Memorandum 1996 (October 2012). ISSN 1874-4850. Available from: http://www.math.utwente.nl/publications
Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

Flexible nurse staffing based on hourly bed census predictions

N. Kortbeek[1234], A. Braaksma[123], C.A.J. Burger[13],
P.J.M. Bakker[2], R.J. Boucherie[13]

[1]Center for Healthcare Operations Improvement and Research (CHOIR),
University of Twente, Drienerlolaan 5, 7500 AE Enschede, The Netherlands

[2]Department of Quality and Process Innovation, Academic Medical Center Amsterdam,
Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

[3]Stochastic Operations Research, Department of Applied Mathematics,University of Twente,
Drienerlolaan 5, 7500 AE Enschede, The Netherlands

[4]n.kortbeek@utwente.nl

Abstract

Workload on nursing wards depends highly on patient arrivals and patient lengths of stay, which are both inherently variable. Predicting this workload and staffing nurses accordingly is essential for guaranteeing quality of care in a cost effective manner. This paper introduces a stochastic method that uses hourly census predictions to derive efficient nurse staffing policies. The generic analytic approach minimizes staffing levels while satisfying so-called nurse-to-patient ratios. In particular, we explore the potential of flexible staffing policies which allow hospitals to dynamically respond to their fluctuating patient population by employing float nurses. The method is applied to a case study of the surgical inpatient clinic of the Academic Medical Center (AMC) Amsterdam. This case study demonstrates the method's potential to study the complex interaction between staffing requirements and several interrelated planning issues such as case mix, care unit partitioning and size, and surgical block planning. Inspired by the numerical results, the AMC decided that this flexible nurse staffing methodology will be incorporated in the redesign of the inpatient care operations during the upcoming years.

Keywords. Probability; workforce planning; inpatient care; nurse-to-patient ratio; float nurse.

1 Introduction

Deploying adequate nurse staffing levels is one of the prime responsibilities of inpatient care facility managers. Nursing staff typically accounts for the majority of hospital budgets [35], which makes that every appearance of overstaffing is scrutinized in times that tight cost-containment efforts are required [22]. At the same time, maintaining appropriate staffing levels is crucial to be able to provide high-quality care. There is a growing body of evidence implicating associations between decreased staffing and higher hospital related mortality and adverse patient events [19, 25], and increased work stress and

burnout among nurses [1, 2]. In this paper, we present an exact method to assist healthcare administrators in ensuring safe patient care, while also maintaining an efficient and cost-effective nursing service.

Workload on nursing wards depends highly on patient arrivals and lengths of stay, which are both inherently variable. Predicting workload, and staffing nurses accordingly, is essential for guaranteeing quality of care in a cost effective manner [6, 10]. Accurate workload predictions require the consideration of the dynamics of surrounding departments, since many patient arrivals at the inpatient care facility originate from the operating theater and the emergency department. In [21], we presented a method to predict bed census by hour in various care units of an inpatient clinic as a function of the operating room block schedule and a cyclic arrival pattern of emergency patients. The stochastic analytic model presented in the current paper takes these predictions of [21] as starting point to determine appropriate nurse staffing levels.

When designing and operating inpatient care services, recognizing the interrelation between various planning decisions, such as case mix, care unit partitioning, and care unit size, is important [21]. In addition, especially for surgical inpatient departments, alignment with the planning of the operating room schedule is beneficial. All these decisions are also intertwined with inpatient care workforce requirements, for example in terms of skill mix, number of full time equivalents, and staffing levels per working shift. In the present paper, we incorporate the tactical decision in [17] referred to as 'staff-shift scheduling' in the integrated modeling framework of [21]. We address the question: for all working shifts during the planning horizon, how many employees should be assigned to each inpatient care unit? These numbers, in turn, provide a guideline for the workforce dimensioning decision on the strategic level.

We explore the potential of flexible staffing policies that allow hospitals to dynamically respond to their fluctuating patient population. This flexibility is achieved by employing a pool of cross-trained nurses, for whom it is only at the start of a shift decided in which specific care units they will work. The commonly applied term for such flexible employees is 'float nurses' [14, 30]. The basic idea behind the possible added value of the introduction of flex pools is the following. Although the inpatient population fluctuates, this fluctuation is to a certain extent predictable, due to its dependance on the operating room schedule and other predictable variability in patient arrivals (e.g., seasonality, day-of-week, and time-of-day effects). This predictable variation can be taken into account when setting the staffing levels for 'dedicated nurses', nurses with a fixed assignment to a care unit. Typically, staffing levels are to be determined a number of weeks in advance, so that individual nurse rosters can be settled timely. Therefore, when only dedicated nurses are employed, the buffer capacity required to protect against random demand fluctuations can lead to regular overstaffing. When two or more care units cooperate by jointly appointing a flexible nurse pool, the variability of these random demand fluctuations balances out due to economies of scale, so that less buffer capacity is required.

Nurse-to-patient ratios are commonly used when determining staffing levels [2, 36]. These ratios indicate how many patients a registered nurse can care for during a shift, taking into account both direct and indirect patient care. Staffing according to nurse-to-patient ratios can be done in two ways. The ratios can be taken as mandatory lower bound, such as in California (USA) and Victoria (Australia), where legal minimums for nurse-to-patient ratios were set for general medical and surgical wards [3, 32]. The advantage of such minimum ratios is that a consistently high level of patient safety is guar-

anteed [19, 22]. The disadvantage, however, is that always all beds need to be staffed, because the possibility that all beds are occupied always exists and, as described, the nurse rosters have to be settled in advance. Therefore, overstaffing is a threat, since there is no flexibility to adjust staffing levels to predicted patient demand. Overcoming this disadvantage, a second version of applying nurse-to-patient ratios exists, which involves taking them merely as guidelines [12]. In that case, the assumption is that there exists slack in the time window within which some indirect patient care tasks can be performed, without having direct negative consequences on patient safety or work stress. As a result, the ratios may sometimes be violated, but not too often, and not too long. In our approach, we combine the advantages of both approaches, by utilizing two nurse-to-patient ratio targets. The first ratio needs to be satisfied at all times, while the second more restrictive ratio only for a certain fraction of time.

Our contribution is a generic exact analytic approach to find the number of nurses to be staffed each working shift that guarantees a desired quality of care reflected by nurse-to-patient ratios, in the most cost-effective manner. The approach directly connects with the bed census prediction method presented in [21], so that alignment of staffing decisions with other interrelated inpatient planning decisions can be achieved, as well as coordination with the operating theater and the emergency department. First, to match nursing capacity with demand predictions, a stochastic mathematical program is formulated to determine optimal staffing levels when only dedicated nurses are employed: the 'fixed staffing policy' model. Next, we formulate a model in which the flex pool with float nurses is introduced, and in which exactly the same quality constraints are satisfied as in the fixed staffing policy model. The formulation of the flexible staffing policy model includes an assignment procedure that prescribes the rules according to which the float nurses are assigned to specific care units at the start of each working shift. Because the flexible staffing model is computationally too expensive to solve to optimality in reasonable time, we present an approximation model, which provides a lower and an upper bound on the staffing requirements.

To illustrate its potential, the method is applied to the same case study as that of [21]. This case study involves the care units in the surgical inpatient clinic of the Academic Medical Center (AMC), serving the specialties traumatology, orthopedics, plastic surgery, urology, vascular surgery, and general surgery. Inspired by the numerical results, the AMC decided that the flexible nurse staffing method will be fully implemented during the upcoming years, as part of the total redesign of its inpatient care services.

This paper is organized as follows. Section 2 provides a review of relevant literature. Section 3 presents the models for the fixed and the flexible staffing policies. Section 4 presents the numerical results, and Section 5 closes the paper with a general discussion.

2 Literature

Capacity planning for nursing staff has received considerable attention from the Operations Research community. The nurse staffing process involves a set of hierarchical decisions over different time horizons with different precision.The first, strategic, decision level is the workforce dimensioning decision which concerns both the number of employees that has to be employed, often expressed in the number of full time equivalents, and the mix in terms of skill categories [16, 23, 26]. The second, tactical, level concerns staff-shift scheduling, which deals with the problem of selecting what shifts are to be worked and

how many employees should be assigned to each shift to meet patient demand [13, 20]. The third, operational offline, decision level concerns the creation of individual nurse timetables with the objective to meet the required shift staffing levels set on the tactical level, while satisfying a complex set of restrictions involving work regulations and employee preferences. This planning step is often referred to as 'nurse rostering' [7, 8]. The fourth, operational online, decision level, concerns the staff schedule reconsideration at the start of a shift. At this level, float nurses are assigned to specific care units [7, 30], and, based on the severity of need, on-call nurses, overtime, and voluntary absenteeism can be used to further align patient care supply and demand [15, 28]. The interdependence of the decision levels must be recognized to bring about systematic nurse staffing improvements. As expressed in the literature review [28], each level is constrained by previous commitments made at higher levels, and by the degrees of flexibility for later correction at lower levels. For a more elaborate exposition of the relevant decisions and considerations involved at each decision level, and a detailed overview of relevant literature, we refer the reader to [17].

The literature has mainly focused on nurse rostering, for example reflected by the survey and classification articles [7, 9, 13]. Although the rostering methods are computationally efficient and very helpful to support practitioners in creating timetables, they generally take required staffing levels as prerequisite information [5, 16]. Incorrect assumptions on the (tactical) required staffing levels, during the (operational offline) rostering process, might therefore result in the necessity to make expensive corrections on the operational online decision level, for instance by additionally hiring temporary staff. Therefore, to be able to provide adequate input for the rostering process, we focus on the tactical decision level, by specifying appropriate 24-hours-a-day-staffing levels divided in shifts (e.g., a day, evening and night shift).

Tactical workforce decision making in healthcare has received less attention. A spreadsheet approach has been presented in [12], to retrospectively fit optimal shift staffing levels on historical census data. Prospectively assessing the impact of alternative interventions is difficult via such approaches, since they lack the flexibility to explicitly model and study the coordination between different inpatient care decision levels, and the alignment with surrounding departments. Simulation studies have shown to be successful in taking a more integral approach (e.g., [15, 16]). The inherent disadvantage of simulation studies is, however, that they are typically context-specific, which limits the generalizability of study outcomes. Analytic, but deterministic, approaches can for example be found in [4, 27, 34]. Stochastic approaches to determine shift staffing levels are available in [10, 35, 36]. None of these references take an integral approach, as the demand distributions underlying the staffing decisions are not based on patient arrival patterns from the operating theater and emergency department.

