Multi-Dimensional Workload Balancing: Creating a Healthy Healthcare System by Balancing Emotional and Operational Load

Nitzan Carmeli[1], Dorit Efrat-Treister[2], Avishai Mandelbaum[1], Ori Plonsky[1], Anat Rafaeli[1], Galit B. Yom-Tov[1]

[1] Faculty of Data and Decision Sciences, Technion—Israel Institute of Technology, [2] Department of Management, Ben-Gurion University of the Negev nitzanyuviler@yahoo.com, tdorit@bgu.ac.il, avim,plonsky,anatr,gality@technion.ac.il

Problem definition: Healthcare systems operate under substantial workload pressures. A major component of this workload is the emotional load placed on employees, which is distinct from the operational load required to complete clinical tasks. Despite its importance, emotional load is rarely incorporated into operational decisions because there is limited data and no accepted methodology to quantify it. As a result, managerial decisions that rely only on operational indicators fail to reflect the full burden placed on staff and may inadvertently harm both employees and patients. **Methodology/results:** We propose a structured methodology for measuring emotional load based on a detailed mapping of the tasks performed by healthcare workers. The approach assigns emotional intensity scores to routine activities, allowing emotional load to be captured alongside traditional operational indicators. Using these measurements, we develop a general mathematical framework that balances total workload, defined as the combination of operational and emotional demands, across groups of service providers. The framework supports managerial decisions related to staffing, patient routing, and system configuration. We apply the methodology and framework to two maternity wards in a large hospital. The analysis shows that emotional load varies meaningfully across tasks and units. **Managerial implications:** Our case study demonstrates that incorporating emotional load substantially changes the recommended distribution of patients between wards. When only operational load is considered, emotional load stays imbalanced. Integrating emotional load into operational decision making enables managers to more accurately assess staff burden, design more equitable work systems, and support employee well-being. This approach improves patient experience and enhances system performance by ensuring that decisions reflect the complete set of demands faced by healthcare workers.

1. Introduction

In a field study (§5), maternity nurses were interviewed about their workload. We were told that perhaps their most challenging task is to manage a fetus in stress, since then a nurse must cater, simultaneously, to two life-threatened patients. Now suppose that the fetus tragically dies: operational load is then drastically reduced; but emotional load is skyrocketing, with potentially long-lasting effects on the nurse involved. In the same vein, here is a quote from a judge, when interviewed about workload: “Everyone is concerned with how many hours I work at the office, but no one cares about how many sleepless hours I suffer through at night”. These two examples constitute a natural motivation and a starting point for the present paper, the conclusion of which is that *emotional load can and should be incorporated into operational decisions*.

Research shows that different aspects of workload affects system performance. Specifically in healthcare, which is our focus here, high occupancy of medical wards may harm patient safety (Kc and Terwiesch 2009, Berry Jaeker and Tucker 2017); long hours and high responsibility cause stress and reduce well-being healthcare employees (Meese et al. 2021); and excessive cognitive load increases employee burnout and reduces patient satisfaction (Ripp 2021). As a general term, we refer to workload as the amount of "effort" that is required from a system's resource. However, workload has been commonly defined by and measured via its operational characteristics. The latter could be, for example, counting (patients, tasks, hours, multitasking level) per resource (bed, nurse), or changeovers across customers; and measures could vary dynamically in time, or be accumulated and averaged over time (Delasay et al. 2019). We refer to these types of measured quantities as *operational load*, which is essential in supporting operational policies, for example workforce staffing (Whitt 2007). Yet, an important aspect of workload, mostly overlooked in Operations Management, is emotional load, namely the emotional burden inflicted on employees by customers, above and beyond operational load.

To elaborate, operational decisions tend to ignore the aspects of emotional load that employees experience (Field et al. 2018). Failing to take into account the influence of emotional dynamics on employee performance could lead to operational decisions that are undesirable for both employees and patients/customers. For example, emotion entails important information for decision-makers and for navigating situations (Van Kleef et al. 2010), hence ignoring its dynamics is likely to lead to sub-optimal decisions and, in turn, to lower employee well-being and patient satisfaction. The first step of incorporating emotional load in an operational decision was carried out by Daw et al. (2023): following the identification that customers' expressed sentiment impacts employee reaction time (Altman et al. 2021), these sentiments were incorporated into operational models of predicting service time and optimizing customer allocation to employees in real-time (routing).

