



Flexible nurse staffing based on hourly bed census predictions



N. Kortbeek^{a,b,c,*}, A. Braaksma^{a,b,c}, C.A.J. Burger^{a,c}, P.J.M. Bakker^b, R.J. Boucherie^{a,c}

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

^b Department of Quality Assurance and Process Innovation, Academic Medical Center Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

^c Stochastic Operations Research, Department of Applied Mathematics, University of Twente, Drienerlolaan 5, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Article history:

Received 10 May 2013

Accepted 3 December 2014

Available online 15 December 2014

Keywords:

Probability

Workforce planning

Inpatient care

Float nurse

Nurse-to-patient ratio

ABSTRACT

Workloads in nursing wards depend highly on patient arrivals and lengths of stay, both of which are inherently variable. Predicting these workloads and staffing nurses accordingly are 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 that 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 Amsterdam (AMC). This case study demonstrates the method's potential to evaluate the complex interaction between staffing requirements and several interrelated planning issues such as case mix, care unit partitioning and size, as well as surgical block planning. Inspired by the quantitative results, the AMC concluded that implementing this flexible nurse staffing methodology will be incorporated in the redesign of the inpatient care operations in the upcoming years.

© 2015 Published by Elsevier B.V.

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 (Wright et al., 2006), which means that every incidence of overstaffing is scrutinized during times when cost-containment efforts are required (Lang et al., 2004). 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 (Kane et al., 2007; Needleman et al., 2002), as well as increased work stress and burnout among nurses (Aiken et al., 2002, 2012). 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 encountered in nursing wards depends heavily on patient arrivals and lengths of stay, both of which are inherently variable. Predicting workloads and staffing nurses accordingly are essential for guaranteeing quality of care in a cost effective manner

(Broyles et al., 2010; de Véricourt and Jennings, 2011). Accurate workload predictions require that the dynamics of surrounding departments are considered, given that many patient arrivals at the inpatient care facility originate from the operating theater and the emergency department. In Kortbeek et al. (2014), we presented a method to predict hourly bed census across 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 as starting points with which we 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 (Hulshof et al., 2012; Kortbeek et al., 2014). In addition, especially for surgical inpatient departments, an alignment with the planning of the operating room schedule is beneficial. All these decisions are also intertwined with inpatient care workforce requirements, such as the skill mix, number of full time equivalents, and staffing levels per working shift. In the present paper, we incorporate the tactical decision that is referred to as 'staff-shift scheduling' in Hulshof et al. (2012) into the integrated modeling framework of Kortbeek et al. (2014). We address the following question: for each working shift during a given planning horizon, how many employees should be assigned to each inpatient care unit? These numbers, in turn, provide a guideline for the decisions regarding the scale of the workforce at the strategic planning level.

* Corresponding author at: Stochastic Operations Research, Department of Applied Mathematics, University of Twente, Drienerlolaan 5, 7500 AE Enschede, The Netherlands. Tel.: +31 534893461; fax: +31 534893069.

E-mail addresses: n.kortbeek@utwente.nl (N. Kortbeek), a.braaksma@utwente.nl (A. Braaksma), burgercaj@gmail.com (C.A.J. Burger), p.j.bakker@amc.nl (P.J.M. Bakker), r.j.boucherie@utwente.nl (R.J. Boucherie).

We explore the potential of flexible staffing policies that allow hospitals to dynamically respond to their fluctuating patient populations. This flexibility is achieved by employing a pool of cross-trained nurses, for whom assignments to specific care units are decided at the start of their shifts. The commonly applied term for such flexible employees is ‘float nurses’ (Gnanlet and Gilland, 2009; Smith-Daniels et al., 1988). The basic rationale underlying the possible benefits of introducing flex pools is the following: although the inpatient population fluctuates, this fluctuation is, to a certain extent, predictable due to its dependence on the operating room schedule and other predictable variability in patient arrivals (e.g., seasonal, day of week, and time of day effects). This predictable variation can be taken into account when determining the staffing levels for ‘dedicated nurses’, which are nurses with a fixed assignment to a care unit. Typically, staffing levels need to be determined a number of weeks in advance, so that individual nurse rosters can be settled in a timely manner. As a result, 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 applied when determining staffing levels (Aiken et al., 2012; Yankovic and Green, 2011). 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 based on nurse-to-patient ratios can be performed in two ways. The ratios can be considered 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 (Aiken et al., 2010; Twigg et al., 2011). The advantage of such minimum ratios is that a consistently high level of patient safety is guaranteed (Kane et al., 2007; Lang et al., 2004). The disadvantage, however, is that all beds need to be continuously staffed because there is always a possibility that all beds are occupied and, as described, the nurse rosters have to be settled in advance. Therefore, overstaffing is a threat because there is little flexibility to adjust staffing levels to the predicted patient demand. To overcome this disadvantage, a second application of nurse-to-patient ratios exists that involves using these ratios merely as guidelines (Elkhuizen et al., 2007). In such a case, the assumption is that there is slack in the time window during which certain indirect patient care tasks can be performed, without having direct negative consequences on patient safety or work stress. As a result, the ratios may at times be violated, but not too often, nor for 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, whereas the second more restrictive ratio must be satisfied for a certain fraction of time.

Our contribution is a generic exact analytic approach to determine the number of nurses to be staffed each working shift that guarantees a desired quality of care, as reflected by nurse-to-patient ratios, in the most cost-effective manner. The approach directly builds upon the bed census prediction method presented in Kortbeek et al. (2014), so that the 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, called the ‘fixed staffing policy model’, is formulated to determine optimal staffing levels when only dedicated nurses are employed. Next, we present a model in which a flex pool with float nurses is introduced, which satisfies precisely the same quality constraints as 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 a 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 a case study that builds on the case study presented in Kortbeek et al. (2014). The case involves the care units in the surgical inpatient clinic of the Dutch university hospital the Academic Medical Center Amsterdam (AMC), which serve the specialties of traumatology, orthopedics, plastic surgery, urology, vascular surgery, and general surgery. Inspired by the quantitative results, the AMC decided that the flexible nurse staffing method will be fully implemented during the upcoming years as part of the global 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

Personnel scheduling in general and capacity planning for nursing staff in specific have received considerable attention from the operations research community, which can be observed from the extensive literature review (Van den Bergh et al., 2013). The nurse staffing process involves a set of hierarchical decisions over different time horizons with different levels of precision. The first strategic level of decision-making is the workforce dimensioning decision which concerns both the number of employees that must be employed and is often expressed as the number of full time equivalents and the mix in terms of skill categories (Harper et al., 2010; Lavieri and Puterman, 2009; Oddo et al., 2009). The second tactical level concerns staff-shift scheduling, which deals with the problem of selecting which shifts are to be worked and how many employees should be assigned to each shift to meet the patient demand (Ernst et al., 2004; Kellogg and Walczak, 2007). The third operational offline decision level concerns the creation of individual nurse timetables, designed 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’ (Burke et al., 2004; Cheang et al., 2003; Chiamonte and Chiamonte, 2008). The fourth operational online decision level concerns the reconsideration of the staff schedule at the start of a shift. At this level, float nurses are assigned to specific care units (Burke et al., 2004; Smith-Daniels et al., 1988), 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 (Griffiths et al., 2005; Pierskalla and Brailer, 1994). The interdependence of the decision levels must be recognized to facilitate systematic improvements in nurse staffing. As expressed in the literature review by Pierskalla and Brailer (1994), each level is constrained by previous commitments made at higher levels, as well as by the degrees of flexibility conserved 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 Hulshof et al. (2012).

The literature has mainly focused on nurse rostering, as reflected by the survey and classification articles by Burke et al. (2004), de Causmaecker and vanden Berghe (2011), and Ernst et al. (2004). 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 (Brandeau et al., 2004; Harper et al., 2010). Incorrect assumptions

regarding the required staffing levels (tactical), during the rostering process (operational offline), might therefore necessitate expensive corrections made on the operational online decision level, for instance, by hiring additional temporary staff. Therefore, 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 into shifts (e.g., a day, evening and night shift).