Workforce flexibility is indicated as a powerful concept in reducing the required size of workforce and increasing job satisfaction [7, 11, 14, 15, 18, 29, 30]. To adequately respond to patient demand variability various types of flexibility are suggested, among which the use of part-time employees, overtime, temporary agency employees, and float nurses. Related to our work are the articles [14, 24] in which the potential of float pools with cross-trained nurses is investigated. Both these references address the aggregate decision which budget of float nurse hours should be available during a given time horizon, and, as such, do not address the level of working shifts. For the assignment strategy of a given number of available float nurses to care units at the start of working shifts the

authors of [31] indicate that formulating such an assignment strategy requires the consideration of three issues: (1) a methodology for the measurement of the severity of need for an additional nurse, (2) a prediction per care unit of that severity of need during an upcoming shift, and (3) development of a technique for the allocation of the available float nurses to care units to meet this need. While [31] focuses on the third issue by developing a branch-and-bound algorithm, our assignment strategy involves the consideration of all three steps.

Staffing according to nurse-to-patient ratios has received attention in the operations research literature in [10, 35, 36]. Both [10] and [35] indicate that in practice, setting the numerical values of the ratios is more based on negotiation than on science. The authors of [35] studied the relation between staffing costs and nurse-to-patient ratios. In this article, also two interesting directions for future research were stated: first, exploring the use of float nurse pools in satisfying nurse-to-patient ratios, and, second, developing models to make scientific recommendations on the numerical values of the ratios. The first issue is addressed in the current study. The second issue has been the focus of [10, 36]. Both these references present a queueing model along which they motivate that the ratios as mandated in California are too rigid. They underline the importance of differentiating ratios with patient mix (reflecting the severity of patients' illnesses and their acuity), and with care unit size. In our study, we focus on determining staffing levels given prespecified nurse-to-patient ratios. Nevertheless, we do want to stress the importance of employing meaningful nurse-to-patient ratios in realizing high-quality staffing.

To conclude, our contribution is an exact stochastic analytic approach, aimed at deriving appropriate staffing levels, including the flexibility of float nurses, using nurse-to-patient ratios, while taking an integrated care chain perspective.

3 Methods

In this section, the staffing models are presented. The staffing models are based on the bed census predictions that are obtained from the model of [21]. In Section 3.1 we first provide an overview of this bed census prediction model. Section 3.2 discusses the requirements that need to be satisfied in setting appropriate staffing levels. Section 3.3 presents the fixed staffing model. Section 3.4 formulates the model to find optimal staffing levels when float nurse pools are applied: the flexible staffing model. Since the flexible model suffers from the curse of dimensionality, we approximate the solution via two models that respectively find upper and lower bounds on the staffing requirements.

3.1 Bed census predictions

The model from [21] predicts the workload at an inpatient care facility of several care units on a time scale of hours due to patients originating from the upstream operating theater and emergency department. The basis for the operating room outflow prediction is the Master Surgery Schedule (MSS). The MSS is a blueprint prescribing which specialty operates in which operating room on which day of the week [33]. The basis for the emergency department outflow prediction is a cyclic random arrival process which we defined in [21] as the Acute Admission Cycle (AAC). Schematically, the approach is as follows. First, the impact of the MSS and the AAC are separately determined and then combined to obtain the overall steady state impact of the repeating cycles. Second, the obtained

demand distributions are translated to bed census distributions. Here we provide a short overview of the prediction model, Appendix A provides a detailed summary.

For the demand predictions, for both elective and acute patients three steps are performed. First, the impact of a single patient type in a single cycle (MSS or AAC) is determined, by which in the second step the impact of all patient types within a single cycle can be calculated. Then, since the MSS and AAC are cyclical, the predictions from the second step are overlapped to find the overall steady state impact of the repeating cycles. The workload predictions for elective and acute patients are combined to find the probability distributions of the number of recovering patients at the inpatient care facility on each unique day in the cycle which we denote as the Inpatient Facility Cycle (IFC). The length of the IFC is the least common multiple of the lengths of the MSS and the AAC.

Patient admission requests may have to be rejected due to a shortage of beds, or patients may (temporarily) be placed in less appropriate units. As a consequence, demand predictions and bed census predictions do not coincide. Therefore, an additional step is required to translate the demand distributions into census distributions. This translation is performed by assuming that after a misplacement the patient is transferred to his preferred care unit when a bed becomes available, where a fixed patient-bed allocation policy is assumed, which prescribes the prioritization of such transfers.

3.2 Staffing requirements

We consider a planning horizon of Q days ($q = 1, \ldots, Q$). Each day is divided in T time intervals ($t = 0, 1, \ldots, T-1$). The set of working shifts is denoted by $\mathcal{T}$, where a shift τ is characterized by its start time b_τ and its length ℓ_τ. Within the time horizon (q, t) is a unique time interval and (q, τ) a unique shift. For notational convenience, $t \geq T$ indicates a time interval on a later day, e.g., $(q, T+5) = (q+1, 5)$. For each of K inpatient care units, with the capacity of unit k being M^k beds, staffing levels have to be determined for each shift (q, τ).

We consider two types of staffing policies: 'fixed' and 'flexible' staffing. Under fixed staffing the number of nurses working in unit k during shift (q, τ), denoted by $s^k_{q,\tau}$, is completely determined in advance. In the flexible case, 'dedicated' staffing levels $d^k_{q,\tau}$ per unit are determined, together with a number of nurses $f_{q,\tau}$ available in a flex pool. The decision to which particular units the float nurses are assigned is delayed until the start of the execution of a shift. We assign float nurses to one and the same care unit for a complete working shift, to avoid many hand-overs, which increase the risk of medical errors. Thus, we obtain staffing levels $s^k_{q,\tau} = d^k_{q,\tau} + f^k_{q,\tau}$, $k = 1, \ldots, K$, where $f^k_{q,\tau}$ denotes the number of float nurses assigned to unit k from the available $f_{q,\tau}$. Taking into account the current bed census and the predictions on patient admissions and discharges, the allocation of the float nurses to care units at the start of a shift is done according to a predetermined assignment procedure. We denote such an assignment procedure by π.

Our goal is to determine the most cost-efficient staffing levels such that certain quality-of-care constraints are satisfied. Since float nurses are required to be cross-trained it is likely that these are more expensive. To be able to differentiate, we therefore consider staffing costs ω_d for each dedicated nurse that is staffed for one shift and ω_f for each flexible nurse. Next, the nurse-to-patient ratio targets during shift (q, τ) are reflected by $r^k_{q,\tau}$, indicating the number of patients a nurse can be responsible for at any point in time. To keep track of the compliance to these targets, we define the concept 'nurse-to-

patient coverage', or shortly 'coverage'. With x^k the number of patients present at unit k at a certain time (q,t), $b_\tau \le t < b_\tau + \ell_\tau$, the coverage is given by $r^k_{q,\tau} \cdot s^k_{q,\tau}/x^k$. Thus, a coverage of one or higher corresponds to the preferred situation.

Starting from the following quality-of-care requirements as prerequisites, we will formulate the fixed and flexible staffing models by which the most cost-effective staffing levels can be found:

(i) **Staffing minimum.** For safety reasons, at least S^k nurses have to be present at care unit k at any time.

(ii) **Coverage minimum.** The coverage at care unit k may never drop below β^k.

(iii) **Coverage compliance.** The long-run fraction of time that the coverage at care unit k is one or higher is at least α^k. We denote the expected compliance at care unit k during shift (q,τ) by $c^k_{q,\tau}(\cdot)$; the arguments of this function depend on which staffing policy is considered.

(iv) **Flexibility ratio.** To ensure continuity of care, at any time, the fraction of nurses at care unit k that are dedicated nurses has to be at least γ^k.

(v) **Fair float nurse assignment.** The policy π, according to which the allocation of the available float nurses to care units at the start of a shift is done, has to be 'fair'. Fair is defined as assigning every next float nurse to the care unit where the expected coverage compliance during the upcoming shift is the lowest.

3.3 Fixed staffing

When only dedicated staffing is allowed, there is no interaction between care units. Therefore, the staffing problem decomposes in the following separate decision problems for each care unit k, and each shift (q,τ):

$$\begin{aligned}
\min \quad z_F = \quad & \omega_d s^k_{q,\tau} && (1)\\
\text{s.t.} \quad & s^k_{q,\tau} \ge S^k && (2)\\
& s^k_{q,\tau} \ge \left\lceil \beta^k \cdot M^k / r^k_{q,\tau} \right\rceil && (3)\\
& c^k_{q,\tau}\left(s^k_{q,\tau}, r^k_{q,\tau}\right) \ge \alpha^k && (4)
\end{aligned}$$

The constraints (2), (3), and (4) reflect requirements (i), (ii), and (iii), respectively. Let $X^k_{q,t}$ be the random variable with bed census distribution $\hat{Z}^k_{q,i}$ counting the number of patients present on care unit k at time (q,t). Then, the coverage compliance in (4) can be calculated as follows:

$$\begin{aligned}
c^k_{q,\tau}\left(s^k_{q,\tau}, r^k_{q,\tau}\right) &= \mathbb{E}\left[\frac{1}{\ell_\tau} \sum_{t=b_\tau}^{b_\tau+\ell_\tau-1} \mathbb{1}\left(X^k_{q,t} \le s^k_{q,\tau} \cdot r^k_{q,\tau}\right)\right]\\
&= \frac{1}{\ell_\tau} \sum_{t=b_\tau}^{b_\tau+\ell_\tau-1} \sum_{x=0}^{s^k_{q,\tau}\cdot r^k_{q,\tau}} \hat{Z}^k_{q,t}(x).
\end{aligned}$$

Observe that $\sum_{x=0}^{s^k_{q,\tau}\cdot r^k_{q,\tau}} \hat{Z}^k_{q,t}(x)$ reflects the probability that with staffing level $s^k_{q,\tau}$ and under ratio $r^k_{q,\tau}$ the nurse-to-patient ratio target is satisfied during time interval $[t, t+1)$. The optimum of (1) is found by choosing the minimum $s^k_{q,\tau}$ satisfying constraints (2) and (3), and increasing it until constraint (4) is satisfied.

3.4 Flexible staffing

The next step is to formulate the flexible staffing model. Note that for requirements (i) and (ii), the constraints are similar to those for fixed staffing. Under the assumption $\omega_d \leq \omega_f$, we can replace $s^k_{q,\tau}$ by $d^k_{q,\tau}$ in (2) and (3). Due to the presence of a flex pool the care units cannot be considered in isolation anymore. Hence, constraint (4) has to be replaced. An assignment procedure has to be formulated that fulfils requirement (v), and this assignment procedure influences the formulation of the constraint for requirement (iii). In addition, a constraint needs to be added for requirement (iv).

For an assignment procedure π that allocates the float nurses to care units at the start of a shift (q,τ), let $g^{\pi}_{q,\tau}(\boldsymbol{d}, f, \boldsymbol{y})$ be the vector of length K denoting the number of float nurses assigned to each care unit, when f flex nurses are available to allocate, the number of staffed dedicated nurses equals $\boldsymbol{d} = (d^1, \ldots, d^K)$, and the census at the different care units at time (q, b_τ) equals $\boldsymbol{y} = (y^1, \ldots, y^K)$. A vector of the type $\boldsymbol{y}$ reflects what we will call a *census configuration*.