Emotional load in service systems can stem from customer behavior and may have a significant impact on systems' performance. For example, customers negative emotions that are targeted toward service employees affect the latter's cognitive functioning and performance (Rafaeli et al. 2012, Ashtar et al. 2021). This type of real-time emotional load was shown to impact employees' efficiency in contact centers much more than classical operational load (Altman et al. 2021). Yet, scarce research has been done to identify and monitor emotional load of healthcare employees and, consequently, offer methods to incorporate its measurements in operational decisions. This paper aims to take first steps in addressing this gap.

The methods for measuring emotional load, reported by Altman et al. (2021) and Daw et al. (2023), aimed at supporting *online* management. This requires technologically mediated services, as in chat or call centers: for example sentiment can be monitored automatically, using text analysis (Yom-Tov et al. 2018) or voice recognition (Wittels et al. 2002) which can be used to prompt managerial intervention in real-time. We, on the other hand, are seeking to support offline management, specifically the design of load balancing (work distribution) protocols in healthcare services, and those are predominantly face-to-face. To this end, measuring emotional load, so that it can be incorporated into operational models, calls for a novel methodology. Indeed, operational decisions must rely on measurements that relate to the measurable, non-abstract aspects of job requirements. To this end, we submit that emotional load is to be measured with reference to the map of activities performed by healthcare employees, as we now elaborate on.

1.1. Narrowing the Gap: Mapping Activities to System-Level Measurements of Emotional Load

We build on Affective Events Theory (Weiss and Cropanzano 1996) to propose a task-level measurement, of the *emotional difficulty* that a task inflicts on a healthcare employee. We combine this theoretical psychological approach with well-established preference elicitation methods (e.g. Wakker and Deneffe 1996), from economics and decision sciences, to measure task-specific *emotional demand* brought about by treatment tasks (§3.1). This approach allows us to aggregate the emotional demands imposed on employees, across the tasks that their job requires over time, and thus measure emotional demands at the customer class and system levels (§3.2).

Our approach reveals that one can distinguish between patient classes, not only in terms of length of stay (LOS) say, but also by the *emotional demand function* as it changes during a patient's hospitalization process (i.e., throughout the patient LOS). In proposing this approach, and to the best of our knowledge, we are the first to offer a concrete measure of emotional demand for healthcare services. Moreover, this measure allows one to define the *total emotional demand*, over the entire LOS, that a specific patient class imposes on service providers. This, in turn, gives rise to *emotional offered-load*, which is a system-level measurement of the total emotional demand required from the system on an average day. Measuring it against the capacity of service providers determines the system-level *emotional load* (§3.3). Our innovative method, for measuring emotional demand, paves the way for incorporating emotional load into operational decisions of face-to-face services.

With the above theory at one's disposal, a practical challenge is to account for emotional offered-load in patients allocation to hospital wards. The need arose in a tertiary 1000-bed Israeli hospital as follows: there are two maternity wards in that hospital, and the nurses, in *both* wards, perceived injustice in the distribution of workload. Specifically, nurses in *each* ward claimed that the workload at the other ward is lower. (Interestingly, no nurse expressed interest in exchanging wards.) This type of perceived injustice — the feeling of imbalance between effort and reward — is known to influence nurses' well-being (Afzali et al. 2017), as well as their health (Topa et al. 2016) and efficiency (Yom-Tov and Rafaeli 2025) — hence the importance of designing a just allocation mechanism. Put in other words, it is impossible that *both* wards experience "higher" operational load. Hence, the perceived injustice must result from other facets of workload allocation; indeed, in Section 3, we show that the emotional load on nurses in Ward B is almost 40% lower than on those in Ward A, while the operational load of Ward A is slightly higher than that of Ward B.

1.2. Balancing Operational and Emotional Loads: Theory and Practice

In Organizational Behavior (OB), *distributive justice* is defined as the fairness of resource distribution among individuals or groups within an organization (Adams 1965). In Operations Research (OR), such justice is referred to as *fair* load-balancing among resources, and algorithms are developed to achieve justice by optimizing various load-metrics as it proxy. (In the sequel, we use the terms *fairness* and *justice*, or fair and just, interchangeably.)

