Generating balanced workload allocations in hospitals

Pieter Smet

KU Leuven, Department of Computer Science
Gebroeders de Smetstraat 1, 9000 Gent, Belgium
pieter.smet@kuleuven.be

Abstract

As pressure on healthcare systems continues to increase, it is becoming more and more important for hospitals to properly manage the high workload levels of their staff. Ensuring a balanced workload allocation between various groups of employees in a hospital has been shown to contribute considerably towards creating sustainable working conditions. However, allocating work to different organizational units in a fair manner is not straightforward when it involves complex decision-making processes. In this paper we set out to balance the workload of heterogeneous hospital wards by optimizing the patient admission scheduling problem. Given the multi-period nature of patient admission scheduling, we introduce a new equity objective that captures both spatial (between hospital wards) and temporal (between days in the planning period) workload balancing. The resulting bi-objective problem is solved using an exact criterion space search algorithm. Our computational study employs problem instances that have been generated based on real-world data. The results demonstrate how spatially and temporally balanced workload allocations can be generated by minimizing the proposed equity objective. Moreover, we analyze sets of non-dominated solutions to gain various insights into the trade-off between schedule cost and workload balance.

Keywords: workload balancing, fairness, patient admission scheduling, bi-objective optimization

1 Introduction

Due to improved healthcare and declining birth rates, the global population is ageing. This demographic shift poses a range of significant challenges for governments, businesses and society in general. One sector that is profoundly affected by this evolution is healthcare. As older people become increasingly reliant on various care services, many health systems are coming under immense pressure. This situation was worsened by the challenges posed by the COVID-19 pandemic, with hospitals now experiencing even higher workloads (World Health Organization, 2022). However, this increase in the demand for healthcare services has not led to a proportional growth in the number of healthcare workers. The negative effects of high workload on the performance and well-being of hospital employees have been well documented. Several studies have confirmed high workload as a predictor for burnout, which in turn may lead to reduced professional efficacy, poor quality of care and employees quitting their job (Dall'Ora et al., 2020). Allocating workload in such a way that it is balanced between different groups of employees is an important measure which hospitals can take to improve working conditions (van den Oetelaar et al., 2020).

This paper introduces an approach for balancing the workload of heterogeneous wards in a hospital by optimizing the patient admission scheduling problem. A set of patients with varying

demands must be admitted within their admission time window to a hospital ward that is capable of providing them with their required care. Depending on the patient's treatment and how far along they are in their recovery, the workload may vary. The quality of an admission schedule is determined based on two criteria: (i) the operational costs related to the delay in admission and utilization of the operating theater, and (ii) how well the workload between different wards is balanced. While algorithms for minimizing operational costs in patient admission scheduling have been introduced in the past (Demeester et al., 2010), generating admission schedules that result in balanced workload allocations has not yet been addressed, despite its practical relevance. Due to the multi-period nature of patient admission scheduling, workload balancing is considered from both a *spatial* and *temporal* perspective. In spatial workload balancing, the goal is to arrive at an admission schedule which results in a similar patient workload allocation for each ward. By contrast, temporal workload balancing requires the workload of a single ward to be consistent across multiple periods (for example across different days in a planning period). In this paper, we consider spatial and temporal workload balancing to be equally important.

A common approach for generating fair or balanced solutions is to use a so-called *inequality measure* to quantify the fairness of an allocation as a single number (Matl et al., 2019). An inequality measure consists of two main components: (i) a workload metric that defines the workload and how to compute it, and (ii) an equity function to transform a vector of workloads into a scalar. We model our problem as a bi-objective problem in which one objective represents the operational costs, while the other objective involves an inequality measure that should be minimized to generate spatially and temporally balanced workload allocations (the equity objective). We consider all problem elements to be deterministic such that all patients and their characteristics are known at the time of solving. Using this approach, we generate a set of non-dominated solutions which we can then analyze to gain insights concerning the trade-off between minimizing operational costs and balancing workload.

The remainder of this paper is organized as follows. Section 2 reviews related literature on workload balancing in three important application domains of operations research. Section 3 defines the bi-objective optimization problem. Section 4 introduces a multi-period equity function and workload metric for spatial and temporal workload balancing. Section 5 presents the algorithm employed to generate solutions for the bi-objective problem. Section 6 discusses the results of a series of computational experiments in which we analyze the algorithm's performance and the solutions generated. Finally, Section 7 concludes the paper and identifies directions for future research.

2 Literature review

Workload balancing has received considerable attention over the past two decades in the academic literature on operations research. Much of this research can be presented within the same framework: activities are assigned to agents (for example, employees or machines) to meet a certain demand, thereby establishing the agents' workload. The definition of balanced workload is problem-specific and varies depending on the specific setting. Our literature review is therefore structured around three major application domains of operations research: manufacturing, logistics and healthcare. We focus on the equity functions and workload metrics that have been proposed to get an overview of how balanced workload allocations are generated for different types of problem. Note that this review is not exhaustive given that workload balancing and fairness have been considered in other applications as well, as discussed by Karsu and Morton (2015); Bektaş and Letchford (2020).

2.1 Manufacturing

Balancing machine workload to remove bottlenecks and reduce idle time is one means by which to improve resource utilization in manufacturing (Schwerdfeger and Walter, 2018). The traditional makespan objective $C_{\max}$ is an example of a commonly used equity objective. On identical parallel machines, minimizing $C_{\max}$ results in balanced workload allocations, as permitted by other constraints. While most studies consider a machine's makespan to be representative of its workload, there is no clear consensus concerning the most appropriate equity function. Ho et al. (2009) minimize the normalized sum of squared workload deviations, which has been shown to be equivalent to minimizing the sum of squared completion times. Ouazene et al. (2014) minimize the difference between the maximal and minimal machine completion times. The relation between these two alternative equity functions was established by Ouazene et al. (2016). Efficient algorithms for these problems and some of their special cases have been developed by Cossari et al. (2012); Schwerdfeger and Walter (2016); Christ et al. (2019). Other equity functions that have been proposed include the relative percentage of imbalance (which is equivalent to minimizing $C_{\max}$) (Rajakumar et al., 2004) and normalized mean difference (Cossari et al., 2013).

Products are often manufactured on assembly lines where workpieces are sent through a sequence of work stations. Each station contributes to the workpiece by carrying out one or more tasks. Designing an assembly line in such a way that the workload of stations is balanced can have several benefits: increasing production output, reducing machine breakdown, improving ergonomics and reducing dissatisfaction among workers (Finco et al., 2020). In many studies, a station's workload is computed as the sum of assigned task durations. Some examples of equity functions that have been proposed for this workload metric include the average squared mean workload deviation (Rachamadugu and Talbot, 1991) and the sum of squared mean workload deviations (Walter et al., 2021). Azizoğlu and Imat (2018) and Walter (2020) investigate an alternative equity objective that minimizes the sum of squared idle times.

2.2 Logistics

The primary objective in many vehicle routing problems (VRPs) concerns operational costs such as the distance driven or route duration. As in many other sectors, drivers are becoming increasingly difficult to recruit due to poor working conditions. One way to remedy this situation is by ensuring a balanced allocation of workload to drivers, which typically improves driver job engagement and quality of service (Liu et al., 2006). The VRP with route balancing was first introduced by Jozefowiez et al. (2002) as a bi-objective problem in which one objective function minimizes total distance while the other minimizes the difference between the longest and shortest route lengths.