Tactical workforce decision making in healthcare has received little attention. A spreadsheet approach has been presented by [Elkhuizen et al. \(2007\)](#), to retrospectively fit optimal shift staffing levels to historical census data. Prospectively assessing the impact of alternative interventions is difficult via such approaches, given that they lack the flexibility to explicitly model and study the coordination between different inpatient care decision levels, including their alignment with surrounding departments. Simulation studies have shown to be successful in taking a more integral approach (e.g. [Griffiths et al., 2005](#); [Harper et al., 2010](#)). The inherent disadvantage of simulation studies is, however, that they are typically context-specific, which limits the generalizability of study outcomes. Analytic yet deterministic approaches can, for example, be found in [Beliën and Demeulemeester \(2008\)](#), [Oddoye et al. \(2007\)](#), and [Walts and Kapadia \(1996\)](#). Stochastic approaches to determine shift staffing levels are available in [de Véricourt and Jennings \(2011\)](#), [Wright et al. \(2006\)](#), and [Yankovic and Green \(2011\)](#). These references do not present an integral care chain approach, given that the demand distributions underlying the staffing decisions are not based on patient arrival patterns from the operating theaters and emergency departments.

Workforce flexibility is considered a powerful concept in reducing the required size of the workforce and increasing job satisfaction ([Burke et al., 2004](#); [Dellaert et al., 2011](#); [Gnanlet and Gilland, 2009](#); [Griffiths et al., 2005](#); [Jack and Powers, 2009](#); [Siferd and Benton, 1992](#); [Smith-Daniels et al., 1988](#); [Stewart et al., 1994](#)). To adequately respond to variability in patient demand, various types of flexibility are suggested, including the use of part-time employees, overtime, temporary agency employees, and float nurses. Related to our work are the studies by [Gnanlet and Gilland \(2009\)](#) and [Li and King \(1999\)](#), which investigate the potential of float pools with cross-trained nurses. Both references address the aggregate decision of which budget of float nurse hours should be available during a given time period, and, as such, they do not address the level of working shifts. Concerning the assignment strategy to place a given number of available float nurses in care units at the start of their shifts, [Trivedi and Warner \(1976\)](#) indicate that formulating such an assignment strategy requires the consideration of three issues: (1) a method for measuring of the urgency of need for an additional nurse; (2) a prediction per care unit of that urgency of need for an upcoming shift; and (3) development of a technique for the allocation of the available float nurses to care units in order to meet this need. Whereas [Trivedi and Warner \(1976\)](#) focus 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, as seen in [de Véricourt and Jennings \(2011\)](#), [Wright et al. \(2006\)](#), and [Yankovic and Green \(2011\)](#). Both [de Véricourt and Jennings \(2011\)](#) and [Wright et al. \(2006\)](#) indicate that in practice, setting the numerical values of the ratios is more based on negotiation than on science. [Wright et al. \(2006\)](#) studied the relation between staffing costs and nurse-to-patient ratios. In this paper, 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 for the numerical values of the ratios. The first issue is addressed in the current study. The second issue has been the focus of [de Véricourt and Jennings \(2011\)](#) and [Yankovic and Green \(2011\)](#). Both of those studies present a queuing model

according to which they motivate that the ratios as mandated in California are too rigid. They underline the importance of differentiating ratios with patient mix (thereby reflecting the severity of patients' illnesses and their acuity) as well as with care unit size. In our study, we focus on determining staffing levels given pre-specified nurse-to-patient ratios. Nevertheless, we do emphasize the importance of employing meaningful nurse-to-patient ratios in realizing high-quality staffing.

To conclude, our contribution of an exact stochastic analytic approach is 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 bed census predictions obtained from the model of [Kortbeek et al. \(2014\)](#). In [Section 3.1](#), we first provide an overview of this bed census prediction model, and in [Section 3.2](#), we discuss the requirements that need to be satisfied in setting appropriate staffing levels. [Section 3.3](#) presents the fixed staffing model, and [Section 3.4](#) formulates the model to find optimal staffing levels when float nurse pools are applied: the flexible staffing model. Because the flexible model suffers from the curse of dimensionality, we approximate the solution via two models that identify upper and lower bounds of the staffing requirements.

3.1. Bed census predictions

The model in [Kortbeek et al. \(2014\)](#) predicts the workload at an inpatient care facility that consists of several care units on a time scale of hours. In this section, we provide a short overview of the prediction model; [Appendix A](#) provides a detailed summary. The model considers a planning horizon of Q days ($q=1, \dots, Q$), in which each day is divided into T time intervals ($t=0, 1, \dots, T-1$). A total number of K inpatient care units are considered ($k=1, \dots, K$), with the capacity of unit k being M^k beds. Probability distributions \hat{Z}_{qt}^k are determined reflecting the total number of patients recovering during each time interval t at each day q on each care unit k , 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 (sub)specialty operates in which operating room on which day of the week ([Van Oostrum et al., 2008](#)). The basis for the emergency department outflow prediction is a cyclic random arrival process that we defined in [Kortbeek et al. \(2014\)](#) as the Acute Admission Cycle (AAC). Schematically, the approach is as follows: first, the impact of the MSS and that of 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 into bed census distributions.

For the demand predictions, three steps are performed for both elective and acute patients. First, the impact of a single patient type in single MSS (time horizon: S days) and AAC (time horizon: R days) cycles is determined; in the second step, the impact of all patient types within individual MSS and AAC cycles can be calculated based on the single patient impact. Then, in the third step, the predictions from the second step are overlapped to determine the overall steady state impact of the repeating cycles (for the MSS and the AAC, separately). Finally, 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 (Q days) is the least common multiple of the lengths of the MSS and the AAC. At this point, the probability distributions $Z_{q,t}^k$ are obtained reflecting the total number of patients that request recovery in care unit $k, k=1, \dots, K$, during time interval $t, t=0, 1, \dots, T-1$, on day $q, q=1, \dots, Q$.

Due to the finite capacities of the care units, 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, the demand predictions $Z_{q,t}^k$ and bed census predictions $\hat{Z}_{q,t}^k$ 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 or her preferred care unit when a bed becomes available. In such a scenario, a fixed patient-bed allocation policy ϕ is assumed that prescribes the prioritization of such transfers.

3.2. Staffing requirements

Corresponding with the bed census prediction model, we consider a planning horizon of Q days ($q=1, \dots, Q$), during which each day is divided into T time intervals ($t=0, 1, \dots, T-1$). 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 $s_{q,\tau}^k$, is completely determined in advance. In the flexible case, ‘dedicated’ staffing levels $d_{q,\tau}^k$ per unit are determined, together with the number of nurses $f_{q,\tau}$ available in a flex pool. The decision regarding the particular units to which 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 frequent handovers, which increase the risk of medical errors. Thus, we obtain staffing levels $s_{q,\tau}^k = d_{q,\tau}^k + f_{q,\tau}^k, k=1, \dots, K$, where $f_{q,\tau}^k$ 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 decided according to a predetermined assignment procedure. We denote such an assignment procedure by π . For both staffing policies we assume shifts to be non-overlapping, and for the flexible policy we assume shifts to be equivalent for each care unit.

Our goal is to determine the most cost-efficient staffing levels such that certain quality-of-care constraints are satisfied. Because float nurses are required to be cross-trained, it is likely that these staff members are more expensive to employ. To be able to differentiate such costs, we therefore consider staffing costs ω_d for each dedicated nurse who 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 $r_{q,\tau}^k$, 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_t^k being the number of patients present at unit k at a certain time $(q, t), b_\tau \leq t < b_\tau + \ell_\tau$, the coverage at that time is given by $r_{q,\tau}^k \cdot s_{q,\tau}^k / x_t^k$. Thus, a coverage of one or higher corresponds to a preferred situation.

Starting from the following quality-of-care requirements as pre-requisites, 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 expectation of the coverage compliance at care unit k during shift (q, τ) by $c_{q,\tau}^k(\cdot)$; the arguments of this function depend on which staffing policy is considered. (Note that ‘coverage compliance’ is a measure defined for a shift, based on the measure ‘coverage’ that is defined for the time periods within that shift).
- (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’. Fairness is defined as assigning each 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) :

$$\min z_F = \omega_d s_{q,\tau}^k \tag{1}$$

$$\text{s.t. } s_{q,\tau}^k \geq S^k \tag{2}$$

$$s_{q,\tau}^k \geq \lceil \beta^k \cdot M^k / r_{q,\tau}^k \rceil \tag{3}$$

$$c_{q,\tau}^k(s_{q,\tau}^k, r_{q,\tau}^k) \geq \alpha^k \tag{4}$$

The constraints (2), (3), and (4) reflect requirements (i), (ii), and (iii), respectively. Let $X_{q,t}^k$ be the random variable with bed census distribution $\hat{Z}_{q,t}^k$ 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:

$$\begin{aligned} c_{q,\tau}^k(s_{q,\tau}^k, r_{q,\tau}^k) &= \mathbb{E} \left[\frac{1}{\ell_\tau} \sum_{t=b_\tau}^{b_\tau + \ell_\tau - 1} \mathbb{1}(X_{q,t}^k \leq s_{q,\tau}^k \cdot r_{q,\tau}^k) \right] \\ &= \frac{1}{\ell_\tau} \sum_{t=b_\tau}^{b_\tau + \ell_\tau - 1} \sum_{x=0}^{s_{q,\tau}^k \cdot r_{q,\tau}^k} \hat{Z}_{q,t}^k(x). \end{aligned}$$