Let π^* denote the assignment procedure that ensures constraint (v). The assignment procedure π^* depends on $\boldsymbol{d}_{q,\tau}$, $f_{q,\tau}$, and $r^k_{q,\tau}, k = 1, \ldots, K$, and therefore also the coverage does. Hence, requirement (v) gives a constraint of the form $c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}) \geq \alpha^k$. But, in addition, assignment procedure π^* depends on the census configuration $\boldsymbol{y}$ at time (q, b_τ), so to be able to calculate the coverage compliance we first need to compute $c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y})$, the coverage compliance given that at the start of shift (q,τ) census configuration $\boldsymbol{y}$ is observed. Then, the coverage compliance is given by:

$$c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}) = \sum_{\boldsymbol{y}} \Big\{ c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y}) \prod_{w=1}^{K} \hat{Z}^w_{q,\tau}(y^w) \Big\}.$$

Using $c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y})$, the assignment policy π^* satisfying requirement (v) is the one that satisfies:

$$g^{\pi^*}_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{y}) = \max_{\left\{f^1_{q,\tau}, \ldots, f^K_{q,\tau} : \sum_k f^k_{q,\tau} = f_{q,\tau}\right\}} \min_k \; c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y}). \tag{5}$$

Applying policy π^* provides $s^k_{q,\tau}(\boldsymbol{y})$, the number of nurses staffed at care unit k if census configuration $\boldsymbol{y}$ is observed at the start of shift (q,τ). Hence, the flexible model is, for each shift (q,τ):

$$\min \quad z_E = \omega_f f_{q,\tau} + \sum\nolimits_k \omega_d d^k_{q,\tau} \tag{6}$$

$$\text{s.t.} \quad d^k_{q,\tau} \geq S^k, \qquad \text{for all } k, \tag{7}$$

$$d^k_{q,\tau} \geq \left\lceil \beta^k \cdot M^k / r^k_{q,\tau} \right\rceil, \qquad \text{for all } k, \tag{8}$$

$$c^k_{q,\tau}(\boldsymbol{d}_{q,t}, f_{q,\tau}, r^k_{q,\tau}) \geq \alpha^k, \qquad \text{for all } k, \tag{9}$$

$$d^k_{q,\tau} \geq \gamma^k \cdot s^k_{q,\tau}(\boldsymbol{y}), \qquad \text{for all } k, \boldsymbol{y}, \tag{10}$$

$$s^k_{q,\tau}(\boldsymbol{y}) = d^k_{q,\tau} + g^{k,\pi^*}_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{y}), \qquad \text{for all } k, \boldsymbol{y}. \tag{11}$$

Constraints (7)–(11) reflect (i)–(v) respectively. Finding the optimum for (6) requires the computation of $c^k_{q,\tau}(\boldsymbol{d}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y})$ by considering every sample path of census configurations during a shift. For realistic instances this is computationally too expensive to find

the optimal solution for $d^1_{q,\tau},\ldots,d^K_{q,\tau},f_{q,\tau}$ in a reasonable amount of time (see Appendix B). Therefore, two approximations are proposed. The first approximation is obtained by deriving the probability distribution for the maximum number of patients present during each shift, and finding the optimal staffing for this maximum census. In this case the number of patients present is overestimated, therefore the required staffing levels are overestimated, and thus we obtain an upper bound on the staffing requirements. In the second approximation we reassign the float nurses to the care units at the start of each time interval. Since this provides more flexibility to align the float nurse allocation to the current census, we obtain an underestimation of the required staffing levels. As such, a lower bound on the actual staffing requirements is found. Finally, comparing the lower and upper bound solutions and the solution for the fixed model, provides us (an approximation of) the optimal solution of the flexible staffing model. To be more specific, the upper bound solution guarantees that the constraints are satisfied in the flexible staffing model. When the lower bound solution coincides with the upper bound or the fixed staffing solution, we are sure to have found the optimal solution. Otherwise the lower bound provides an error bound.

Upper bound model. Based on the observed maximum census configuration $\boldsymbol{x} = (x^1,\ldots,x^K)$ during a shift, let π^{up} be the assignment policy that allocates the nurses from the flex pool to the care units where the number of nurses short is the highest:

$$g^{\pi^{up}}_{q,\tau}(\boldsymbol{d}_{q,\tau},f_{q,\tau},\boldsymbol{x}) = \max_{\left\{f^1_{q,\tau},\ldots,f^K_{q,\tau}\,:\,\sum_k f^k_{q,\tau}=f_{q,\tau}\right\}} \min_k \frac{r^k_{q,\tau}\cdot(d^k_{q,\tau}+f^k_{q,\tau})-x^k}{r^k_{q,\tau}}.$$

Let $\hat{W}^k_{q,\tau}(x)$ be the probability that during shift (q,τ) the maximum census level that occurs at care unit k is x patients. These probabilities are derived by analogy with the derivation of $\hat{Z}^k_{q,\tau}(x)$ in [21] (for details see Appendix C). To obtain the upper bound, for $b_\tau \le t < b_\tau + \ell_\tau$, we approximate the original distribution $\hat{Z}^k_{q,t}(x)$ by $\hat{W}^k_{q,\tau}(x)$. Let $\bar{X}^k_{q,\tau}$ be the random variable with distribution $\hat{W}^k_{q,\tau}$ that counts the maximum number of patients on care unit k during shift (q,τ). To see that this approximation leads to an upper bound on the required staffing levels, observe that $\bar{X}^k_{q,\tau} \ge X^k_{q,t}$, for $b_\tau \le t < b_\tau + \ell_\tau$, so that for every time interval of a shift the census is overestimated, and thus staffing requirements are overestimated.

Since we use the same census distribution in every time interval during a shift, the coverage compliance over a shift $\bar{c}^k_{q,\tau}(\boldsymbol{d}_{q,\tau},f_{q,\tau},r^k_{q,\tau})$ is calculated by:

$$\bar{c}^k_{q,\tau}\left(\boldsymbol{d}_{q,\tau},f_{q,\tau},r^k_{q,\tau}\right) = \sum_{\boldsymbol{x}}\left\{\mathbb{1}\left(x^k \le r^k_{q,\tau}\cdot s^k_{q,\tau}(\boldsymbol{x})\right)\cdot\prod_{w=1}^{K}\hat{W}^w_{q,\tau}(x^w)\right\},$$

where $s^k_{q,\tau}(\boldsymbol{x})$ is the number of nurses staffed at care unit k for shift (q,τ) under assignment policy π^{up}, when the maximum observed census configuration is $\boldsymbol{x}$. Summarizing, for each shift (q,τ), we have:

$$\min \quad z_U = \omega_f f_{q,\tau} + \sum\nolimits_k \omega_d d^k_{q,\tau} \tag{12}$$

$$\text{s.t.} \quad d^k_{q,\tau} \ge S^k, \qquad \text{for all } k, \tag{13}$$

$$d^k_{q,\tau} \ge \left\lceil \beta^k \cdot M^k / r^k_{q,\tau} \right\rceil, \qquad \text{for all } k, \tag{14}$$

$$\overline{c}^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}) \geq \alpha^k, \qquad \text{for all } k, \tag{15}$$

$$d^k_{q,\tau} \geq \gamma^k \cdot s^k_{q,t}(\boldsymbol{x}), \qquad \text{for all } k, \boldsymbol{x}, \tag{16}$$

$$s^k_{q,\tau}(\boldsymbol{x}) = d^k_{q,\tau} + g^{k,\pi^{up}}_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{x}), \qquad \text{for all } k, \boldsymbol{x}. \tag{17}$$

The optimum of (12) is found by first finding the solution space for $d^k_{q,\tau}, k = 1, \ldots, K$, using constraints (13) and (14), and the optimal solution of the fixed staffing model, and, second, the solution space for $f_{q,\tau}$ using constraint (16). Next, complete enumeration over the obtained solution space is applied, which can be done quickly for realistically sized instances.

Lower bound model. For the lower bound model, we assume that we are allowed to reconsider the nurse-to-care-unit assignment at the start of every time interval. To observe that this relaxation leads to a lower bound on staffing requirements, note that with a given number of nurses, a higher coverage compliance can be achieved than in the original model. The assignment procedure π^{low} is executed at the start of each time interval, and the coverage compliance can thus be calculated per time interval. The coverage compliance over a shift $\underline{c}^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau})$ can then be calculated by:

$$\underline{c}^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}) = \frac{1}{\ell_\tau} \sum_{t=b_\tau}^{b_\tau+\ell_\tau-1} \sum_{\boldsymbol{x}} \left\{ \mathbb{1}\left(x^k \leq r^k_{q,\tau} \cdot s^k_{q,t}(\boldsymbol{x})\right) \cdot \prod_{w=1}^{K} \hat{Z}^w_{q,t}(x^w) \right\}.$$

where $s^k_{q,t}(\boldsymbol{x})$ is the number of nurses staffed at care unit k for time interval $[t, t+1)$ on day q under assignment policy π^{low}, when census configuration $\boldsymbol{x}$ is observed at time (q, t).

Since π^{low} is executed every time interval, it is based on the census configuration at the start of that time interval. A nurse from the flex pool gets staffed on the unit where the number of nurses short is the highest:

$$g^{\pi^{low}}_{q,t}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{x}) = \max_{\left\{f^1_{q,t}, \ldots, f^K_{q,t} : \sum_k f^k_{q,t} = f_{q,\tau}\right\}} \min_k \frac{r^k_{q,\tau} \cdot (d^k_{q,\tau} + f^k_{q,t}) - x^k}{r^k_{q,\tau}}.$$

As a result, for each shift (q, τ), we have:

$$\min\ z_L = \omega_f f_{q,\tau} + \sum\nolimits_k \omega_d d^k_{q,\tau} \tag{18}$$

$$\text{s.t. } d^k_{q,\tau} \geq S^k, \qquad \text{for all } k, \tag{19}$$

$$d^k_{q,\tau} \geq \left\lceil \beta^k \cdot M^k / r^k_{q,\tau} \right\rceil, \qquad \text{for all } k, \tag{20}$$

$$\underline{c}^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}) \geq \alpha^k, \qquad \text{for all } k, \tag{21}$$

$$d^k_{q,\tau} \geq \gamma^k \cdot s^k_{q,t}(\boldsymbol{x}), \qquad b_\tau \leq t < b_\tau + \ell_\tau, \text{ for all } k, \boldsymbol{x}, \tag{22}$$

$$s^k_{q,t}(\boldsymbol{x}) = d^k_{q,\tau} + g^{k,\pi^{low}}_{q,t}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{x}), \qquad b_\tau \leq t < b_\tau + \ell_\tau, \text{ for all } k, \boldsymbol{x}. \tag{23}$$

The optimum of (18) is found by first finding the solution space for $d^k_{q,\tau}, k = 1, \ldots, K$, using constraints (19) and (20), and the optimal solution of the fixed staffing model, and,

second, the solution space for $f_{q,\tau}$ using constraint (22). Next, complete enumeration over the obtained solution space is applied, which can be done quickly for realistically sized instances.