There is a vast literature, in both research and practice, about load-balancing among servers, but it all refers to balancing operational-load indicators (see §2 below, for an overview of relevant literature). We thus acknowledge the importance of balancing operational load, but also posit that balancing emotional load can be important. Therefore, we develop here a mathematical formulation for balancing *multiple dimensions of load* between agent groups (§4). This formulation is based on the theory of Resource-Driven Activity-Networks (RANs) for modeling service systems (Momčilović et al. 2022). Here every customer joins the system with a set of demands that may change over time, and every agent brings to the system a set of capabilities that may also change over time. Service is carried out by matching demands with capabilities. We use RANs to formulate a load-balancing optimization problem that aims to simultaneously balance, continuously over time, both operational and emotional load of two or more medical wards.

In a case study, based on data collected from the above-mentioned two maternity wards, we implemented our methods, for measuring emotional demand profiles of patients (§3) and for optimizing load-balancing (§5). The current routing at the hospital exhibits a close-to-balanced operational load, namely bed occupancy (operational offered-load per bed). This routing, however, creates a significantly unbalanced emotional load, namely emotional offered-load per nurse. Our RAN formulation, in contrast, gives rise to an implementable balancing of multiple types of load, simultaneously, both on average and dynamically over time.

1.3. Contributions

To summarize, we make the following contributions:

- Developing a measurement for task-specific emotional demand that is additive and thus can be aggregated by adopting an economic preference elicitation approach. Specifically, we transform self-report survey responses to direct measurements of the task-level emotional demand per minute of operation. (See details in Section 3.1.)
- Developing methods to construct an emotional demand profile for each customer class. Such a profile distinguishes the average emotional demand of one customer class from that of another customer class. Therefore, this method fits as a tool for the solution of long-term planning problems. The method is broader and more generalizable than emotional load proxies that have been used for real-time monitoring of emotional load, for example sentiment analysis (Altman et al. 2021). (See details in Section 3.2.)
- Conceptualizing emotional offered-load, in concert with and supplementing the classical operational offered-load (Whitt 2013); this is made concrete by summing up emotional demands of all tasks and patients at the system level. (See details in Section 3.3.)
- Developing a mathematical formulation for multi-dimensional load-balancing in a multi-class, multi-server, time-varying environment, both in steady state (§4.1) and dynamically over time (§4.2); therefore, extending current literature on balancing operational load, that ignores time-variability and heterogeneity of customers and resources alike (see details in Section 2).
- Testing the developed measurements on real data, gathered in wards of a large tertiary hospital, and thus substantializing the impact that the present research, and its consequent methods, can have in practice. (See details in Sections 3 and 5.)

1.4. Overview

In Section 2, we review existing research-based methods for load balancing. In Section 3, we propose a novel measure of emotional demand, at the task, customer, and system levels. In Section 4, we develop mathematical models for balancing multi-dimensional load among groups of employees, both averaging over time (§4.1) and dynamically (§4.2). In Section 5, we test our methodology with data from two maternity wards, which underscores the difference between balancing operational load and balancing multi-dimensional load. We conclude, in Section 6 with, what we view as worthwhile directions for future research.

2. OR Literature on Load Balancing

As already mentioned, OR distributive justice, or fair load-balancing, is achieved by designing algorithms that distribute workload fairly among resources (e.g., servers, employees, wards). In fact, OR papers optimize various workload metrics as proxies for optimal fairness, also under different assumptions regarding the service environment. Atar (2008) proposed the longest-idle-server-first (LISF) policy, to balance workload among statistically identical servers. Armony and Ward (2010) extended this work and showed that the longest-weighted-idle-server-first (LWISF) routing policy outperforms LISF. Atar et al. (2011) extended the single server to pools of servers and proposed the longest-idle-pool-first policy, which routes a customer to the pool with the longest cumulative idleness among the available pool. Closest to our environment, Mandelbaum et al. (2012) aimed at equalizing, jointly in steady-state, bed turnover-rates and bed occupancy levels; this was achieved via a randomized most-idle policy, and tested on patient routing in hospitals—from the emergency department to internal wards. Ward and Armony (2013) considered heterogeneous customer classes, aiming to minimize cost subject to idleness fairness constraints. Similarly, Do and Shunko (2020) maximize the expected arrival rate, while several fairness performance measures (such as the variance of the server-level arrival rate) are improved. Daw et al. (2023) propose a prediction-based routing that balances residual workload of servers. Finally, Yom-Tov and Rafaeli (2025) implemented a round-robin routing procedure in hospitals, showing that fairness is indeed important and load-balancing reduces patients' LOS.