Observe that $\sum_{x=0}^{s_{q,\tau}^k \cdot r_{q,\tau}^k} \hat{Z}_{q,t}^k(x)$ reflects the probability that with staffing level $s_{q,\tau}^k$ and under ratio $r_{q,\tau}^k$ the nurse-to-patient ratio target is satisfied during time interval $[t, t+1)$. The optimum of (1) is found by choosing the minimum $s_{q,\tau}^k$ 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 $s_{q,\tau}^k$ by $d_{q,\tau}^k$ 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 fulfills 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 $g_{q,\tau}^\pi(\mathbf{d}, f, \mathbf{y}) = (g_{q,\tau}^{1,\pi}(\mathbf{d}, f, \mathbf{y}), \dots, g_{q,\tau}^{K,\pi}(\mathbf{d}, f, \mathbf{y}))$ be the vector 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 $\mathbf{d} = (d^1, \dots, d^K)$, and the census at the different care units at time (q, b_τ) equals $\mathbf{y} = (y^1, \dots, y^K)$. A vector of the type \mathbf{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 $\mathbf{d}_{q,\tau}, f_{q,\tau}$, and $r_{q,\tau}^k, k = 1, \dots, K$, and therefore the coverage as well. Hence, requirement (v) gives a constraint of the form $c_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k) \geq \alpha^k$. However, assignment procedure π^* depends on the census configuration \mathbf{y} at time (q, b_τ) , so calculation of the coverage compliance first requires the computation of $c_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k; \mathbf{y})$, which describes the coverage compliance, given that at the start of shift (q, τ) census configuration \mathbf{y} is observed. Then, the coverage compliance is given by

$$c_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k) = \sum_{\mathbf{y}} \left\{ c_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k; \mathbf{y}) \prod_{w=1}^K \hat{Z}_{q,b_\tau}^w(\mathbf{y}^w) \right\}.$$

Using $c_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k; \mathbf{y})$, the assignment policy π^* satisfying requirement (v) is the one that satisfies

$$g_{q,\tau}^{\pi^*}(\mathbf{d}_{q,\tau}, f_{q,\tau}, \mathbf{y}) = \operatorname{argmax}_{\{(f_{q,\tau}^1, \dots, f_{q,\tau}^K) : \sum_k f_{q,\tau}^k = f_{q,\tau}\}} \min_k c_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k; \mathbf{y}). \quad (5)$$

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

$$\min z_E = \omega_f f_{q,\tau} + \omega_d \sum_k d_{q,\tau}^k \quad (6)$$

$$\text{s.t. } d_{q,\tau}^k \geq S^k \quad \text{for all } k, \quad (7)$$

$$d_{q,\tau}^k \geq \lceil \beta^k \cdot M^k / r_{q,\tau}^k \rceil \quad \text{for all } k, \quad (8)$$

$$c_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k) \geq \alpha^k \quad \text{for all } k, \quad (9)$$

$$d_{q,\tau}^k \geq \gamma^k \cdot s_{q,\tau}^k(\mathbf{y}) \quad \text{for all } k, \mathbf{y}, \quad (10)$$

$$s_{q,\tau}^k(\mathbf{y}) = d_{q,\tau}^k + g_{q,\tau}^{k,\pi^*}(\mathbf{d}_{q,\tau}, f_{q,\tau}, \mathbf{y}) \quad \text{for all } k, \mathbf{y}. \quad (11)$$

Constraints (7)–(11) reflect (i)–(v), respectively. Finding the optimum for (6) requires the computation of $c_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k; \mathbf{y})$ 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 $d_{q,\tau}^1, \dots, d_{q,\tau}^K, f_{q,\tau}$ 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 then finding the optimal staffing for this maximum census. In this case, the number of patients present is overestimated, and subsequently the required staffing levels are overestimated; 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 instead of at the start of each shift. Because 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 with (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 also provides an error bound.

Upper bound model: Based on the observed maximum census configuration $\mathbf{x} = (x^1, \dots, x^K)$ during a shift, let π^{up} be the assignment policy that allocates the nurses from the flex pool to the care units in which the nurse deficiency is the highest:

$$g_{q,\tau}^{\pi^{up}}(\mathbf{d}_{q,\tau}, f_{q,\tau}, \mathbf{x}) = \operatorname{argmax}_{\{f_{q,\tau}^1, \dots, f_{q,\tau}^K : \sum_k f_{q,\tau}^k = f_{q,\tau}\}} \min_k \frac{r_{q,\tau}^k \cdot (d_{q,\tau}^k + f_{q,\tau}^k) - x^k}{r_{q,\tau}^k}.$$

Let $\hat{W}_{q,\tau}^k(x)$ 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 $\hat{Z}_{q,\tau}^k(x)$ in Kortbeek et al. (2014) (for details see Appendix C). To obtain the upper bound, for $b_\tau \leq t < b_\tau + \ell_\tau$, we approximate the original distributions $\hat{Z}_{q,t}^k(x)$ by $\hat{W}_{q,\tau}^k(x)$. Let $\bar{X}_{q,\tau}^k$ be the random variable with distribution $\hat{W}_{q,\tau}^k$ that reflects 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 $\bar{X}_{q,\tau}^k \geq X_{q,t}^k$, for $b_\tau \leq t < b_\tau + \ell_\tau$, so that for every time interval of a shift the census is overestimated, and thus staffing requirements are overestimated.

Because we use the same census distribution in every time interval during a shift, the coverage compliance over a shift $\bar{c}_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k)$ is calculated by

$$\bar{c}_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k) = \sum_{\mathbf{x}} \left\{ \mathbb{1}(x^k \leq r_{q,\tau}^k \cdot s_{q,\tau}^k(\mathbf{x})) \cdot \prod_{w=1}^K \hat{W}_{q,\tau}^w(x^w) \right\},$$

where $s_{q,\tau}^k(\mathbf{x})$ 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 \mathbf{x} . Summarizing, for each shift (q, τ) , we have

$$\min z_U = \omega_f f_{q,\tau} + \sum_k \omega_d d_{q,\tau}^k \quad (12)$$

$$\text{s.t. } d_{q,\tau}^k \geq S^k \quad \text{for all } k, \quad (13)$$

$$d_{q,\tau}^k \geq \lceil \beta^k \cdot M^k / r_{q,\tau}^k \rceil \quad \text{for all } k, \quad (14)$$

$$\bar{c}_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k) \geq \alpha^k \quad \text{for all } k, \quad (15)$$

$$d_{q,\tau}^k \geq \gamma^k \cdot s_{q,\tau}^k(\mathbf{x}) \quad \text{for all } k, \mathbf{x}, \quad (16)$$

$$s_{q,\tau}^k(\mathbf{x}) = d_{q,\tau}^k + g_{q,\tau}^{k,\pi^{up}}(\mathbf{d}_{q,\tau}, f_{q,\tau}, \mathbf{x}) \quad \text{for all } k, \mathbf{x}. \quad (17)$$

The optimum of (12) is identified by first finding the feasible solution space for $d_{q,\tau}^k, k = 1, \dots, K$, using constraints (13) and (14). Second, the feasible solution space for $f_{q,\tau}$ is found using constraint (16) as well as the optimal solutions of the k underlying separate fixed staffing models. Next, complete enumeration over the obtained feasible solution space is applied, which can be done quickly for realistic situations.

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 $c_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k)$ can then be

calculated by

$$\underline{c}_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k) = \frac{1}{\ell_\tau} \sum_{t=b_\tau}^{b_\tau + \ell_\tau - 1} \sum_{\mathbf{x}} \left\{ \mathbb{1}(x^k \leq r_{q,\tau}^k \cdot s_{q,t}^k(\mathbf{x})) \cdot \prod_{w=1}^K \hat{z}_{q,t}^w(x^w) \right\}$$

where $s_{q,t}^k(\mathbf{x})$ 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 \mathbf{x} is observed at time (q, t) .

