can lead to similar declines in quality of patient care and in an increase in the rate of medical errors. Thus, an accurate understanding of the nature of nurse absenteeism and the use of a model that accurately incorporates this phenomenon in determining appropriate staffing levels is imperative to assuring high quality patient care.

5. Appendix

Below we present outlines of the proofs of the analytical results. Detailed proofs are available from the authors.

Proof of Proposition 1. Note that $W_{\text{ave}}(y+1)-W_{\text{ave}}(y)=w+w_e\left(L(y+1,\gamma_{\text{ave}})-L(y,\gamma_{\text{ave}})\right)=w+w_e\Delta L(y,\gamma_{\text{ave}})$. We will establish the result of the proposition by showing that $\Delta L(y,\gamma_{\text{ave}})\leq 0$ and
$\Delta^2 L(y,\gamma_{\text{ave}})=L(y+2,\gamma_{\text{ave}})-2L(y+1,\gamma_{\text{ave}})+L(y,\gamma_{\text{ave}})\geq 0$. First, note that $q(k+1)-q(k)=-\sum_{n=k+1}^{\infty}p_T(n)\leq 0$ and $q(k+2)-2q(k+1)+q(k)=p_T(k+1)\geq 0$. Then, $\Delta L(y,\gamma_{\text{ave}})=\sum_{k=0}^{y+1}q(k)(p(k;y+1,\gamma_{\text{ave}})-p(k;y,\gamma_{\text{ave}}))$, where we have used $p(y+1;y,\gamma_{\text{ave}}))\equiv 0$. Next, using

$$p(k;y+1,\gamma_{\text{ave}})=(1-\gamma_{\text{ave}})p(k-1;y,\gamma_{\text{ave}})+\gamma_{\text{ave}}p(k;y,\gamma_{\text{ave}}),k=0,\ldots,y \tag{32}$$

we get

$$\sum_{k=0}^{y+1}q(k)(p(k;y+1,\gamma_{\text{ave}})-p(k;y,\gamma_{\text{ave}}))=(1-\gamma_{\text{ave}})\sum_{k=0}^{y}\left(q(k+1)-q(k)\right)p(k;y,\gamma_{\text{ave}})\leq 0, \tag{33}$$

where we have used $\sum_{k=0}^{y+1}q(k)p(k-1;y,\gamma_{\text{ave}})=\sum_{k=0}^{y+1}q(k+1)p(k;y,\gamma_{\text{ave}})$. Note that $p(k-1;y,\gamma_{\text{ave}})\equiv 0$. From $p(y+2;y,\gamma_{\text{ave}})\equiv 0$ and (32), we get

$$\Delta^2 L(y,\gamma_{\text{ave}})=(1-\gamma_{\text{ave}})^2\left(\sum_{k=0}^{y+2}\left(q(k+1)-q(k)\right)\left(p(k-1;y,\gamma_{\text{ave}})-p(k;y,\gamma_{\text{ave}})\right)\right). \tag{34}$$

Note that $\sum_{k=0}^{y+2}(q(k+1)-q(k))\,p(k-1;y,\gamma_{\text{ave}}) = \sum_{k=0}^{y+2}(q(k+2)-q(k+1))\,p(k;y,\gamma_{\text{ave}})$ where we have used $p(-1;y,\gamma_{\text{ave}}) \equiv p(y+1;y,\gamma_{\text{ave}}) \equiv p(y+2;y,\gamma_{\text{ave}}) \equiv 0$. Thus, we obtain for $\Delta^2 L(y,\gamma_{\text{ave}}) = (1-\gamma_{\text{ave}})^2\left(\sum_{k=0}^{y}(q(k+2)-2q(k+1)+q(k))\,p(k;y,\gamma_{\text{ave}})\right) \geq 0$. Further, note that $\Delta L(y,\gamma_{\text{ave}}) \geq -\frac{w}{w_e}$ is equivalent to $\sum_{k=0}^{y}F_T(k)p(k;y,\gamma_{\text{ave}}) \geq 1-\frac{w}{w_e(1-\gamma_{\text{ave}})}$. Now, consider $F_T^1(k)$ and $F_T^2(k)$ such that $F_T^1(k) \geq F_T^2(k)$ for any $k \in \mathcal{N}^+$. Then, $\Delta L^1(y,\gamma_{\text{ave}}) \geq \Delta L^2(y,\gamma_{\text{ave}})$ for any $y \in \mathcal{N}^+$ and, respectively, $y_{\text{ave}}^{*,1} = \min\left(y \in \mathcal{N}^+ | \Delta L^1(y,\gamma_{\text{ave}}) \geq -\frac{w}{w_e}\right) \leq \min\left(y \in \mathcal{N}^+ | \Delta L^2(y,\gamma_{\text{ave}}) \geq -\frac{w}{w_e}\right) = y_{\text{ave}}^{*,2}$. □

Proof of Proposition 2. First, we introduce $G(y,\gamma) = \sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma)$ and note that

$$G(y+1,\gamma) = (1-\gamma)\sum_{k=1}^{y+1}(1-F_T(k))\,p(k-1;y,\gamma) + \gamma\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma), \tag{35}$$

where we have used (32). Then,

$$G(y+1,\gamma) \leq (1-\gamma)\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma) + \gamma\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma) = G(y,\gamma). \tag{36}$$

Further,

$$\begin{aligned}\frac{\partial G}{\partial \gamma} &= y\left(\sum_{k=0}^{y-1}(1-F_T(k))\,\frac{(y-1)!}{k!(y-1-k)!}\left(\gamma^{y-1-k}(1-\gamma)^k\right)\right)\\ &\quad - y\left(\sum_{k-1=0}^{y-1}(1-F_T(k))\,\frac{y!}{(k-1)!(y-1-(k-1))!}\left(\gamma^{y-1-(k-1)}(1-\gamma)^{k-1}\right)\right)\\ &= y\left(\sum_{k=0}^{y-1}(F_T(k+1)-F_T(k))\,p(k;y-1,\gamma)\right) \geq 0.\end{aligned} \tag{37}$$