We note that the above papers aim at balancing operational load in steady-state. We thus extend previous work by offering methods for load-balancing, that take into account both operational and emotional load — or, for that matter, any number of load-types — allowing them to vary over time. Indeed, our framework accommodates many complex aspects of healthcare systems that have been neglected previously. Specifically, ignored aspects include: (a) *time-varying dynamics* (in contrast to steady-state): e.g. daily variation in patient arrival rates and service times; (b) *heterogeneous demand*: e.g. different patient classes require different service times, namely varying levels of commitment from medical staff, or different emotional load levels; and (c) *heterogeneous capacity*: e.g. service may require multiple types of resources (e.g., bed, physician, nurse), and hospital wards vary in their available resources, meaning the size of their server group, such as number of beds and staffing levels. The latter may also change over time (e.g., between shifts).

Theoretically, we build on OR papers that emphasizes the need to model service systems at the task level. For example, Yom-Tov and Mandelbaum (2014) and Campello et al. (2017) showed how staffing should take into account task-level demand. Daw et al. (2023) showed how routing, in contact centers, must account for dynamic message-level information. In this vein, we model dynamic emotional-load measures from the task-level up, thus creating an emotional load profile throughout the duration of a patient's stay. To be concrete, our optimization framework builds on Momčilović et al. (2022), who develop Resource-driven Activity Networks (RANs): these are mathematical models of complex operations, in which service requirements are characterized at the task level, and each task may involve multiple resource types. As such, patients require, in addition to a bed, also emotional capacity of employees during their stay. To the best of our knowledge, we are the first to implement an analytical framework, specifically RANs, to support a solution of the customer-routing problem that accounts for emotional load of servers. (In our case study (§5), customers are patients and servers are beds and nurses.)

3. Measuring Emotional Demand/Load

In this section, we develop a method for measuring emotional difficulty and demand at the task level (§3.1), followed by emotional demand at the patient level (§3.2). The latter measure is then aggregated to form the emotional offered-load at the ward level (§3.3), which supports managerial decisions at ward-resolution; e.g. load-balancing across wards, as we do here. We implement our method on data collected via surveys of the nursing staff, at two maternity wards of a partner hospital. The surveys deliberately separate between three patient classes — Regular Birth (denoted RB), C-Section Birth (CS), High-Risk Pregnancy (HR) — in order to capture three different profiles of emotional demand, as they evolve during a patient's LOS (see Figure 1, noting how the three profiles vary over time, and how they differ from each other; there is no need now to dwell into more details — these will be provided later, below the figure.)

In line with *Affective Events Theory* (AET) (Weiss and Cropanzano 1996), our working assumption is that events experienced by employees during their work trigger emotional reactions, which then impact employees' job performance and satisfaction. Hence, argues AET, assessments of emotional load should be based on the emotional difficulty of specific *work events*. In our study, we aim at solving an offline design problem: We seek to determine which patient should be routed to which ward so that load is fairly balanced. For such a goal, it is natural to focus on work events that can be anticipated in advance (before the patient is routed) and to create, hence first define and measure, the *average anticipated* emotional demand profile, brought by each patient class into the system. To this end, we define a set of common work tasks performed on each patient type (e.g. patient admission or medication distribution), and base our measures of emotional difficulty and demand on the events that typically occur during each common task a nurse performs. This is in contrast to stochastic real-time realization of work-events, e.g. customers anger, as analyzed in Altman et al. (2021).

3.1. Measuring Emotional Demand at the Task Level

The *emotional difficulty of a task* is the mental and/or emotional hardship that the task inflicts, which reflects how difficult, irritating, or annoying the task is. We measure it via surveys of employees (nurses), who rate the task's emotional difficulty on a scale of 1 (easiest) to 7 (most difficult). From these data, we identify the emotionally easiest and most difficult tasks. (See Appendix A.2 for details of the survey, including the rational behind its design. See Appendix A.1 for the resulting lists of tasks, and Appendix B for the corresponding scores of emotional difficulty.)