Since π^{low} is executed at 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 nurse deficiency is the highest:

$$g_{q,t}^{\pi^{\text{low}}}(\mathbf{d}_{q,\tau}, f_{q,\tau}, \mathbf{x}) = \underset{\{f_{q,t}^1, \dots, f_{q,t}^k, \dots, f_{q,t}^K: \sum_k f_{q,t}^k = f_{q,\tau}\}}{\text{argmax}} \min_k \frac{r_{q,\tau}^k \cdot (d_{q,\tau}^k + f_{q,t}^k) - x^k}{r_{q,\tau}^k}$$

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

$$\min z_L = \omega_f f_{q,\tau} + \sum_k \omega_d d_{q,\tau}^k \quad (18)$$

$$\text{s.t. } d_{q,\tau}^k \geq S^k \quad \text{for all } k, \quad (19)$$

$$d_{q,\tau}^k \geq [\beta^k \cdot M^k / r_{q,\tau}^k] \quad \text{for all } k, \quad (20)$$

$$\underline{c}_{q,\tau}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k) \geq \alpha^k \quad \text{for all } k, \quad (21)$$

$$d_{q,\tau}^k \geq \gamma^k \cdot s_{q,t}^k(\mathbf{x}), \quad b_\tau \leq t < b_\tau + \ell_\tau \quad \text{for all } k, \mathbf{x}, \quad (22)$$

$$s_{q,t}^k(\mathbf{x}) = d_{q,\tau}^k + g_{q,t}^k(\mathbf{d}_{q,\tau}, f_{q,\tau}, \mathbf{x}), \quad b_\tau \leq t < b_\tau + \ell_\tau \quad \text{for all } k, \mathbf{x}. \quad (23)$$

The optimum of (18) is found by first finding the feasible solution space for $d_{q,\tau}^k, k = 1, \dots, K$, using constraints (19) and (20). Second, the feasible solution space for $f_{q,\tau}$ is found using constraint (22), and the optimal solutions $d_{q,\tau}^{k,*}$ of the k underlying separate fixed staffing models. Next, complete enumeration over the obtained feasible 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. In this section, we discuss how the solutions of the fixed model, as well as 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 should be selected 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 bounds coincide so that the optimal solution is found. When $z_L < z_U$, but $z_L = z_F$, the optimal solution is also found because, 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 has been identified; it is then 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 uncertain whether constraints (7)–(11) are satisfied; therefore, we never select this solution. In addition, when $z_F = z_U$, as a tie breaker, 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 models, respectively.

We now provide an overview of the different cases:

- $z_L = z_F = z_U$: The optimal solution is found; if $\min_k \bar{c}_{q,\tau}^k(\cdot) \geq \min_k c_{q,\tau}^k(\cdot)$, S_U is selected as the best staffing configuration, otherwise S_F .
- $z_L = z_U < z_F$: The optimal solution is found; S_U is selected.
- $z_L = z_F < z_U$: The optimal solution is found; S_F is selected.
- $z_L < z_F = z_U$: Uncertain whether the optimal solution is found; if $\min_k \bar{c}_{q,\tau}^k(\cdot) \geq \min_k c_{q,\tau}^k(\cdot)$, S_U is selected, otherwise S_F . The bound on the error margin is $z_U - z_L$.
- $z_L < z_U < z_F$: Uncertain whether the optimal solution is found; S_U is selected; the error bound is $z_U - z_L$.
- $z_L < z_F < z_U$: Uncertain whether the optimal solution is found; S_F is selected; the error bound is $z_F - z_L$.

4. Quantitative results

This section presents the experimental results. The case study entails six surgical specialties of the university hospital AMC, which together have 104 beds in operation. The entire hospital has 20 operating rooms, and 30 inpatient departments, with a total of 1000 beds. Building on the case study presented in Kortbeek et al. (2014), the practical potential of the staffing methodology will be illustrated by returning to a selection of the interventions presented in Kortbeek et al. (2014), which were formulated to improve the efficiency of the inpatient care service operations in terms of productivity of the inpatient beds. In addition, we formulate two additional interventions. Section 4.1 describes additional information on the case study. Section 4.2 presents the interventions to be considered. Before presenting the numerical results in Section 4.4, in Section 4.3, we validate our approximation approach by investigating the distance between the upper and the lower bound solutions.

All methods were coded with the Embarcadero Delphi XE programming language and tested on an Intel 2.4 GHz PC with 3.42 GB of RAM. For a given shift, the required staffing levels can be computed within a few seconds.

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 above-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.

The physical building is such that units A and B are physically adjacent (Floor I), as are units C and D (Floor II). For these specialties, we have historical data available over 2009–2010 on 3498 (5025) elective (acute) admissions, with an average length-of-stay (LOS) of 4.85 days (see Table 1). Currently, no cyclic MSS is applied. Each time, roughly six weeks in advance the MSS is determined for a period of four weeks. The capacities of units A, B, C, and D are 32, 24, 24, and 24 beds, respectively. The utilizations over 2009–2010 were 53.2%, 55.6%, 54.4%, and 60.6%, respectively (which includes some patients from other specialties that were placed in these care units).

Working days are divided into 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 indicate the times that nurses are responsible for direct patient care. Around these time intervals, the working shifts also incorporate time for patient handovers,

Table 1
Overview of historical data 2009–2010.

Specialty	Acronym	Care unit	Elective admissions	Acute admissions	Average LOS (in days)	Load ^a (# patients)
General surgery	GEN	A	611	901	3.31	6.88
Urology	URO	A	818	1157	3.68	9.99
Vascular surgery	VAS	B	257	634	8.30	10.16
Plastic surgery	PLA	B	639	288	2.29	2.91
Traumatology	TRA	C	337	1200	5.88	12.41
Orthopedics	ORT	D	836	845	6.23	14.38

^a Load = Expected number of patient arrivals per day * Average LOS.

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 [Elkhuizen et al., 2007](#)). The yearly cost per FTE, including all costs and bonuses, is roughly €53,000.

The nurse-to-patient ratio targets prescribed by the board of the AMC for the care units of interest 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 they are occupied, and no float nurse pools are employed. Thus, for example, for a care unit size of 24 beds and staffing ratio of 1:4, the number of dedicated nurses to staff is always 6. A scarcity of nursing capacity frequently leads to the expensive hiring of temporary nurses from external agencies, as well as 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.

4.2. Interventions

To illustrate the potential of the presented staffing methodology for the case study, we will return to a selection of the interventions that we presented in [Kortbeek et al. \(2014\)](#) and formulate two additional interventions. For self-containment of the present paper, in this section, we first provide a summary of the selected previously considered interventions (Interventions (0), (1), (3), (4), (5); and not (2) and (6)) and then introduce the two new interventions (Interventions (7) and (8)). The following cases were considered in [Kortbeek et al. \(2014\)](#):

- (0) *Base case*: To assess the effects of the interventions, we first evaluated the performance of a base case scenario, which is the situation that most closely resembles current practice. The base case involved the current bed capacities and misplacements between care units A and B (Floor I), and between units C and D (Floor II).
- (1) *Rationalize bed requirements*: Because the current numbers of beds are a result of historical development, we determined whether the number of beds can be reduced to achieve a higher bed utilization while a certain quality-of-service level is guaranteed. To this end, we considered rejection probabilities not exceeding 5%, 2.5%, and 1%, with the outcome that a significant reduction in the number of beds is possible.

- (3) *Change operational process*: This intervention predicted the potential impact of two changes in the operational process. First, admitting all elective patients on the day of surgery, since admitting patients the day before surgery is generally induced by logistical reasons. Second, stimulating discharges to take place before noon, to reduce census peaks during midday hours. It was shown that, compared to Intervention (1), the number of beds can be further decreased, and the number of patients treated per bed per day can be significantly increased.
- (4) *Balance MSS*: The realized MSS created artificial demand variability. This intervention estimated the potential of a cyclic MSS that is designed with the purpose to balance bed census and showed that both the midweek peak and the weekend dip can be cleared to a large extent, which results in distinct efficiency gains.
- (5) *Combination (1), (3), and (4)*: By combining Interventions (1), (3), and (4), we demonstrated that a reduction of the number of beds by 20% is possible, as well as an increase of the number of patients treated per bed per day by roughly 25%.

For a complete specification of these interventions and the corresponding results, we refer the reader to [Kortbeek et al. \(2014\)](#). In applying the two staffing models with respect to these interventions, we will use the bed census distributions that were obtained by running the prediction model with input parameters based on the historical data from the year 2010. Because the management of the hospital agreed upon a service level norm of rejection probabilities < 2.5%, in the present paper we focus on the bed census predictions that correspond to this particular service level requirement. Based on the initial intention of the AMC, for Interventions (0)–(5), 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. Finally, we test the restrictiveness of this assumption by evaluating the impact of the following two additional interventions:

- (7) *Centralized flex pool*: This intervention involves the merging of the two flex pools into one flex pool that 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).
- (8) *Merging care units*: Finally, this intervention merges care units A and B, and care units C and D. The two remaining care units share one flex pool. Possible economies-of-scale effects are tested in Intervention (8a) for the situation of Intervention (1), and in Intervention (8b) for that of Intervention (5).