The optimality condition (14) can be re-written as $y_{\text{ave}}^* = \min\left(y \in \mathcal{N}^+ | G(y,\gamma_{\text{ave}}) \leq \frac{w_r-w_n}{w_e} + \frac{w_n}{w_e(1-\gamma_{\text{ave}})}\right)$. Since y is a discrete variable and γ_{ave} is a continuous parameter, there exist a set of values $\gamma_{\text{ave}}^i \in [0,1]$, $i=1,...,I_{\max}$, with $\gamma_{\text{ave}}^1 = 0$ and $\gamma_{\text{ave}}^{I_{\max}} = 1$, with the optimal $y_{\text{ave}}^*(\gamma_{\text{ave}})$ remaining constant in each interval $(\gamma_{\text{ave}}^i, \gamma_{\text{ave}}^{i+1})$, $i=1,...,I_{\max}-1$, and exhibiting finite jumps at γ_{ave}^i, i.e., $y_{\text{ave}}^*(\gamma_{\text{ave}}^i-0) \neq y_{\text{ave}}^*(\gamma_{\text{ave}}^i+0)$, $i=2,...,I_{\max}-1$. Note that the sign of the difference $y_{\text{ave}}^*(\gamma_{\text{ave}}^i+0) - y_{\text{ave}}^*(\gamma_{\text{ave}}^i-0)$ is completely determined by the sign of

$$H(\gamma_{\text{ave}}^i) = \frac{\partial G\left(\min\left(y_{\text{ave}}^*(\gamma_{\text{ave}}^i-0), y_{\text{ave}}^*(\gamma_{\text{ave}}^i+0)\right), \gamma_{\text{ave}}^i\right)}{\partial \gamma_{\text{ave}}^i} - \frac{w_n}{w_e(1-\gamma_{\text{ave}}^i)^2}. \tag{38}$$

Indeed, suppose that $y_{\text{ave}}^*(\gamma_{\text{ave}}^i+0) > y_{\text{ave}}^*(\gamma_{\text{ave}}^i-0)$. Then, we should have $G(y_{\text{ave}}^*(\gamma_{\text{ave}}^i-0), \gamma_{\text{ave}}^i-0) \leq \frac{w_r-w_n}{w_e} + \frac{w_n}{w_e(1-(\gamma_{\text{ave}}^i-0))}$ and $G(y_{\text{ave}}^*(\gamma_{\text{ave}}^i-0), \gamma_{\text{ave}}^i+0) > \frac{w_r-w_n}{w_e} + \frac{w_n}{w_e(1-(\gamma_{\text{ave}}^i+0))}$. Combining these two expressions, we get

$$G(y_{\text{ave}}^*(\gamma_{\text{ave}}^i-0), \gamma_{\text{ave}}^i+0) - G(y_{\text{ave}}^*(\gamma_{\text{ave}}^i-0), \gamma_{\text{ave}}^i-0) > \frac{w_n}{w_e(1-(\gamma_{\text{ave}}^i+0))} - \frac{w_n}{w_e(1-(\gamma_{\text{ave}}^i-0))}, \tag{39}$$

or $\frac{\partial G(y_{\text{ave}}^*(\gamma_{\text{ave}}^i-0), \gamma_{\text{ave}}^i-0)}{\partial \gamma_{\text{ave}}} > \frac{w_n}{w_e(1-(\gamma_{\text{ave}}^i-0))^2}$. Similarly, $y_{\text{ave}}^*(\gamma_{\text{ave}}^i+0) < y_{\text{ave}}^*(\gamma_{\text{ave}}^i-0)$ implies that $\frac{\partial G(y_{\text{ave}}^*(\gamma_{\text{ave}}^i+0), \gamma_{\text{ave}}^i-0)}{\partial \gamma_{\text{ave}}} < \frac{w_n}{w_e(1-(\gamma_{\text{ave}}^i-0))^2}$. Note that according to (37) $\frac{\partial G(y^*, \gamma_{\text{ave}}^i)}{\partial \gamma_{\text{ave}}^i} =$

$y^*\left(\sum_{k=0}^{y^*-1} p_T(k+1)p(k;y^*-1,\gamma^i_{\text{ave}})\right)$, so that (38) can be expressed as $H(\gamma^i_{\text{ave}}) = \hat{y}\left(\sum_{k=0}^{\hat{y}-1} p_T(k+1)p(k;\hat{y}-1,\gamma^i_{\text{ave}})\right) - \frac{w_n}{w_e(1-\gamma^i_{\text{ave}})^2}$, where $\hat{y} = \min\left(y^*_{\text{ave}}(\gamma^i_{\text{ave}}-0), y^*_{\text{ave}}(\gamma^i_{\text{ave}}+0)\right)$. Note that for $\gamma^i_{\text{ave}} \to 0$, $y^*_{\text{ave}}(\gamma^i_{\text{ave}}) \to y^*_0$, as expressed in (10), so that $y^*_0 = \left\lceil F_T^{-1}\left(1-\frac{w_r}{w_e}\right)\right\rceil$. Then, $H(\gamma^i_{\text{ave}} \to 0) = \left(\left\lceil F_T^{-1}\left(1-\frac{w_r}{w_e}\right)\right\rceil\right) p_T\left(\left\lceil F_T^{-1}\left(1-\frac{w_r}{w_e}\right)\right\rceil\right) - \frac{w_n}{w_e}$. Thus, $\left(\left\lceil F_T^{-1}\left(1-\frac{w_r}{w_e}\right)\right\rceil\right) p_T\left(\left\lceil F_T^{-1}\left(1-\frac{w_r}{w_e}\right)\right\rceil\right) \geq \frac{w_n}{w_e}$ implies that there exist γ^l_{ave} such that y^*_{ave} is a non-decreasing function of γ_{ave} for $\gamma_{\text{ave}} \leq \gamma^l_{\text{ave}}$. On the other hand, $\gamma^i_{\text{ave}} \geq \frac{w_e-w_r}{w_n+w_e-w_r}$ implies that $y^*_{\text{ave}}(\gamma^i_{\text{ave}}) = 0$, and for $\gamma^i_{\text{ave}} \to \frac{w_e-w_r}{w_n+w_e-w_r}$, $y^*_{\text{ave}}(\gamma^i_{\text{ave}}) \to 0$, and $H\left(\gamma^i_{\text{ave}} \to \frac{w_e-w_r}{w_n+w_e-w_r}\right) = -\frac{(w_n+w_e-w_r)^2}{w_e w_n} < 0$. Thus, there exist γ^u_{ave} such that y^*_{ave} is a non-increasing function of γ_{ave} for $\gamma_{\text{ave}} \geq \gamma^u_{\text{ave}}$. □

