esmt european school of management and technology

July 3, 2008

ESMT
WORKING PAPER

ESMT No. 08-005

Nurse-To-Patient Ratios in Hospital Staffing: A Queueing Perspective

Francis de Véricourt, ESMT
Otis B. Jennings, Duke University

ISSN 1866-3494

Abstract

Nurse-To-Patient Ratios in Hospital Staffing: A Queueing Perspective[1]

Author(s): Francis de Véricourt, ESMT
Otis B. Jennings, Duke University

The immediate motivation of this paper is California Bill AB 394, legislation which mandates fixed nurse-to-patient staffing ratios as a means to address the current crisis in the quality of health care delivery. Modeling medical units as closed queueing systems, we seek to determine whether or not ratio policies are effective at managing nurse workload. Our many-server asymptotic results suggest that ratio policies cannot provide consistently high service quality across medical units of different sizes. As a remedy, we recommend policies that deviate from the restrictive linear nature of ratio policies, employing the "square root rule" commonly used to staff large service systems. Under some quality of care assumptions, our policies exhibit a type of "super" pooling effect, in which, for large systems, the requisite workforce is significantly smaller than the nominal patient load.

Keywords: queueing system, health care, public policy, nursing, staffing, many-server limit theorems

[1] The authors are grateful to Suri Gurumurthi for his work on the numerical examples. We also thank Galit Yom-Tov for assisting with the current presentation of the paper.

Contact: Francis de Véricourt; ESMT, Schlossplatz 1, 10178 Berlin, Germany; Tel: +49 (0)30 212 31-1291, Email: devericourt@esmt.org.

Submitted to
manuscript v.2

Nurse-To-Patient Ratios in Hospital Staffing: a Queueing Perspective

Francis de Véricourt
European School of Management and Technology, Berlin, Germany, fdv1@duke.edu

Otis B. Jennings
Fuqua School of Business, Duke University, Durham, USA, otisj@duke.edu

The immediate motivation of this paper is California Bill AB 394, legislation which mandates fixed nurse-to-patient staffing ratios as a means to address the current crisis in the quality of health care delivery. Modeling medical units as closed queueing systems, we seek to determine whether or not ratio policies are effective at managing nurse workload. Our many-server asymptotic results suggest that ratio policies cannot provide consistently high service quality across medical units of different sizes. As a remedy, we recommend policies that deviate from the restrictive linear nature of ratio policies, employing the "square root rule" commonly used to staff large service systems. Under some quality of care assumptions, our policies exhibit a type of "super" pooling effect, in which, for large systems, the requisite workforce is significantly smaller than the nominal patient load.

Key words: Queueing System, Health Care, Public Policy, Nursing, Staffing, Many-Server Limit Theorems

1. Introduction

In 1999, California introduced the nation's first law mandating nurse-to-patient ratios in hospitals, Bill AB 394. The legislation, implemented in 2004, specifies the minimum number of nurses that should be staffed for each hospital unit, given the current number of patients therein. The minimum number of nurses is a fixed, unit-specific fraction of the number of patients. For instance, according to AB 394, at least one nurse for every four patients should now be present at any time in any pediatrics unit within California. This legislation has inspired other states to consider establishing similar requirements. Further, two House bills in a recent Congressional session proposed regulating nurse staffing among Medicare-participating hospitals (see for instance Spetz 2005). The rationale for implementing these ratios stems from the association between nurse workload and clinical outcomes (see Aiken et al. 2002, Needleman and Buerhaus 2003). The purpose of the mandated nurse-to-patient ratios is to provide a consistently high level of patient safety throughout the state by providing adequate nurse staffing levels.

To explore the effectiveness of ratio policies and alternatives thereto, we explicitly model the dynamics within a single medical unit as a closed, multi-server queueing system. The performance of the system is based on the *probability of excessive delay*, the relative frequency with which the delay between the onset of patient neediness and the provision of care from a nurse exceeds a given time threshold. We derive many-server asymptotic results and identify novel staffing regimes, each of which allow staffing according to a pre-specified probability of excessive delay. For specific time thresholds, our results demonstrate that policies based on fixed nurse-to-patient ratios can provide consistent quality of care across medical units of different sizes, a desirable property from a decision maker point of view. However, the resulting probabilities of excessive delay always exceed 50% in these cases, ostensibly an undesirable frequency. To remedy this problem, effective staffing policies should deviate from threshold-specific nurse-to-patient ratios by factors which take into account the total number of patients present in the unit. These specific ratios are, in many cases, smaller than the ones mandated by AB 394. Moreover, these additional factors, variants of the square root

rule, take into account congestion due to variability in patient needs, the absence of which, in our opinion, is the primary shortcoming of ratio policies.

So far, the existing body of clinical research has not succeeded in formulating evidence-based guidelines for setting staffing levels. Essentially, current research data are actually not rich enough (Clark 2005). One of the main issues, as noted by Needleman and Buerhaus (2003), concerns the lack of documentation of "failures" in the nursing process and the impact of such failures on patient outcomes. So California policy is based primarily on the recommendations of the different stakeholders – such as the Californian Department of Health Services, the nurse unions and hospital administrations – whose suggestions differ significantly. As a result, setting ratio levels has been based more on negotiation than science. Not surprisingly then, the mandated ratios currently in place have sparked active debate throughout both academic and public sectors, yet no viable methodology is in place to resolve the disputes. In light of these difficulties, we believe that the normative methodology advanced in this paper constitutes a compelling approach for deriving structural results and suggesting future empirical research for the nurse staffing problem.

To our knowledge no empirical studies have explored the impact of the frequency of excessive delays on patient care and safety. Nevertheless, we argue that excessive delays are akin to possible adverse events. Indeed, the link between workload and quality of care is often thought of in terms of adverse events. For instance, in evaluating how hospitals improve quality of care, Tucker and Edmondson (2003) identify failures that occur in care delivery processes, such as tasks that are unnecessarily or incompletely performed. Delays also give rise to unfinished tasks, either because nurses fail to remember them later or because they abandon them in order to take care of more urgent procedures. Unfinished care has been identified as a strong factor impacting the quality of nursing care (Sochalski 2004). Further, it is known that delaying certain medical procedures can endanger patient health. For instance, the medical guidelines for certain myocardial infarctions recommend the immediate administration of aspirin (ACC/AHA 2002). More generally, even though many existing empirical studies that analyze the relationship between delay and quality of care do not concern nursing activities, this link is generally framed in terms of time constraints in the medical field. An example is the angioplasty procedure, which significantly cuts a patient's risk of dying when performed within 90 minutes of a heart attack (see ACC/AHA 2001, Sternberg 2006). [1] Hence, in the absence of a uniformly agreed upon performance metric with clear links to nurse staffing, we believe that time thresholds and associated frequencies of excessive delay are the most natural and relevant measures in this setting.

Time thresholds might typically be small in certain units (as in ICU) or long in others (geriatric services). However, it is worth noting that we do not claim to know what the acceptable time threshold and noncompliance frequency should be. Neither do we determine the precise values of the parameters describing our model (average nurse service times, frequency of patient needs, etc.). Avoiding these simplifications enables us to argue in the most general terms that, *whatever constitutes an excessive delay and tolerance thereof, mandating nurse-to-patient ratios cannot ensure uniform quality of care across all hospitals.*

Our queueing model seeks to represent the workload experienced by nurses over time in a medical unit. Traditionally, medical units have been modelled as multi-server, open-loop, $M/M/s$ queues, where the arrival process represents the stream of incoming patients and the service time captures the average length of stay. Studies of these models provide, for instance, the number of required beds to achieve availability targets in a given unit (see Green 2004, for overviews of queueing models used in capacity planning and management of hospitals). On the other hand, the nursing activities taking place in the unit are determined by the patients who require different services and

[1] Even when no empirical evidences exist, the medical community seems to intuitively make use of the notion of time threshold as exemplified by the concept of "golden hour" which commonly characterizes the urgent need for the care of trauma patients (see Lerner and Moscati 2001).

the nurses who deliver these services. Hence, during a time period when the number of patients is relatively constant, the dynamics of the system are better described by a closed queueing system – i.e. one with a finite population of patients, an $M/M/s//n$ queue – where each patient alternates between requiring assistance and not.

The staffing ratios mandated by AB 394 are minimum in that they should be adjusted upward locally at individual hospitals to account for case mix (i.e., the collection of patients and their respective acuity levels) as well as nurses with varied skill levels. In our model we assume that nurses and patients are homogeneous. Policymakers may want to assume the lowest possible acuity level for all patients in the unit and that the nursing staff is equally qualified. Then likewise, as with the legislated ratios, our model suggests minimums staffing levels that should be adjusted upwards by hospitals to account for locally realized discrepancies in acuity and nurse skill level.

As the number of patients in the unit varies over time, we assume that the number of nurses is quickly adjusted. Bill AB 394 (see also CDHS 2005) makes a similar assumption when it explicitly specifies that "... the ratios must be maintained 'at all times' ". Approaches used by hospitals to accommodate this requirement include cross-training nurses who float among units, temporarily employing nurses from external agencies etc. Although such strategies raise other issues related to costs and safety (see for instance Alonso-Echanove et al. 2003), we follow AB 394 and assume that providing the desired staffing level at any time is feasible. Moreover, we assume that between significant changes in the number of patients and nurses, the system reaches steady-state relatively quickly. Similar assumptions are common in hospital capacity planning (Green 2004) or call center staffing (Gans et al. 2003).