Several other equity functions for VRPs have been proposed over the years. Examples include minimizing the longest route length (Reiter and Gutjahr, 2012), minimizing route length in lexicographical order (Saliba, 2006), minimizing the sum of deviations from a mean or target route length (Halvorsen-Weare and Savelsbergh, 2016), and minimizing the range of route lengths in two stages (Sartori et al., 2021). For an extensive discussion of these different equity functions and their properties, we refer interested readers to Matl et al. (2018).

Workload in VRPs is typically defined as route length in terms of distance or duration. Duration-based equity objectives are similar to their distance-based counterparts, but they may prove more challenging to compute. This is especially the case when considering flexible departure times. Matl et al. (2019) investigate how alternative workload metrics based on demand/load and the number of stops/customers affect the obtained workload allocations. One of their main conclusions is that the trade-off between cost and balance primarily depends on the workload metric and not on the equity function.

In periodic VRPs, a driver can experience different workloads in each period. Existing

approaches to obtain balanced allocations include minimizing the difference between the largest and smallest number of customers served during the complete planning period (Gulczynski et al., 2011) and minimizing the duration of the longest route in the complete planning period (Liu et al., 2013). Wang et al. (2022) solve a dynamic stochastic periodic VRP in which workload should be balanced within each period as well as over the complete planning period. Periodic balancing is realized by minimizing the maximum route length, while global balancing is achieved by minimizing the difference between the highest and lowest route revenue.

2.3 Healthcare

Healthcare workers are considered a scarce and costly resource in hospitals. It is thus unsurprising that ensuring sustainable working conditions for nurses and physicians is considered a high priority for hospital managers. Several studies have shown how fairness (or the lack thereof) has a strong correlation with absenteeism (De Boer et al., 2002) and pessimistic decision making (Robbins et al., 2012). In contrast to manufacturing and logistics, there is much more variation in which inequality measures are used to arrive at fair schedules in healthcare environments.

In physician and nurse scheduling, a frequently employed fairness criterion relates to the preferences of the medical and nursing staff. Gross et al. (2019) propose an approach to balance the number of granted requests over long periods of time. Based on the value of a fairness indicator, the weights associated with the requests are updated to ensure each physician has approximately the same number of granted requests in the long term. Other desirable properties associated with fair staff schedules include an even distribution of working hours and on-call services (Stolletz and Brunner, 2012; Fügener et al., 2015) as well as an even distribution of violations of constraints associated with sustainable working conditions (Martin et al., 2013). Examples of equity functions include minimizing the maximum allocated value, minimizing the difference between the largest and smallest allocated value and minimizing expressions based on quadratic sums.

Mullinax and Lawley (2002) solve a patient-to-nurse assignment problem in which the goal is to balance nurse workload. Each patient is associated with a workload, depending on their specific medical condition, and the proposed algorithm minimizes the difference between the highest and lowest nurse workload. Schaus et al. (2009) point out flaws in this approach and instead suggest to minimize the L_2 norm of the nurses' workload. Mandelbaum et al. (2012) use queuing theory to address the flow of patients from the emergency department to the various wards of a hospital. To ensure a balanced workload allocation among the wards, two metrics are proposed that take into account bed occupancy levels and bed turnover rates. Mandelbaum et al. (2012) highlight how only considering bed occupancy is not sufficient as some wards may typically have shorter patient stays, resulting in high bed turnover. Because workload associated with a patient is typically higher during their first days of their stay, bed occupancy alone is an unsuitable metric for how fair an allocation is.

3 Problem description

We consider a set of patients who must be admitted to the hospital for urgent or elective surgery. Both sets of patients must also remain in the hospital for a recovery period after surgery. Each patient is characterized by their required surgical procedure and their earliest and latest day of admission. Patients who are admitted for elective surgery have some flexibility when it comes to scheduling their surgery, as it typically does not involve a medical emergency. Urgent or emergency patients have less flexibility and usually have a narrow time window for their admission. Each surgical procedure belongs to a surgical discipline and is characterized by an expected surgery duration and recovery length of stay (LOS). We consider a set of wards where each ward is associated with its own bed capacity, nursing staff and specialization(s) in surgical

disciplines. In addition to one major specialization, a ward may also treat patients in other disciplines for which it is less well-equipped or for which it has less trained personnel. Treating patients for these so-called *minor specializations* places increased strain on the ward's staff, thereby increasing their perceived workload. We also consider the hospital's operating theater (OT) consisting of one or more operating rooms. The OT's master surgery schedule (MSS) is given and defines the amount of available OT time for each surgical discipline on each day.

The goal is to assign each patient to a ward that has a major or minor specialization in the patient's surgical discipline and also has available bed capacity. Moreover, a feasible admission date must be determined that falls within the patient's admission time window. We assume that a patient cannot change wards while they are in the hospital. The quality of a solution is determined by the following three aspects:

- *Admission delays*: while each patient has an admission time window during which they must be admitted to the hospital, it is preferred to admit a patient as early as possible within this time window. However, postponement of an admission to a later date in the time window is sometimes inevitable due to capacity constraints. In such cases the length of time between the first possible day and the actual day of admission should be as short as possible. In other words, delays in the patients' admission should be minimized.

- *OT utilization*: we assume patients are scheduled to undergo surgery on the day they are admitted to the hospital. The available OT time for each surgical discipline is limited on each day, as determined by the MSS. When scheduling patient admissions, the available OT time for the different surgical disciplines should be respected as much as possible. OT over- and undertime should therefore be minimized.

- *Workload balancing*: given the scarcity of hospital staff, it is vital to provide sustainable working conditions. To do so, we balance workload both between wards and within individual wards over time. Large imbalances in workload between wards should be avoided by allocating the work more fairly. Moreover, avoiding large fluctuations of the workload of an individual ward over time is also import to ensure stable working conditions and to avoid over- or understaffing.

The problem is modeled as a bi-objective optimization problem in which the first objective function concerns the schedule costs, defined as the weighted sum of admission delays and OT utilization. The second objective function is an equity objective for spatial and temporal workload balancing and will be further discussed in the next section. Table 3 provides an overview of the notation we use throughout the remainder of the paper.

4 Multi-period workload balancing

We consider an allocation of work spatially balanced if each agent has a similar workload. Spatial workload balancing in the context of patient admission scheduling is challenging as the schedule is constructed for a long planning period such as a week, while the workload should be balanced with respect to shorter periods such as each day. If balancing with respect to these shorter periods is ignored and instead the total workload accumulated over the planning period is balanced, individual imbalanced days may cancel each other out. By way of example, Figure 1 shows two possible workload allocations of three wards (W1, W2 and W3) over four days (D1, D2, D3 and D4). In both solutions, the wards have the same total workload (12). Figure 1a shows the desired solution in which the workload is spatially balanced between wards on each day. Figure 1b, on the other hand, shows a solution in which the total workload is the same for all wards, but there are days on which there is a considerable imbalance in ward workload. These imbalanced days negate each other, resulting in an identical total workload for all wards.

Symbol	Description
W	set of wards in the hospital, indexed by w
D	set of days in the planning period, indexed by d
P	set of patients, indexed by p
S	set of surgical disciplines, indexed by s
$P_{sd} \subseteq P$	set of patients of discipline s that can be admitted on day d
$P_w \subseteq P$	set of patients that can be admitted to ward w based on their surgical discipline
$W_p \subseteq W$	set of wards to which patient p can be admitted based on their surgical discipline
b_w	number of beds available in ward w
q_{sd}	OT capacity for discipline s on day d as determined by the MSS, in minutes
$f_p \in D$	earliest possible admission date of patient p
$e_p \in D$	latest possible admission date of patient p
$D_p^A \subseteq D$	set of all possible admission days of patient p, defined as the set $\{f_p, \ldots, e_p\}$
l_p	expected recovery length of stay of patient p, in days
u_p	expected surgery duration of patient p, in minutes
W^{UT}	weight of OT under-utilization in the objective function
W^{OT}	weight of OT overtime in the objective function
W^{WAIT}	weight of patient admission delay in the objective function

Table 1: Overview of the used notation.