4.3. 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 for the base case scenario. We consider a planning horizon of one year, during which no cyclic MSS was used; we thus have to staff $365 \times 3 = 1095$ unique working shifts.

For our set of test instances, [Table 2](#) provides an overview of the considered parameter settings. We vary over the following variables: the (relative) staffing cost for float nurses, the nurse-to-patient ratios, 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. [Table 3](#) displays the results. The overall result is that in 94.0% of the cases the

Table 2
Input parameter settings of the test instances for care units $k \in \{A, B, C, D\}$.

Parameter	Description	Value
<i>Fixed</i>		
Q	Planning horizon in days	365
T	Number of time intervals per day	24
$ \mathcal{T} $	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
<i>To be varied</i>		
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}
$(r_{q,1}^k, r_{q,2}^k, r_{q,3}^k)$	Nurse-to-patient ratio targets	{(4, 6, 10), (4, 6, 8), (5, 5, 10)}

Table 3
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 ($r_{q,\tau}^k$)
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%

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 solutions 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 counter-intuitive. 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 3.4, we described how to find error bounds on the deviation from the optimal objective value in case one is not sure whether the optimum has been found. For a given shift, let z^* denote the objective value of the selected staffing configuration. We calculate the deviation of the obtained solution from the lower bound solution in percentages as $(z^* - z_L) / z_L \cdot 100\%$. Fig. 1 shows a histogram of these deviations per shift, for the 6.0% of shifts for which it is not sure whether the optimum has been found. The average maximum deviation for non-optimal shifts is 8.1%. 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, Fig. 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 approximation approach via bounds on the staffing levels, approaches optimal performance for our case study.

4.4. Case study results

In this section, we present the results for the case study on the interventions described in Section 4.2. We investigate both the

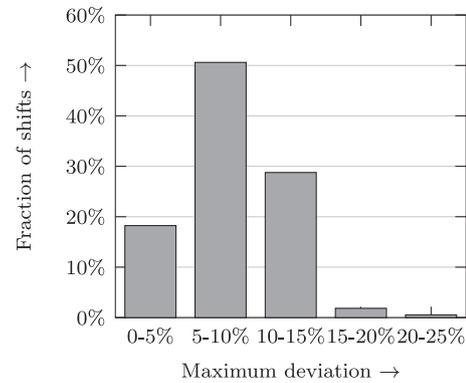


Fig. 1. Distribution of the relative deviation of the obtained solution, z^* , from the lower bound solution, z_L (non-optimal shifts, $n=147,426$).

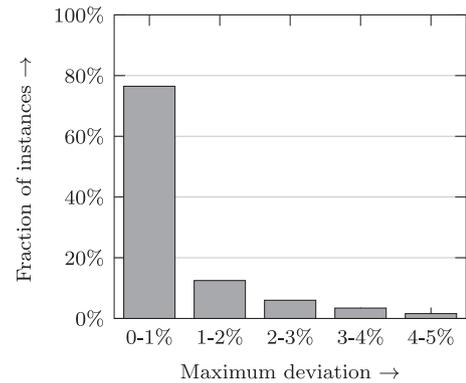


Fig. 2. Distribution of the error bound on total staffing costs (all instances, $n=2250$).

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 a care unit capacity of M^k beds at unit k , under the full staffing policy, $\lceil M^k / r_{q,\tau}^k \rceil$ nurses are required at all times.

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 nurses are equally expensive, i.e., $\omega_d = \omega_f$. In addition to the fixed input as displayed in Table 2, the board of the AMC has chosen to deploy the following quality of care requirements: nurse-to-patient ratios $r_{q,1}^k = 4$, $r_{q,2}^k = 6$, $r_{q,3}^k = 10$, 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., $\gamma^k = 0.67$.

The detailed results are displayed in Tables 4 and 5. Table 6 provides an overview of the results for the various interventions and includes the calculation of the productivity measure of the number of patients treated per employed FTE per year.

Base case: First, we evaluate the performance of the base case scenario (see Table 4). In the flexible staffing policy, two flex pools are installed, one on each floor; we therefore present the results per floor. 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.

The number of FTEs required is calculated by summing the total number of staffed nurse hours and dividing by the 1525.7 direct nursing hours that one FTE has available. Note that in this calculation we do not include scheduling restrictions that might

Table 4

The numerical results for the base case (Floor I: 56 beds, 56.7% utilization; Floor II: 48 beds, 58.6% utilization; with the **FTE**– $\Delta\%$ relative to full staffing).

Intervention	Floor	Full staffing FTE (#)	Fixed staffing			Flexible staffing			
			Average coverage	FTE		Error bound (%)	Average coverage	FTE (float)	
				(#)	($\Delta\%$)			(#)	($\Delta\%$)
<i>Base case</i>									
$\alpha=0.85$	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

Table 5

The numerical results for the various interventions (with the **FTE**– $\Delta\%$ relative to full staffing).

Intervention	Capacity(# beds)	Utilization (%)	Full staffing FTE (#)	Fixed staffing			Flexible staffing			
				Average coverage	FTE		Average coverage	FTE (float)		
					(#)	($\Delta\%$)		(#)	($\Delta\%$)	
<i>1. Rationalize bed requirements</i>										
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	
<i>3. Change operational process</i>										
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	
<i>4. Balance MSS</i>										
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	
<i>5. Combination (1), (3) and (4)</i>										
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	
<i>7a. Combination (1) and centralized flex pool</i>										
Floors I & II	88	67.9	90.7	0.99	83.1	–8.4	0.98	80.2 (9.5)	–11.5	
<i>7b. Combination (5) and centralized flex pool</i>										
Floors I & II	83	68.1	90.7	0.98	81.3	–10.3	0.98	77.4 (8.6)	–14.6	
<i>8a. Combination (7a) and merge care units</i>										
Floors I & II	88	67.9	84.9	0.97	74.7	–12.1	0.96	73.8 (9.7)	–13.1	
<i>8b. Combination (7b) and merge care units</i>										
Floors I & II	83	68.1	83.3	0.97	72.0	–13.5	0.97	71.5 (9.6)	–14.1	

be involved when assigning individual nurses to working shifts. Therefore, at a particular inpatient clinic, the number of FTEs to hire might need to be larger than the displayed number of FTEs required, depending on the local labour regulations and nurse rostering practice.

For both the fixed and the flexible staffing models, it turns out that the realized coverage compliance is, on average, much higher than the minimum requirement. This result occurs because when the coverage compliance constraint is slightly violated, an additional nurse needs to be staffed, which significantly increases the coverage compliance because this nurse can care for $r_{q,\tau}^k$ 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 of the fixed model). The additional value of employing float nurses is

case dependent, and in most cases, the value is higher with increasing α^k due to the increasing gap with the minimum coverage requirement set by β^k .

Interventions (1), (3), (4), and (5): Intervention (1) rationalizes the care unit dimensions. Table 5 shows that fixed staffing with $\alpha^k = 0.90$ reduces nursing capacity requirements by 8–9% compared to full staffing, and flexible staffing yields an additional 1% reduction. Table 6 indicates the gain against current practice: 22.6% reduction in FTE requirements, with a simultaneous increase of staff productivity by 26.5%.

Intervention (3) focuses on changes in the operational process that shorten the average lengths of stay. 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, as well as higher productivity.

Intervention (4) intends to decrease the artificial demand variability by designing a balanced cyclic MSS. Note that due to the integrality of the number of scheduled operating room blocks, the resulting MSS has slightly increased patient demand.

Table 6
FTE and productivity results for all interventions (with both the FTE- $\Delta\%$ and the productivity- $\Delta\%$ relative to full staffing in the base case).

Intervention	Full staffing				Fixed staffing				Flexible staffing			
	FTE		Productivity		FTE		Productivity		FTE		Productivity	
	(#)	($\Delta\%$)	(#/yr)	($\Delta\%$)	(#)	($\Delta\%$)	(#/yr)	($\Delta\%$)	(#)	($\Delta\%$)	(#/yr)	($\Delta\%$)
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.