We study the probability of excessive delay in a medical unit by letting the numbers of patients and nurses approach infinity. To our knowledge, this constitutes the first many-server asymptotic analysis of health care related issues. Such an approach seems nevertheless well suited for policy making as it allows describing the structure of efficient staffing policies in a simple way. Pioneering many-server asymptotic results for open queueing systems are due to Halfin and Whitt (1981), who identified the so-called *quality- and efficiency-driven (QED)* staffing regime, a notion we extend to closed queueing systems and generalize. For recent results on staffing many-server queues see Mandelbaum and Zeltyn (2006) and references within.

In all previous papers, staffing rules in the QED regime can only achieve a pre-specified probability of excessive delay for which the delay threshold is asymptotically null. If avoiding any wait is essential in intensive care units, positive delays appear more reasonable in other units, such as oncology for instance.

In this paper, we identify new staffing regimes that use nursing resources efficiently while providing good quality of care, where good quality is defined as addressing patient needs within a time threshold T. Because it is indexed by the service quality parameter T, we refer to these as the family of QED(T) regimes (Mandelbaum and Zeltyn 2006, have independently proposed equivalent regimes for open queues with abandonement). de Véricourt and Jennings (2006a) consider the special case where $T = 0$. Baron and Milner (2006), a contemporaneous paper, has independently identified similar staffing regimes for open queueing systems with customer abandonment.

We present the basic model in Section 2. Our many-server analysis and the formulation of the optimal nurse staffing problem are presented in Section 3. Section 4 contains our main insights on Bill AB 394. We show through simulations that our main results are robust to some of our key assumptions in Section 5. We conclude by discussing our results and suggesting future analytical and empirical research.

2. The Queueing Model

We model a medical unit where s nurses serve n patients as an $M/M/s//n$ closed queueing system. Patients exist in one of two states: *stable* and *needy*. Stable patients become needy after an exponentially distributed *activation time* with mean $1/\lambda$. Needy patients are served by the nurses on a

FCFS basis and the duration of service is exponentially distributed with mean $1/\mu$. Once treated, a patient becomes stable again, until she needs other care procedures. We assume that nurses work as a team so that any one of them can serve any patient. In some medical units however, a nurse is assigned to a specific set of patients. Our model remains relevant in such settings, provided that nurses whose patients temporarily require less care will help their busier colleagues, as is often the case in practice.

The exponential distribution assumption is common in hospital capacity planning (Green 2004) or policy making (Liu and Wein 2005). The main reason for this assumption is mathematical tractability. The wide range of possible patient needs and nursing tasks also suggests a high degree of variability in the health care process, which is consistent with the high coefficient of variation of the exponential distribution. Nonetheless, we show in Section 5 that our main results do not change with non-exponential service times.

The quantity $r \equiv \lambda/(\lambda+\mu)$ represents the long run fraction of time a patient would spend in the needy state if her needs were always immediately addressed by a nurse when they arose. Naturally, the quantity rn is the nominal patient load (also referred to as the total offered workload). The principal metric of concern is *the probability of excessive delay*, the likelihood that a needy patient's waiting period before getting access to a nurse is longer than the time threshold $T \geq 0$ that delineates between acceptable and excessive delays. For the special case threshold for which any delay is excessive ($T=0$), this performance metric is simply referred to as the *probability of delay*. Letting $\rho \equiv \lambda/\mu = r/\bar{r}$ with $\bar{r} \equiv 1-r$, the steady state probability distribution of N, the number of needy patients in the system, is given by (see for instance Kleinrock 1975):

$$\pi_k = \begin{cases} \pi_0 \binom{n}{k} \rho^k & \text{for } k = 0, \cdots, s; \\ \pi_0 \binom{n}{k} \frac{k!}{s!} s^{s-k} \rho^k & \text{for } k = s+1, \cdots, n, \end{cases} \tag{1}$$

where π_0 is a normalizing constant. A patient who becomes needy when there are already $k \geq s$ other needy patients will experience an in-queue random waiting time that follows an Erlang distribution with $(k-s+1)$ stages, each with rate $s\mu$. The probability that this Erlang-distributed random variable is greater than T is $e^{-s\mu T}\sum_{j=0}^{k-s}(s\mu T)^j/(j!)$. Let W denote the steady state, in-queue waiting time for a hypothetical newly needy patient.

The probability of excessive delay for a system with n patients and s nurses is denoted $p_n(s,T)$, or simply $p_n(T)$ when no confusion is possible, and is a point along the function $p_n(s,\cdot)$. Letting $\lambda_k \equiv \lambda(n-k)$ denote the activation rate when k out of n patients are needy (and suppressing the parameter n except when necessary), we obtain

$$p_n(s,T) = \sum_{k=s}^{n} \frac{\lambda_k \pi_k}{\sum_{i=0}^{n} \lambda_i \pi_i} P(W > T | N = k) = e^{-s\mu T} \sum_{k=s}^{n} \frac{(n-k)\pi_k}{\sum_{i=0}^{n}(n-i)\pi_i} \sum_{j=0}^{k-s} \frac{(s\mu T)^j}{j!}. \tag{2}$$

3. Probability of Excessive Delay

In this section, we investigate families of staffing rules, where the staffing level is a function of the number of patients ($n \mapsto s_n$). In particular, we derive a staffing rule that takes the form of a ratio policy adjusted by a square root term and provides a consistent service level across units of different size.

The following proposition, our main result, provides necessary and sufficient conditions under which the probabilities of excessive delay, associated with a sequence of staffing levels, indexed by the number of patients has a non-degenerate limit ϵ, as the number of patients goes to infinity. This result is used later to derive an accurate approximation for the optimal number of nurses in the staffing problem.

PROPOSITION 1. *For any given $T>0$, the probability of excessive delay $p_n(s_n,T)$ has a non-degenerate limit $\epsilon\in(0,1)$ if and only if*

$$\left(\frac{s_n}{n}-\hat{r}_T\right)\sqrt{n}\to\beta,\quad as\quad n\to\infty, \tag{3}$$

where

$$\hat{r}_T=\frac{r}{1+r\mu T}, \tag{4}$$

for some $\beta\in(-\infty,\infty)$, with

$$\epsilon=\Phi\left(-\frac{\beta}{\hat{r}_T}\sqrt{\frac{1+r\mu T}{\bar{r}+r\mu T}}\right). \tag{5}$$

Proof. The probability of excessive delay can be written as

$$p_n(s_n,T)=\left(\sum_{k=0}^{n}\pi_k(n-k)\right)^{-1}\left(\sum_{k=s_n}^{n}\pi_k(n-k)e^{-s_n\mu T}\sum_{j=0}^{k-s_n}\frac{(s_n\mu T)^j}{j!}\right)=\frac{D_n}{B_n}\left(1+\frac{A_n}{B_n}\right)^{-1},$$

where

$$A_n\equiv\sum_{k=0}^{s_n-1}\binom{n}{k}(n-k)\rho^k,\qquad B_n\equiv\sum_{k=s_n}^{n-1}\frac{n!}{(n-k-1)!s_n!}s_n^{s_n}\left(\frac{\rho}{s_n}\right)^k$$

and

$$D_n\equiv\sum_{k=s_n}^{n-1}\frac{n!}{(n-k-1)!}\frac{s_n^{s_n}}{s_n!}\left(\frac{\rho}{s_n}\right)^k e^{-s_n\mu T}\sum_{j=0}^{k-s_n}\frac{(s_n\mu T)^j}{j!}.$$

The quantity A_n can be expressed as

$$A_n=n(1+\rho)^{n-1}\sum_{k=0}^{s_n-1}\binom{n-1}{k}r^k\bar{r}^{n-1-k}=n(1+\rho)^{n-1}P(X_n\le s_n-1),$$

where for each n, X_n is a binomial random variable with parameters $n-1$ and r. Similarly, one can express B_n as

$$\begin{aligned}B_n&=\frac{n!}{s_n!}s_n^{s_n}\left(\frac{\rho}{s_n}\right)^{n-1}e^{s_n/\rho}\sum_{k=0}^{n-s_n-1}\left(\frac{s_n}{\rho}\right)^k\frac{1}{k!}e^{-s_n/\rho}\\&=\frac{n!}{s_n!}s_n^{s_n}\left(\frac{\rho}{s_n}\right)^{n-1}e^{s_n/\rho}\,P(Y_n\le n-s_n-1),\end{aligned}$$

where Y_n follows a Poisson distribution with parameter s_n/ρ, and D_n as

$$\begin{aligned}D_n&=\frac{n!}{s_n!}s_n^{s_n}\left(\frac{\rho}{s_n}\right)^{n-1}e^{s_n/\rho}\left(\sum_{k=0}^{n-s_n-1}\left(\sum_{j=0}^{n-s_n-k-1}\left(\frac{s_n}{\rho}\right)^k\frac{1}{k!}\frac{(s_n\mu T)^j}{j!}e^{-(s_n/\rho+s_n\mu T)}\right)\right)\\&=\frac{n!}{s_n!}s_n^{s_n}\left(\frac{\rho}{s_n}\right)^{n-1}e^{s_n/\rho}\,P(Z_n^1+Z_n^2\le n-s_n-1),\end{aligned}$$

where Z_n^1 and Z_n^2 are independent Poisson random variables with parameters s_n/ρ and $s_n\mu T$, respectively. It follows that

$$D_n=\frac{n!}{s_n!}s_n^{s_n}\left(\frac{\rho}{s_n}\right)^{n-1}e^{s_n/\rho}\,P(Z_n\le n-s_n-1),$$

where $Z_n \equiv Z_n^1 + Z_n^2$ is a Poisson random variable with parameter $s_n/\rho + s_n\mu T$.