	D1	D2	D3	D4	Total
W1	3	1	1	7	12
W2	3	1	1	7	12
W3	3	1	1	7	12

(a) Spatial balance with respect to individual days

	D1	D2	D3	D4	Total
W1	2	6	1	3	12
W2	7	2	1	2	12
W3	3	0	3	6	12

(b) Spatial balance with respect to the complete planning period

Figure 1: Spatially balanced workload allocations.

A related problem is to find temporally balanced solutions for a multi-period problem. In temporally balanced solutions, the workload of an individual ward is comparable on each day of the planning period. Similarly to the scenario outlined for spatial balancing, imbalanced wards may cancel each other out on different days. Figure 2 illustrates this scenario with two example solutions. While both solutions have the same total workload on each day (9), there are considerable differences when examining the daily workload of each individual ward. The solution shown in Figure 2a is the desired solution in which the workload of each ward is perfectly temporally balanced. Figure 2b, on the other hand, shows large variations in a ward's workload on different days, even though the total workload on each day is the same.

	D1	D2	D3	D4
W1	3	3	3	3
W2	1	1	1	1
W3	5	5	5	5
Total	9	9	9	9

(a) Temporal balance with respect to individual wards

	D1	D2	D3	D4
W1	3	5	3	1
W2	2	2	5	3
W3	4	2	1	5
Total	9	9	9	9

(b) Temporal balance with respect to workload aggregated over all wards

Figure 2: Temporally balanced workload allocations

Note that neither of the solutions shown in Figure 2 are spatially balanced. Likewise, neither of the solutions in Figure 1 are temporally balanced. A solution which is both perfectly spatially and temporally balanced has the same workload for each ward on each day, as shown in the

solution of Figure 3.

	D1	D2	D3	D4	Total
W1	3	3	3	3	12
W2	3	3	3	3	12
W3	3	3	3	3	12
Total	9	9	9	9	

Figure 3: Example of a spatially and temporally balanced workload allocation

We are interested in patient admission schedules that result in both spatially and temporally balanced ward workload. Formally, let $\mathbf{X}$ be an $|W| \times |D|$ workload allocation matrix representing the workload allocation of $|W|$ wards on $|D|$ days. Let x_{wd} be the workload of ward w on day d. An inequality measure $I(\mathbf{X})$ is a function that returns a value based on an allocation matrix $\mathbf{X}$ that quantifies the spatial and temporal workload imbalance. In the following subsections, the two defining characteristics of $I(\mathbf{X})$ are discussed: the equity function and the workload metric.

4.1 Equity function

We propose a variant of the well-known *min max* equity function that minimizes the maximum workload. As discussed in Section 2, many other equity functions have been proposed in the academic literature which compare the workload of individual agents to the mean workload (such as variance or standard deviation) or which employ a pairwise comparison of all agents' workload (such as the Gini coefficient). While these functions have their own advantages, they are generally difficult for practitioners to interpret or may result in solutions with undesirable properties such as artificially increased workload. Moreover, they are often non-linear, which makes them less suitable for integer linear programming algorithms. Equation (1) provides our proposed equity function.

$$\texttt{max}\,(\mathbf{X}) = \max_{d \in D} \left(\max_{w \in W} x_{wd} \right) \tag{1}$$

Matl et al. (2018) list several criteria that an inequality measure should satisfy. While most of these affect optimized allocations in a small way, there are two that are considered particularly important: the Pigou-Dalton (PD) transfer principle and monotonicity. The PD principle concerns shifting workload from one agent to another. Such a transfer is called progressive if it benefits the agent with the highest workload, and regressive otherwise. The PD principle states that if a new allocation can be reached by a finite series of progressive transfers, the inequality measure should not be higher than before the transfer. In order to define monotonicity, let $\mathbf{X}'$ be formed as follows: $x'_{wd} = x_{wd} + \delta_{wd}$ for at least one w, d in $\mathbf{X}$. If $\delta_{wd} \geq 0$ for all w, d with at least one strict inequality, then $I(\mathbf{X}') \geq I(\mathbf{X})$. It is well known that the weak version of both the PD principle and monotonicity hold for the `max` function in single-period problems (Matl et al., 2018). It can be easily seen that this is also true for the $\texttt{max}\,(\mathbf{X})$ function with respect to spatial and temporal balancing.

To enable workload balancing, the problem must have sufficient decision flexibility: it should be possible to not only shift workload between agents, but also between periods. In the context of patient admission scheduling, this is true if patients can be assigned to different wards and their admission dates can be shifted. The degree to which workload balancing is possible therefore depends on two characteristics of the problem: ward specialization and patient flexibility. The former refers to the number of wards that share specializations, while the latter refers to the number of days on which a patient can be admitted. Note that at a minimum, the ability to engage in workload balancing requires at least two wards that share a specialization or at least one patient who can be admitted on two different days.

4.2 Workload metric

The workload of a ward is determined by the care required by the admitted patients, and depends on their specific surgical procedure in addition to whether the ward is well-equipped and has sufficiently-trained personnel to provide the necessary treatment during recovery. In our problem, the workload associated with a patient depends on whether they are admitted to a ward with a major or minor specialization in their surgical discipline. If the ward has a major specialization in a patient's surgical discipline, their workload will be lower than if the patient is admitted to a ward where it is a minor specialization.

Typically, the workload associated with a patient is not constant in time (Mandelbaum et al., 2012). For example, a patient will likely require the highest amount of care on the day they are admitted to a ward immediately following surgery. The amount of care they require will then gradually decrease over time.

The maximum capacity of a ward in terms of workload is primarily determined by its staff. Depending on the number of employees, their skills and their rosters, a ward is capable of carrying out a specific amount of work. In this paper we assume the maximum workload of a ward is constant. In other words, at each moment in time we assume that a similarly skilled team of employees is present in the ward. Depending on the ward's staff composition, this workload capacity may vary. To account for the non-identical nature of wards, workload is normalized by means of reference distribution.

Formally, the normalized workload is defined as follows. Let $\bar{AD}_{wd}$ be the set of patients who are being treated in ward w on day d. For each patient $p \in \bar{AD}_{wd}$, let θ_{pwt} be their associated workload in ward w on the t-th day of admission and let d'_p correspond to their admission date. The maximum workload of ward w is denoted as β_w. Equation (2) defines the normalized workload of ward w on day d. Note that due to the heterogeneous nature of wards in terms of their maximum workload and the applied normalization, the proposed workload metric is a variable-sum metric.

$$x_{wd} = \frac{1}{\beta_w} \sum_{p \in \bar{AD}_{wd}} \theta_{pw(d-d'_p)} \tag{2}$$

5 Solution approach

To solve the bi-objective problem defined in Section 3, we use the balanced box method together with integer programming to solve the subproblems. Section 5.1 introduces an integer programming formulation of the problem. The balanced box method is then reviewed in Section 5.2.