Proof of Corollary 1. Under the discrete uniform demand distribution specified by (16), the sum in the expression for the optimal staffing level (14) becomes

$$\sum_{k=0}^{y} F_T(k)p(k;y,\gamma_{\text{ave}}) = \begin{cases} \frac{y(1-\gamma_{\text{ave}})+1}{T_{\max}+1}, & \text{for } y \leq T_{\max}, \\ \frac{1}{T_{\max}+1}\sum_{k=0}^{T_{\max}}(k+1)p(k;y,\gamma_{\text{ave}}) + \sum_{k=T_{\max}+1}^{y} p(k;y,\gamma_{\text{ave}}), & y \geq T_{\max}+1. \end{cases} \tag{40}$$

Note that for $y = T_{\max}$, (40) becomes $\frac{T_{\max}(1-\gamma_{\text{ave}})+1}{T_{\max}+1} = 1 - \frac{\gamma_{\text{ave}}T_{\max}}{T_{\max}+1} \geq 1 - \frac{w}{w_e(1-\gamma_{\text{ave}})}$, as long as $\frac{w}{w_e} \geq \gamma_{\text{ave}}(1-\gamma_{\text{ave}})\frac{T_{\max}}{T_{\max}+1}$. The supremum of the right-hand side of this expression is $\frac{1}{4}$ (for $\gamma_{\text{ave}} = 0.5$ and $T_{\max} \to \infty$), so that this expression is implied by $\frac{w}{w_e} \geq \frac{1}{4}$. Thus, under this condition, the optimal staffing level does not exceed $T_{\max}$ and, consequently, $y^*_{\text{ave}} = \min\left(y \in \mathcal{N}^+ \mid \frac{y(1-\gamma_{\text{ave}})+1}{T_{\max}+1} \geq 1 - \frac{w}{w_e(1-\gamma_{\text{ave}})}\right) = \left\lceil\left(\frac{T_{\max}}{1-\gamma_{\text{ave}}} - \frac{T_{\max}+1}{(1-\gamma_{\text{ave}})^2}\frac{w}{w_e}\right)\right\rceil$. Further, differentiating the expression under the "ceiling" function on the right-hand side with respect to γ_{ave}, we get $\frac{1}{(1-\gamma_{\text{ave}})^3}\left(\left(T_{\max} - (T_{\max}+1)\left(\frac{w_r-w_n}{w_e}\right)\right)(1-\gamma_{\text{ave}}) - 2(T_{\max}+1)\frac{w_r}{w_e}\right)$, which is non-negative (non-positive) if and only if $\gamma_{\text{ave}} \leq \gamma^{\text{u}}_{\text{ave}}$ ($\gamma_{\text{ave}} \geq \gamma^{\text{u}}_{\text{ave}}$). □

Proof of Proposition 3. Using $L(y,\gamma(y)) = \sum_{k=0}^{y} q(k)p(k;y,\gamma(y))$, we have $W(y+1) - W(y) = w_r - (w_r - w_n)\Delta a(y) + w_e\Delta L(y,\gamma(y))$, where $\Delta L(y,\gamma(y)) = L(y+1,\gamma(y+1)) - L(y,\gamma(y))$, and $\Delta a(y) = a(y+1) - a(y)$. We proceed by identifying sufficient conditions for $\Delta L(y,\gamma(y)) \leq 0$, $\Delta^2 L(y,\gamma(y)) = \Delta L(y+1,\gamma(y+1)) - \Delta L(y,\gamma(y)) \geq 0$, and $(w_r - w_n)\Delta^2 a(y) = (w_r - w_n)(\Delta a(y+1) - \Delta a(y)) \leq 0$. Since $\gamma(y)$ is continuous and twice differentiable, it follows immediately that $a(y)$ is also continuous and twice differentiable therefore a sufficient condition for $(w_r - w_n)(\Delta a(y+1) - \Delta a(y)) \leq 0$ is given by $(w_r - w_n)\frac{d^2}{dy^2}a(y) \leq 0$. Now, $\Delta L(y,\gamma(y)$ can be written as

$$\sum_{k=0}^{y+1} q(k)p(k;y+1,\gamma(y+1)) - \sum_{k=0}^{y} q(k)p(k;y,\gamma(y+1)) + \sum_{k=0}^{y} q(k)p(k;y,\gamma(y+1)) - \sum_{k=0}^{y} q(k)p(k;y,\gamma(y)). \tag{41}$$

We are now going to consider separately the first two and the last two terms in equation (41) and show that both are non-positive. The first two terms can be expressed as $-(1-\gamma(y+$