For each sequence of the random variables $\{X_n\}$, $\{Y_n\}$ and $\{Z_n\}$, we generate a central limit theorem. We center and rescale to obtain

$$P(X_n \le s_n - 1) = P\left(\frac{X_n - (n-1)r}{\sqrt{(n-1)r\bar{r}}} \le \frac{s_n - 1 - (n-1)r}{\sqrt{(n-1)r\bar{r}}}\right), \tag{6}$$

$$P(Y_n \le n - s_n - 1) = P\left(\frac{Y_n - s_n/\rho}{\sqrt{s_n/\rho}} \le \frac{n - 1 - s_n/r}{\sqrt{s_n/\rho}}\right) \tag{7}$$

and

$$P(Z_n \le n - s_n - 1) = P\left(\frac{Z_n - s_n(\frac{1}{\rho} + \mu T)}{\sqrt{s_n(\frac{1}{\rho} + \mu T)}} \le \frac{n - 1 - s_n(\frac{1}{r} + \mu T)}{\sqrt{s_n(\frac{1}{\rho} + \mu T)}}\right). \tag{8}$$

The sequences $(X_n - (n-1)r)/\sqrt{(n-1)r\bar{r}}$, $(Y_n - s_n/\rho)/\sqrt{s_n/\rho}$ and $(Z_n - s_n(\frac{1}{\rho} + \mu T))/\sqrt{s_n(\frac{1}{\rho} + \mu T)}$ each converge in distribution to standard normal random variables (i.e., with mean zero and variance one). From (3), the staffing level as a function of n is $s_n = rn/(1 + r\mu T) + o(n)$. It follows that $[s_n - 1 - (n-1)r]/\sqrt{(n-1)r\bar{r}} \to -\infty$ as $n \to \infty$. By (6) one can conclude that

$$\lim_{n\to\infty} P(X_n \le s_n - 1) = 0. \tag{9}$$

Likewise, $(n - 1 - s_n/r)/\sqrt{s_n/\rho} \to \infty$ as $n \to \infty$, so that by (7),

$$\lim_{n\to\infty} P(Y_n \le n - s_n - 1) = 1. \tag{10}$$

Finally, by the conditions provided for s_n, we have $s_n/n \to \hat{r}_T$ and $(1 - s_n/(n\hat{r}_T))\sqrt{n} \to -\beta/\hat{r}_T$. It follows that, with $\rho = r/\bar{r}$

$$\begin{aligned}\frac{n - 1 - s_n(\frac{1}{r} + \mu T)}{\sqrt{s_n(\frac{1}{\rho} + \mu T)}} &= \frac{-1/\sqrt{n} + (1 - s_n/(n\hat{r}_T))\sqrt{n}}{\sqrt{s_n/n(\frac{1}{\rho} + \mu T)}} \\ &\to -\frac{\beta}{r}\sqrt{\frac{(1 + r\mu T)^3}{\bar{r} + r\mu T}}\end{aligned}$$

as $n \to \infty$, and by (8) we have

$$\lim_{n\to\infty} P(Z_n \le n - s_n - 1) = \Phi\left(-\frac{\beta}{r}\sqrt{\frac{(1 + r\mu T)^3}{\bar{r} + r\mu T}}\right). \tag{11}$$

The ratio A_n/B_n can be written as

$$\frac{A_n}{B_n} = C_n \frac{P(X \le s_n - 1)}{P(Y \le n - s_n - 1)},$$

where

$$C_n = \frac{s_n!}{(n-1)!}\left(\frac{s_n}{\rho}\right)^{n-1}\frac{(1+\rho)^{n-1}}{s_n^{s_n}}e^{-s_n/\rho}.$$

Following the analysis in the proof of Proposition 1 of de Véricourt and Jennings (2006a) along with (3), we have $C_n \to \bar{r}\left(\frac{1}{r} + \mu T\right)e^{-\beta^2/2r^2}$ as $n \to \infty$. The relevant feature is that the limit of C_n

is finite. It follows then by (9) and (10) that $A_n/B_n \to 0$. Finally, notice that $D_n/B_n = P(Z_n \leq n - s_n - 1)/P(Y_n \leq n - s_n - 1)$ so that by (10) and (11),

$$\lim_{n\to\infty} p_n(s_n, T) = \Phi\left(-\frac{\beta}{r}\sqrt{\frac{(1+r\mu T)^3}{\bar{r}+r\mu T}}\right).$$

We show the "only if" part of the equivalence by contradiction. Let $f(\epsilon) \equiv -\hat{r}_T\,\Phi^{-1}(\epsilon)\sqrt{\frac{\bar{r}+r\mu T}{1+r\mu T}}$. Suppose that $p_n(s_n, T)$ has a limit $\epsilon \in (0,1)$ and that $\beta \neq f(\epsilon)$ is a (possibly infinite) limit point of $\{(s_n/n - \hat{r}_T)\sqrt{n}\}$. Assume for now that $\beta > f(\epsilon)$. Construct a sequence $\{s'_n\}$ such that $s'_n \leq s_n$ and $(s'_n/n - \hat{r}_T)\sqrt{n} \to \beta' = \min((\beta + f(\epsilon))/2, f(\epsilon) + 1)$, as $n \to \infty$. Notice that $f(\epsilon) < \beta' < \infty$, which implies $\alpha > f^{-1}(\beta') = \Phi\left(-\frac{\beta'}{\hat{r}_T}\sqrt{\frac{1+r\mu T}{\bar{r}+r\mu T}}\right)$. Since $s'_n \leq s_n$, $p_n(s'_n, T)) \geq p_n(s_n, T)$. However, taking the limit of both sides yields $f(\beta') \geq \alpha$, a contradiction. A similar argument shows that $\beta < f^{-1}(\alpha)$ is also impossible. Hence, the convergence $a_n \to \alpha \in (0,1)$ implies $\{(s_n/n - r)\sqrt{n}\}$ has a unique finite limit as well. □

In our framework, the nurse staffing problem consists in finding the minimum staffing levels s_n that guarantee a bound ϵ on the probability of excessive delay $p_n(s_n, T)$ for any n. Given T and ϵ, a staffing rule $n \mapsto s_n$ is said to be *asymptotically optimal* if $lim_{n\to\infty} p_n(s_n, T) = \epsilon$. For positive T, Proposition 1 suggests the following policy to solve this minimization problem:

$$s_n^* = \lceil \hat{r}_T n + \beta_T \sqrt{n} \rceil \tag{12}$$

with

$$\beta_T = -\hat{r}_T\,\Phi^{-1}(\epsilon)\sqrt{\frac{\bar{r}+r\mu T}{1+r\mu T}},$$

where $\lceil x \rceil$ is equal to the smallest integer larger or equal to x.

Asymptotically as the number of patients goes to infinity, for staffing policies obeying (12), all servers are busy and the probability of conforming to the service quality threshold T is $1-\epsilon$. In other words, the policy in (12) can be categorized as operating in the QED(T) regime. For time thresholds $T > 0$, the asymptotic distribution of the steady state waiting time is normal. When $T = 0$ however, the distribution is a truncated normal. Using Proposition 1 of de Véricourt and Jennings (2006a)), we can find the optimal staffing policy when $T = 0$: choose β_0 such that

$$\epsilon = \left(1 + e^{-\beta_0^2/2r^2}\sqrt{r}\frac{\Phi\left(\frac{\beta_0}{\sqrt{r\bar{r}}}\right)}{\Phi\left(\frac{-\beta_0}{r\sqrt{\bar{r}}}\right)}\right)^{-1} \tag{13}$$

and use this quantity in (12). Note that $\hat{r}_0 = r$ from (4) in this case. A classic refinement to (12) involves slightly modifying the round up procedure as follows (see Browne and Whitt 1995, Feller 1971),

$$\tilde{s}_n^* = \lceil \hat{r}_T n + \beta_T \sqrt{n} + 1/2 \rceil. \tag{14}$$

We have compared our heuristics (12) and (14) (using (13) when $T = 0$ and (5) when $T > 0$) with the optimal policy for $n = 2, 4, \ldots, 10, 15, \ldots, 25, 50, 100, 200, 500$, $r = 0.1, 0.25, 0.9$, $T = 0, 1/\mu, 2/\mu$ and $\epsilon = 1\%, 5\%, 10\%$ (which correspond to 432 numerical experiments). For two cases only, the difference in staffing levels was equal to 2. For all other cases, the difference was less than or equal to one. Overall, the refined policy $\tilde{s}_n^*$ worked slightly better. Further, in roughly 90% of the cases the relative error in the staffing levels was less than 10%. These errors occur when staffing levels are low (typically less than 3), for which a difference of one unit becomes significant. In short, our numerical results show that the heuristics perform well.

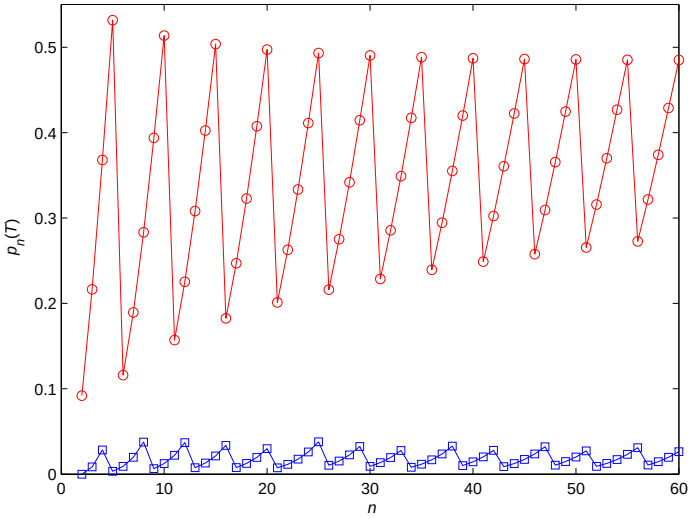

Figure 1 **Probability of Excessive Delay for $s_n = \lceil (1/5)n \rceil$ (above) and $\tilde{s}_n^* = \lceil (1/5)n + 0.3\sqrt{n} + 1/2 \rceil$ (bellow), for $r = 1/4$ and $T = 1/\mu$**

Figure 1 compares the performance of heuristic (12) with a ratio policy when $\epsilon = 5\%$, $r = 1/4$ and $T = 1/\mu$ for which $\hat{r}_T = 1/5$ and $\beta_T = 0.3$. The sharp oscillations in performance can be explained by the stepwise increase in the staffing level that occurs when $\hat{r}_T n$ is rounded up. Obviously, the ratio policy does not guarantee the target ϵ for all n while our heuristic consistently provides a probability of excessive delay less than 5%. This observation is discussed in detail and generalized in the next section.