Flexible staffing levels. The upper and lower bound models were formulated to be able to find, or otherwise approximate, the optimal solution of the flexible staffing model. Here, we discuss how the solutions of the fixed model, and the upper and lower bound models, can be used to select the best staffing configuration. Two questions need to be answered: (1) did we find the optimal solution for the flexible staffing model, and, (2) which staffing configuration to select as the best solution?

Let us first discuss question (1). Observe that $z_L \leq z_U$ and $z_L \leq z_F$. When $z_L = z_U$ the upper and lower bound coincide so that the optimal solution is found. When $z_L < z_U$, but $z_L = z_F$, the optimal solution is also found, since in this case we are sure that flexible staffing cannot improve upon fixed staffing. In other cases, we are not sure whether or not the optimal solution is found; then, it is of interest to identify a bound on the distance between the optimal and the obtained solution.

The consideration involved when answering question (2) is to select the solution with the lowest optimal objective value, while it assures that the constraints (7)–(11) of the flexible staffing model are satisfied. For the solution of the lower bound model we are not sure whether or not constraints (7)–(11) are satisfied, therefore we never select this solution. In addition, when $z_F = z_U$, as tiebreaker, we choose the solution that achieves the highest minimum coverage compliance.

Let us denote with S_F, S_U, and S_L the optimal staffing configurations in the fixed, upper, and lower bound model respectively. We now provide an overview of the different cases:

(a) $z_L = z_U = z_F$. The optimal solution is found; if $\min_k \overline{c}^k_{q,\tau}(\cdot) \geq \min_k c^k_{q,\tau}(\cdot)$, S_U is selected as the best staffing configuration, otherwise S_F.

(b) $z_L = z_U < z_F$. The optimal solution is found; S_U is selected.

(c) $z_L = z_F < z_U$. The optimal solution is found; S_F is selected.

(d) $z_L < z_F = z_U$. Not sure whether or not the optimal solution is found; if $\min_k \overline{c}^k_{q,\tau}(\cdot) \geq \min_k c^k_{q,\tau}(\cdot)$, S_U is selected, otherwise S_F. The bound on the error margin is $z_U - z_L$.

(e) $z_L < z_U < z_F$. Not sure whether or not the optimal solution is found; S_U is selected; the error bound is $z_U - z_L$.

(f) $z_L < z_F < z_U$. Not sure whether or not the optimal solution is found; S_F is selected; the error bound is $z_F - z_L$.

4 Numerical results

This section presents the experimental results. The numerical results in this section are based on the case study as presented in [21]. Section 4.1 describes additional information on the case study with respect to staffing. Section 4.2 validates our approximation approach by investigating the distance between the upper and the lower bound solutions. Finally, Section 4.3 illustrates the practical potential of our methodology by returning to a selection of the interventions presented in [21] and formulating two additional interventions.

4.1 Case study description

The following specialties are taken into account: traumatology (TRA), orthopedics (ORT), plastic surgery (PLA), urology (URO), vascular surgery (VAS), and general surgery (GEN). In the present setting, the patients of the mentioned specialties are admitted to four different inpatient care departments. On floor I, care unit A houses GEN and URO, and unit B VAS and PLA. On floor II, care unit C houses TRA, and unit D ORT.

Working days are divided in three shifts: the day shift (8:00–15:00), the evening shift (15:00–23:00), and the night shift (23:00–8:00). These time intervals do indicate the times that nurses are responsible for direct patient care. Around these time intervals, the working times of the day and evening shift also incorporate time for patient handovers, indirect patient care, and professional development. At all times there should be at least two nurses present at each care unit. According to agreements on working conditions for nurses in all university hospitals in the Netherlands, the contractual number of annual working hours per full time equivalent (FTE) is 1872. The number of hours that one FTE can be employed for direct nursing care, after deduction of time reserved for professional development, holiday hours, and sick leave, is 1525.7 on average (also see [12]). The yearly cost per FTE including all costs and bonuses is roughly €50,000.

The nurse-to-patient ratio targets prescribed by the board of the AMC for the studied care units are 1:4 during the day shifts, 1:6 during the evening shifts, and 1:10 during the night shifts. The current staffing practice is based on the number of beds in service, independent of whether these are occupied or not, and no float nurse pools are employed. Thus, for example, for a care unit size of 24 beds and staffing ratio 1:4, the number of dedicated nurses to staff is always 6. Scarcity of nursing capacity frequently leads to expensive hiring of temporary nurses from external agencies, and to undesirable ad hoc bed closings. Also, the prescribed staffing levels cannot always be realized in practice. As a result, the inpatient care units experience a lack of consistency in the delivered quality of nursing care.

Table 1: Input parameter settings of the test instances for care units $k \in \{A, B, C, D\}$.

Parameter	*Description*	*Value*
Fixed		
Q	Planning horizon in days	365
T	Number of time intervals per day	24
$\lvert\mathcal{T}\rvert$	Number of shift types	3
(b_1, b_2, b_3)	Shift start times	$(8, 15, 23)$
(ℓ_1, ℓ_2, ℓ_3)	Shift durations	$(7, 8, 9)$
S^k	Minimum staffing levels	2
ω_d	Staffing cost dedicated nurse	1
Variable		
ω_f	Staffing cost float nurse	$\{1, 1.25, 1.5\}$
α^k	Minimum coverage compliance	$\{0.75, 0.80, 0.85, 0.90, 0.95\}$
β^k	Minimum coverage	$\{0.5, 0.6, 0.7, 0.8, 0.9\}$
γ^k	Minimum fraction of dedicated nurses	$\{0.5, 0.6, 0.7, 0.8, 0.9\}$
$(r^k_{q,1}, r^k_{q,2}, r^k_{q,3})$	Nurse-to-patient ratio targets	$\{(4, 6, 10), (4, 6, 8), (5, 5, 10)\}$

4.2 Quality of the bounds

To investigate the performance of the approximation approach for flexible staffing, we test the fixed, the upper, and the lower bound models on a variety of parameter settings. The bed census distributions as were obtained with the prediction model of [21] for the base case for the year 2010 are taken as input for the three staffing models. Based on the intention of the AMC, we assume that two float nurse pools are created: one serving care units A and B on floor I, and one serving care units C and D on floor II. During the planning horizon of a year, during which no cyclical Master Surgery Schedule (MSS) was used, we thus have to staff $365 \times 3 = 1095$ unique working shifts.

For our set of test instances, Table 1 provides an overview of the considered parameter settings. We vary over the (relative) staffing cost for float nurses, the coverage compliance threshold, the minimum coverage requirement, and the minimum dedicated nurse fraction. In addition, three different staffing ratio configurations are considered. We evaluate 2250 instances, together containing 2,463,750 working shifts to be staffed.

For each of the evaluated shifts, we recorded whether the optimum for the flexible staffing model was found or not. Table 2 displays the results. The overall result is that in 94.0% of the cases the optimum is found. In addition, the following effects can be observed. The optimum is found more often when flexible staffing is less attractive (which is reflected by increasing β^k and γ^k). Also, the minimum staffing levels $S^k = 2$ make that for night shifts the fixed and flexible solution generally coincide. Therefore, the optimum is almost always found for these shifts. For decreasing α^k the optimum is found more often, which may seem counterintuitive. However, for lower α^k the minimum coverage requirement given by β^k becomes decisive, which reduces the attractiveness of float nurses.

At the end of Section 3.2, we described how to find error bounds on the deviation from the optimal objective value in case one is not sure whether or not the optimum is found. Figure 1 zooms in on the 6.0% of shifts for which this holds; it shows a histogram of the deviations per shift of the obtained solution from the lower bound solution. The average maximum deviation for non-optimal shifts is 8.1%. It can be observed that on an individual shift level, the deviation can be substantial, because of the inherent integrality of the number of nurses that can be staffed. By displaying the error bound on the total staffing cost per instance, Figure 2 shows that the impact of these deviations on the overall performance is small: on average the obtained total staffing costs are within 0.6% of the optimum. We conclude that the approximative approach via bounds on the staffing levels, performs nearly optimal for our case study.

Table 2: The percentage of shifts for which the optimal solution is found (ceteris paribus).

Shift type (τ)		*Float nurse cost* (ω_f)		*Nurse-to-patient ratios* ($r^k_{q,\tau}$)	
day	87.3%	1.00	94.2%	4,6,8	93.8%
evening	94.9%	1.25	93.6%	4,6,10	93.9%
night	99.9%	1.50	94.3%	5,5,10	94.3%
Coverage compliance (α^k)		*Coverage minimum* (β^k)		*Flexibility ratio* (γ^k)	
0.75	96.4%	0.50	82.9%	0.50	91.0%
0.80	95.4%	0.60	89.2%	0.60	91.0%
0.85	94.2%	0.70	98.3%	0.70	91.4%
0.90	93.1%	0.80	99.6%	0.80	96.6%
0.95	90.9%	0.90	100.0%	0.90	100.0%

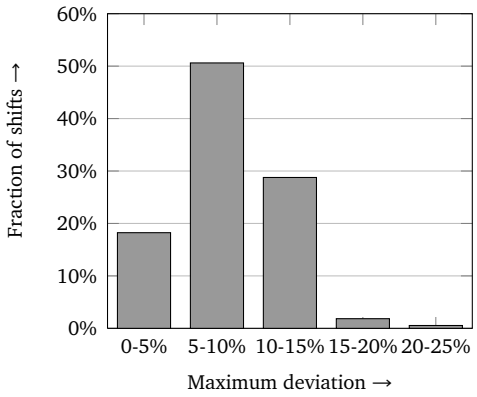

Figure 1: Distribution of the distance between the obtained solution and the lower bound solution (non-optimal shifts, $n = 147,426$).

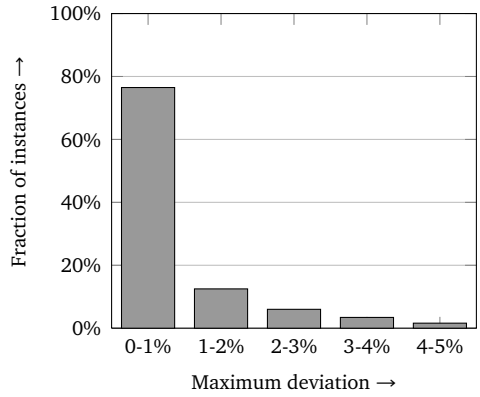

Figure 2: Distribution of the error bound on total staffing costs (all instances, $n = 2250$).

4.3 Case study results

To illustrate the potential of the presented staffing methodology for the case study, we return to a selection of the interventions that we presented in [21], which were formulated to improve the efficiency of the inpatient care service operations in terms of productivity of the inpatient beds.

We investigate both the value of aligning staffing levels with bed census predictions and of employing float nurses, by comparing the results of the fixed and flexible staffing models with the current staffing policy, which we refer to as 'full staffing'. With M^k the capacity of care unit k in the number of beds, under the full staffing policy always $\lceil M^k / r^k_{q,\tau} \rceil$ nurses are required.