5.1 Integer programming formulation

For each $p \in P$, $w \in W_p$ and $d \in D_p^A$, let y_{pwd} be a binary variable which equals one if patient p is admitted to ward w on day d. A continuous variable x_{wd} equals the workload of ward w on day d, for each $w \in W$ and $d \in D$. A continuous variable z equals the maximum workload over all wards and days. Finally, for each $s \in S$ and $d \in D$, two continuous variables u_{sd} and v_{sd} equal the amount of OT under- and overtime, respectively, for surgical discipline s on day d. The integer programming problem can now be formulated as follows:

$$\min \sum_{s \in S} \sum_{d \in D} \left(W^{\mathrm{OT}} v_{sd} + W^{\mathrm{UT}} u_{sd} \right) + W^{\mathrm{WAIT}} \sum_{p \in P} \sum_{w \in W_p} \sum_{d=f_p+1}^{e_p} (d - f_p) y_{pwd} \tag{3}$$

$$\min z \tag{4}$$

$$s.t. \sum_{w \in W_p} \sum_{d \in D_p^A} y_{pwd} = 1 \qquad \forall p \in P \qquad (5)$$

$$\sum_{p \in P_{sd}} \sum_{w \in W_p} u_p y_{pwd} + u_{sd} - v_{sd} = q_{sd} \qquad \forall s \in S, d \in D \qquad (6)$$

$$\frac{1}{\beta_w} \sum_{p \in P_w} \sum_{d'=d-l_p+1}^{d} \theta_{pw(d-d'+1)} y_{pwd'} + v_{dw} = x_{wd} \qquad \forall w \in W, d \in D \qquad (7)$$

$$x_{wd} \leq z \qquad \forall w \in W, d \in D \qquad (8)$$

$$\sum_{p \in P_w} \sum_{d'=\max(f_p, d-l_p+1)}^{\min(|D|-l_p, d)} y_{pwd'} \leq b_w - h_{dw} \qquad \forall w \in W, d \in D \qquad (9)$$

$$y_{pwd} \in \{0,1\} \qquad \forall p \in P, w \in W_p, d \in D_p^A \qquad (10)$$

$$x_{wd} \geq 0 \qquad \forall w \in W, d \in D \qquad (11)$$

$$z \geq 0 \qquad (12)$$

$$u_{sd} \geq 0 \qquad \forall w \in W \qquad (13)$$

$$v_{sd} \geq 0 \qquad \forall w \in W \qquad (14)$$

Objective function (3) minimizes the total cost of a schedule defined as the weighted sum of OT over- and under-utilization and patient admission delays. Objective function (4) optimizes workload balancing by minimizing the equity objective $\texttt{max}\,(\mathbf{X})$. Constraints (5) ensure each patient is admitted to a feasible ward within their admission time window. Constraints (6) model the available OT capacity for each surgical discipline and ensure the under- and overtime variables are set correctly. Constraints (7) compute the normalized workload of each ward on each day. Note that the workload carried over from the preceding planning period is also taken into account here. Let v_{dw} be the workload on day d in ward w that originates from patients who were admitted in the preceding planning period, but who have not yet been discharged by day d. Constraints (8) link the z and x_{wd} variables. Constraints (9) ensure the wards' maximum bed capacity is never exceeded. These constraints use parameter h_{wd} to represent the number of beds occupied in ward w on day d by patients who were admitted in the preceding planning period, but who have not yet been discharged by day d. Constraints (10)-(14) enforce bounds on the decision variables.

5.2 Balanced box method

We use the balanced box method proposed by Boland et al. (2015) to generate the complete set of non-dominated solutions. This criterion space search algorithm maintains a diverse set of non-dominated solutions throughout its execution, which is iteratively updated by solving a series of single-objective problems. The algorithm begins by constructing a rectangular area in the criterion space (a so-called *box*) defined by two lexicographically optimal solutions. This initial box is then subdivided into a lower and an upper box. The algorithm proceeds by searching for non-dominated solutions in both boxes by again solving a series of single-objective problems. Each time a non-dominated solution is identified, the box in which it was found is updated by removing the areas that have become dominated. This process is repeated until all boxes have been explored and it is guaranteed that no additional non-dominated solutions exist.

To identify a new non-dominated solution in a box, a lexicographic objective optimization problem is solved. This is achieved by first optimizing one objective function, then adding a constraint to the model restricting the value of this objective function to the optimal value, and finally optimizing the second objective function. In our implementation, the single-objective optimization problems are solved using a commercial integer programming solver. Warm starts

are employed such that the solver starts from a given non-empty solution. Note that while the warm start solutions may be infeasible due to the constraint added after optimizing the first objective function, our preliminary computational experiments demonstrated that it reduces overall computation times compared to providing no warm start.

6 Computational study

To confirm that minimizing the `max` function results in spatial and temporal workload balance, a series of experiments is conducted and analyzed. Moreover, we also investigate the trade-off between schedule cost and workload balance. Section 6.1 details the data used and how problem instances are generated. Section 6.2 discusses the computation time required by the proposed method. Section 6.3 analyzes whether spatial and temporal workload balance is achieved and discusses the role of ward specialization in workload balancing. Finally, Section 6.4 discusses the trade-off between schedule cost and workload balance.

6.1 Instance generation and experimental setup

Parameter	Description
$\mathcal{W}$	number of wards
$\mathcal{D}$	number of days in the planning period
$\mathcal{M}$	number of minor specializations per ward
$\mathcal{B}$	bed shortage probability
$\mathcal{R}$	ratio urgent to non-urgent patients

Table 2: Parameters used during instance generation.

The problem instances used in the computational study have been generated based on historical data from a large hospital in Belgium using the procedure described by Vancroonenburg et al. (2019). The available data concerns several admissions with a surgical pathway. More specifically, the data describes several surgical disciplines, carried out procedures per discipline during a period of one year, procedure durations and recovery LOS per procedure. For each procedure, a log-normal distribution was fitted to both its duration and recovery LOS. Finally, this data also describes the MSS. For more details concerning this data, we refer interested readers to Vancroonenburg et al. (2019).

Given the parameters in Table 2, an instance is generated as follows. We assume each ward has one unique major specialization. The instance generation process begins by randomly choosing $\mathcal{W}$ surgical disciplines. Subsequently, $\mathcal{M}$ randomly chosen minor specializations are assigned to each ward. Note that the minor specializations of a ward always correspond to the major specializations of other wards.

For each surgical discipline s, patients are generated according to a Poisson process with arrival rate $\lambda_s = q_s^{\max}/E[u_s]$, where $E[u_s]$ is the expected duration of a surgery in discipline s and $q_s^{\max} = \max_{d \in D} q_{sd}$ the maximum available OT time for discipline s over all days in the planning period. For each patient, a surgical procedure within the discipline is chosen according to their relative frequencies. A patient's recovery LOS and surgery duration are determined based on the specific surgical procedure. With a probability of $\mathcal{R}$, the patient's latest admission date is set to their earliest admission date, thereby defining them as an emergency patient. Otherwise, the latest admission date is set to a date at least one day after their earliest admission date and at most seven days after. The base-level workload associated with a patient is set to 1. This base-level workload is increased with a random percentage chosen uniformly between 5% and 15% if a patient is admitted to a ward with a minor specialization in the their surgical discipline rather

than to a ward with major specialization. Moreover, we assume a linear decrease of workload over the patient's stay, beginning at 120% of the patient's workload on the first day of their admission and ending at 80%.

The number of beds per ward is generated using an Erlang loss queuing model (De Bruin et al., 2010). Let μ_s be the mean recovery LOS of the ward's major specialization, Equation (15) determines the probability B_b that there will be a bed shortage if the ward has b beds. Given a probability $\mathcal{B}$, the required number of beds to meet $\mathcal{B}$ is determined by enumerating b until $B_b \leq \mathcal{B}$. The maximum workload of a ward is also set to the final value of b.

$$B_b = \frac{(\lambda_s \mu_s)^b / b!}{\sum_{k=0}^{b} (\lambda_s \mu_s)^k / k!} \tag{15}$$

In order to generate realistic scenarios, we assume a number of patients is already present in the wards at the start of the planning period. The procedure used to generate these patients is the same as the one described above. However, each of the preexisting patients is assigned to the ward whose major specialization corresponds to their surgical discipline.