Therefore, its impact on staffing requirements is not directly evident. However, its impact is revealed by the outcomes of Intervention (5) (the combination between Interventions (1), (3), and (4)), which outperform all previous configurations on the productivity measure. As an illustration, the effect of staffing levels following bed census demand patterns, including the differences between fixed and flexible staffing therein, is visualized in Fig. 3. Also, in this figure, average demand is displayed for day shifts in a 4-week period as the average bed census divided by the applied nurse-to-patient ratios. It signals that the high variability in bed census implies that the number of nurses to be staffed, to guarantee the coverage compliance on the nurse-to-patient ratios, is considerably higher than average demand. It is a clear indication of the savings potential of increasing the predictability of demand for nursing staff by balancing bed census.

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, based on our results we cannot deduce general rules-of-thumb for the potential of float nurses. The outcomes for each particular care unit are a complex interplay between care unit sizes, nurse-to-patient ratios, and the shapes of the bed census distributions.

Intervention (7): Intervention (7a) evaluates the impact of a 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 Interventions (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): Intervention (8a) merges care units A and B and care units C and D for the situation of Intervention (1), and Intervention (8b) does the same for that of Intervention (5). The two remaining care units, Floor I and Floor II, share one flex pool. The implementation of this intervention would require a renovation of the building. The positive outcomes of this intervention indicate that it is worthwhile to consider this renovation to benefit from the economies-of-scale effect. The economies-of-scale effect manifests 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, thereby 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.

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, given that hospital managers lack the tools to base current staffing decisions on information about future patient demand. Since patient safety is jeopardized when medical care units are understaffed, a 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 about 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 Kortbeek et al. (2014) and the staffing models from the present 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 us 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.

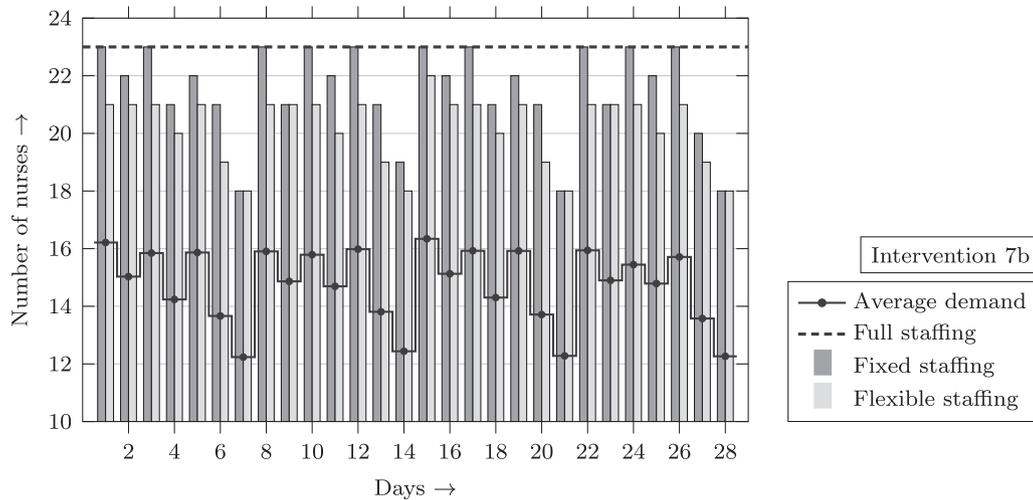


Fig. 3. Total staffing levels for day shifts during the 4-week period starting on Monday January 25 (the average demand pattern shows the average census divided by ratios $r_{q,1}^k$).

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 insights have been obtained from this case study that we believe are worthwhile to consider when studying other inpatient clinics. When working with nurse-to-patient ratios, our case study suggests that care units should be sufficiently large to avoid efficiency losses due to the lack of granularity in the values of the ratios. Next, it suggests that 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, it supports that flexible staffing is beneficial also in case it does not reduce capacity requirements, 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 because the AMC has 30 inpatient departments and staffing costs account for 66% of the total expenses in the AMC. We have applied our staffing models to data from several care units, and we presented results from four of them 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. This collaboration resulted in the joint conclusion that substantial efficiency gains are possible, while improving upon the adherence to nurse-to-patient ratio targets.

Based on the outcomes of both studies, the bed census prediction model presented in Kortbeek et al. (2014) and the subsequent flexible staffing method presented in this paper are embraced by the AMC as valuable instruments to support the resource capacity planning of its inpatient care services. The decision-making process on which specific interventions to apply in practice and the subsequent implementation phase will take place during the upcoming years embedded in a hospital-wide improvement program. What is clear at this point in time is that the staffing policies that are currently applied in the AMC will be revised and formalized along the lines of the presented method, and float nurse pools will be installed.

To fully exploit the potential of the staffing method, which is the intention of the AMC, a user-friendly decision support tool (DSS) based on bed census prediction and staffing models is required. The

prediction model relies on data which 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. In addition, integration with the nurse rostering software is a prerequisite. As a next step in achieving practical impact, we are currently in the process of developing such a tool.

Acknowledgments

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 (Grant no. 08140).

Appendix A. Detailed summary bed census prediction model

This appendix provides a summary of the hourly bed census prediction model of Kortbeek et al. (2014).

A.1. 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 into T time intervals. Therefore, we have time points $t=0, \dots, T$, 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 $s \in \{1, \dots, S\}$, a (sub)specialty j can be assigned to an available operating room i , $i \in \{1, \dots, I\}$. The OR block at operating room i on day s is denoted by $b_{i,s}$, and is possibly divided into a morning block $b_{i,s}^M$ and an afternoon block $b_{i,s}^A$, 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, $k \in \{0, 1, \dots, C^j\}$. If an OR block is divided into a morning OR block and an afternoon OR block, c_M^j and c_A^j represent the utilization probability distributions, respectively. Such shared OR blocks are not explicitly included in our formulation, given

that these can be modeled as two separate (fictitious) operating rooms.

Admissions: With probability e_n^j , $n \in \{-1, 0\}$, 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 $w_{n,t}^j$. 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 $w_{0,t}^j = 0$ for $t = \vartheta_j + 1, \dots, T - 1$.

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

Single surgery block: In this first step, we consider a single specialty j operating in a single OR block. We compute the probability $h_{n,t}^j(x)$ 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 $n = 1, \dots, L^j$. Therefore, we calculate $h_{n,t}^j(x)$ as follows:

$$h_{n,t}^j(x) = \begin{cases} a_{n,t}^j(x) & \text{if } n = -1 \text{ and } n = 0, t \leq \vartheta_j, \\ d_{n,t}^j(x) & \text{if } n = 0, t > \vartheta_j \text{ and } n = 1, \dots, L^j, \end{cases}$$

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

Single MSS cycle: Next, we consider a single MSS in isolation. From the distributions $h_{n,t}^j$, we can determine the distributions $H_{m,t}$, the discrete distributions for the total number of recovering patients at time t on day m , $m \in \{0, 1, 2, \dots, S, S + 1, S + 2, \dots\}$, 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 $H_{s,t}^{SS}$, which are the steady state probability distributions of the number of recovering patients at time t on day s of the cycle, $s \in \{1, \dots, S\}$.

A.2. Demand predictions for acute patients

Model input: The demand predictions for acute patients are 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 , $p = 1, \dots, P$, 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, $n \in \{0, \dots, L^j\}$. Given that a patient is discharged at day n , the probability of being discharged in time interval $[t, t + 1)$ is given by $\tilde{m}_{n,t}^j$. By definition, $\tilde{m}_{0,t}^j = 0$ for $t \leq \theta$.

Single patient type: In this first step we consider a single patient type j . We compute the probability $g_{n,t}^j(x)$ 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 $n = 1, \dots, L^j$. Therefore, we

calculate $g_{n,t}^j(x)$ as follows:

$$g_{n,t}^j(x) = \begin{cases} \tilde{a}_t^j(x) & \text{if } n = 0, t = \theta, \\ \tilde{d}_{n,t}^j(x) & \text{if } n = 0, t > \theta \text{ and } n = 1, \dots, L^j, \end{cases}$$

where $\tilde{a}_t^j(x)$ represents the probability that x patients are admitted in time interval $[t, t + 1)$ on day $n = 0$, and $\tilde{d}_{n,t}^j(x)$ 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 $g_{n,t}^j(x)$, we can determine the distributions $G_{w,t}$, the distributions for the total number of recovering patients at time t on day w , $w \in \{1, \dots, R, R + 1, R + 2, \dots\}$, 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 $G_{r,t}^{SS}$, the steady state probability distributions of the number of recovering patients at time t on day r of the cycle, $r \in \{1, \dots, R\}$.