4. A Queueing Model Perspective on California Bill AB 394

In this section we present a queueing perspective of California Bill AB 394, to evaluate the effectiveness of ratio policies in providing high quality of care consistently across all medical units and hospital sizes. Section 4.1 focuses on the probability of excessive delay. Nurse burnouts and average delays are briefly discussed in Sections 4.2 and 4.3, respectively.

4.1. Inconsistent Quality of Care

The main idea when deriving the collection of staffing ratios mandated by California Bill AB 394 was to evaluate, for each hospital unit, the proportion of nursing time a patient requires during a typical shift (controlling for patients' acuity levels and nurses' skills). Because quantitative data are lacking, the interest groups and research teams who framed Bill AB 394, and ultimately set the ratio values, relied heavily on expert panels comprised primarily of highly qualified and experienced registered nurses and nurse administrators (see CDHS 2003, IHSEP 2001). One should note that this proportion of time is precisely the same as the load factor r in our queueing framework. That is, when the number of patients is n, the mandated ratio policy sets the number of nurses s_n equal to the nominal patient load rn: $s_n^R = \lceil rn \rceil$. For example, the mandated ratio is 1/4 for pediatrics, which also provides an estimate of r for these medical units. We refer to this staffing rule as the *nominal ratio policy*. Within the spectrum of ratio policies of the form $\lceil \gamma n \rceil$, the nominal ratio policy is the one such that $\gamma = r$. The following result captures the asymptotic behavior of both nominal and more general ratio policies.

PROPOSITION 2. *Ratio policies can be asymptotically optimal only when $\epsilon \geq 50\%$. In particular, the nominal ratio policy mandated by AB 394 is asymptotically optimal only for $T = 0$ and $\epsilon =$*

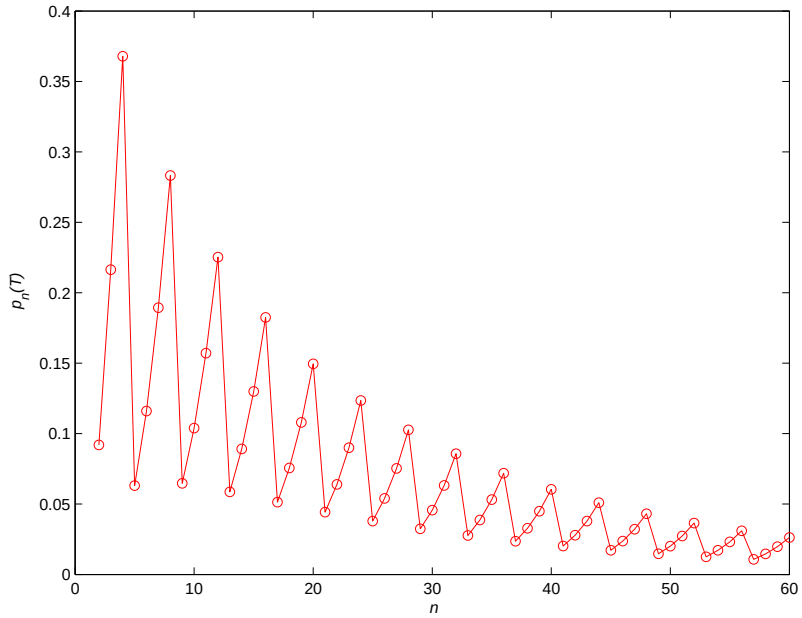

Figure 2 **Probability of Excessive Delay for** $s_n = \lceil (1/4)n \rceil$, $r = 1/4$, $T = 1/\mu$

$1/(1+\sqrt{r}) > 50\%$. *When* $T > 0$, *only one ratio policy of the form* $\hat{r}_T n$ *can be asymptotically optimal, in which case* $\epsilon = 50\%$.

Proof. These results follow from Proposition 1 above and Proposition 1 of de Véricourt and Jennings (2006b). □

This result together with the good performance of heuristic (12) suggest that when a ratio policy performs consistently across all hospital sizes, it necessarily performs consistently poorly, providing a probability of excessive delay greater than 50%. Further, when $T > 0$ only one ratio policy can yield a consistent (and high) probability of excessive delay, but it is not the nominal ratio policy mandated by California legislation. This is illustrated by Figure 1 (curve above) which depicts the performance of $s_n = \lceil \hat{r}_T n \rceil$ (with $\hat{r}_T = 1/5$ when $r = 1/4$ and $T = 1/\mu$). Ignoring round up effects, this policy seems to provide consistent but poor quality of care (with a probability of excessive delay reaching 50% as expected). By contrast, Figure 2 depicts the performance of $s_n = \lceil rn \rceil$, the nominal ratio policy required by AB 394. For large n, the policy provides reasonable probability of excessive delay (less than 5%). However, this performance is not consistent across all n. Indeed, suppose that the threshold time is strictly positive, the nominal ratio policy is used for staffing, and the target probability is reasonable: $\epsilon << 50\%$. Equation (12) suggests that the nominal staffing policy exceeds the target probability, say $\epsilon = 5\%$, for small hospital units (i.e. for $n < \hat{n}$, where $\hat{n}$ is such that $r\hat{n} = \hat{r}_T \hat{n} + \beta_T \sqrt{\hat{n}}$) and unnecessarily overstaffs for large units. This is illustrated in Figure 2. Ratio policies can also result in quality of service that worsens with unit size. If the ratio is below $\hat{r}_T$, which corresponds to $\beta = -\infty$, Proposition 1 states that the quality of service converges to 1, as illustrated in Figure 3.

The previous discussion sheds light on pooling effects under the optimal staffing rule. When unit size and the staffing levels are assumed to take real values (i.e. we ignore round-up effects), s_n^* can easily be shown to be concave for reasonable values of ϵ (i.e., $\epsilon > 50\%$), which suggests that the optimal staffing level exhibits usual pooling effects. However, the effects are more dramatic than just concavity. Recall that rn represents the nominal patient load, that is, the long run cumulative fraction of nursing time required in the unit provided nurses are always available when patients become needy. We deduce that for large systems ($n > \hat{n}$), s_n^* actually specifies a number of nurses *less* than the baseline rn; moreover, the deviation from the baseline is order n, a phenomenon we

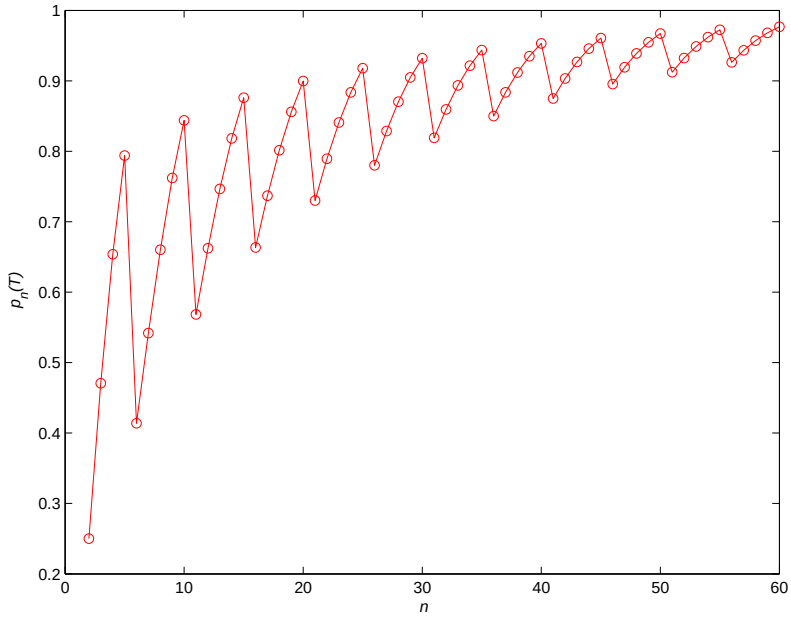

Figure 3 **Probability of Excessive Delay for** $s_n = \lceil (1/5)n \rceil$**,** $r = 1/4$**,** $T = 0$

refer to as the super pooling effect. This constitutes a significant deviation from traditional staffing problems, where a common point of departure is covering at minimum the offered load of the system. Note that super pooling effects also appear in open queues with abandonment. When such systems are staffed in the $QED(T)$ regime, the service rate is intentionally less than the arrival rate and the excess workload is accommodated through customer abandonment (Mandelbaum and Zeltyn 2006). In our case, the effect is due to forcing patients to wait for service, a process which decreases the carried load.

Our model and analysis are designed to ensure a safe environment for patients. This requires choosing the right values for T and ϵ and staffing accordingly. Notice that we make no assumptions about what these specific values should be; we leave this task to the policy makers. Of course, when T and ϵ are not well chosen, patients are not seen in a timely fashion and develop complications more frequently. For such out of control situations the advantage of the super pooling effect is lost.