Measures of emotional demand must be additive since we aim to aggregate the emotional demands of tasks to the patient level and then to the ward level. The raw ratings of emotional difficulty do not have this property. For example, it is probably false to assume that two tasks rated by the nurses as having an emotional difficulty of 3 (on a scale of 1 to 7) are necessarily more emotionally demanding than one task rated as having an emotional difficulty of 5 on the same scale. But the emotional difficulty ratings provide *relative* differences between emotional difficulties of the tasks, and are helpful by identifying the tasks that the nurses on average perceive to be the least and the most emotionally demanding.

To translate these emotional difficulty ratings to a quantitative scale with interpretable, additive units, we perform in the second step a process adapted from well-established *trade-off methods*, in economics and decision sciences (e.g., Wakker and Deneffe 1996). In these trade-off methods, designed to elicit preferences in scenarios where direct measurement is challenging (Fishburn 1967, Tversky et al. 1988, Wakker and Deneffe 1996, Johnson 1974), people are asked to report how much of one thing that they value (here, time) they are willing to sacrifice in order to gain another thing of value (here, emotional ease), such that they will be indifferent between the two states. This process thus allows for translating emotional ease/difficulty to units of time, an additive measure that nurses can easily relate to, and has been used extensively in previous applications of the trade-off method (e.g., Dolan et al. 1996).

Specifically, we asked nurses to choose between performing the most emotionally demanding task for a certain time period and performing the least emotionally demanding task for a longer time period. By iteratively changing the additional time spent on the least emotionally demanding task, we were able to identify the point of indifference. This allows us to measure how much extra time performing the least emotionally demanding task is "worth" the possibility to avoid the extra emotional difficulty associated with the most emotionally demanding task. Using this process, we concluded, for example, that (on average) nurses calibrated the hardest task of "conversations with family members of high-risk patients" as 1.64 times more emotionally difficult per minute than the easiest task of "shift handoff". We term this ratio the task's *emotional factor*, and denote it by F_i. For the above example, we assigned $F_{HR-family} = 1.64$, and concluded that the range of emotional factors in our case-study is $F_i \in [1, ..., 1.64]$, in "easiest-emotional-task minutes per minute". (See more details in Appendix A.) Next, for each patient class, the other tasks were assigned F_i scores according to their relative ratings from the survey. Accordingly, every task has an emotional factor, F_i, which represents the emotional difficulty required to perform one minute of this task in "easiest-emotional-task minutes per minute" units.

Finally, the *emotional demand* of a task i, ED_i, is computed as the product of the task's emotional factor with its duration (denoted T_i) in minutes: $ED_i = T_i \cdot F_i$. The emotional demand is thus measured in "easiest-emotional-task minutes". For the easiest task, it is equal to the task's actual duration, because the easiest task's emotional factor equals 1 (i.e, $ED_{shift-handoff} = T_{shift-handoff} \cdot 1 = T_{shift-handoff}$). For the other tasks, however, the emotional demand is larger than their duration and reflects their perceived duration if their emotional difficulty had been equal to the easiest task. Appendix B provides a detailed list of the emotional factors and demands, computed for each of the tasks per patient class.

To summarize, the task-level emotional demand, defined in this section, is measured in *units of time*, which will be here either minutes or days, depending on the context. The trade-off method we propose enables us to calculate the emotional demand of a specific task relative to the emotional demand of the *emotionally-easiest task*. The use of time-units to measure emotional demand of tasks allows us, in the next section, to aggregate emotional demands to the patient level, and ultimately to compute the emotional offered load at the ward level.

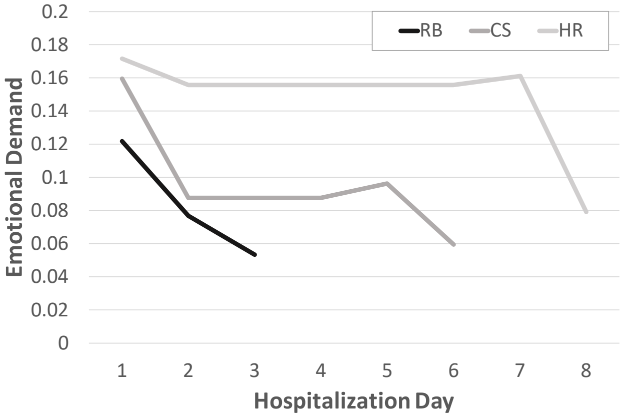
3.2. Aggregating Emotional Demand to Patient and Ward Levels.