$1))\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma(y+1))$ which is non-positive. Next we examine the second two terms, which can be expressed as

$$\sum_{k=0}^{y} q(k)\int_{y}^{y+1}\frac{\partial p(k;y,\gamma(s))}{\partial s}ds=\int_{y}^{y+1}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y}q(k)\frac{\partial p(k;y,\gamma(s))}{\partial\gamma(s)}. \tag{42}$$

Note that

$$\begin{aligned}
&\sum_{k=0}^{y} q(k)\frac{\partial p(k;y,\gamma(s))}{\partial\gamma(s)}=\sum_{k=0}^{y}q(k)\left(\left(\frac{y!}{k!(y-k)!}\right)(y-k)\gamma(s)^{y-k-1}(1-\gamma(s))^{k}-k\gamma(s)^{y-k}(1-\gamma(s))^{k-1}\right)\\
&=\sum_{k=0}^{y-1}q(k)\left(\frac{y!}{k!(y-k-1)!}\right)\gamma(s)^{y-k-1}(1-\gamma(s))^{k}-\sum_{k=1}^{y}q(k)\left(\frac{y!}{(k-1)!(y-k)!}\right)\gamma(s)^{y-k}(1-\gamma(s))^{k-1}\\
&=y\sum_{k=0}^{y-1}q(k)p(k;y-1,\gamma(s))-y\sum_{k=1}^{y}q(k)p(k-1;y-1,\gamma(s))=y\sum_{k=0}^{y-1}(1-F_T(k))\,p(k;y-1,\gamma(s)),
\end{aligned} \tag{43}$$

so that (42) becomes $\sum_{k=0}^{y}q(k)(p(k;y,\gamma(y+1))-p(k;y,\gamma(y)))=y\int_{y}^{y+1}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y-1}(1-F_T(k))\,p(k;y-1,\gamma(s))$. A sufficient condition for this last expression to be negative is $\frac{d\gamma(s)}{ds}\le 0$. Thus, $\Delta L(y,\gamma(y))\le 0$. Further, consider $\Delta^2 L(y,\gamma(y))$ expressed as

$$\begin{aligned}
&-(1-\gamma(y+2))\sum_{k=0}^{y+1}(1-F_T(k))\,p(k;y+1,\gamma(y+2))+(1-\gamma(y+1))\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma(y+1))\\
&+(y+1)\int_{y+1}^{y+2}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma(s))-y\int_{y}^{y+1}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y-1}(1-F_T(k))\,p(k;y-1,\gamma(s)).
\end{aligned} \tag{44}$$

Focusing on the first line in (44), we get

$$\begin{aligned}
&-(1-\gamma(y+2))\sum_{k=0}^{y+1}(1-F_T(k))\,p(k;y+1,\gamma(y+2))+(1-\gamma(y+1))\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma(y+1))\\
&=(1-\gamma(y+2))\left(\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma(y+2))-\sum_{k=0}^{y+1}(1-F_T(k))\,p(k;y+1,\gamma(y+2))\right)\\
&+\sum_{k=0}^{y}(1-F_T(k))\left((1-\gamma(y+1))p(k;y,\gamma(y+1))-(1-\gamma(y+2))p(k;y,\gamma(y+2))\right).
\end{aligned} \tag{45}$$

The second line above is equivalent to $(1-\gamma(y+2))^2\left(\sum_{k=0}^{y}p_T(k+1)p(k;y,\gamma(y+2))\right)$ and is non-negative. the last line in (45) is equal to $-\sum_{k=0}^{y}(1-F_T(k))\int_{y+1}^{y+2}\frac{\partial((1-\gamma(s))p(k;y,\gamma(s)))}{\partial s}ds$ or $-\sum_{k=0}^{y}(1-F_T(k))\int_{y+1}^{y+2}\frac{\partial((1-\gamma(s))p(k;y,\gamma(s)))}{\partial\gamma(s)}\frac{d\gamma(s)}{ds}ds$. Focusing on the second line in (44), we obtain

$$\begin{aligned}
&(y+1)\int_{y+1}^{y+2}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma(s))-y\int_{y+1}^{y+2}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y-1}(1-F_T(k))\,p(k;y-1,\gamma(s))\\
&+y\int_{y+1}^{y+2}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y-1}(1-F_T(k))\,p(k;y-1,\gamma(s))-y\int_{y}^{y+1}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y-1}(1-F_T(k))\,p(k;y-1,\gamma(s))\\
&=(y+1)\int_{y+1}^{y+2}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma(s))-y\int_{y+1}^{y+2}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y-1}(1-F_T(k))\,p(k;y-1,\gamma(s))
\end{aligned}$$

$$
\begin{aligned}
&+ y\int_{y+1}^{y+2} ds \sum_{k=0}^{y-1} (1-F_T(k)) \left[\frac{d\gamma(s)}{ds} p(k;y-1,\gamma(s)) - \frac{d\gamma(s-1)}{ds} p(k;y-1,\gamma(s-1))\right] \\
&= \int_{y+1}^{y+2} ds \frac{d\gamma(s)}{ds} \sum_{k=0}^{y} (1-F_T(k)) \left((y+1)p(k;y,\gamma(s)) - yp(k;y-1,\gamma(s))\right) \\
&+ y\int_{y+1}^{y+2} ds \int_{s-1}^{s} d\xi \left[\sum_{k=0}^{y-1} (1-F_T(k)) \frac{\partial}{\partial \xi}\left[\frac{d\gamma(\xi)}{d\xi} p(k;y-1,\gamma(\xi))\right]\right]. \qquad (46)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\int_{y+1}^{y+2} ds \frac{d\gamma(s)}{ds} \sum_{k=0}^{y} (1-F_T(k)) \left(-\frac{\partial(1-\gamma(s))p(k;y,\gamma(s))}{\partial\gamma(s)} + (y+1)p(k;y,\gamma(s)) - yp(k;y-1,\gamma(s))\right) \\
&+ y\int_{y+1}^{y+2} ds \int_{s-1}^{s} d\xi \left[\sum_{k=0}^{y-1} (1-F_T(k)) \frac{\partial}{\partial \xi}\left[\frac{d\gamma(\xi)}{d\xi} p(k;y-1,\gamma(\xi))\right]\right] \\
&= -2\int_{y+1}^{y+2} \frac{ds}{\gamma(s)} \frac{d\gamma(s)}{ds} \sum_{k=0}^{y} (1-F_T(k)) p(k;y,\gamma(s)) \left(y(1-\gamma(s)) - k - \gamma(s)\right) \\
&+ y\int_{y+1}^{y+2} ds \int_{s-1}^{s} d\xi \sum_{k=0}^{y-1} (1-F_T(k)) \frac{d^2\gamma(\xi)}{d\xi^2} + y\int_{y+1}^{y+2} ds \int_{s-1}^{s} d\xi \\
&\times \left(\frac{d\gamma}{d\xi}\right)^2 \frac{1}{\gamma(\xi)(1-\gamma(\xi))} \left[\sum_{k=0}^{y-1} (1-F_T(k)) p(k;y-1,\gamma(\xi))((y-1)(1-\gamma(\xi)) - k)\right]. \qquad (47)
\end{aligned}
$$