The intended AMC practice will be that registered nurses will alternately be rostered as a dedicated or float nurse. Therefore, we consider the case in which dedicated and float staff members are equally expensive, i.e., $\omega_d = \omega_f$. In addition to the fixed input as displayed in Table 1, the board of the AMC has chosen to deploy the following quality of care requirements: nurse-to-patient ratios $r^k_{q,1} = 4$, $r^k_{q,2} = 6$, $r^k_{q,3} = 10$, minimum coverage $\beta^k = 0.70$, coverage compliance $\alpha^k = 0.90$, and at least two out of three nurses should be dedicated nurses, i.e., $\gamma^k = 0.67$. Compared to [21], we formulate two additional interventions, and we do not consider interventions 2 and 6. For a complete specification of

Table 3: The numerical results for the base case (Floor I: 56 beds, 56.7% utilization; Floor II: 48 beds, 58.6% utilization; with the FTE-Δ% relative to full staffing).

Intervention	*Floor*	*Full staffing*	*Fixed staffing*			*Flexible staffing*			
		FTE	Average	FTE		Error bound	Average	FTE (float)	
		(#)	coverage	(#)	(Δ%)	(%)	coverage	(#)	(Δ%)
Base case									
$\alpha = 0.85$	I	57.7	0.96	44.8	-22.2	0.4	0.96	44.7 (1.7)	-22.4
	II	48.3	0.96	38.9	-19.5	0.0	0.95	38.8 (2.0)	-19.7
$\alpha = 0.90$	I	57.7	0.98	46.0	-20.3	0.8	0.97	45.7 (2.7)	-20.8
	II	48.3	0.97	40.0	-17.3	0.1	0.97	39.6 (2.8)	-18.0
$\alpha = 0.95$	I	57.7	0.99	47.9	-16.9	1.4	0.99	47.4 (4.6)	-17.8
	II	48.3	0.99	42.5	-12.1	0.4	0.99	41.1 (4.3)	-14.9

the base case scenario, and intervention 1, 3, 4, and 5, we refer the reader to [21]. The detailed results are presented in Tables 3 and 4. Table 5 provides an overview of the results for the various interventions. It includes the calculation of the following productivity measure: the number of patients treated per employed FTE per year.

Base Case. First, we evaluate the performance of the base case scenario, the situation that most closely resembles current practice. The results are displayed in Table 3. In the flexible staffing policy, two flex pools are installed, one on each floor, we therefore present the results per floor. The number of FTEs required is calculated by adding up the total number of staffed nurse hours and dividing by the 1525.7 direct nursing hours that one FTE has available. For the base case we show three values for the coverage compliance threshold ($\alpha^k = \{0.85, 0.90, 0.95\}$), to illustrate the effect of this quality-of-care constraint on required nursing capacity.

For both the fixed and the flexible staffing model, it turns out that the realized coverage compliance is on average much higher than the minimum requirement. This is due to the fact that when the coverage compliance constraint is slightly violated, an additional nurse needs to be staffed, which significantly increases the coverage compliance since this nurse can care for $r^k_{q,\tau}$ patients. Although full staffing ensures a coverage compliance of 100%, it frequently overstaffs care units. It is clear that the acceptance of slight coverage reductions (still realizing average coverage compliances higher than 95%), allows managers to better match care supply and demand, thereby realizing efficiency gains of 12–22%. The largest gain is achieved by the staffing based on census predictions (see results fixed model). The additional value of employing float nurses is case dependent, and in most cases higher with increasing α^k, due to the increasing gap with the minimum coverage requirement set by β^k.

Interventions 1,3,4,5. Intervention 1 rationalized the care unit dimensions based on the requirement of rejection probabilities not exceeding 1%, 2.5%, and 5%. We focus on the outcomes for 2.5%; this is the threshold selected by the AMC to be implemented in practice. Table 4 shows that fixed staffing with $\alpha^k = 0.9$ reduces nursing capacity requirements by 8–9% compared to full staffing, and flexible staffing yields an additional 1% reduction. Table 5 indicates the gain against current practice: 22.6% reduction in FTE requirements, with a simultaneous increase of staff productivity by 26.5%.

Table 4: The numerical results for the various interventions (with the FTE-Δ% relative to full staffing).

Intervention	*Capacity* (# beds)	*Utilization* (%)	*Full staffing* FTE (#)	*Fixed staffing* Average coverage	FTE (#)	(Δ%)	*Flexible staffing* Average coverage	FTE (float) (#)	(Δ%)
1. Rationalize bed requirements									
Floor I	48	66.1	48.1	0.99	43.8	-8.9	0.98	43.3 (6.2)	-9.9
Floor II	40	70.1	42.6	0.99	39.3	-7.8	0.98	38.7 (5.2)	-9.1
3. Change operational process									
Floor I	45	63.4	48.1	0.98	41.8	-13.0	0.98	41.6 (4.4)	-13.5
Floor II	39	68.3	42.6	0.98	38.4	-9.9	0.98	37.2 (6.9)	-12.7
4. Balance MSS									
Floor I	46	71.3	48.1	0.99	45.7	-5.0	0.99	44.9 (7.8)	-6.7
Floor II	40	71.5	44.5	0.98	40.9	-8.2	0.98	39.6 (6.1)	-11.0
5. Combination (1), (3) and (4)									
Floor I	44	66.9	48.1	0.98	42.4	-11.7	0.98	41.8 (6.4)	-13.1
Floor II	39	69.5	42.6	0.98	38.8	-8.8	0.98	38.1 (4.6)	-10.6
7a. Combination (1) and centralized flex pool									
Floors I & II	88	67.9	90.7	0.99	83.1	-8.4	0.98	80.2 (9.5)	-11.5
7b. Combination (5) and centralized flex pool									
Floors I & II	83	68.1	90.7	0.98	81.3	-10.3	0.98	77.4 (8.6)	-14.6
8a. Combination (7a) and merge care units									
Floors I & II	88	67.9	84.9	0.97	74.7	-12.1	0.96	73.8 (9.7)	-13.1
8b. Combination (7b) and merge care units									
Floors I & II	83	68.1	83.3	0.97	72.0	-13.5	0.97	71.5 (9.6)	-14.1

Intervention 3 focused on changes in the operational process by: (a) decreasing lengths of stay by admitting all elective patients on the day of surgery, and (b) reducing afternoon census peaks by encouraging discharges to take place before noon. The reduction of demand and its variability lowered the number of beds required. Here we see that our staffing methodology also translates this into significantly lower staff requirements, and higher productivity.

Intervention 4 intended to decrease artificial demand variability by designing a cyclical Master Surgery Schedule (MSS) with the purpose to balance bed census. Recall that due to the integrality of the number of scheduled operating room blocks, the resulting MSS slightly increased patient demand. Therefore, its impact on staffing requirements is not directly visible. However, its impact is revealed by the outcomes on the fifth intervention (the combination between interventions 1, 3, and 4) which outperforms all previous configurations on the productivity measure. As an illustration, the effect of staffing levels following bed census demand patterns and the difference between fixed and flexible staffing therein are visualized in Figure 3.

Finally, let us state two general insights. First, note that under the old (full) staffing policy, a reduction in the number of beds not always translates into a reduction in staffing requirements. This is the case when the number of beds does not decrease to a capacity level such that it crosses a level that is a multiple of one of the nurse-to-patient ratios. Second, from our results we cannot deduce general rules-of-thumb for the potential of float nurses. The outcomes for each particular care unit are a

Table 5: FTE and productivity results for all interventions (with both the FTE-Δ% and the productivity-Δ% relative to full staffing in the base case).

Intervention	*Full staffing*				*Fixed staffing*				*Flexible staffing*			
	FTE		Productivity		FTE		Productivity		FTE		Productivity	
	(#)	(Δ%)	(#/yr)	(Δ%)	(#)	(Δ%)	(#/yr)	(Δ%)	(#)	(Δ%)	(#/yr)	(Δ%)
Base case	106.0	-	42.3	-	85.9	-18.9	52.2	+23.3	85.3	-19.5	52.6	+24.2
(1)	90.7	-14.4	48.5	+14.5	83.1	-21.6	52.9	+25.0	82.1	-22.6	53.5	+26.5
(3)	90.7	-14.4	48.4	+14.4	80.2	-24.3	54.7	+29.4	78.7	-25.7	55.8	+31.8
(4)	92.6	-12.6	48.6	+14.8	86.5	-18.4	52.0	+22.8	84.5	-20.3	53.2	+25.8
(5)	90.7	-14.4	49.6	+17.2	81.3	-23.3	55.3	+30.7	79.8	-24.7	56.3	+33.0
(7a)	90.7	-14.4	48.5	+14.5	83.1	-21.6	52.9	+25.0	80.2	-24.3	54.8	+29.5
(7b)	90.7	-14.4	49.6	+17.2	81.3	-23.3	55.3	+30.7	77.4	-27.0	58.1	+37.2
(8a)	84.9	-19.9	51.7	+22.3	74.7	-29.5	58.8	+39.0	73.8	-30.3	59.5	+40.7
(8b)	83.3	-21.4	54.0	+27.6	72.0	-32.0	62.4	+47.5	71.5	-32.5	62.8	+48.5

Productivity: number of patients treated per employed FTE per year

complex interplay between care unit sizes, nurse-to-patient ratios, and the shapes of the bed census distributions.

Interventions 7 and 8. The first additional intervention involves the merging of the two flex pools into one flex pool which serves all four care units. Intervention 7a evaluates the impact of this centralized flex pool for the situation of intervention 1, and intervention 7b for that of intervention 5. Naturally, for the full and fixed staffing policies the outcomes for intervention 7a and 7b coincide with 1 and 5 respectively, due to the unchanged care unit sizes and bed census distributions. With the flexible staffing policy, the additional flexibility of having four instead of two allocation options for each float nurse pays off: an additional saving of around 1.5–2.5 FTEs can be realized, in conjunction with an additional productivity increase of 3–4%.

Intervention 8 merges care units A and B, and care units C and D (intervention 8a for the situation of intervention 1, and intervention 8b for that of intervention 5). The two remaining care units, floor I and floor II, share one flex pool. This intervention is hard to implement, because it would imply the necessity of thorough renovation of the building. Although fictitious on the short-term, the positive outcomes for this intervention show that it is worthwhile considering. The economies-of-scale effect shows itself in various ways. First, larger care unit sizes reduce the occurrence of overstaffing due to staffing levels that have to be rounded upwards as a result of the nurse-to-patient ratios. Second, the relative variation in bed census decreases, making it easier to align staffing levels with patient demand, which is expressed by the results for the fixed staffing model. Third, in this case the minimum staffing levels of $S^k = 2$ per care unit only need to be satisfied for two care units, which often results in decreased staffing requirements during night shifts. Finally, it can be observed that the additional value of employing float nurses is lower for larger care unit sizes, again due to the decreasing relative census variation.