All experiments are carried out on an AMD Ryzen 9 5950X processor with 64GB RAM. Gurobi 9.5 is used to solve the integer programming problems with default settings and configured to use 16 threads. No time limit is imposed. The x_{wd} variables in model (3)-(14) are scaled with a factor $1e4$ and modeled as integer variables to avoid numerical issues with the solver.

6.2 Algorithm performance

Before analyzing the solutions generated, we will first discuss the computational performance of the algorithm described in Section 5. The data set used consists of 250 problem instances generated using the procedure described in Section 6.1. The instances consist of $\mathcal{W} = 4$ wards, a planning period of $\mathcal{D} = 7$ days, an emergency patient ratio of $\mathcal{R} = 0.3$ and a bed shortage probability of $\mathcal{B} = 0.1$. The resulting number of patients to be admitted in the planning period ranges from 64 to 142. Of each instance, four variants are generated with a different number of minor specializations per ward ($\mathcal{M} \in \{0, 1, 2, 3\}$), resulting in a data set consisting of 1000 unique problem instances. If each ward has three minor specializations ($\mathcal{M} = 3$) in addition to their unique major specialization, there is actually no longer any difference between the wards and all patients can be admitted to all wards. By contrast, if a ward has no minor specialization ($\mathcal{M} = 0$), a patient can only be admitted to one ward. The instances will be made publicly available upon acceptance of the manuscript.

Figure 4 shows the run time of the algorithm in seconds for different values of $\mathcal{M}$. The run time on an instance is included in the plot if the algorithm terminated successfully for all values of $\mathcal{M}$. For $\mathcal{M} = 0$, the algorithm finds the optimal solutions for each instance very quickly. The mean computation time is 4.4 seconds, while the maximum recorded time is 25.9 seconds. However, as the number of minor specializations per ward increases, the required computation time also increases. For instances with $\mathcal{M} = 3$, the mean computation time is 1011.8 seconds and the maximum recorded run time is 2718.9 seconds.

6.3 Achieving spatial and temporal workload balance

In this section we will investigate whether the proposed equity function $\texttt{max}(\mathbf{X})$ can be used to generate solutions with both temporal and spatial workload balance. The 1000 problem instances described in Section 6.2 are now solved as single-objective problems. For each instance, two solutions are compared: a cost-optimal solution obtained by minimizing Equation (3) and a balance-optimal solution obtained by minimizing Equation (4). For each solution, the variance in temporal and spatial workload balance is computed per ward and per day. More precisely, the temporal variance for ward w is computed as $1/|D| \sum_{d \in D} (x_{wd} - \mu_w)^2$, with μ_w the average

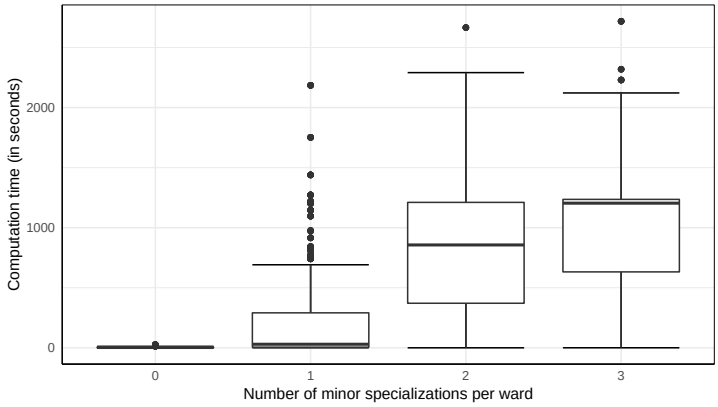

Figure 4: Run time of the algorithm for different values of $\mathcal{M}$.

workload of ward w over all days in D. The spatial variance on each day is computed in a similar fashion. Figure 5 plots these values for different values of $\mathcal{M}$.

The first conclusion from Figure 5 is that minimizing $\texttt{max}\,(\mathbf{X})$ indeed results in temporally and spatially balanced solutions. Both the temporal and spatial variances are generally small in the balance-optimal solutions. By contrast, the cost-optimal solutions are not always spatially and/or temporally balanced. This imbalance is particularly acute for larger values of $\mathcal{M}$. For $\mathcal{M} = 0$, both objective functions result in similar variances.

The second observation concerns only the balance-optimal solutions. For these solutions, the temporal variance remains more or less constant across different values of $\mathcal{M}$. By contrast, the spatial variance decreases as $\mathcal{M}$ increases. This demonstrates how ward specialization or generalization does not contribute much to balancing workload across time in individual wards, while it does have clear impact on spatial workload balancing.

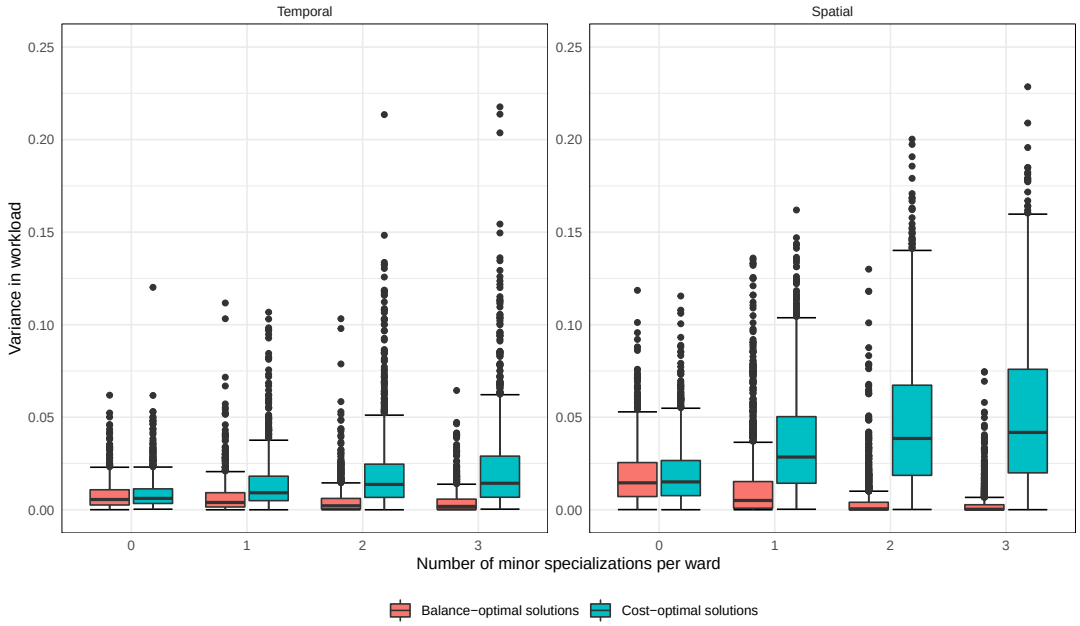

Figure 5: Temporal and spatial variance of ward workload.

By way of example, Figure 6 plots the daily ward workload in the balance-optimal and cost-optimal solutions for a randomly selected problem instance. The workload allocation in the balance-optimal solution is both spatially and temporally balanced. Spatial balance can be

identified by the fact that the wards' graphs remain relatively close to each other throughout the planning period. Temporal balance in the balance-optimal solution is confirmed by the horizontal stability of the wards' graphs. This indicates that for a given ward, workload fluctuates very little over time. By contrast, in the cost-optimal solution, workload is clearly spread out (spatial imbalance), but also fluctuates greatly over time for individual wards (temporal imbalance).