As a next step, in order to gain a precise estimation of the operational and emotional offered loads that each patient class imposes on the system, we need to determine the frequency and timing of these activities during the LOS of each patient class. In general, we considered three types of tasks:

1. *Protocol tasks*—these are tasks that depend on the time that passed since a patient was admitted to the ward. (For example, 4 hours after admission, the patient must get medications; 6 hours after admission, her vital signs must be measured; and discharge instructions are provided on the last day of hospitalization.)
2. *Ward work profile tasks*—these are tasks that are carried out as part of the ward schedule, at specific times during the day, and are performed as long as a patient is hospitalized. (For example, nurses' rounds at the beginning/end of every shift.)
3. *Continuous treatment tasks*—these are ongoing tasks that are done throughout the day, and do not depend on the specific patient time in the system, nor on the time of day.

Based on this mapping of the three types of activities, we constructed the emotional demand functions during the course of patient hospitalization at a daily resolution, per each patient class (see Figure 1). For the purpose of current demonstration, we assume that the duration of patient hospitalization is its average duration: 2.75, 5.67, and 7.5 days respectively, for RB, CS and HR patients.

Figure 1 Emotional Demand in "Easiest-emotional-task-days" Units, During Patient Hospitalization, for each Patient Class



We now observe that every patient class has a different emotional demand function, both in volume and in shape. Furthermore, the emotional demand is not constant during hospitalization. Rather, the first day of hospitalization comprises more emotional demand than subsequent days;

and demands during the last day of hospitalization are lower, because last-days in the ward are not full days, and patients are in a relatively good condition when discharged. The total measured emotional demand for each patient over an average LOS duration is 0.252, 0.579, and 1.19 easiest-emotional-task days for RB, CS, and HR patients, respectively (see the summary in Table 1).

Table 1 **Summary Statistics by Patient Class**

	Regular Birth (RB)	C-section (CS)	High Risk (HR)
Daily arrival rate	6.15	3.33	1.67
Average LOS (days)	2.75	5.67	7.5
Total emotional demand (easiest-emotional-task days)	0.252	0.578	1.19

Finally, we define *emotional offered load* in a manner analogous to the well known notion of *operational offered load*, which represents the total operational demand that patients impose on the system. Hence, we define *emotional offered load* as *the total emotional demand that patients impose on the system.*

3.3. Calculating Offered Load: Operational & Emotional, Stationary & Time-Varying.

Classical queueing theory acknowledges operational offered-load as being either *stationary* or *time-varying* (Whitt 2013). The former is calculated simply from long-run averages: e.g. offered-load by a specific customer-type to a server (in time-units of work per time-unit) is this type's average arrival-rate multiplied by average service-duration. Time-varying offered-load is more subtle to calculate since, at a specific time t, it must accumulate work that arrived prior to t and is yet to be completed at or after t. Analogously, in Section 4 we develop measures of both stationary and time-varying emotional offered-load, by basing their calculation on patients arrival rate and patients emotional demand, for each patient class.

For example, Table 2 depicts *stationary* calculations from our hospital data for each ward; it is based on the hospital routing policy by which all High-Risk patients are sent to Ward A, as well as 52% of the RB patients and 20% of the CS patients; the rest of the patients are allocated to Ward B — see Figure 2. This table also presents two *relative* load measures: *operational load* (total operational offered load divided by the number of beds in the ward) as an operational measure, and *emotional load* (total emotional offered load divided by the number of nurses at the ward) as a behavioral measure. Fairness is then measured by the difference of these load measures across wards. Our analyses reveals that operational load is almost balanced (~ 2% difference), but there are large gaps in emotional load between the wards (~ 40% difference). Moreover, load imbalance is in opposing directions — Ward A has a slightly lower operational load and a much higher emotional load than Ward B. In the next section, we rectify this unfair state of affairs (mathematically), by jointly balancing the two load types.