4.2. Nurse Burnout

Nurse burnout and fatigue are important factors contributing to the national nursing shortage and frequent turnover that lead to understaffing of medical units (Wright et al. 2006). These factors can also affect overall patient satisfaction (Vahey et al. 2004). Burnout and the total workload experienced by nurses are usually managed by adequately scheduling shifts (which should also guarantee the staffing levels specified by the queueing model). In particular, these shifts should limit nurse working hours, allow for enough breaks and consider individual preferences (see Rogers et al. 2004). In fact, some hospitals offer flexible shifts with long recovery periods in order to retain nurses (Brooks 2000, Richardson et al. 2003, Cline et al. 2003).

Nonetheless, in conjunction with efficient scheduling systems, legislators may also want to limit the utilization rates experienced by nurses. In this case, the choice of a staffing rule $n \mapsto s_n$ should guarantee that

$$u_n \leq \kappa \tag{15}$$

where $\kappa \leq 1$ is a pre-specified constraint and $u_n \equiv E[B_n]/s_n$ denotes the long run utilization rate of a nurse, with B_n denoting the number of busy nurses in steady state. Based on a simple fluid approximation, we ignore round-up effects and take $E[B_n] \approx n - \overline{r}/r s_n$ (from Corollary 1 in de Véricourt

and Jennings 2006a, with $B_n = s_n - I_n$ where I_n is the number of idled servers). It follows that (15) becomes $s_n \geq \gamma_\kappa n$ with

$$\gamma_\kappa = \frac{r}{r\kappa + \overline{r}}. \tag{16}$$

This suggests that a γ_κ-ratio policy can keep the utilization rate of a nurse at a value around κ across unit sizes. However, the resulting staffing levels may not provide timely services, especially when $\gamma_\kappa n < s_n^*$ that is for $\sqrt{n} \leq \beta_T/(\gamma_\kappa - \hat{r}_T)$. The staffing rule in (12) can then be adjusted as follows,

$$\check{s}_n^* = \begin{cases} \hat{r}_T\, n + \beta_T \sqrt{n} & \text{if} \quad \sqrt{n} \leq \frac{\beta_T}{\gamma_\kappa - \hat{r}_T} \\ \gamma_\kappa\, n & \text{otherwise,} \end{cases} \tag{17}$$

and should provide a good alternative. On the other hand, the nominal ratio policy mandated by the law, $s_n = rn$, corresponds in our framework to $\kappa = 1$ and nurses are highly utilized (recall that we ignore round-up effects). This is not surprising since rn represents the nominal patient load. Note also that the modified policy $\check{s}_n^*$ does not display the super pooling effect. This effect appears for n such that the constraint on the utilization rate is binding (since $\hat{r}_T \leq r \leq \gamma_\kappa$).

4.3. Other Metrics: Average Delays

We argue in this paper that the probability of excessive delay is a natural choice to evaluate and determine nurse staffing rules. In this section, we briefly explore situations where staffing levels should guaranty a pre-specified average delay τ.

In this case, fluid approximations suggests the following staffing rules $n \mapsto s_n = rn/(1 + \mu r \tau)$ (see Corollary 1 in de Véricourt and Jennings (2006a)). This justifies the use of ratio policies when quality of care is measured in terms of average delays. The corresponding ratios depend on the pre-specified expected delays. This is however not the case for the nominal ratios required by AB 394. In fact, the California legislation results in staffing too many nurses (since $\hat{r_\tau} < r$). Further, Proposition 1 suggests that under this policy 50% of the medical needs will be delayed more than τ, which does not seem reasonable for many medical units.

The previous fluid approximation is appropriate when targets on average delays are not too small. For values of τ close to zero, more refined staffing rules can be derived based on the QED(0) regime (i.e., when the time threshold $T = 0$). In this regime, the limit of the average delay W_n when $n \to \infty$ is degenerate. On the other hand, for staffing rules such that $(s_n/n - r)\sqrt{n} \to \theta$, the inflated delay $\tilde{W}_n = \sqrt{n} W_n$ converges to a diffusion process with steady-state expectation $E[\tilde{W}]$ equal to

$$E[\tilde{W}] = \frac{\epsilon(\theta)}{\mu r \Phi\left(\frac{-\theta}{r\sqrt{\overline{r}}}\right)} \int_0^{+\infty} y \phi\left(y + \frac{\theta}{r\sqrt{\overline{r}}}\right) dy \tag{18}$$

where $\epsilon(\theta) = \left(1 + e^{-\theta^2/2r^2} \sqrt{r} \Phi\left(\frac{\theta}{\sqrt{r\overline{r}}}\right) / \left(\Phi\left(\frac{-\theta}{r\sqrt{\overline{r}}}\right)\right)\right)^{-1}$ (see Theorems 2, 4 and Section 6 in de Véricourt and Jennings 2006a). Given τ, the target on expected delay, this approach leads to staffing rule $s_n = rn + \theta_{\tau,n}\sqrt{n}$ where $\theta_{\tau,n}$ is such that $E[\tilde{W}] = \mu r \tau \sqrt{n}$. It is then worth noting that s_n is not a ratio policy. In particular, the nominal ratio policy mandated by the law (for which $\theta_{\tau,n} = 0$) generates average delays equal to $2/(\mu r(1 + \sqrt{r})) \int_0^{+\infty} y\phi(y) dy \sqrt{n}$ which are unit-size dependent.

In any case, we find the use of average delays in nurse staffing problematic. Rules based on this objective may allow frequent excessively long delays, which could have dire consequences for patients. More generally, the probability of excessive delay enables decoupling medical patient safety (defined by T) from public policy choices regarding the acceptable risk levels in medical units and their associated costs (driven by ϵ).

5. Robustness

The previous analysis is based on several assumptions that simplify the reality of nurse-staffing problems in hospitals. Using simulations, we show in this section that our findings are robust in general when these assumptions are relaxed.

We consider situations where service times are not exponentially distributed, patients' acuity levels are non-homogeneous and service times depend on patients' delays. For each of these cases, the optimal staffing rule exhibits a structure consistent with (12) which indicates that the main insights of Section 4.1 still hold.

Other extensions are possible. For instance, the experience levels of nurses can be non-homogeneous in medical units. Hospitals also hire temporary staff and pool nurses from different wards. These situations raise questions concerning the mix of nurses that needs to be staffed. This also brings up issues related to coordination of care and learning effects. Further, decision rules need to be specified in order to dynamically designate the available nurse that should assist a needy patient. In any case, mandated ratios[2] do not directly address these issues which fall outside the scope of our study.

5.1. General Service Times

As mentioned earlier, our assumption that service times are exponentially distributed is common in hospital capacity planning and makes the mathematical analysis tractable. Nevertheless, we show in the following that the structure of the optimal staffing rule remains consistent with (12) when distributions are not exponential.

We consider service times following Erlang and hyperexponential distributions with coefficients of variation less or greater than one, respectively. The optimal staffing levels s_n^* are then generated through simulations and are compared to the staffing rule $s_n = \hat{r}_T n + \beta\sqrt{n}$ for different values of the parameters. We do not expect the linear part of the optimal policy to be affected by the variance of the service time and therefore keep the linear coefficient of s_n equal to $\hat{r}_T$ as defined in (4). On the other hand, we need to adjust the square root coefficient β. Since no analytical formula is available to set the value of this parameter, we resort to the original definition of β in (3) and take $\beta \approx (s^*/n - \hat{r}_T)\sqrt{n}$ for large values of n.

Tables 2 and 3 (in the appendix) indicate the optimal staffing levels s_n^* for different parameters where C_s^2 denotes the squared coefficient of variation of the corresponding distribution. For all of these cases, $r = 1/4$ and the average service time $1/\mu$ is equal to one. Table 2 corresponds to service times following Erlang distributions with number of stages η and rate δ (with mean $\eta/\delta = 1$). The coefficient of variation is less then one and is equal to $1/\delta$. Similarly, Table 3 corresponds to service times generated by the random variable of the form $U \times X$, where U is bernoulli(p) and X is exponential with mean $1/p$ (so that the mean of UX is equal to one). The resulting coefficient of variation is then equal to $(2-p)/p$ which is larger than one.

Tables 4 and 5 (in the appendix) indicate the difference $s_n^* - s_n$ corresponding to Tables 2 and 3, respectively. The square root term β of s_n is evaluated at $n = 100$. For instance, when $\epsilon = 5\%$, $C_s^2 = 0.5$ and $T = 0$, the optimal staffing level for $n = 100$ is equal to 33 which leads to $\beta = 3.1$ using (3).

Tables 4 and 5 show that the absolute value of the difference $s_n^* - s_n$ is always less than or equal to one. This confirms the non-linear structure of the optimal staffing rule. It also follows that our main results and insights still hold. For instance, the ratio policy $\lceil n/4 \rceil$ mandated by the law staffs too few nurses when $T = 0$, even for low C_s^2, unless ϵ is large. These results are consistent with Proposition 2. Table 1 briefly illustrates this point and reports the values of $s_n^* - \lceil n/4 \rceil$ for different values of n and ϵ. On the other hand, the optimal policy staffs below the nominal load when T and

[2] The law requires the ratios to be adjusted to account for nurse's skills levels but without specifics.

n are large, showing a super-pooling effect even when the coefficient of variation is less than one. For instance, the optimal staffing level for $\epsilon = 1\%$, $n = 20$, $T = 10$ and $C_s^2 = 0.1$ is equal to $s_n^* = 2$ which is less than the offered load $rn = 5$.