A sufficient condition for the second line of (47) to be non-negative is $\frac{d^2\gamma(\xi)}{d\xi^2} \geq 0$. Since $\frac{d\gamma(s)}{ds} \leq 0$, a sufficient condition for the first line of (47) to be non-negative is that, for given y,

$$
\sum_{k=0}^{y} (1-F_T(k)) p(k;y,\gamma) \left(y(1-\gamma) - k - \gamma\right) \geq 0, \qquad (48)
$$

for any $\gamma \in [\gamma(y+2), \gamma(y+1)]$, and for the third line of (47) is that for given y, $\sum_{k=0}^{y-1} (1-F_T(k)) p(k;y-1,\gamma) \left((y-1)(1-\gamma) - k\right) \geq 0$, for any $\gamma \in [\gamma(y), \gamma(y+1)]$. This last condition is equivalent to the condition, for given y, $\sum_{k=0}^{y} (1-F_T(k)) p(k;y,\gamma) \left(y(1-\gamma) - k\right) \geq 0$, for any $\gamma \in [\gamma(y+1), \gamma(y+2)]$. We will next derive sufficient a condition for (48). Note that

$$
\begin{aligned}
&\sum_{k=0}^{y} (1-F_T(k)) p(k;y,\gamma) \left(y(1-\gamma) - k - \gamma\right) = \sum_{k=0}^{y-2} (1-F_T(k)) p(k;y,\gamma) \left(y(1-\gamma) - k - \gamma\right) \\
&+ (1-F_T(y-1)) \gamma y (1-\gamma)^{y-1} (1-(y+1)\gamma) + (1-F_T(y)) (1-\gamma)^y (-(y+1)\gamma), \qquad (49)
\end{aligned}
$$

so that, for $(y+1)\gamma \leq 2$, $y(1-\gamma) - k - \gamma \geq 0$ for all $k = 0, ..., y-2$, and

$$
\begin{aligned}
&\sum_{k=0}^{y-2} (1-F_T(k)) p(k;y,\gamma) \left(y(1-\gamma) - k - \gamma\right) \geq (1-F_T(y-2)) \sum_{k=0}^{y-2} p(k;y,\gamma) \left(y(1-\gamma) - k - \gamma\right) \\
&= (1-F_T(y-2)) \left(\sum_{k=0}^{y} p(k;y,\gamma) \left(y(1-\gamma) - k - \gamma\right) - \gamma y (1-\gamma)^{y-1} (1-(y+1)\gamma) - (1-\gamma)^y (-(y+1)\gamma)\right)
\end{aligned}
$$

$$= (1-F_T(y-2))\left(-\gamma-\gamma y(1-\gamma)^{y-1}(1-(y+1)\gamma)-(1-\gamma)^y(-(y+1)\gamma)\right). \quad (50)$$

Thus, the expression in (49) is non-negative if $(F_T(y)-F_T(y-2))(1-\gamma)^y(y+1)+(F_T(y-1)-F_T(y-2))\,y(1-\gamma)^{y-1}((y+1)\gamma-1)\geq 1-F_T(y-2)$. The left-hand side of this expression can be re-arranged as

$$\begin{aligned}
&(F_T(y)-F_T(y-2))(1-\gamma)^y(y+1)+(F_T(y-1)-F_T(y-2))\,y(1-\gamma)^{y-1}((y+1)\gamma-1)\\
&=(p_T(y)+p_T(y-1))(1-\gamma)^y(y+1)+p_T(y-1)y(1-\gamma)^{y-1}((y+1)\gamma-1)\\
&=p_T(y)(1-\gamma)^y(y+1)+p_T(y-1)(1-\gamma)^y\left(1+\frac{y^2\gamma}{1-\gamma}\right)\geq(1-\gamma)^y\left(p_T(y)(y+1)+p_T(y-1)\right). \quad (51)
\end{aligned}$$

Thus, $(F_T(y)-F_T(y-2))(1-\gamma)^y(y+1)+(F_T(y-1)-F_T(y-2))\,y(1-\gamma)^{y-1}((y+1)\gamma-1)\geq 1-F_T(y-2)$ as long as $(1-\gamma)^y\left(p_T(y)(y+1)+p_T(y-1)\right)\geq 1-F_T(y-2)$, or $\gamma\leq 1-\left(\frac{1-F_T(y-2)}{(y+1)p_T(y)+p_T(y-1)}\right)^{1/y}$. Combining this expression with $(y+1)\gamma\leq 2$ and $\gamma\in[\gamma(y+2),\gamma(y+1)]$, and noting that for $\gamma=0$ the monotonicity of $\Delta L(y,\gamma(y))$ is assured, we obtain the final sufficient condition

$$\gamma(y+1)\leq\min\left(\frac{2}{y+1},1-\min\left(1,\left(\frac{1-F_T(y-2)}{(y+1)p_T(y)+p_T(y-1)}\right)^{1/y}\right)\right)=\min\left(\frac{2}{y+1},\gamma_T(y+1)\right). \quad (52)$$