Table 2 **Operational and Emotional Measurements in Current State by Medical Ward (Stationary Measures)**

	Ward A				Ward B			
	RB	CS	HR	Total	RB	CS	HR	Total
Arrival rate per day	3.20	0.67	1.67	5.54	2.95	2.66	0	5.62
Operational offered load	8.79	3.78	12.53	25.10	8.12	15.10	0	23.22
Emotional offered load	0.807	0.387	1.987	3.18	0.743	1.537	0	2.28

	Ward A	Ward B	Difference
Arrival rate per day	5.53	5.62	-1.5%
Operational offered load	25.10	23.22	8.1%
Emotional offered load	3.18	2.28	39.5%
Number of beds	32	29	
Nurse hours per day	72	72	
Average number of nurses in ward	3	3	
Fairness measures:			
Operational load	0.78	0.8	-2.1%
Emotional load	1.06	0.76	39.5%

4. Balancing Multi-Dimensional Load Among Multiple Resources

In this section, we develop and analyze two mathematical network formulations of multi-dimensional load balancing: static formulation, which seeks a balance in the long-run (§4.1); and dynamic (§4.2), which balances loads across wards at all times (though load-values may vary in time). Dynamic balancing is more challenging to solve for than static balancing, yet it is worth the effort, as it offers superior performance. The static formulation, while more intuitive, then serves as an informative initial point for our dynamic algorithm. Further details are presented in the sequel.

Our modeling framework is that of Resource-driven Activity Networks, or RANs for short (Momčilović et al. 2022); and our RANs, static and dynamic, will model hospital wards, in which capacitated resources (e.g. nurses, beds, patients) collaborate on activities (e.g. hospitalization), that are carried out over time according to state-dependent itineraries. The creation of the RAN models is guided by Section 3: each patient "brings" a specific multi-dimensional demand vector, which is created by aggregating demands at the lower tasks level. Demand in dynamic models is a time-varying function over a patient's hospitalization, as in Figure 1. In static models, demand is a constant, calculated by summing up (calculating area under) the dynamic demand.

In a specific RAN application, demand values, as well as the values of other RAN primitives, are inferred from the system that originates the RAN. Here, RAN models a system of two maternity wards, each subject to a two-dimensional demand (operational, emotional). As mentioned, the values of its primitives are inferred from the data described in Section 3. The RAN is then analyzed and finally, in Section 5, it helps improve upon present load balancing, notably without additional resources.

4.1. Static Load Balancing

Consider an open network with K classes of customers, to be served by M server pools. Let the scalar λ_k represent the (static) arrival rate of class-k customers; in vector form, λ_k is the kth coordinate of a K-dimensional arrival-rate vector $\lambda_{K\times 1} = [\lambda_k]^T$, or λ for short (when mathematically convenient).

Each customer class is characterized by a vector of L demand types (here operational or emotional demand imposed on servers; and it could be also cognitive demand required for service, which is not accounted for here). These demands are formalized by a matrix $V_{L\times K}$, or V for short, of which element $V_{l,k}$ is the *average* amount of the lth demand type associated with a *single* class-k customer.

Demands require resources (e.g. beds, emotional) to perform them; and these resources, that are of finite capacity (e.g. number of beds, emotional capacity), are associated with the servers. Thus, L demand types require L resource types to perform them, which is formalized by a matrix $C_{L\times M}$, or C for short: its element $C_{l,m}$ is the *average* amount of capacity of server pool m, available to accommodate demand of type l. This way, an analogy is drawn between physical capacity and emotional capacity: to elaborate, *physical resource capacity* quantifies the maximal server time-units, available to meet operational demand (e.g. bed-days or nurse-hours, available to accommodate hospitalization- or treatment-demand of patients); analogically, *emotional resource capacity* indicates the maximal emotional capability of employees to handle emotional demand.

We now formalize the concept of a plan for load balancing. A plan determines the dynamics of our system, and it is formalized as follows: a *plan* X is a $K \times M$ matrix, of which the (k, m)-element $X_{k,m}$ is the rate (number per unit of time) of class k customers that are routed to service at server pool m. A feasible plan must then satisfy the following constraints (in vector- or matrix-form):

$$Xe \le \lambda, \quad VX \le C, \quad X \ge 0, \tag{1}$$

where e is a vector of length M, with all its coordinates being 1's . The first constraint in (1) is an arrival-rate constraint—the number of services of class k customers, per unit of time, is bounded by the arrival rate of class k customers. The second constraint in (1) is a capacity constraint that folds, in matrix form, $L \times M$ scalar constraints—one per each demand type per each server pool: constraint (l, m) ensures that the total offered-load l, imposed on server pool m (namely $\sum_k V_{l,k} X_{k,m}$, which is total demand arriving per time-unit), does not exceed the corresponding available capacity (namely $C_{k,m}$).