Table 1 **Performance of the Ratio Policy: $s_n^* - \lceil n/4 \rceil$ for $T = 0$ and $C_s^2 = 0.1$**

ϵ	n=5	n=10	n=20	n=30	n=100
1%	2	3	5	6	11
5%	2	2	4	4	8
10%	1	2	3	3	7
30%	1	1	2	2	4

5.2. Non-Homogeneous Acuity Levels

In general, patients' needs depend on their acuity levels. However when setting public policy, the possible combinations of patient states are too vast to fully accommodate with specific guidelines. Instead, policy makers may want to set nurse-to-patient ratios assuming either the lowest or the highest possible acuity level for all patients in the unit. In fact, California Bill AB394 specifically advises that, in order to account for case mix and patient acuity levels, hospitals use the mandated minimum staffing ratio recommendations in conjunction with staffing policies and procedures developed locally at the hospitals. We have considered so far the case of homogeneous patients for which such adjustments are not required. Nonetheless, we explore in the following how our result will change when patients present in the unit have different acuity levels.

Specifically, we consider a situation where two types of patients are present in the system. In our framework, a patient's acuity level is represented by her corresponding activation rate. Hence, we assume that patients of the two types generate needs at different activation rates λ_1 and λ_2, respectively, with $\lambda_1 \leq \lambda_2$. Note that acuity levels can also affect the service times, but we do not expect the results to significantly change and for the sake of simplicity we set $1/\mu = 1$ regardless of the patient's type. We denote by ν (resp. $1-\nu$) the proportion of type 1 patients (resp. type 2) in the unit. For n patients in the system, the total offered load is then equal to rn where $r = \nu\lambda_1/(\lambda_1+\mu) + (1-\nu)\lambda_1/(\lambda_1+\mu)$ (which constitutes a natural generalization of the definition of r in Section 2 and therefore uses the same notation when no confusion is possible).

Table 6 indicates the optimal staffing levels s_n^* evaluated through simulations for different values of ν. For all of these cases, we set $\lambda_1 = 0.1$ and determine λ_2 such that $r = 1/4$ (for instance when $\nu = 50\%$, we have $\lambda_2 = 0.692$). Table 7 reports then the difference $s_n^* - s_n$, where s_n^* is found in Table 6 and s_n is of the form $s_n = \hat{r}_T n + \beta\sqrt{n}$, with $\hat{r}_T$ defined in (4) using $r = 1/4$. Since no analytical result exists, β is estimated at $n = 100$ with $\beta \approx (s^*/n - \hat{r}_T)\sqrt{n}$. Table 6 confirms the non-linear structure of the optimal staffing rule and the main insights of Section 4.1 hold. In particular, an analysis similar to Section 5.1 easily shows that ratio policies systematically understaff the system for $T = 0$ unless ϵ is large. Further, the optimal policy staffs below the nominal load when T and n are large, showing a super-pooling effect (this is for example the case with $n = 20$, $\epsilon = 1\%$, $T = 10/\mu$ and $v = 50\%$ where $s_n^* = 3 < rn = 5$).

5.3. Delay-Dependent Service Times

The vast majority of staffing models in health care settings consider delay-independent service times. (For recent reviews of the use of queueing models in health care settings, see for instance Green (2004), Preater (2002b,a), Singh (2006).) A sample of such models is Tucker et al. (1998) which determines the need to activate a backup OR team during the night shift at a Level II trauma center. Another recent example is Green et al. (2006) which studies staffing levels in Emergency

Departments. The absence of delay-independent service times in all these studies is justified in general by the fact that reasonable delays in access to care should not affect patients' treatment times.

To illustrate this point, we relax our assumption that service times do not depend on delays. In the absence of an existing proper framework to model this relationship (to the best of our knowledge), we present two approaches which appear natural and general enough for our purposes. In our first model, a patient has k potential needs which may occur when she is waiting or in a stable condition. The activation time of a particular need follows an exponential distribution with rate λ/k, so that the activation rate of the first issue is equal to λ when the patient's condition is stable. When assisting a patient, nurses must handle one issue at a time. The resulting service time follows an Erlang distribution with rate μ and a delay-dependent number of stages equal to the number of activated needs. In our second case, we assume that the potential number of issues is infinite and follows a Poisson distribution with rate λ. For instance, a nurse serving a patient who has not received care for the last t units of time, must treat λt issues on average. Again the distribution of the service time is Erlang with rate μ and a delay-dependent number of stages equal to the number of needs that occurred before service. We also denote by $s_n^{(k)}$, the optimal staffing level when a patient can potentially accumulate k needs. In particular $s_n^{(1)}$ corresponds to the base case of this paper.

Table 8 (in the appendix) indicates the optimal staffing levels for different values of k, while Table 9 reports the difference $s_n^{(1)} - s_n^{(k)}$, which measures the impact of delay-dependent service times. For most reasonable cases (that is for $n \leq 30$, $T < 10/\mu$ and $\epsilon \leq 30\%$ in our study), delay-dependent service times have a negligible impact on the optimal staffing levels. This is because the optimal staffing rule dimensions the system in such a way that nurses can assist patients before additional needs accumulate. By contrast, issues may build up under mandated ratio policies when the resulting staffing level is too low, as described in Section 4.1. Patients also accumulate needs when the values of T and ϵ are not appropriately chosen. This is clearly the case in our numerical results for $T = 10/\mu$ and $\epsilon = 67\%$, although the effect appears only for large n ($n \geq 30$). In fact, the time threshold T should precisely be selected so as to avoid this accumulation of issues or even a dramatically worse event to occur that would require additional services or a transfer to the ICU. This interpretation reinforces the role of the second parameter ϵ to limit the frequency of unreasonable delays.

6. Conclusion

In this paper, we argue that closed queueing models and the probability of excessive delay constitute a good framework to tackle nurse staffing problems. Using new many-server asymptotic results, we have shown that ratio policies (nominal or not), as mandated by California Bill AB 394, cannot provide consistently low probability of excessive delay across units of different sizes. Ratio policies can sometimes provide consistent quality levels across hospital sizes, a desirable trait from a policy making point of view. (One should note the contrast with staffing proportionally to the nominal load in open queues, which never displays this consistency.) However, the corresponding probability of delay is always larger than 50% and this is likely to result in poor quality of care. Further, some fixed ratio policies may perform well over the typical unit size range ($n < 100$), but the resulting level of care will then be inconsistent.

This inconsistency can only be observed when accounting for the randomness inherent to the care delivery processes. However, to the best of our knowledge, variability and congestion have been systematically ignored in both the debate surrounding nurse-to-patient ratios and the supporting empirical studies; for instance, The Institute of Medicine never mentions these issues in its recent nurse-staffing recommendations, see Greiner and Knebel (2004). As an alternative to ratio policies, we also develop a policy which provides staffing rules that 1) are easy to implement, 2) are

consistent across unit size, and 3) can achieve pre-specified limits on the probability of excessive delay. Furthermore, our results do not make any assumptions about what this target or the system parameters should be.

Heretofore, the tasks of understanding and measuring patient needs, in terms of total nursing time, and understanding what provides quality service has been tightly coupled. For instance, the general belief implicit in ratio policy arguments is that improving patient outcomes is as simple as increasing nursing ratios (Aiken et al. 2002). This paper helps to unravel the two issues. Namely, one should understand patient needs (calculating r) independently of defining how quickly patients must be attended to (determining T and ϵ). We feel that our approach can help better frame the policy debate and further statistical studies on patient outcomes. In particular, empirical studies should examine how the time threshold T affects the quality of care in different types of medical units. Variability and congestion in the care delivery process should also be quantified. Finally, public policies and the society as a whole need to decide what constitutes an acceptable level of safety through the choice of ϵ while considering resulting costs.

There are several ways our model can be extended. For instance, accounting for a heterogeneous workforce (such as the combination of licensed and registered nurses) can be achieved by considering mean service times which are skill-level specific. This raises the question of how to dynamically dispatch nurses with different experience levels when patients are in need. Finally, following the underlying hypothesis of AB 394, we have assumed that the number of nurses in a medical unit could be easily adjusted as the number of patient changes. To fully understand the associated transient effects, one should add patient arrivals and random lengths of stay to our model.

Acknowledgments

The authors are grateful to Suri Gurumurthi for his work on the numerical examples. We also thank Galit Yom-Tov for assisting with the current presentation of the paper.

References

ACC/AHA. 2001. *Guidelines for Percutaneous Coronary Intervention (Revision of the 1993 PTCA Guidelines)*. American College of Cardiology.

ACC/AHA. 2002. *Guideline Update for the Management of Patients With Unstable Angina and Non-ST-Segment Elevation Myocardial Infarction*. American College of Cardiology.

Aiken, LH., SP. Clarke, DM. Sloane, J. Sochalski, JH. Silber. 2002. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA* **288**(16) 1987–93.

Alonso-Echanove, J., JR Edwards, MJ Richards, P. Brennan, RA Venezia, J. Keen, V. Ashline, K. Kirkland, E. Chou, M. Hupert, et al. 2003. Effect of nurse staffing and antimicrobial-impregnated central venous catheters on the risk for bloodstream infections in intensive care units. *Infect Control Hosp Epidemiol* **24**(12) 916–25.

Baron, O., J. Milner. 2006. Staffing to maximize profit for call centers with alternative service level agreements. Submitted to Operations Research.

Brooks, I. 2000. Nurse retention: moderating the ill-effects of shiftwork. *Human Resource Management Journal* **10**(4) 16–31.

Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. J. Dshalalow, ed., *Advances in Queueing: Theory, Methods, and Open Problems*. CRC Press, Boca Raton, FL, 463–480.

CDHS. 2003. *Final Statement of Reasons, on Hospital Nurse Staff Ratios and Quality of Care*. California Department of Health Services. Report on AB394.

CDHS. 2005. *DA/DM memorandum following the California Superior Court Decision of March 14, 2005*. California Department of Health Services. Http://www.dhs.ca.gov/lnc/ntp/default.htm.