Given that $\Delta L(y,\gamma(y))$ is a monotone function of y if $\gamma(y)$ is a non-increasing convex function of y and if (52) is satisfied, we establish the monotonicity of the optimal staffing level y^* with respect to changes in $\frac{w_r}{w_e}$ and $F_T(k)$ following the same arguments used in the proof of Proposition 1. □

Proof of Lemma 1. Note that for $\alpha,\beta\geq 0$, the function $a(x)=x\gamma(x)=\frac{x}{1+e^{\alpha+\beta x}}$ defined on continuous set $x\geq 0$ has a unique global maximum x^* which satisfies the first-order optimality condition $e^{-\alpha}=e^{\beta x^*}(\beta x^*-1)$. Thus, the maximum value for this function can be expressed as $x^*\gamma(x^*)=\frac{1}{\beta}\frac{\beta x^*}{1+e^{\alpha+\beta x^*}}=\frac{\beta x^*-1}{\beta}$. This last expression does not exceed 2 if and only if $\beta x^*\leq 2\beta+1$, which is equivalent to $e^{-\alpha}\leq 2\beta e^{2\beta+1}\Leftrightarrow 2\beta e^{\alpha+2\beta+1}\geq 1$. This condition ensures that the maximum of $y\gamma(y)$ cannot exceed 2 for all integer values of y as well. Further, $\frac{d^2a}{dy^2}=\beta^2\gamma(y)(1-\gamma(y))(\beta(1-2\gamma(y))-2)$, which is always non-positive for $\beta\leq 2$. □

Proof of Proposition 4. Under (24), the optimal staffing level y^* satisfies (23), which can be re-expressed as $y^*=\min\left(y\in\mathcal{N}^+|w_r(1-\gamma(y))+w_n\gamma(y)-(w_r-w_n)y\left(\gamma(y+1)-\gamma(y)\right)+w_e\Delta L(y,\gamma(y))\geq 0\right)$, where the last two terms inside the bracket are

$$\begin{aligned}
&-(w_r-w_n)y\left(\gamma(y+1)-\gamma(y)\right)-w_e(1-\gamma(y+1))\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma(y+1))\\
&+w_e y\int_y^{y+1}ds\frac{d\gamma(s)}{ds}\sum_{k=0}^{y-1}(1-F_T(k))\,p(k;y-1,\gamma(s))=-w_e(1-\gamma(y+1))\sum_{k=0}^{y}(1-F_T(k))\,p(k;y,\gamma(y+1))
\end{aligned}$$

$$+\, y\int_{y}^{y+1} ds\frac{d\gamma(s)}{ds}\left[w_e\left(\sum_{k=0}^{y-1}\left(1-F_T(k)\right)p(k;y-1,\gamma(s))\right)-(w_r-w_n)\right]. \tag{53}$$

Thus, $w_r(1-\gamma(y^*))+w_n\gamma(y^*)-w_e(1-\gamma(y^*+1))\sum_{k=0}^{y^*}\left(1-F_T(k)\right)p(k;y^*,\gamma(y^*+1))+y^*\int_{y^*}^{y^*+1}ds\frac{d\gamma(s)}{ds}\left[w_e\left(\sum_{k=0}^{y^*-1}\left(1-F_T(k)\right)p(k;y^*-1,\gamma(s))\right)-(w_r-w_n)\right]\geq 0$ or

$$\begin{aligned}&\sum_{k=0}^{y^*}\left(1-F_T(k)\right)p(k;y^*,\gamma(y^*+1))\\ &\leq\frac{w_r-(w_r-w_n)\gamma(y^*)+y^*\int_{y^*}^{y^*+1}ds\frac{d\gamma(s)}{ds}\left[w_e\left(\sum_{k=0}^{y^*-1}\left(1-F_T(k)\right)p(k;y^*-1,\gamma(s))\right)-(w_r-w_n)\right]}{w_e(1-\gamma(y^*+1))},\end{aligned} \tag{54}$$

Now, consider an element of E, y^{ML}, which satisfies $y^{\mathrm{ML}}=\min\left(y\in\mathcal{N}^+|\sum_{k=0}^{y}\left(1-F_T(k)\right)p(k;y,\gamma(y^{\mathrm{ML}}))\leq\frac{w_r-(w_r-w_n)\gamma(y^{\mathrm{ML}})}{w_e\left(1-\gamma(y^{\mathrm{ML}})\right)}\right)$, so that $\sum_{k=0}^{y^{\mathrm{ML}}}\left(1-F_T(k)\right)p(k;y^{\mathrm{ML}},\gamma(y^{\mathrm{ML}}))\leq\frac{w_r-(w_r-w_n)\gamma(y^{\mathrm{ML}})}{w_e\left(1-\gamma(y^{\mathrm{ML}})\right)}$. Below we will show by contradiction that, if $a(y^*+1)<a(y^*)$, then $y^{\mathrm{ML}}\leq y^*$. Suppose that $y^{\mathrm{ML}}>y^*\Leftrightarrow y^{\mathrm{ML}}\geq y^*+1$. This, in turn, implies that $\gamma(y^{\mathrm{ML}})\leq\gamma(y^*+1)$, and that

$$\sum_{k=0}^{y^*}\left(1-F_T(k)\right)p(k;y^*,\gamma(y^{\mathrm{ML}}))>\frac{w_r-(w_r-w_n)\gamma(y^{\mathrm{ML}})}{w_e\left(1-\gamma(y^{\mathrm{ML}})\right)}. \tag{55}$$