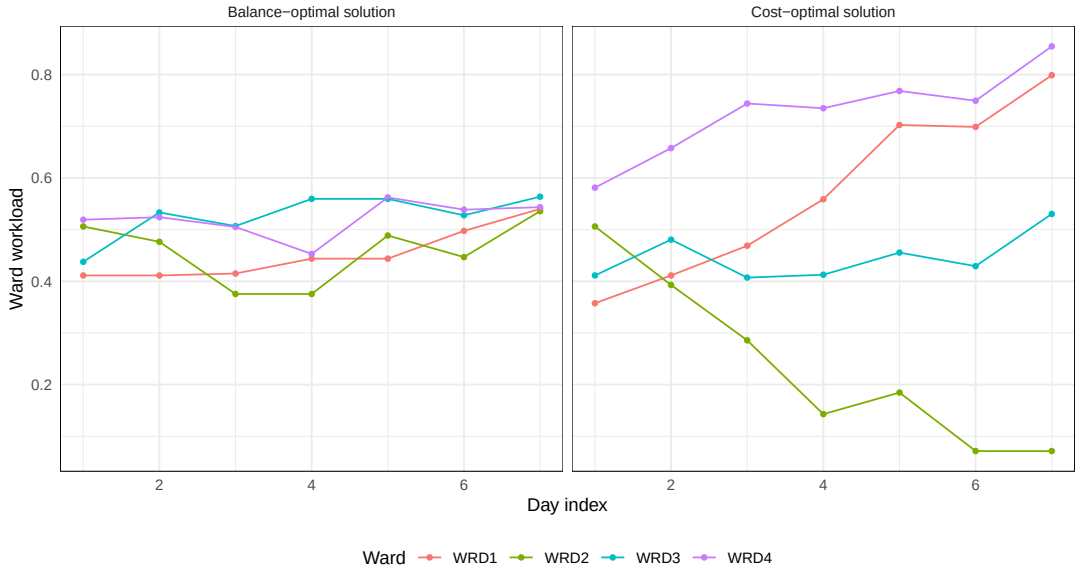

Figure 6: Workload allocation of $\mathcal{W} = 4$ wards in a planning period of $\mathcal{D} = 7$ days and $\mathcal{M} = 1$.

Figure 7 shows how workload balance changes when the number of minor specializations per ward is increased. Values are plotted for both the cost-optimal and balance-optimal solutions. In cost-optimal solutions, increasing $\mathcal{M}$ produces solutions with worse workload balance. This can be seen by the increasing values of $\texttt{max}\,(\mathbf{X})$ when increasing the number of minor specializations per ward. The worst workload allocation in terms of balance is obtained for $\mathcal{M} = 3$, which resulted in a mean value of $\texttt{max}\,(\mathbf{X}) = 1.04$. In the balance-optimal solutions, more balanced allocations of workload are possible when wards are capable of treating multiple disciplines as minor specializations. For $\mathcal{M} = 0$, the mean value of $\texttt{max}\,(\mathbf{X})$ is 0.74, which decreases to 0.59 for $\mathcal{M} = 2$ and $\mathcal{M} = 3$.

Improvements in workload balance can be achieved by making wards more general and by admitting patients to wards with a minor specialization. Typically, the number of patients in wards with minor specializations increases as wards have more minor specializations. While not explicitly accounted for in the current problem, these wards may incur additional hidden costs. Facilitating care for additional surgical disciplines may require cross-trained staff or the use of floating staff. Moreover, it requires physicians to visit multiple wards, which may lead to unnecessary delays concerning when the patient is discharged.

6.4 Trade-off schedule cost and workload balance

To analyze the trade-off between workload balance and schedule cost, the 1000 problem instances described in Section 6.2 are now solved as bi-objective problems using the algorithm detailed in Section 5. Figure 8 provides histograms showing the number of non-dominated solutions found for the instances when the value of $\mathcal{M}$ is varied. The largest number of non-dominated solutions is generated when wards have no or just one minor specialization. For $\mathcal{M} = 0$ and $\mathcal{M} = 1$, 421 and 479 non-dominated solutions are generated, respectively, compared to 346 and

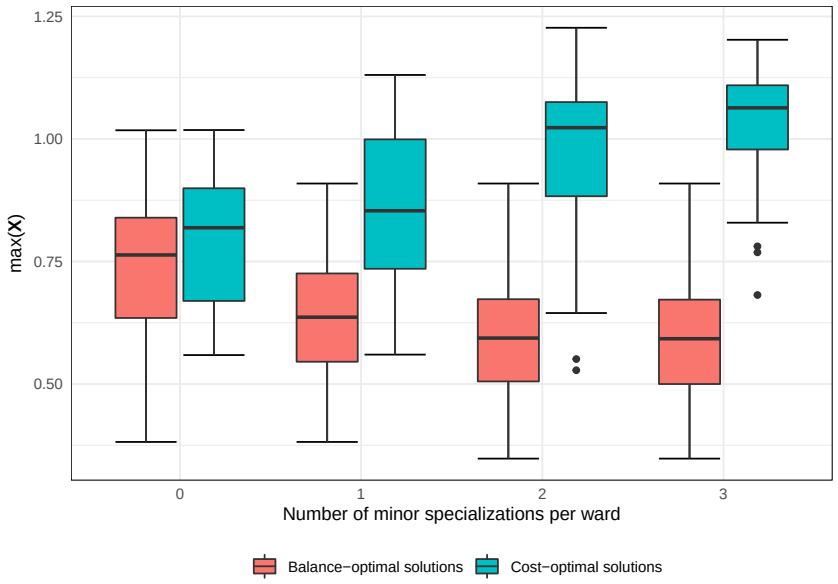

Figure 7: Impact on workload balance when varying the number of minor specializations per ward.

285 solutions when $\mathcal{M} = 2$ and $\mathcal{M} = 3$. This is somewhat surprising, as these settings have less decision flexibility compared to larger values of $\mathcal{M}$. Admission can typically only be shifted in time, as there are few or no choices to make concerning which ward a patient should be admitted to. This result indicates that the scenarios represented by the problem instances are relatively flexible in terms of admission time windows.

For $\mathcal{M} = 0$ and $\mathcal{M} = 1$, most problem instances only have a single non-dominated solution. These instances indicate scenarios in which there is no trade-off between cost and workload balance. In other words, the cost-optimal solution is also optimal in terms of workload balance. For larger values of $\mathcal{M}$ this is no longer true, as most instances then have two extreme solutions. In general, the number of non-dominated solutions is relatively small, which is characteristic of min-max equity functions (Matl et al., 2018).