Clark, S. 2005. The policy implications of staffing-outcomes research. *JONA* **35**(1) 17–19.

Cline, D., C.A. Reilly, C. CCNS, J.F. Moore. 2003. What's behind RN turnover? *Nursing Management (Springhouse)* **34**(10) 50–53.

de Véricourt, F., Otis Jennings. 2006a. Dimensioning large-scale membership services. Duke University, submitted to Operations Research.

de Véricourt, F., Otis Jennings. 2006b. Supplement to "nurse-to-patient ratios in hospital staffing: a queuing perspective". Tech. rep., Duke University.

Feller, W. 1971. An introduction to probability theory and its applications. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1971, 3rd ed.* .

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Mgmt* **5** 79–141.

Green, Linda. 2004. *Chapter 3. Capacity Planning in Hospitals*. Kluwer Academic Publishers. In Handbook of Operations Research/Management Science Applications in Health Care.

Green, L.V., J. Soares, J.F. Giglio, R.A. Green. 2006. Using Queueing Theory to Increase the Effectiveness of Emergency Department Provider Staffing. *Academic Emergency Medicine* **13**(1) 61.

Greiner, A., E. Knebel. 2004. Keeping patients safe: Transforming the work environment of nurses [Committee on the Work Environment for Nurses and Patient Safety, Board on Health Care Services, Institute of Medicine of the National Academies].

Halfin, S., W. Whitt. 1981. Heavy-traffc limits for queues with many exponential servers. *Operations Research* **29** 567–588.

IHSEP. 2001. *AB 384: California and the Demand for Safe and Effective Nurse to Patient Staffing Ratios*. Institute for Health & Socio-Economic Policy. California.

Kleinrock, L. 1975. *Queueing Systems*, vol. 1. John Wiley, New York.

Lerner, E.B., R.M. Moscati. 2001. The Golden Hour: Scientific Fact or Medical" Urban Legend"? *Academic Emergency Medicine* **8**(7) 758–760.

Liu, Y., L. Wein. 2005. A queueing analysis to determine how many additional beds are needed for the detention and removal of illegal aliens. Working Paper.

Mandelbaum, A., S. Zeltyn. 2006. Staffing many-server queueus with impatient customers: constraint satisfaction in call centers. Submitted to Operations Research.

Needleman, J., P. Buerhaus. 2003. Nurse staffing and patient safety: current knowledge and implications for action. *International Journal for Quality in Health Care* **15**(4) 275–277.

Preater, J. 2002a. A bibliography of queues in health and medicine. *Health Care Management Science* **5** 283–283.

Preater, J. 2002b. Queues in Health. *Health Care Management Science* **5**(4) 283–283.

Richardson, A., N. Dabner, S. Curtis. 2003. Literature Review Twelve-hour shift on ITU: a nursing evaluation. *Nursing in Critical Care* **8**(3) 103.

Rogers, A.E., W.T. Hwang, L.D. Scott, L.H. Aiken, D.F. Dinges. 2004. The working hours of hospital staff nurses and patient safety. *Health Affairs* **23**(4) 202–212.

Singh, V. 2006. Use of Queuing Models in Health Care. Working paper, University of Arkansas for Medical Sciences.

Sochalski, J. 2004. Is more better?: the relationship between nurse staffing and the quality of nursing care in hospitals. *Med Care.* **42**(2) II67–73.

Spetz, J. 2005. Public policy and nurse staffing: What approach is best. *JONA* **35**(1) 14–16.

Sternberg, S. 2006. Hospitals too slow on heart attacks - study finds only 35% react quickly enough. *US Today* Nov. 11.

Tucker, AL, AC Edmondson. 2003. Why hospitals don't learn from failures: organizational and psychological dynamics that inhibit system change. *Calif Manage Rev.* **45** 55–72.

Tucker, J.B., J.E. Barone, J. Cecere, R.G. Blabey, C.K. Rha. 1998. Using Queueing Theory to Determine Operating Room Staffing Needs. *The Journal of Trauma: Injury, Infection, and Critical Care* **45**(1) 192.

Vahey, D.C., L.H. Aiken, D.M. Sloane, S.P. Clarke, D. Vargas. 2004. Nurse burnout and patient satisfaction. *Medical Care* **42**(2) 57–66.

Wright, P.D., K.M. Bretthauer, M.J. Cote. 2006. Reexamining the Nurse Scheduling Problem: Staffing Ratios and Nursing Shortages. *Decision Sciences* **37**(1).

7. Appendix

Table 2 **Optimal Staffing levels for $C_s^2 \leq 1$**

		n=5				n=10				n=20				n=30				n=100			
		μT				μT				μT				μT				μT			
	C_s^2	0	1	2	10	0	1	2	10	0	1	2	10	0	1	2	10	0	1	2	10
ϵ=1%	1	4	3	3	1	7	5	4	2	10	7	6	3	14	9	8	4	36	25	21	9
	0.5	4	3	2	1	7	4	3	2	10	7	5	2	14	9	7	3	36	24	20	9
	0.2	4	3	2	1	7	4	3	2	10	6	5	2	14	9	7	3	36	24	20	9
	0.1	4	3	2	1	6	4	3	1	10	6	5	2	14	8	7	3	36	23	19	8
ϵ=5%	1	4	3	2	1	6	4	3	2	9	6	5	2	12	8	7	3	33	24	20	9
	0.5	4	2	2	1	6	4	3	2	9	6	5	2	12	8	7	3	33	23	19	9
	0.2	4	2	2	1	6	4	3	1	9	6	5	2	12	8	6	3	33	23	19	8
	0.1	4	2	2	1	5	3	3	1	9	6	5	2	12	8	6	3	33	23	19	8
ϵ=10%	1	3	2	2	1	5	4	3	1	8	6	5	2	11	8	7	3	32	23	19	9
	0.5	3	2	2	1	5	3	3	1	8	6	5	2	11	8	6	3	32	22	19	8
	0.2	3	2	2	1	5	3	3	1	8	5	4	2	11	8	6	3	32	22	19	8
	0.1	3	2	2	1	5	3	3	1	8	5	4	2	11	7	6	3	32	22	18	8
ϵ=30%	1	2	2	2	1	4	3	2	1	7	5	4	2	10	7	6	3	29	21	19	8
	0.5	2	2	1	1	4	3	2	1	7	5	4	2	10	7	6	3	29	21	18	8
	0.2	3	2	1	1	4	3	2	1	7	5	4	2	10	7	6	3	29	21	18	8
	0.1	3	2	1	1	4	3	2	1	7	5	4	2	10	7	6	3	29	21	18	8

Table 3 **Optimal Staffing levels for $C_s^2 \geq 1$**

		n=5				n=10				n=20				n=30				n=100			
		μT				μT				μT				μT				μT			
	C_s^2	0	1	2	10	0	1	2	10	0	1	2	10	0	1	2	10	0	1	2	10
ϵ=1%	1	4	3	3	1	7	5	4	2	10	7	6	3	14	9	8	4	36	25	21	9
	3	4	4	3	2	6	5	5	2	10	8	7	3	14	11	9	4	36	27	23	11
	9	4	4	4	2	6	6	5	3	10	9	8	4	14	12	11	5	36	31	26	13
ϵ=5%	1	4	3	2	1	6	4	3	2	9	6	5	2	12	8	7	3	33	24	20	9
	3	3	3	2	1	5	4	4	2	9	7	6	3	12	9	8	4	33	25	21	10
	9	3	3	3	2	5	5	4	2	9	8	7	4	12	10	9	4	33	28	24	11
ϵ=10%	1	3	2	2	1	5	4	3	1	8	6	5	2	11	8	7	3	32	23	19	9
	3	3	3	2	1	5	4	3	2	8	6	5	3	11	9	7	3	31	24	20	9
	9	3	3	3	2	5	4	4	2	8	7	6	3	11	9	8	4	31	26	22	10
ϵ=30%	1	2	2	2	1	4	3	2	1	7	5	4	2	10	7	6	3	29	21	19	8
	3	2	2	2	1	4	3	3	1	7	5	4	2	9	7	6	3	28	22	18	8
	9	2	2	2	1	3	3	2	1	6	5	4	2	9	7	6	3	27	22	19	8

Table 4 $s_n^* - s_n$ **for** $C_s^2 \leq 1$

		n=5				n=10				n=20				n=30			
		μT				μT				μT				μT			
	C_s^2	0	1	2	10	0	1	2	10	0	1	2	10	0	1	2	10
ϵ=1%	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
	0.5	0	1	0	0	1	0	0	0	0	1	0	-1	0	0	0	-1
	0.2	0	1	0	0	1	0	0	0	0	0	0	-1	0	0	0	-1
	0.1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ϵ=5%	1	0	1	0	0	0	0	0	0	0	0	0	-1	0	-1	0	-1
	0.5	0	0	0	0	0	1	0	0	0	0	0	-1	0	0	0	-1
	0.2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	-1	0
	0.1	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	-1	0
ϵ=10%	1	0	0	0	0	0	1	0	-1	-1	0	0	-1	-1	0	0	-1
	0.5	0	0	0	0	0	0	0	0	-1	1	0	0	-1	0	-1	0
	0.2	0	0	0	0	0	0	0	0	-1	0	-1	0	-1	0	-1	0
	0.1	0	0	0	0	0	0	0	0	-1	0	0	0	-1	-1	0	0
ϵ=30%	1	-1	0	0	0	0	0	-1	0	0	0	-1	0	0	0	-1	0
	0.5	-1	0	-1	0	0	0	-1	0	0	0	0	0	0	0	0	0
	0.2	0	0	-1	0	0	0	-1	0	0	0	0	0	0	0	0	0
	0.1	0	0	-1	0	0	0	-1	0	0	0	0	0	0	0	0	0