A.3. 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 $H_{s,t}^{SS}$ and $G_{r,t}^{SS}$. 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 :

$$Z_{q,t} = H_{q \bmod S + S \cdot \lfloor \frac{q}{S} \rfloor, t}^{SS} \otimes G_{q \bmod R + R \cdot \lfloor \frac{q}{R} \rfloor, t}^{SS},$$

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 , $k \in \{1, \dots, K\}$, and V^k the set of acute patient types j that are (preferably) admitted to unit k . Then, the demand distribution for unit k , $Z_{q,t}^k$, can be calculated by exclusively considering the patients in U^k and V^k .

A.4. Bed census predictions

We translate the demand distributions $Z_{q,t}^k$ into bed census distributions $\hat{Z}_{q,t}^k$, $k = 1, \dots, K$, 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 $\mathbf{x} = (x_1, \dots, x_K)$ a bed census vector $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_K)$, 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 the allocation policy ϕ . Let $\hat{Z}_{q,t}^k$ denote the marginal distribution of the census at unit k given by distribution $\hat{Z}_{q,t}$. With a care unit capacity of M^k beds at unit k , we obtain

$$\hat{Z}_{q,t}(\hat{\mathbf{x}}) = \left(\hat{Z}_{q,t}^1(\hat{x}_1), \dots, \hat{Z}_{q,t}^K(\hat{x}_K) \right) = \sum_{\{\mathbf{x} | \hat{\mathbf{x}} = \phi(\mathbf{x})\}} \left\{ \prod_{k=1}^K Z_{q,t}^k(x_k) \right\}. \quad (24)$$

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

capacities M^1 and M^2 :

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

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 in the second unit.

Appendix B. Complexity of the flexible staffing model

This appendix investigates the complexity of the calculations involved in solving the flexible staffing model, formulated by Eqs. (6)–(11). The complexity is such that the computation time inhibits the evaluation of realistically sized instances. This limitation 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 $d_{q,\tau}^1, \dots, d_{q,\tau}^K, f_{q,t}$. For every census configuration \mathbf{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 $O(N_y \cdot N_f \cdot N_c \cdot K)$, where N_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 $c_{q,\tau}^k(d_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k; \mathbf{y})$, which has to be done for all K care units.

Since the census range for care unit k is $\{0, \dots, M^k\}$, with $\hat{M} = \max_k M^k$, we have $N_y = (\hat{M} + 1)^K$. Second, in counting the number of possible allocations of $f_{q,\tau}$ nurses over K care units, 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 of a patient cohort (as also introduced in Kortbeek et al., 2014): a cohort is a group of patients originating from a single instance of an OR block (electives) or admission time interval (acute patients). We use the indicator χ to refer to a cohort. As specified in Kortbeek et al. (2014), all patients of one cohort χ 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, because at such a point the maximum amount of information on possible admissions and discharges can be 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 care unit k . For notational convenience we introduce the function $v_{q,t}^\chi$ as $v_{q,t}^\chi = h_{q,t}^\chi$ for the elective patients, and $v_{q,t}^\chi = g_{q,t}^\chi$ for acute patient types. In addition, for each patient cohort, we define for $b_\tau \leq t < b_\tau + \ell_\tau$ the conditional distribution $v_{q,t}^{\chi, z_\chi}$, with $v_{q,t}^{\chi, z_\chi}(x_\chi)$ the probability that x_χ patients of cohort χ 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_χ . Then, the coverage compliance, given that census configuration \mathbf{y} is observed at the start of shift (q, τ) , is

$$c_{q,\tau}^k(d_{q,\tau}, f_{q,\tau}, r_{q,\tau}^k; \mathbf{y}) = \sum_{z_1, \dots, z_\Phi} \sum_{\substack{\chi \in \mathcal{W}^k \\ k=1, \dots, K}} \left\{ \prod_{\chi=1}^{\Phi} v_{q,b_\tau}^\chi(z_\chi) \right\}$$

$$\frac{1}{\ell_\tau} \sum_{t=b_\tau}^{b_\tau + \ell_\tau + 1} \sum_{x_\chi: \forall \chi \in \mathcal{W}^k} \mathbb{1} \left(\sum_{\chi \in \mathcal{W}^k} x_\chi \leq r_{q,\tau}^k \cdot s_{q,\tau}^k(\mathbf{y}) \right) \left\{ \prod_{\chi} v_{q,t}^{\chi, z_\chi}(x_\chi) \right\}$$

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 $d_{q,\tau}^1, \dots, d_{q,\tau}^K, f_{q,t}$ is of the order:

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

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

Appendix C. Derivation maximum census

In this appendix, $\hat{W}_{q,\tau}^k$ is derived, which represents 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 $t \in \{(q, b_\tau), \dots, (q, b_\tau + \ell_\tau - 1)\}$ the number of patients of this cohort reaches its maximum.

We first determine for each cohort χ , the probability distribution $w_{q,\tau}^\chi$ for the maximum number of patients of this cohort present during shift (q, τ) . Because all patients of one cohort are preferably placed on the same care unit, to obtain the probability distribution $W_{q,\tau}^k$ for the maximum demand for unit k during shift (q, τ) , we take the discrete convolution over the distributions $w_{q,\tau}^\chi$ relevant to unit k . Finally, from the maximum demand distribution $W_{q,\tau}^k$, the maximum census distribution $\hat{W}_{q,\tau}^k$ is obtained by applying the same transformation as was done for $Z_{q,\tau}^k$ and $\hat{Z}_{q,\tau}^k$ in Eq. (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 $\Delta^{\text{MSS}}(q, n)$:

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

Also, note that the definition of the cohorts implies that the combination of day q and cohort χ uniquely defines the number of days the patients of this cohort is already present after surgery; let us denote this value by $N(\chi, q)$. For elective patients, $w_{q,\tau}^\chi$ is defined if \exists an operating room i such that $\chi \in b_{i, \Delta^{\text{MSS}}(q, N(\chi, q))}$, and it can be calculated as follows:

$$w_{q,\tau}^\chi = \begin{cases} h_{N(\chi, q), b_\tau}^\chi & \text{if } N(\chi, q) = 1, \dots, L^\chi, \\ h_{0, b_\tau}^\chi & \text{if } N(\chi, q) = 0, \vartheta_\chi < b_\tau, \\ h_{0, \vartheta_\chi}^\chi & \text{if } N(\chi, q) = 0, b_\tau \leq \vartheta_\chi < b_\tau + \ell_\tau, \\ h_{0, b_\tau + \ell_\tau - 1}^\chi & \text{if } N(\chi, q) = 0, \vartheta_\chi \geq b_\tau + \ell_\tau, \\ h_{-1, b_\tau + \ell_\tau - 1}^\chi & \text{if } N(\chi, q) = -1, b_\tau + \ell_\tau \leq T, \\ h_{-1, T + \vartheta_\chi}^\chi & \text{if } N(\chi, q) = -1, b_\tau + \ell_\tau > T, \vartheta_\chi < b_\tau + \ell_\tau - T, \\ h_{-1, b_\tau + \ell_\tau - 1}^\chi & \text{if } N(\chi, q) = -1, b_\tau + \ell_\tau > T, \vartheta_\chi \geq b_\tau + \ell_\tau - T. \end{cases}$$

Acute patients: Let $\Delta^{\text{AAC}}(q, n)$ 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:

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

Recall that an acute patient type is identified by (p, r, θ) . Observe that an acute patient cohort χ is specified by the combination of a patient type j and a specific admission day. For acute patients, the

combination of day q and cohort χ again uniquely defines the number of days the patients of this cohort is already present; let us denote this value by $M(\chi, q)$. 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 $w_{q,\tau}^\chi$ is calculated by

$$w_{q,\tau}^\chi = \begin{cases} g_{M(\chi,q),b_\tau}^\chi & \text{if } M(\chi, q) = 1, \dots, L^\chi, \chi \text{ such that } \Delta^{AAC}(q, M(\chi, q)) = r, \\ g_{0,b_\tau}^\chi & \text{if } M(\chi, q) = 0, \theta < b_\tau, \chi \text{ such that } \Delta^{AAC}(q, M(\chi, q)) = r, \\ g_{0,\theta}^\chi & \text{if } M(\chi, q) = 0, b_\tau \leq \theta < b_\tau + \ell_\tau, \chi \text{ such that } \\ & \Delta^{AAC}(q, M(\chi, q)) = r, \\ g_{0,\theta}^\chi & \text{if } M(\chi, q) = 0, b_\tau + \ell_\tau > T, \theta < b_\tau + \ell_\tau - T, \chi \text{ such that } \\ & \Delta^{AAC}((q+1) \bmod Q + Q \cdot \mathbb{1}_{((q+1) \bmod Q = 0)}, M(\chi, q)) = r. \end{cases}$$