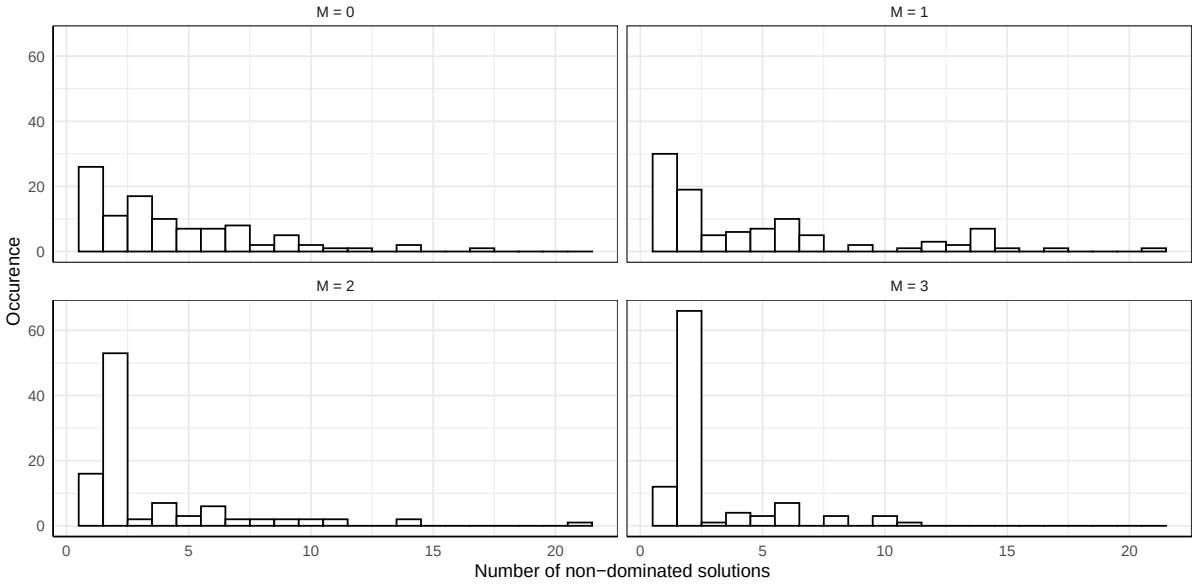

Figure 8: Histogram of the number of generated non-dominated solutions.

Figure 9 shows the Pareto fronts of three randomly selected problem instances for different values of $\mathcal{M}$. For instance 1, the set of non-dominated solutions is the same for $\mathcal{M} = 0$ and $\mathcal{M} = 1$. For larger values of $\mathcal{M}$, only two non-dominated solutions exist. Similar observations can be made for the other two instances. Comparing the Pareto fronts consisting of three or more solutions, it is evident that they do not share the same shape. For $\mathcal{M} = 0$, instance 3 exhibits a favorable trade-off between schedule cost and workload balance. Either of the two objectives can be improved without considerably worsening the other objective. However, this desirable property no longer holds for larger values of $\mathcal{M}$. Instances 1 and 2 never exhibit such a favorable trade-off, regardless of the number of minor specializations per ward.

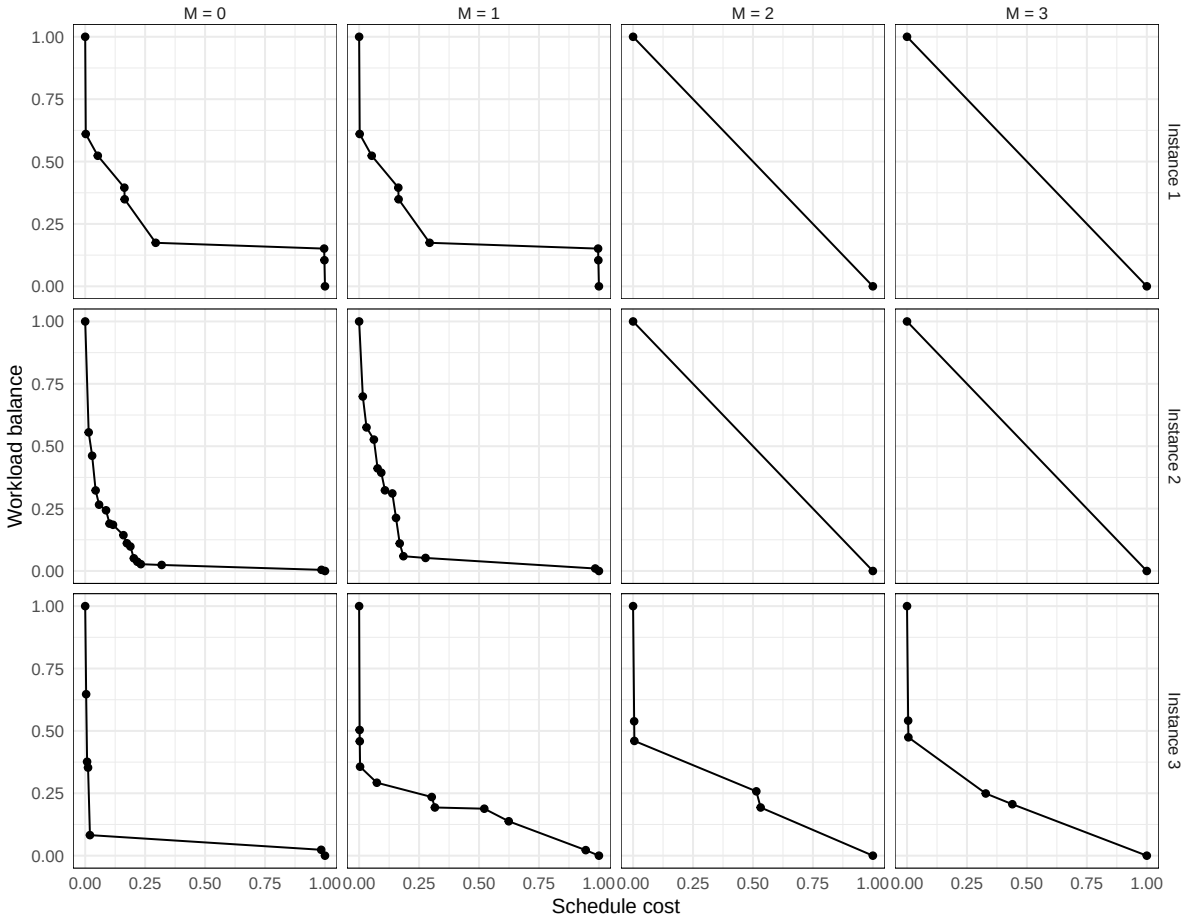

Figure 9: Pareto fronts of three randomly selected problem instances for different values of $\mathcal{M}$. The solutions' objective values are normalized on a scale ranging from 0 to 1.

7 Conclusions and future work

Given the increasing pressure on healthcare systems, it is increasingly important to ensure sustainable working conditions for hospital employees. In this paper, we explored how optimized patient admission scheduling can contribute towards better workload balancing between wards in a hospital. We introduce the concepts of spatial and temporal balance to formalize the requirements of fairly allocated and consistent workloads over time in hospitals. We propose a suitable equity function and workload metric that are minimized as part of a bi-objective model which is solved using a criterion space search algorithm in combination with integer programming.

Our computational study confirmed that the proposed model can produce solutions that are both spatially and temporally balanced. Moreover, the computation time required to generate a set of non-dominated solutions with respect to workload balance and schedule cost is relatively low, enabling the proposed method to be used in practice if patients' admissions are scheduled at regular intervals. This could be achieved by using a batch scheduling policy. A series of experiments also demonstrated how a more balanced workload allocation can be obtained by increasing the generality of wards. By increasing the number of minor specializations per ward there is more decision flexibility, enabling the generation of more balanced solutions. Finally, we analyzed the trade-off between workload balance and schedule costs. The results demonstrated that most problem instances allow for a set of non-dominated solutions from which a decision maker can choose a final solution.

While our paper considered a static, deterministic patient admission scheduling problem, it would be worthwhile investigating whether compromise solutions could be generated when considering dynamic patient arrivals and stochastic problem elements such as recovery LOS or surgery duration. This modeling change would increase the computational challenge considerably, and, as such, heuristics may prove necessary to obtain non-dominated solutions. Alternatively, future research could also explore alternative approaches to obtain a balanced workload allocation. For example, by adjusting staffing levels when demand peaks are unavoidable. By employing a suite of complementary approaches, it is possible that the workload can be allocated in much more balanced way.

Acknowledgments

This research received funding from the Flemish Government under the Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen. Editorial consultation provided by Luke Connolly (KU Leuven).