Table 5 $s_n^* - s_n$ **for** $C_s^2 \geq 1$

		n=5				n=10				n=20				n=30			
		μT				μT				μT				μT			
	C_s^2	0	1	2	10	0	1	2	10	0	1	2	10	0	1	2	10
ϵ=1%	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
	3	0	1	0	0	0	0	1	0	0	0	0	-1	0	1	0	-1
	9	0	0	1	0	0	0	0	0	0	0	0	-1	0	-1	0	-1
ϵ=5%	1	0	1	0	0	0	0	0	0	0	0	0	-1	0	-1	0	-1
	3	-1	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0
	9	-1	0	0	0	-1	0	0	0	0	0	0	0	0	-1	-1	-1
ϵ=10%	1	0	0	0	0	0	1	0	-1	-1	0	0	-1	-1	0	0	-1
	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
	9	0	0	0	1	0	0	0	0	0	0	0	0	0	-1	0	0
ϵ=30%	1	-1	0	0	0	0	0	-1	0	0	0	-1	0	0	0	-1	0
	3	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	0	0
	9	0	0	0	0	-1	0	-1	0	0	0	-1	0	0	-1	-1	0

Table 6 Optimal Staffing levels with Different Acuity Levels

		n=10				n=20				n=30				n=100			
		μT				μT				μT				μT			
	ν	0	1	2	10	0	1	2	10	0	1	2	10	0	1	2	10
ϵ=1%	0%	7	5	4	2	10	7	6	3	14	9	8	4	36	25	21	9
	20%	6	5	4	2	10	7	6	3	14	9	8	3	36	25	20	9
	50%	6	4	3	2	10	7	5	3	13	9	7	3	35	24	19	9
	80%	5	4	3	2	9	6	5	2	12	8	6	3	33	21	17	7
ϵ=5%	0%	6	4	3	2	9	6	5	2	12	8	7	3	33	24	20	9
	20%	5	4	3	2	9	6	5	2	12	8	7	3	33	23	19	9
	50%	5	4	3	1	9	6	5	2	12	8	7	3	33	22	18	8
	80%	5	3	3	1	8	5	4	2	11	7	6	3	30	20	15	7
ϵ=10%	0%	5	4	3	1	8	6	5	2	11	8	7	3	32	23	19	9
	20%	5	3	3	1	8	6	5	2	11	8	6	3	31	23	19	8
	50%	5	3	3	1	8	5	4	2	11	8	6	3	31	22	18	8
	80%	4	3	2	1	7	5	4	2	10	7	5	3	29	19	15	7
ϵ=30%	0%	4	3	2	1	7	5	4	2	10	7	6	3	29	21	18	8
	20%	4	3	2	1	7	5	4	2	10	7	6	3	29	21	17	8
	50%	4	3	2	1	7	5	4	2	9	7	5	3	28	20	16	7
	80%	3	2	2	1	6	4	3	2	9	6	5	2	27	17	14	6

Table 7 $s_n^* - s_n$ for Different Acuity Levels

		n=10				n=20				n=30			
		μT				μT				μT			
	ν	0	1	2	10	0	1	2	10	0	1	2	10
ϵ=1%	0%	1	1	0	0	0	0	0	0	0	0	0	0
	20%	0	1	1	0	0	0	1	0	0	0	1	-1
	50%	0	0	0	0	0	1	0	0	0	0	0	-1
	80%	-1	1	1	1	0	1	1	0	0	1	0	0
ϵ=5%	0%	0	0	0	0	0	0	0	-1	0	-1	0	-1
	20%	-1	1	0	0	0	0	0	-1	0	0	0	-1
	50%	-1	1	0	0	0	1	1	0	0	0	1	0
	80%	0	1	1	0	0	1	1	0	0	1	1	0
ϵ=10%	0%	0	1	0	-1	-1	0	0	-1	-1	0	0	-1
	20%	0	0	0	0	0	0	0	0	0	0	-1	0
	50%	0	0	0	0	0	0	0	0	0	0	0	0
	80%	0	1	0	0	0	1	1	0	0	1	0	0
ϵ=30%	0%	0	0	-1	0	0	0	0	0	0	0	0	0
	20%	0	0	0	0	0	0	0	0	0	0	0	0
	50%	0	1	0	0	0	1	0	0	-1	1	0	0
	80%	-1	0	1	0	0	1	0	1	0	1	1	0

Table 8 Optimal Staffing levels with Delay-Dependent Service Times

		n=5				n=10				n=30				n=100			
		μT				μT				μT				μT			
	K	0	1	2	10	0	1	2	10	0	1	2	10	0	1	2	10
ϵ=1%	1	4	3	3	1	7	5	4	2	14	9	8	4	36	25	21	9
	2	4	3	3	2	7	5	4	2	14	10	8	5	36	26	24	15
	10	4	3	3	2	7	5	4	3	14	10	9	8	36	28	26	23
	Poisson	4	3	3	2	7	5	4	3	14	10	9	8	36	28	27	26
ϵ=10%	1	3	2	2	1	5	4	3	3	11	8	7	3	32	23	19	9
	2	3	2	2	1	5	4	3	2	11	8	7	5	32	25	22	14
	10	3	2	2	2	5	4	3	3	11	9	8	7	32	26	25	23
	Poisson	3	2	2	2	5	4	3	3	11	9	9	8	32	27	26	26
ϵ=30%	1	2	2	2	1	4	3	2	1	10	7	6	3	29	21	18	8
	2	2	2	2	1	4	3	3	2	10	8	7	4	29	24	21	13
	10	2	2	2	2	4	3	3	3	10	8	8	7	29	25	25	22
	Poisson	2	2	2	2	4	3	3	3	10	8	8	8	29	26	26	26
ϵ=67%	1	2	1	1	1	3	2	2	1	8	6	5	2	26	20	16	7
	2	2	1	1	1	3	3	2	2	8	7	6	4	26	22	20	12
	10	2	1	1	1	3	3	3	3	9	8	7	7	26	25	24	22
	Poisson	2	2	2	2	3	3	3	3	9	8	8	8	27	26	26	26

Table 9 $s_n^{(1)} - s_n^{(k)}$ for Delay-Dependent Service Times

		n=5				n=10				n=30				n=100			
		μT				μT				μT				μT			
	k	0	1	2	10	0	1	2	10	0	1	2	10	0	1	2	10
ϵ=1%	2	0	0	0	-1	0	0	0	0	0	-1	0	-1	0	-1	-3	-6
	10	0	0	0	-1	0	0	0	-1	0	-1	-1	-4	0	-3	-5	-14
	Poisson	0	0	0	-1	0	0	0	-1	0	-1	-1	-4	0	-3	-6	-17
ϵ=10%	2	0	0	0	0	0	0	0	1	0	0	0	-2	0	-2	-3	-5
	10	0	0	0	-1	0	0	0	0	0	-1	-1	-4	0	-3	-6	-14
	Poisson	0	0	0	-1	0	0	0	0	0	-1	-2	-5	0	-4	-7	-17
ϵ=30%	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	10	0	0	0	-1	0	0	0	-1	0	0	-1	-3	0	-1	-4	-9
	Poisson	0	0	0	-1	0	0	0	-1	0	0	-1	-4	0	-2	-5	-13
ϵ=67%	2	0	0	0	0	0	-1	0	-1	0	-1	-1	-2	0	-2	-4	-5
	10	0	0	0	0	0	-1	-1	-2	-1	-2	-2	-5	0	-5	-8	-15
	Poisson	0	-1	-1	-1	0	-1	-1	-2	-1	-2	-3	-6	-1	-6	-10	-19

ESMT Working Papers

	ESMT No.	Competence Center
Nurse-To-Patient Ratios in Hospital Staffing: A Queueing Perspective Francis de Véricourt, ESMT Otis B. Jennings, Duke University	08-005	Management and Technology
Critical Mass Michał Grajek, ESMT Tobias Kretschmer, Ludwig-Maximilians-Universität München	08-004	Management and Technology
The Rhythm of the Deal: Negotiation as a Dance Erik H. Schlie Mark A. Young, Rational Games, Inc.	08-003	Leadership
Legacy Effects in Radical Innovation: A Study of European Internet Banking Erik H. Schlie, ESMT Jaideep C. Prabhu, Tanaka Business School, Imperial College London Rajesh K. Chandy, Carlson School of Management, University of Minnesota	08-002	Management and Technology
Upsetting Events and Career Investments in the Russian Context Konstantin Korotov, ESMT Svetlana Khapova, ESMT Visiting Professor and Assistant Professor at VU University Amsterdam	08-001	Leadership
Ambiguity Aversion and the Power of Established Brands A. V. Muthukrishnan, Hong Kong University of Science and Technology Luc Wathieu, ESMT	07-005	Management and Technology
Accelerated Development of Organizational Talent Konstantin Korotov, ESMT	07-004	Leadership
Usage and Diffusion of Cellular Telephony, 1998-2004 Michał Grajek, ESMT Tobias Kretschmer, Ludwig-Maximilians-Universität München	07-003	European Competitiveness
Estimating Level Effects in Diffusion of a New Technology: Barcode Scanning at the Checkout Counter Jonathan Beck, Humboldt Universität zu Berlin Michał Grajek, ESMT Christian Wey, Technische Universität, Berlin	07-002	European Competitiveness
Estimating Network Effects and Compatibility in Mobile Telecommunication Michał Grajek, ESMT	07-001	European Competitiveness

ESMT
European School of Management
and Technology GmbH

ESMT Campus
Schlossplatz 1
10178 Berlin
Phone: +49 (0) 30 212 31-1279 (publications)

www.esmt.org