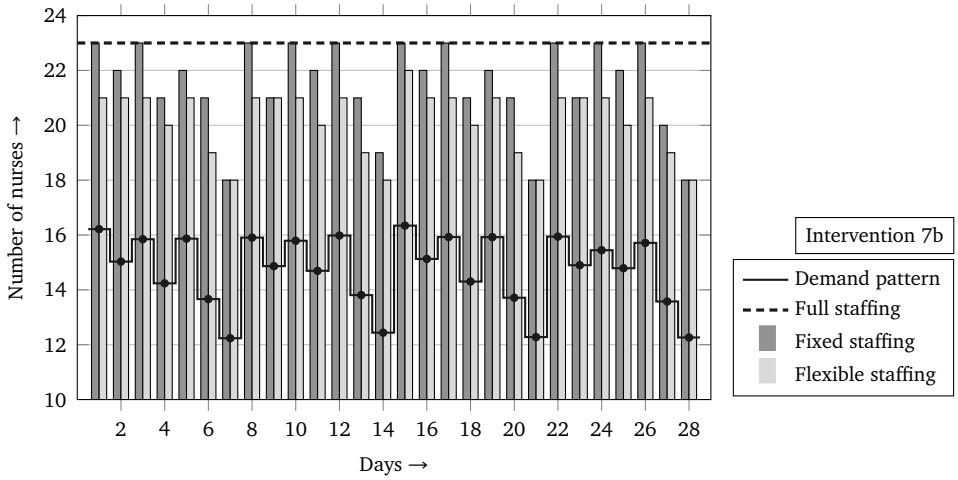

Figure 3: Staffing levels for day shifts on Floor I during the 4-week period starting on Monday January 25 (the demand pattern shows the average census divided by ratios $r^k_{q,1}$).

5 Discussion

Rising healthcare costs and increasing nurse shortages make cost-effective nurse staffing of utmost importance. In many hospitals, staffing levels are a result of historical development, as hospital managers lack the tools to base staffing decisions on information about future patient demand. Since patient safety is jeopardized when medical care units are understaffed, scarcity of nursing capacity can lead to expensive hiring of nurses from external agencies and to undesirable ad hoc bed closings. In this paper, we have presented a generic analytical method that can quantitatively support decision making on required staffing levels in inpatient care facilities. We have demonstrated its potential with a case study of the AMC, for which we have shown that by achieving coherence between patient demand and staffing supply simultaneous cost reductions and quality of care improvements are possible.

The combined application of the bed census prediction model from [21] and the staffing models from the current paper enables hospital administrators to gain insight into the value of integrated decision making. The interrelation between decisions such as case mix, care unit partitioning, care unit size, and admission/discharge times is made explicit. Because the demand prediction model incorporates the operating room block schedule and the patient arrival pattern from the emergency department, the presented methodology also facilitates alignment between the design and operations of the inpatient care facility and its surrounding departments. With this integrated framework, staffing effectiveness can be attained in three steps. First, the method can help to reduce artificial variability of bed occupancies, for example by adjusting the operating room schedule. Second, by predicting the bed census distributions and determining staffing levels for dedicated nurses accordingly, the predictive part of the remaining variability can be anticipated. Third, to be able to effectively respond to random variability, adequately sized float nurse pools can be created.

Staffing requirements are the result of a complex interaction between care unit sizes, nurse-to-patient ratios, the bed census distributions, and the quality-of-care requirements.

The optimal configuration strongly depends on the particular characteristics of a specific case under study. Nonetheless, several general insights have been obtained. When working with nurse-patient-ratios, care units should be sufficiently large, to avoid efficiency losses due to the lack of granularity in the values of the ratios. Next, under the premise that the costs per float nurse remain unchanged, the more care units float nurse pools can serve, the more effective they are. Finally, also when it does not reduce capacity requirements, flexible staffing staffing is beneficial since it enhances the adherence to the nurse-to-patient ratio targets.

The case study of the AMC provides an example of how the methodology can be applied in practice. Due to both economic and medical developments, the AMC is forced to reorganize the operations of the inpatient services during the upcoming years. Nurse staffing is high on the agenda, since the AMC has 30 inpatient departments, staffing costs account for 66% of the total expenses in the AMC, and one full-time registered nurse yearly costs around €50,000. We have applied our staffing models on data of several care units; for four of them we presented the results in this paper. The formulations of all interventions and the eventual parameter settings are the results of close cooperation between operations researchers and hospital managers from different levels within the organization. It has resulted in the joint conclusion that efficiency gains are possible, while improving upon the adherence to nurse-to-patient ratio targets. As a result, the AMC decided that the flexible nurse staffing method will be fully implemented during the upcoming years.

The development of a user-friendly decision support system (DSS) based on our method will be a next step in achieving practical impact. Our model relies on data that is easily extractable from typical hospital management systems. This makes it possible to automate the process of collecting the required input parameters to run the model. Integration with the hospital management system, visualization of the results, and the possibility to run what-if scenarios will be desired specifications of the DSS. We believe that the adoption of such a system by healthcare administrators of inpatient care services can result in more cost-effective resource capacity planning and control decisions.

Appendix

A Detailed summary bed census prediction model

This appendix provides a summary of the hourly bed census prediction model of [21].

Demand predictions for elective patients

Model input. The demand predictions for elective patients will be based on the following input parameters.

Time. An MSS is a repeating blueprint for the surgical schedule of S days. Each day is divided in T time intervals. Therefore, we have time points $t = 0, \ldots, T$, in which $t = T$ corresponds to $t = 0$ of the next day. For each single patient, day n counts the number of days before or after surgery, i.e., $n = 0$ indicates the day of surgery.

MSS utilization. For each day $s \in \{1, \ldots, S\}$, a (sub)specialty j can be assigned to an available operating room i, $i \in \{1, \ldots, I\}$. The OR block at operating room i on day s is denoted by $b_{i,s}$, and is possibly divided in a morning block $b^m_{i,s}$ and an afternoon block $b^a_{i,s}$, if an OR day is shared. The discrete distributions c^j represent how specialty j utilizes an OR block, i.e., $c^j(k)$ is the probability of k surgeries performed in one block, $k \in \{0, 1, ..., C^j\}$. If an OR block is divided in a morning OR block and an afternoon OR block, c^j_M and c^j_A represent the utilization probability distributions respectively. Such shared OR blocks are not explicitly included in our formulation, since these can be modeled as two separate (fictitious) operating rooms.

Admissions. With probability e^j_n, $n \in \{-1, 0\}$, a patient of type j is admitted on day n. Given that a patient is admitted on day n, the time of admission is described by the probability distribution $w^j_{n,t}$. We assume that a patient who is admitted on the day of surgery is always admitted before or at time ϑ_j; therefore, we have $w^j_{0,t} = 0$ for $t = \vartheta_j + 1, \ldots, T - 1$.

Discharges. $P^j(n)$ is the probability that a type j patient stays n days after surgery, $n \in \{0, \ldots, L^j\}$. Given that a patient is discharged on day n, the probability of being discharged in time interval $[t, t + 1)$ is given by $m^j_{n,t}$. We assume that a patient who is discharged on the day of surgery is discharged after time ϑ_j, i.e., $m^j_{0,t} = 0$ for $t = 0, \ldots, \vartheta_j$.

Single surgery block. In this first step we consider a single specialty j operating in a single OR block. We compute the probability $h^j_{n,t}(x)$ that n days after carrying out a block of specialty j, at time t, x patients of the block are still in recovery. Note that admissions can take place during day $n = -1$ and during day $n = 0$ until time $t = \vartheta_j$. Discharges can take place during day $n = 0$ from time $t = \vartheta_j + 1$ and during days $n = 1, \ldots, L^j$. Therefore, we calculate $h^j_{n,t}(x)$ as follows:

$$h^j_{n,t}(x) = \begin{cases} a^j_{n,t}(x) & \text{if } n = -1 \text{ and } n = 0, t \leq \vartheta_j, \\ d^j_{n,t}(x) & \text{if } n = 0, t > \vartheta_j \text{ and } n = 1, \ldots, L^j, \end{cases}$$

where $a^j_{n,t}(x)$ represents the probability that x patients are admitted until time t on day n, and $d^j_{n,t}(x)$ is the probability that x patients are still in recovery at time t on day n.

Single MSS cycle. Now, we consider a single MSS in isolation. From the distributions $h^j_{n,t}$, we can determine the distributions $H_{m,t}$, the discrete distributions for the total number of recovering patients at time t on day m, $m \in \{0,1,2,\ldots,S,S+1,S+2,\ldots\}$, resulting from a single MSS cycle.

Steady state. In this step, the complete impact of the repeating MSS is considered. The distributions $H_{m,t}$ are used to determine the distributions $H^{SS}_{s,t}$, the steady state probability distributions of the number of recovering patients at time t on day s of the cycle, $s \in \{1,\ldots,S\}$.

Demand predictions for acute patients

Model input. The demand predictions for acute patients will be based on the following input parameters.

Time. The AAC is the repeating cyclic arrival pattern of acute patients with a length of R days. For each single patient, day n counts the number of days after arrival.

Admissions. An acute patient type is characterized by patient group p, $p = 1,\ldots,P$, arrival day r and arrival time θ, which is for notational convenience denoted by type $j = (p,r,\theta)$. The Poisson arrival process of patient type j has arrival rate λ^j.

Discharges. $P^j(n)$ denotes the probability that a type j patient stays n days, $n \in \{0,\ldots,L^j\}$. Given that a patient is discharged at day n, the probability of being discharged in time interval $[t,t+1)$ is given by $\tilde{m}^j_{n,t}$. By definition, $\tilde{m}^j_{0,t} = 0$ for $t \leq \theta$.

Single patient type. In this first step we consider a single patient type j. We compute the probability $g^j_{n,t}(x)$ that on day n at time t, x patients are still in recovery. Admissions can take place during time interval $[\theta,\theta+1)$ on day $n = 0$ and discharges during day $n = 0$ after time θ and during days $n = 1,\ldots,L^j$. Therefore, we calculate $g^j_{n,t}(x)$ as follows:

$$g^j_{n,t}(x) = \begin{cases} \tilde{a}^j_t(x) & \text{if } n = 0, t = \theta, \\ \tilde{d}^j_{n,t}(x) & \text{if } n = 0, t > \theta \text{ and } n = 1,\ldots,L^j, \end{cases}$$

where $\tilde{a}^j_t(x)$ represents the probability that x patients are admitted in time interval $[t,t+1)$ on day $n = 0$, and $\tilde{d}^j_{n,t}(x)$ is the probability that x patients are still in recovery at time t on day n.

Single cycle. Now, we consider a single AAC in isolation. From the distributions $g^j_{n,t}(x)$, we can determine the distributions $G_{w,t}$, the distributions for the total number of recovering patients at time t on day w, $w \in \{1,\ldots,R,R+1,R+2,\ldots\}$, resulting from a single AAC.

Steady state. In this step, the complete impact of the repeating AAC is considered. The distributions $G_{w,t}$ are used to determine the distributions $G^{SS}_{r,t}$, the steady state probability distributions of the number of recovering patients at time t on day r of the cycle, $r \in \{1,\ldots,R\}$.