Now, since for $\forall s\in[y^*,y^*+1]$, it follows that $s\leq y^{\mathrm{ML}}$ and $\gamma(y^{\mathrm{ML}})\leq\gamma(s)$, we can use (36), (37) and (55) to get

$$\begin{aligned}&\sum_{k=0}^{y^*-1}\left(1-F_T(k)\right)p(k;y^*-1,\gamma(s))\geq\sum_{k=0}^{y^*}\left(1-F_T(k)\right)p(k;y^*,\gamma(s))\\ &\geq\sum_{k=0}^{y^*}\left(1-F_T(k)\right)p(k;y^*,\gamma(y^{\mathrm{ML}}))>\frac{w_r-(w_r-w_n)\gamma(y^{\mathrm{ML}})}{w_e\left(1-\gamma(y^{\mathrm{ML}})\right)}.\end{aligned} \tag{56}$$

Further, since $\frac{d\gamma(y)}{dy}\leq 0$, (54) implies that

$$\begin{aligned}&\frac{w_r-(w_r-w_n)\gamma(y^{\mathrm{ML}})}{w_e\left(1-\gamma(y^{\mathrm{ML}})\right)}<\sum_{k=0}^{y^*}\left(1-F_T(k)\right)p(k;y^*,\gamma(y^*+1))\\ &\leq\frac{w_r-(w_r-w_n)\gamma(y^*)+y^*\int_{y^*}^{y^*+1}ds\frac{d\gamma(s)}{ds}\left[w_e\left(\sum_{k=0}^{y^*-1}\left(1-F_T(k)\right)p(k;y^*-1,\gamma(s))\right)-(w_r-w_n)\right]}{w_e(1-\gamma(y^*+1))}\\ &\leq\frac{w_r-(w_r-w_n)\gamma(y^*)+y^*\left(\gamma(y^*+1)-\gamma(y^*)\right)\left(\frac{w_n}{1-\gamma(y^{\mathrm{ML}})}\right)}{w_e\left(1-\gamma(y^*+1)\right)}\\ &\leq\frac{w_r-(w_r-w_n)\gamma(y^*+1)+y^*\left(\gamma(y^*+1)-\gamma(y^*)\right)\left(\frac{w_n}{1-\gamma(y^{\mathrm{ML}})}\right)}{w_e\left(1-\gamma(y^*+1)\right)}.\end{aligned} \tag{57}$$

(57) is equivalent to $\frac{\gamma(y^{\mathrm{ML}})}{1-\gamma(y^{\mathrm{ML}})}\leq\frac{\gamma(y^*+1)}{1-\gamma(y^*+1)}+\frac{y^*\left(\gamma(y^*+1)-\gamma(y^*)\right)\left(\frac{1}{1-\gamma(y^{\mathrm{ML}})}\right)}{(1-\gamma(y^*+1))}$, and $\gamma(y^{\mathrm{ML}})\left(1-\gamma(y^*+1)\right)\leq\gamma(y^*+1)\left(1-\gamma(y^{\mathrm{ML}})\right)+y^*\left(\gamma(y^*+1)-\gamma(y^*)\right)$, so that $\gamma(y^{\mathrm{ML}})\leq\gamma(y^*+1)+y^*\left(\gamma(y^*+1)-\gamma(y^*)\right)$. Note that this contradicts $\gamma(y^{\mathrm{ML}})\geq 0$. □

References

Aiken, L. H., S.P. Clarke, D.M. Sloane, J. Sochalski, J.H. Silber. 2002. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of American Medical Association* **288** 1987–1993.

Allenby, G. M., P. E. Rossi. 1991. There is no aggregation bias: Why macro logit models work. *Journal of Business & Economic Statistics* **9**(1) pp. 1–14.

Bassamboo, A., R. S. Randhawa, A. Zeevi. 2010. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* **56**(10) 1668–1686.

Bollapragada, S., T. E. Morton. 1999. Myopic heuristics for the random yield problem. *Operations Research* **47**(5) 713–722.

Bryant, C., G. Fairbrother, P. Fenton. 2000. Professional issues: The relative influence of personal and workplace descriptors on stress. *British Journal of Nursing* **9** 876–880.

Cachon, G., A.G. Kök. 2007. Implementation of the newsvendor model with clearance pricing: how to (and how not to) estimate a salvage value. *Manufacturing & Service Operations Management* **9**(3) 276–290.

California Department of Health. 2004. Nurse-to-patient staffing ratio regulations. http://www.cdph.ca.gov/services/DPOPP/regs/Documents/R-37-01_Regulation_Text.pdf, last checked on January 23, 2011.

Cho, S.H., S. Ketefian, V.H. Barkauskas. 2003. The effects of nurse staffing on adverse outcomes, morbidity, mortality, and medical costs. *Nursing Research* **52**(2) 71–79.

Ciarallo, F.W., R. Akella, T. Morton. 1994. A periodic review, production planning model with uncertain capacity and uncertain demand-optimality of extended myopic policies. *Management Science* **40** 320–332.

Cooper, W.L., T. Homem de Mello, A.J. Kleywegt. 2004. Models of the spiral down effect in revenue management. *Operations Research* **54**(5) 968–987.

Darr, W., G. Johns. 2008. Work strain, health, and absenteeism: A meta-analysis. *Journal of Occupational Health Psychology* **13**(4) 293–318.

Dwyer, D. J., D. C. Ganster. 1991. The effects of job demands and control on employee attendance and satisfaction. *Journal of Organizational Behavior* **12**(7) pp. 595–608.

Easton, F.F., J.C. Goodale. 2005. Schedule recovery: Unplanned absences in service operations. *Decision Sciences* **36**(3) 459–488.