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

References

- Aiken, L.H., Clarke, S.P., Sloane, D.M., Sochalski, J., Silber, J.H., 2002. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *J. Am. Med. Assoc.* 288 (16), 1987–1993.
- Aiken, L.H., Sermeus, W., van den Heede, K., Sloane, D.M., Busse, R., McKee, M., Bruyneel, L., Rafferty, A.M., Griffiths, P., Moreno-Casbas, M.T., Tishelman, C., Scott, A., Brzostek, T., Kinnunen, J., Schwendimann, R., Heinen, M., Zikos, D., Strömberg, Sjetne, I., Smith, H.L., Kutney-Lee, A., 2012. Patient safety, satisfaction, and quality of hospital care: cross sectional surveys of nurses and patients in 12 countries in Europe and the United States. *Br. Med. J.* 344 (3), 1–14.
- Aiken, L.H., Sloane, D.M., Cimiotti, J.P., Clarke, S.P., Flynn, L., Seago, J.A., Spetz, J., Smith, H.L., 2010. Implications of the California nurse staffing mandate for other states. *Health Serv. Res.* 45 (4), 904–921.
- Beliën, J., Demeulemeester, E., 2008. A branch-and-price approach for integrating nurse and surgery scheduling. *Eur. J. Oper. Res.* 189 (3), 652–668.
- Brandea, M.L., Sainfort, F., Pierskalla, W.P. (Eds.), 2004. *Operations Research and Health Care: a Handbook of Methods and Applications*, International Series in Operations Research & Management Science, vol. 70. Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Broyles, J.R., Cochran, J.K., Montgomery, D.C., 2010. A statistical Markov chain approximation of transient hospital inpatient inventory. *Eur. J. Oper. Res.* 207 (3), 1645–1657.
- Burke, E.K., de Causmaecker, P., vandenBerghe, G., van Landeghem, H., 2004. The state of the art of nurse rostering. *J. Sched.* 7 (6), 441–499.
- Cheang, B., Li, H., Lim, A., Rodrigues, B., 2003. Nurse rostering problems—a bibliographic survey. *Eur. J. Oper. Res.* 151 (3), 447–460.
- Chiaromonte, M.V., Chiaromonte, L.M., 2008. An agent-based nurse rostering system under minimal staffing conditions. *Int. J. Prod. Econ.* 114 (2), 697–713.
- de Causmaecker, P., vanden Berghe, G., 2011. A categorisation of nurse rostering problems. *J. Sched.* 14 (1), 3–16.
- de Véricourt, F., Jennings, O.B., 2011. Nurse staffing in medical units: a queueing perspective. *Oper. Res.* 59 (6), 1320–1331.
- Dellaert, N., Jeunet, J., Mincovics, G., 2011. Budget allocation for permanent and contingent capacity under stochastic demand. *Int. J. Prod. Econ.* 131 (1), 128–138.
- Elkhuizen, S.G., Bor, G., Smeenk, M., Klazinga, N.S., Bakker, P.J.M., 2007. Capacity management of nursing staff as a vehicle for organizational improvement. *BMC Health Serv. Res.* 7 (1), 196–205.
- Ernst, A.T., Jiang, H., Krishnamoorthy, M., Sier, D., 2004. Staff scheduling and rostering: a review of applications, methods and models. *Eur. J. Oper. Res.* 153 (1), 3–27.
- Gnanlet, A., Gilland, W.G., 2009. Sequential and simultaneous decision making for optimizing health care resource flexibilities. *Decis. Sci.* 40 (2), 295–326.
- Griffiths, J.D., Price-Lloyd, N., Smithies, M., Williams, J.E., 2005. Modelling the requirement for supplementary nurses in an intensive care unit. *J. Oper. Res. Soc.* 56 (2), 126–133.
- Harper, P.R., Powell, N.H., Williams, J.E., 2010. Modelling the size and skill-mix of hospital nursing teams. *J. Oper. Res. Soc.* 61 (5), 768–779.
- Hulshof, P.J.H., Kortbeek, N., Boucherie, R.J., Hans, E.W., Bakker, P.J.M., 2012. Taxonomic classification of planning decisions in health care: a review of the state of the art in OR/MS. *Health Syst.* 1 (2), 129–175.
- Jack, E.P., Powers, T.L., 2009. Volume flexible strategies in health services: a research framework. *Prod. Oper. Manag.* 13 (3), 230–244.
- Kane, R.L., Shamiyan, T.A., Mueller, C., Duval, S., Wilt, T.J., 2007. The association of registered nurse staffing levels and patient outcomes: systematic review and meta-analysis. *Med. Care* 45 (12), 1195–1204.
- Kellogg, D.L., Walczak, S., 2007. Nurse scheduling: from academia to implementation or not? *Interfaces* 37 (4), 355.
- Kortbeek, N., Braaksma, A., Smeenk, F.H., Bakker, P.J., Boucherie, R.J., 2014. Integral resource capacity planning for inpatient care services. *J. Oper. Res. Soc.* <http://dx.doi.org/10.1057/jors.2014.67>.
- Lang, T.A., Hodge, M., Olson, V., Romano, P.S., Kravitz, R.L., 2004. Nurse–patient ratios: a systematic review on the effects of nurse staffing on patient, nurse employee, and hospital outcomes. *J. Nurs. Adm.* 34 (7–8), 326–337.
- Lavieri, M.S., Puterman, M.L., 2009. Optimizing nursing human resource planning in British Columbia. *Health Care Manag. Sci.* 12 (2), 119–128.
- Li, L.L.X., King, B.E., 1999. A healthcare staff decision model considering the effects of staff cross-training. *Health Care Manag. Sci.* 2 (1), 53–61.
- Needleman, J., Buerhaus, P., Mattke, S., Stewart, M., Zelevinsky, K., 2002. Nurse-staffing levels and the quality of care in hospitals. *New Engl. J. Med.* 346 (22), 1715–1722.
- Oddoye, J.P., Jones, D.F., Tamiz, M., Schmidt, P., 2009. Combining simulation and goal programming for healthcare planning in a medical assessment unit. *Eur. J. Oper. Res.* 193 (1), 250–261.
- Oddoye, J.P., Yaghoobi, M.A., Tamiz, M., Jones, D.F., Schmidt, P., 2007. A multi-objective model to determine efficient resource levels in a medical assessment unit. *J. Oper. Res. Soc.* 58 (12), 1563–1573.
- Pierskalla, W.P., Brailer, D.J., 1994. Applications of operations research in health care delivery. In: Pollock, S.M., Rothkopf, M.H., Barnett, A. (Eds.), *Operations Research and the Public Sector*, Handbooks in OR & MS, vol. 6. North-Holland, Amsterdam, The Netherlands, pp. 469–505.
- Sifer, S.P., Benton, W.C., 1992. Workforce staffing and scheduling: hospital nursing specific models. *Eur. J. Oper. Res.* 60 (3), 233–246.
- Smith-Daniels, V.L., Schweikhart, S.B., Smith-Daniels, D.E., 1988. Capacity management in health care services: review and future research directions. *Decis. Sci.* 19 (4), 889–919.
- Stewart, B.D., Webster, D.B., Ahmad, S., Matson, J.O., 1994. Mathematical models for developing a flexible workforce. *Int. J. Prod. Econ.* 36 (3), 243–254.
- Trivedi, V.M., Warner, D.M., 1976. A branch and bound algorithm for optimum allocation of float nurses. *Manag. Sci.* 22 (9), 972–981.
- Twigg, D., Duffield, C., Bremner, A., Rapley, P., Finn, J., 2011. The impact of the nursing hours per patient day (NHPPD) staffing method on patient outcomes: a retrospective analysis of patient and staffing data. *Int. J. Nurs. Stud.* 48 (5), 540–548.
- Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., De Boeck, L., 2013. Personnel scheduling: a literature review. *Eur. J. Oper. Res.* 226 (3), 367–385.
- Van Oostrum, J.M., van Houdenhoven, M., Hurink, J.L., Hans, E.W., Wullink, G., Kazemier, G., 2008. A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR Spectr.* 30 (2), 355–374.
- Walts, L.M., Kapadia, A.S., 1996. Patient classification system: an optimization approach. *Health Care Manag. Rev.* 21 (4), 75.
- Wright, P.D., Bretthauer, K.M., Côté, M.J., 2006. Reexamining the nurse scheduling problem: staffing ratios and nursing shortages. *Decis. Sci.* 37 (1), 39–70.
- Yankovic, N., Green, L.V., 2011. Identifying good nursing levels: a queueing approach. *Oper. Res.* 59 (4), 942–955.