Demand predictions per care unit

To determine the complete demand distribution of both elective and acute patients, we need to combine the steady state distributions $H^{SS}_{s,t}$ and $G^{SS}_{r,t}$. In general, the MSS cycle

and AAC are not equal in length, i.e., $S \neq R$. This has to be taken into account when combining the two steady state distributions. Therefore, we define the new IFC length $Q = LCM(S,R)$, where the function LCM stands for *least common multiple*. Let $Z_{q,t}$ be the probability distribution of the total number of patients recovering at time t on day q during a time cycle of length Q:

$$Z_{q,t} = H^{SS}_{q \bmod S+S\cdot\mathbb{1}_{(q \bmod S=0)},t} \otimes G^{SS}_{q \bmod R+R\cdot\mathbb{1}_{(q \bmod R=0)},t},$$

where $\otimes$ denotes the discrete convolution function. Let U^k be the set of specialties j whose operated patients are (preferably) admitted to unit k, $k \in \{1,\ldots,K\}$, and V^k the set of acute patient types j that are (preferably) admitted to unit k. Then, the demand distribution for unit k, $Z^k_{q,t}$, can be calculated by exclusively considering the patients in U^k and V^k.

Bed census predictions

We translate the demand distributions $Z^k_{q,t}$ into bed census distributions $\hat{Z}_{q,t}$, $k = 1,\ldots,K$, the distributions of the number of patients present in each unit k at time t on day q. To this end, we require an allocation policy ϕ that uniquely specifies from a demand vector $\mathbf{x} = (x_1,\ldots,x_K)$ a bed census vector $\hat{\mathbf{x}} = (\hat{x}_1,\ldots,\hat{x}_K)$, in which x_k and $\hat{x}_k$ denote the demand for unit k and the bed census at unit k, respectively. Let $\phi(\cdot)$ be the function that executes allocation policy ϕ. Let $\hat{Z}^k_{q,t}$ denote the marginal distribution of the census at unit k given by distribution $\hat{Z}_{q,t}$. With M^k the capacity of unit k in number of beds, we obtain

$$\hat{Z}_{q,t}(\hat{\mathbf{x}}) = \big(\hat{Z}^1_{q,t}(\hat{x}_1),\ldots,\hat{Z}^K_{q,t}(\hat{x}_K)\big) = \sum_{\{\mathbf{x}|\hat{\mathbf{x}}=\phi(\mathbf{x})\}} \Big\{\prod_{k=1}^{K} Z^k_{q,t}(x_k)\Big\}. \tag{24}$$

We do not impose restrictions on the allocation policy ϕ other than specifying a unique relation between demand $\mathbf{x}$ and census configuration $\hat{\mathbf{x}}$. Recall that the underlying assumption is that a patient is transferred to his preferred unit when a bed becomes available. The policy ϕ also reflects the priority rules that are applied for such transfers. As an illustration, we present an example for an inpatient care facility with two care units of capacity M^1 and M^2 respectively:

$$\phi(\mathbf{x}) = \begin{cases} (x_1, x_2) & \text{if } x_1 \leq M_1, x_2 \leq M_2, \\ (M_1, \min\{x_2 + (x_1 - M_1), M_2\}) & \text{if } x_1 > M_1, x_2 \leq M_2, \\ (\min\{x_1 + (x_2 - M_2), M_1\}, M_2) & \text{if } x_1 \leq M_1, x_2 > M_2, \\ (M_1, M_2) & \text{if } x_1 > M_1, x_2 > M_2. \end{cases} \tag{25}$$

Under this policy patients are assigned to their bed of preference if available, and are otherwise misplaced to the other unit if beds are available there.

B Complexity of the flexible staffing model

This appendix investigates the complexity of the calculations involved in solving the flexible staffing model, formulated by equations (6)–(11). The complexity is such that the computation time inhibits the evaluation of realistically sized instances. This is mainly due

to the large number of census configurations that has to be evaluated to identify the float nurse assignment procedure π^* satisfying the maximization (5). This assignment procedure is involved in constraint (11).

Consider shift (q,τ). Let us investigate the complexity of determining π^* for a given availability of dedicated and float nurses, i.e, for given $d^1_{q,\tau},\ldots,d^K_{q,\tau},f_{q,t}$. For every census configuration $\boldsymbol{y}$ that can possibly be observed at the start of the shift, the assignment (5), to be used in (11), needs to be found. This is of order:

$$O(N_{\boldsymbol{y}}\cdot N_f\cdot N_c\cdot K),$$

where $N_{\boldsymbol{y}}$ denotes the maximum number of possible census configurations at the start of the shift, N_f the number of possible allocations of the $f_{q,\tau}$ available float nurses, and N_c the complexity of the calculations involved in evaluating the coverage compliance $c^k_{q,\tau}(\boldsymbol{d}_{q,\tau},f_{q,\tau},r^k_{q,\tau};\boldsymbol{y})$, which has to be done for all K wards.

Since the census range for ward k is $\{0,\ldots,M^k\}$, with $\hat{M}=\max_k M^k$, we have:

$$N_{\boldsymbol{y}}=(\hat{M}+1)^K.$$

Second, counting the number of possible allocations of $f_{q,\tau}$ nurses over K wards, we have:

$$N_f=\binom{f_{q,\tau}+K-1}{K-1}$$

This leaves us to determine N_c. To this end, we make use of the concept patient *cohort* (as also introduced in [21]): a cohort is a group of patients originating from a single instance of an OR block (electives) or admission time interval (acute patients). As specified in [21], all patients of one cohort are preferably placed on the same care unit. The best coverage is realized when for each patient cohort at the start of the shift it is observed how many patients are present, since in that manner the maximum amount of information on possible admissions and discharges is taken into consideration. Let Φ denote the total number of patient cohorts present during shift (q,τ), and $\mathcal{W}^k$ the set of patient cohorts admitted to ward k. For notational convenience we introduce the function $v^i_{q,t}$ as $v^i_{q,t}=h^i_{q,t}$ for the elective patients, and $v^i_{q,t}=g^i_{q,t}$ for acute patient types. In addition, for each patient cohort, we define for $b_\tau\le t<b_\tau+\ell_\tau$ the conditional distribution $v^{i,z_i}_{q,t}$, with $v^{i,z_i}_{q,t}(x_i)$ the probability that x_i patients of cohort i are present at the start of time interval (q,t), given that at the start of shift (q,τ) the number of patients present of this cohort was z_i. Then, the coverage compliance given that census configuration $\boldsymbol{y}$ is observed at the start of shift (q,τ) is:

$$\begin{aligned}c^k_{q,\tau}(\boldsymbol{d}_{q,\tau},f_{q,\tau},r^k_{q,\tau};\boldsymbol{y})=&\\ \sum_{\substack{z_1,\ldots,z_\Phi:\\ \sum_{i\in\mathcal{W}^k}z_i=\boldsymbol{y}^k,\\ k=1,\ldots,K}}&\left\{\prod_{i=1}^{\Phi}v^i_{q,b_\tau}(z_i)\cdot\right.\\ &\left.\frac{1}{\ell_\tau}\sum_{t=b_\tau}^{b_\tau+\ell_\tau+1}\sum_{x_i:\forall i\in\mathcal{W}^{\|}}\mathbb{1}\Big(\sum_{i\in\mathcal{W}^k}x_i\le r^k_{q,\tau}\cdot s^k_{q,\tau}(\boldsymbol{y})\Big)\Big\{\prod_i v^{i,z_i}_{q,t}(x_i)\Big\}\right\}.\end{aligned}$$

The first summation involves maximally $(\hat{M}+1)^\Phi$ combinations, the second summation ℓ_τ combinations, and the third $\hat{M}+1$. Therefore, we have

$$N_c=(\hat{M}+1)^{\Phi+1}\cdot\ell_\tau.$$

To conclude, the complexity of determining π^* for given $d^1_{q,\tau},\ldots,d^K_{q,\tau},f_{q,t}$ is of the order:

$$O(N_{\mathbf{y}}\cdot N_f\cdot N_c\cdot K)=O\left(\left(\hat{M}+1\right)^{K+\Phi+1}\cdot\binom{f_{q,\tau}+K-1}{K-1}\cdot\ell_\tau\right),$$

which for real-world instances is both in terms of memory and computation time too large to find the optimal $d^1_{q,\tau},\ldots,d^K_{q,\tau},f_{q,t}$.

C Derivation maximum census

In this appendix, $\hat{W}^k_{q,\tau}$ is derived, the probability distribution of the maximum census at care unit k during shift (q,τ). For each patient cohort and each shift (q,τ), we need to determine at which of the time points $t\in\{(q,b_\tau),\ldots,(q,b_\tau+\ell_\tau-1)\}$ the number of patients of this cohort reaches its maximum.

We first determine for each cohort i, the probability distribution $w^i_{q,\tau}$ for the maximum number of patients of this cohort present during shift (q,τ). Since all patients of one cohort are preferably placed on the same care unit, to obtain the probability distribution $W^k_{q,\tau}$ for the maximum demand for unit k during shift (q,τ), we take the discrete convolution over the distributions $w^i_{q,\tau}$ relevant to unit k. Finally, from the maximum demand distribution $W^k_{q,\tau}$, the maximum census distribution $\hat{W}^k_{q,\tau}$ is obtained by applying the same transformation as was done for $Z^k_{q,\tau}$ and $\hat{Z}^k_{q,\tau}$ in equation (24).