Erickson, R.J., L. Nichols, C. Ritter. 2000. Family influences on absenteeism: Testing an expanded process model. *Journal of Vocational Behavior* **57** 246–272.

Fadiloglu, M., E. Berk, M. C. Gurbuzc. 2008. Supplier diversification under binomial yield. *Operations Research Letters* **36** 539–542.

Fry, M.J., M.J. Magazine, U.S. Rao. 2006. Firefighter staffing including temporary absences and wastage. *Operations research* **54**(2) 353.

Gerchak, Y., M. I. Henig. 1994. A flexible conceptualization of random yield and its implications for source selection. *Proceedings of the First Conference of ORSA Technical Section on Manufacturing Management, Carnegie Mellon University* 133–139.

Greene, W.H. 2005. *Econometric analysis*. Pearson Education.

Grosfeld-Nir, A., Y. Gerchak. 2004. Multiple lotsizing in production to order with random yields: Review of recent advances. *Annals of Operations Research* **126** 43–69.

Gupta, D., W. L. Cooper. 2005. Stochastic comparisons in production yield management. *Operations Research* **53**(2) 377–384.

Harrison, J. M., A. Zeevi. 2005. A Method for Staffing Large Call Centers Based on Stochastic Fluid Models. *Manufacturing service and Operations Management* **7**(1) 20–36.

Hill, J.M., E.L. Trist. 1955. Changes in accidents and other absences with length of service: A further study of their incidence and relation to each other in an iron and steel works. *Human Relations* **8**(2) 121.

Hobfoll, S.E. 1989. Conservation of resources. *The American Psychologist* **44**(3) 513.

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

Kristensen, T. S. 1991. Sickness absence and work strain among Danish slaughterhouse workers: An analysis of absence from work regarded as coping behaviour. *Social Science & Medicine* **32** 15–27.

McVicar, A. 2003. Workplace stress in nursing: a literature review. *Journal of Advanced Nursing* **44** 633–642.

Needleman, J., P. Buerhaus, S. Mattke. 2002. Nurse-staffing levels and patient outcomes in hospitals. *New England Journal of Medicine* **346**(22) 1715–1722.

Nevidjon, B., J. I. Erickson. 2001. The nursing shortage: Solutions for the short and long term. *Online Journal of Issues in Nursing* .

Parkes, K. R. 1982. Occupational stress among student nurses: A natural experiment. *Journal of Applied Psychology* **67**(6) 784–796.

Pierskalla, W.P., H. Miller, G. Rath. 1976. Nurse scheduling using mathematical programming. *Operations Research* **24** 857–870.

Powell, S. G., K.L. Schultz. 2004. Throughput in serial lines with state-dependent behavior. *Management Science* **50** 1095–1105.

Rauhala, A., M. Kivimäki, L. Fagerström, M. Elovainio, M. Virtanen, J. Vahtera, A.-K. Rainio, K. Ojaniemi, J. Kinninen. 2007. What degree of work overload is likely to cause increased sickness absenteeism among nurses? Evidence from the RAFAELA patient classification system. *Journal of Advanced Nursing* **57**(3) 286–295.

Schultz, K.L., D.C. Juran, J.W. Boudreau. 1999. The effects of low inventory on the development of productivity norms. *Management Science* **45** 1664–1678.

Schultz, K.L., D.C. Juran, J.W. Boudreau, L.J. Thomas J.O. McClain. 1998. Modeling and worker motivation in JIT production systems. *Management Science* **44** 1595–1607.

Seago, J.A. 2001. Chapter 39 in "Making health care safer: A critical analysis of patient safety practices". *Evidence Report/Technology Assessment Number 43, AHRQ Publication No. 01-E058, Agency for Healthcare Research and Quality* .

Shamian, J., L. O'Brien-Pallas, D. Thomson, C. Alksnis, M.S. Kerr. 2003. Nurse absenteeism, stress and workplace injury: What are the contributing factors and what can/should be done about it? *The International Journal of Sociology and Social Policy* **23**(8/9) 81–103.

Smulders, P.G.W., F.J.N. Nijhuis. 1999. The job demands-job control model and absence behaviour: results of a 3-year longitudinal study. *Work & Stress* **13**(2) 115–131.

Statistics Canada. 2008. Work absence rates. http://www.statcan.gc.ca/pub/71-211-x/71-211-x2009000-eng.pdf, last checked on January 23, 2011.

Steers, R.M., S.R. Rhodes. 1978. Major influences on employee attendance: a process model. *Journal of Applied Psychology* **63**.

The Victorian Department of Health. 2007. http://www.health. vic.gov.au/__data/assets/pdf_file/0008/356696/nurses-public-sector-eba-2004-2007-_ag840794-2.pdf, last checked on january 23, 2011.

Tummers, G.E.R., P.P.M. Janssen, A. Landeweerd, I. Houkes. 2001. A comparative study of work characteristics and reactions between general and mental health nurses: a multi-sample analysis. *Journal of Advanced Nursing* **36**(1) 151–162.

Unruh, L., L. Joseph, M. Strickland. 2007. Nurse absenteeism and workload: Negative effect on restraint use, incident reports and mortality. *Journal of Advanced Nursing* **60**(6) 673–681.

US Bureau of Labor Statistics. 2008. Industry injury and illness data. http://www.bls.gov/iif/oshwc/osh/os/osnr0032.pdf, last checked on January 23, 2011.

Whitt, W. 2006. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* **15**(1) 88–102.

Yang, S., J. Yang, L. Abdel-Malek. 2007. Sourcing with random yields and stochastic demand: A newsvendor approach. *Computers and Operations Research* **34** 3682–3690.

Yankovic, N., L.V. Green. 2011. Identifying good nursing levels: A queueing approach. *Operations Research* **59**(4) 942–955.

Yano, C.A., H.A. Lee. 1995. Lot sizing with random yields: a review. *Operations Research* **43**(2) 311–334.