References

Azizoğlu M, Imat S (2018) Workload smoothing in simple assembly line balancing. Computers & Operations Research 89:51–57

Bektaş T, Letchford AN (2020) Using ℓp-norms for fairness in combinatorial optimisation. Computers & Operations Research 120:104975

Boland N, Charkhgard H, Savelsbergh M (2015) A criterion space search algorithm for biobjective integer programming: The balanced box method. INFORMS Journal on Computing 27(4):735–754

Christ Q, Dauzere-Peres S, Lepelletier G (2019) An iterated min–max procedure for practical workload balancing on non-identical parallel machines in manufacturing systems. European Journal of Operational Research 279(2):419–428

Cossari A, Ho JC, Paletta G, Ruiz-Torres AJ (2012) A new heuristic for workload balancing on identical parallel machines and a statistical perspective on the workload balancing criteria. Computers & Operations Research 39(7):1382–1393

Cossari A, Ho JC, Paletta G, Ruiz-Torres AJ (2013) Minimizing workload balancing criteria on identical parallel machines. Journal of Industrial and Production Engineering 30(3):160–172

Dall'Ora C, Ball J, Reinius M, Griffiths P (2020) Burnout in nursing: a theoretical review. Human resources for health 18(1):1–17

De Boer EM, Bakker AB, Syroit JE, Schaufeli WB (2002) Unfairness at work as a predictor of absenteeism. Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior 23(2):181–197

De Bruin AM, Bekker R, Van Zanten L, Koole G (2010) Dimensioning hospital wards using the Erlang loss model. Annals of Operations Research 178(1):23–43

Demeester P, Souffriau W, De Causmaecker P, Vanden Berghe G (2010) A hybrid tabu search algorithm for automatically assigning patients to beds. Artificial Intelligence in Medicine 48(1):61–70

Finco S, Battini D, Delorme X, Persona A, Sgarbossa F (2020) Workers' rest allowance and smoothing of the workload in assembly lines. International Journal of Production Research 58(4):1255–1270

Fügener A, Brunner JO, Podtschaske A (2015) Duty and workstation rostering considering preferences and fairness: a case study at a department of anaesthesiology. International Journal of Production Research 53(24):7465–7487

Gross CN, Brunner JO, Blobner M (2019) Hospital physicians can't get no long-term satisfaction–an indicator for fairness in preference fulfillment on duty schedules. Health Care Management Science 22(4):691–708

Gulczynski D, Golden B, Wasil E (2011) The period vehicle routing problem: New heuristics and real-world variants. Transportation Research Part E: Logistics and Transportation Review 47(5):648–668

Halvorsen-Weare EE, Savelsbergh MW (2016) The bi-objective mixed capacitated general routing problem with different route balance criteria. European Journal of Operational Research 251(2):451–465

Ho JC, Tseng TLB, Ruiz-Torres AJ, López FJ (2009) Minimizing the normalized sum of square for workload deviations on m parallel processors. Computers & Industrial Engineering 56(1):186–192

Jozefowiez N, Semet F, Talbi EG (2002) Parallel and hybrid models for multi-objective optimization: Application to the vehicle routing problem. In: International conference on parallel problem solving from nature, Springer, pp 271–280

Karsu Ö, Morton A (2015) Inequity averse optimization in operational research. European journal of operational research 245(2):343–359

Liu CM, Chang TC, Huang LF (2006) Multi-objective heuristics for the vehicle routing problem. International Journal of Operations Research 3(3):173–181

Liu R, Xie X, Garaix T (2013) Weekly home health care logistics. In: 2013 10th IEEE international conference on networking, sensing and control (ICNSC), IEEE, pp 282–287

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

Martin S, Ouelhadj D, Smet P, Vanden Berghe G, Özcan E (2013) Cooperative search for fair nurse rosters. Expert Systems with Applications 40(16):6674–6683

Matl P, Hartl RF, Vidal T (2018) Workload equity in vehicle routing problems: A survey and analysis. Transportation Science 52(2):239–260

Matl P, Hartl RF, Vidal T (2019) Workload equity in vehicle routing: The impact of alternative workload resources. Computers & Operations Research 110:116–129

Mullinax C, Lawley M (2002) Assigning patients to nurses in neonatal intensive care. Journal of the operational research society 53(1):25–35

van den Oetelaar WFJM, van Rhenen W, Stellato RK, Grolman W (2020) Balancing workload of nurses: Linear mixed effects modelling to estimate required nursing time on surgical wards. Nursing Open 7(1):235–245

Ouazene Y, Yalaoui F, Chehade H, Yalaoui A (2014) Workload balancing in identical parallel machine scheduling using a mathematical programming method. International Journal of Computational Intelligence Systems 7(sup1):58–67

Ouazene Y, Yalaoui F, Yalaoui A, Chehade H (2016) Theoretical analysis of workload imbalance minimization problem on identical parallel machines. In: Asian Conference on Intelligent Information and Database Systems, Springer, pp 296–303

Rachamadugu R, Talbot B (1991) Improving the equality of workload assignments in assembly lines. The International Journal of Production Research 29(3):619–633

Rajakumar S, Arunachalam V, Selladurai V (2004) Workflow balancing strategies in parallel machine scheduling. The International Journal of Advanced Manufacturing Technology 23(5):366–374

Reiter P, Gutjahr WJ (2012) Exact hybrid algorithms for solving a bi-objective vehicle routing problem. Central European Journal of Operations Research 20(1):19–43

Robbins JM, Ford MT, Tetrick LE (2012) Perceived unfairness and employee health: a meta-analytic integration. Journal of Applied Psychology 97(2):235

Saliba S (2006) Heuristics for the lexicographic max-ordering vehicle routing problem. Central European Journal of Operations Research 14(3):313–336

Sartori C, Smet P, Vanden Berghe G (2021) Efficient duration-based workload balancing for interdependent vehicle routes. In: 21st Symposium on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2021), Dagstuhl Open Access Series in Informatics, vol 96, pp 1–15

Schaus P, Hentenryck PV, Régin JC (2009) Scalable load balancing in nurse to patient assignment problems. In: International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research, Springer, pp 248–262

Schwerdfeger S, Walter R (2016) A fast and effective subset sum based improvement procedure for workload balancing on identical parallel machines. Computers & Operations Research 73:84–91

Schwerdfeger S, Walter R (2018) Improved algorithms to minimize workload balancing criteria on identical parallel machines. Computers & Operations Research 93:123–134

Stolletz R, Brunner JO (2012) Fair optimization of fortnightly physician schedules with flexible shifts. European Journal of Operational Research 219(3):622–629

Vancroonenburg W, De Causmaecker P, Vanden Berghe G (2019) Chance-constrained admission scheduling of elective surgical patients in a dynamic, uncertain setting. Operations Research for Health Care 22:100196

Walter R (2020) A note on "workload smoothing in simple assembly line balancing". Computers & Operations Research 113:104803

Walter R, Schulze P, Scholl A (2021) SALSA: Combining branch-and-bound with dynamic programming to smoothen workloads in simple assembly line balancing. European Journal of Operational Research 295(3):857–873

Wang Y, Zhao L, Savelsbergh M, Wu S (2022) Multi-period workload balancing in last-mile urban delivery. Transportation Science

World Health Organization (2022) Statement on the thirteenth meeting of the international health regulations (2005) emergency committee regarding the coronavirus disease (COVID-19) pandemic. https://www.who.int/news/item/18-10-2022-statement-on-the-thirteenth-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-coronavirus-disease-(covid-19)-pandemic, accessed on November 16, 2022