Elective patients. For each combination of a day q in the Inpatient Facility Cycle (IFC), and a number of days after surgery n, there is a unique corresponding day in the Master Surgery Schedule (MSS). We denote this day by $\Delta^{MSS}(q,n)$:

$$\Delta^{MSS}(q,n)=\begin{cases}(q-n)\bmod S+\mathbb{1}_{\left((q-n)\bmod S=0\right)}\cdot S & \text{if } -1\le n<q,\\ (q-n)+\left[((n-q)\operatorname{div}S)+1\right]\cdot S & \text{if } q\le n\le L^i.\end{cases}$$

Also, note that by definition of the cohorts, the combination of day q and cohort i uniquely defines the number of days the patients of this cohort are already present after surgery; let us denote this value by $N(i,q)$. For elective patients, $w^i_{q,\tau}$ can be calculated as follows. For all i such that $\exists i$ such that $i\in b_{i,\Delta^{MSS}(q,N(i,q))}$:

$$w^i_{q,\tau}=\begin{cases}h^i_{N(i,q),b_\tau} & \text{if } N(i,q)=1,\ldots,L^i,\\ h^i_{0,b_\tau} & \text{if } N(i,q)=0,\vartheta_i<b_\tau,\\ h^i_{0,\vartheta_i} & \text{if } N(i,q)=0,b_\tau\le\vartheta_i<b_\tau+\ell_\tau,\\ h^i_{0,b_\tau+\ell_\tau-1} & \text{if } N(i,q)=0,\vartheta_i\ge b_\tau+\ell_\tau,\\ h^i_{-1,b_\tau+\ell_\tau-1} & \text{if } N(i,q)=-1,b_\tau+\ell_\tau\le T,\\ h^i_{-1,T+\vartheta_i} & \text{if } N(i,q)=-1,b_\tau+\ell_\tau>T,\vartheta_i<b_\tau+\ell_\tau-T,\\ h^i_{-1,b_\tau+\ell_\tau-1} & \text{if } N(i,q)=-1,b_\tau+\ell_\tau>T,\vartheta_i\ge b_\tau+\ell_\tau-T.\end{cases}$$

Acute patients. Let $\Delta^{AAC}(q,n)$ be the admission day in the Acute Admission Cycle (AAC) of an acute patient type present on a given day q in the IFC, and which is at its n-th day after admission:

$$\Delta^{AAC}(q,n)=\begin{cases}(q-n)\bmod R+\mathbb{1}_{\left((q-n)\bmod R=0\right)}\cdot R & \text{if } 0\le n<q,\\ (q-n)+\left[((n-q)\operatorname{div}R)+1\right]\cdot R & \text{if } q\le n\le L^i.\end{cases}$$

Recall that an acute patient type is identified by (p, r, θ). Observe that an acute patient cohort i is specified by the combination of a patient type j and a specific admission day. Also for acute patients, the combination of day q and cohort i uniquely defines the number of days the patients of this cohort are already present; let us denote this value by $M(i,q)$. During shift (q,τ), for an acute patient cohort the maximum demand is obtained at its admission time interval if this lies within (q,τ), otherwise it is obtained at the start of the shift. Hence, for acute patients $w^i_{q,\tau}$ is calculated by:

$$w^i_{q,\tau} = \begin{cases} g^i_{M(i,q),b_\tau} & \text{if } M(i,q) = 1,\dots,L^i, i \text{ such that } \Delta^{AAC}(q, M(i,q)) = r, \\ g^i_{0,b_\tau} & \text{if } M(i,q) = 0, \theta < b_\tau, i \text{ such that } \Delta^{AAC}(q, M(i,q)) = r, \\ g^i_{0,\theta} & \text{if } M(i,q) = 0, b_\tau \le \theta < b_\tau + \ell_\tau, i \text{ such that} \\ & \Delta^{AAC}(q, M(i,q)) = r, \\ g^i_{0,\theta} & \text{if } M(i,q) = 0, b_\tau + \ell_\tau > T, \theta < b_\tau + \ell_\tau - T, i \text{ such that} \\ & \Delta^{AAC}\left((q+1) \bmod Q + Q \cdot \mathbb{1}_{((q+1) \bmod Q = 0)}, M(i,q)\right) = r. \end{cases}$$

Finally $W^k_{q,\tau}$, $k = 1,\dots,K$, is obtained by taking the discrete convolution over the distributions $w^i_{q,\tau}$ relevant to unit k, and $\hat{W}^k_{q,\tau}$, $k = 1,\dots,K$, is obtained from $W^k_{q,\tau}$, by applying the transformation as presented in equation (24).

Acknowledgement

This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs.

References

[1] L.H. Aiken, S.P. Clarke, D.M. Sloane, J. Sochalski, and J.H. Silber. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of the American Medical Association*, 288(16):1987–1993, 2002.

[2] L.H. Aiken, W. Sermeus, K. van den Heede, D.M. Sloane, R. Busse, M. McKee, L. Bruyneel, A.M. Rafferty, P. Griffiths, M.T. Moreno-Casbas, C. Tishelman, A. Scott, T. Brzostek, J. Kinnunen, R. Schwendimann, M. Heinen, D. Zikos, I. Strømseng Sjetne, H.L. Smith, and A. Kutney-Lee. Patient safety, satisfaction, and quality of hospital care: cross sectional surveys of nurses and patients in 12 countries in Europe and the United States. *British Medical Journal*, 344(3):1–14, 2012.

[3] L.H. Aiken, D.M. Sloane, J.P. Cimiotti, S.P. Clarke, L. Flynn, J.A. Seago, J. Spetz, and H.L. Smith. Implications of the California nurse staffing mandate for other states. *Health Services Research*, 45(4):904–921, 2010.

[4] J. Beliën and E. Demeulemeester. A branch-and-price approach for integrating nurse and surgery scheduling. *European Journal of Operational Research*, 189(3):652–668, 2008.

[5] M.L. Brandeau, F. Sainfort, and W.P. Pierskalla, editors. *Operations research and health care: a handbook of methods and applications*, volume 70 of *International Se-*

ries in Operations Research & Management Science. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[6] J.R. Broyles, J.K. Cochran, and D.C. Montgomery. A statistical Markov chain approximation of transient hospital inpatient inventory. *European Journal of Operational Research*, 207(3):1645–1657, 2010.

[7] E.K. Burke, P. de Causmaecker, G.V. Berghe, and H. van Landeghem. The state of the art of nurse rostering. *Journal of Scheduling*, 7(6):441–499, 2004.

[8] B. Cheang, H. Li, A. Lim, and B. Rodrigues. Nurse rostering problems–a bibliographic survey. *European Journal of Operational Research*, 151(3):447–460, 2003.

[9] P. de Causmaecker and G. vanden Berghe. A categorisation of nurse rostering problems. *Journal of Scheduling*, 14(1):3–16, 2011.

[10] F. de Véricourt and O.B. Jennings. Nurse staffing in medical units: a queueing perspective. *Operations Research*, 59(6):1320–1331, 2011.

[11] N. Dellaert, J. Jeunet, and G. Mincsovics. Budget allocation for permanent and contingent capacity under stochastic demand. *International Journal of Production Economics*, 131(1):128–138, 2011.

[12] S.G. Elkhuizen, G. Bor, M. Smeenk, N.S. Klazinga, and P.J.M. Bakker. Capacity management of nursing staff as a vehicle for organizational improvement. *BMC health services research*, 7(1):196–205, 2007.

[13] A.T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier. Staff scheduling and rostering: a review of applications, methods and models. *European Journal of Operational Research*, 153(1):3–27, 2004.

[14] A. Gnanlet and W.G. Gilland. Sequential and simultaneous decision making for optimizing health care resource flexibilities. *Decision Sciences*, 40(2):295–326, 2009.

[15] J.D. Griffiths, N. Price-Lloyd, M. Smithies, and J.E. Williams. Modelling the requirement for supplementary nurses in an intensive care unit. *Journal of the Operational Research Society*, 56(2):126–133, 2005.

[16] P.R. Harper, N.H. Powell, and J.E. Williams. Modelling the size and skill-mix of hospital nursing teams. *Journal of the Operational Research Society*, 61(5):768–779, 2010.

[17] P.J.H. Hulshof, N. Kortbeek, R.J. Boucherie, E.W. Hans, and P.J.M. Bakker. Taxonomic classification of planning decisions in health care: a review of the state of the art in OR/MS. *Health Systems*, 1(2):1–47, 2012.

[18] E.P. Jack and T.L. Powers. Volume flexible strategies in health services: a research framework. *Production and Operations Management*, 13(3):230–244, 2009.

[19] R.L. Kane, T.A. Shamliyan, C. Mueller, S. Duval, and T.J. Wilt. The association of registered nurse staffing levels and patient outcomes: systematic review and meta-analysis. *Medical Care*, 45(12):1195–1204, 2007.

[20] D.L. Kellogg and S. Walczak. Nurse scheduling: from academia to implementation or not? *Interfaces*, 37(4):355, 2007.

[21] N. Kortbeek, A. Braaksma, H.F. Smeenk, P.J.M. Bakker, and R.J. Boucherie. Integral resource capacity planning for inpatient care services based on hourly bed census

predictions. Memorandum 1990, Department of Applied Mathematics, University of Twente, Enschede, The Netherlands, 2012.

[22] T.A. Lang, M. Hodge, V. Olson, P.S. Romano, and R.L. Kravitz. Nurse-patient ratios: a systematic review on the effects of nurse staffing on patient, nurse employee, and hospital outcomes. *Journal of Nursing Administration*, 34(7-8):326–337, 2004.

[23] M.S. Lavieri and M.L. Puterman. Optimizing nursing human resource planning in British Columbia. *Health Care Management Science*, 12(2):119–128, 2009.

[24] L.L.X. Li and B.E. King. A healthcare staff decision model considering the effects of staff cross-training. *Health Care Management Science*, 2(1):53–61, 1999.

[25] J. Needleman, P. Buerhaus, S. Mattke, M. Stewart, and K. Zelevinsky. Nurse-staffing levels and the quality of care in hospitals. *New England Journal of Medicine*, 346(22):1715–1722, 2002.

[26] J.P. Oddoye, D.F. Jones, M. Tamiz, and P. Schmidt. Combining simulation and goal programming for healthcare planning in a medical assessment unit. *European Journal of Operational Research*, 193(1):250–261, 2009.

[27] J.P. Oddoye, M.A. Yaghoobi, M. Tamiz, D.F. Jones, and P. Schmidt. A multi-objective model to determine efficient resource levels in a medical assessment unit. *Journal of the Operational Research Society*, 58(12):1563–1573, 2007.

[28] W.P. Pierskalla and D.J. Brailer. Applications of operations research in health care delivery. In S.M. Pollock, M.H. Rothkopf, and A. Barnett, editors, *Operations research and the public sector*, volume 6 of *Handbooks in OR & MS*, pages 469–505. North-Holland, Amsterdam, The Netherlands, 1994.

[29] S.P. Siferd and W.C. Benton. Workforce staffing and scheduling: hospital nursing specific models. *European Journal of Operational Research*, 60(3):233–246, 1992.

[30] V.L. Smith-Daniels, S.B. Schweikhart, and D.E. Smith-Daniels. Capacity management in health care services: review and future research directions. *Decision Sciences*, 19(4):889–919, 1988.

[31] V.M. Trivedi and D.M. Warner. A branch and bound algorithm for optimum allocation of float nurses. *Management Science*, 22(9):972–981, 1976.

[32] D. Twigg, C. Duffield, A. Bremner, P. Rapley, and J. Finn. The impact of the nursing hours per patient day (NHPPD) staffing method on patient outcomes: a retrospective analysis of patient and staffing data. *International Journal of Nursing Studies*, 48(5):540–548, 2011.

[33] J.M. van Oostrum, M. van Houdenhoven, J.L. Hurink, E.W. Hans, G. Wullink, and G. Kazemier. A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR Spectrum*, 30(2):355–374, 2008.

[34] L.M. Walts and A.S. Kapadia. Patient classification system: an optimization approach. *Health Care Management Review*, 21(4):75, 1996.

[35] P.D. Wright, K.M. Bretthauer, and M.J. Côté. Reexamining the Nurse Scheduling Problem: Staffing Ratios and Nursing Shortages. *Decision Sciences*, 37(1):39–70, 2006.

[36] N. Yankovic and L.V. Green. Identifying good nursing levels: a queuing approach. *Operations research*, 59(4):942–955, 2011.