Remark 1. Note the implicit assumption that customers of class k incur the same demand, regardless of the server pool they are routed to — thus the servers, across pools, are assumed statistically identical in their capabilities. If this would not have been the case, one could define M demand matrices, where the matrix $V^m_{L\times K}$ represents demands associated with server pool m. Then capacity constraints would take the form $V^m X_{[:,m]} \le C_{[:,m]}$, $m = 1, ..., M$, where $[:, m]$ denotes the m-th column of a matrix. △

Plans determine the proportion of customers from each class to be sent to each of the server pools; for a specific plan X, these proportions are given by

$$P_{k,m} = \frac{X_{k,m}}{\sum\limits_{i=1}^{M} X_{k,i}}.$$

Our goal is to determine these proportions (determine the system design), striving for an optimal/fair balance of the multi-dimensional load across the server pools.

There could be multiple ways to achieve fair balance, depending on what is deemed fair. One way is to minimize the absolute difference in proportional load between all server groups. Formally, this would entail defining β^l_m as the proportional load—the total offered load from type l on server pool m, divided by the pool capacity for this load type. Then, a load balancing goal is to minimize the sum of absolute differences (SAD), denoted d_{SAD}, over feasible plans X. Hence the objective is

$$\min_X d_{SAD} = \min_X \sum_{\substack{i<j \\ i,j\in\{1,2,...,M\}}} \sum_{l=1}^{L} \left|\beta^l_i - \beta^l_j\right|, \tag{2}$$

$$\text{where } \beta^l_m = \frac{\sum\limits_{k=1}^{K} V_{l,k} X_{k,m}}{C_{l,m}}.$$

A second option is to minimize staffing costs and add constraints in which we bound the multi-dimensional load difference to be less than some threshold τ. For example, if we assume that the first row of C represents the number of servers in each of the server pools, and that staffing costs are equal for all server types, then the objective function is

$$\min_X \sum_{m=1}^{M} C_{1,m},$$

and we add the following constraints to the general constraint set in (1):

$$\left|\beta_i^l - \beta_j^l\right| \le \tau, \quad \forall l = 1, \ldots, L, \forall i, j \in \{1, \ldots, M\}, i < j,$$

where τ is our tolerable threshold value for the load difference (e.g., 5%); and there are, of course, many other options to characterize fair balance.

4.2. Dynamic Load Balancing

Dynamic here stands for *varying in time*. Time variability may result from temporal pattern-changes in arrivals or in customers' demand during their sojourn time (LOS) — we accommodate both with time-varying offered-loads and capacities, or more specifically time-varying model primitives. Let $\Lambda_k(t)$ represent the total number of arrivals of class-k customers during time interval $[0, t]$, $t \ge 0$. (One can think in terms of arrival rates $\lambda_k(t), t \ge 0$, in which case $\Lambda_k(t) = \int_0^t \lambda_k(u)du$.) In a vector form, $\Lambda_k(t)$ is the kth coordinate of a K-dimensional cumulative arrival-rate vector $\Lambda(t)$.

As in the static case, each customer class is characterized by the L demand types they impose on the system. However, in the dynamic case, the demand from each type may vary during the customer length of stay in the system. We note that, in some cases, the demand that a customer imposes on the servers depends on the *proportional* progress of the service (e.g. half-way to the end), as was shown in Yom-Tov et al. (2018). Alternatively, in some cases, the demand a customer imposes depends on the *absolute* time of the customer in the system.

Let $G_k(\cdot)$ be the CDF of the service duration of a class k customer, and denote by $\bar{G}_k(t)$ the survival function of this duration: $\bar{G}_k(\cdot) = 1 - G_k(\cdot)$. We denote by $L_k^l\left(\bar{G}_k(t)\right)$ the type-l demand that a customer from class k imposes on the servers, when the remaining proportion of service is $\bar{G}_k(t)$. Here $L_k^l(t)$ is the type-l demand that a customer from class k imposes on the servers, after it has been in service for t units of time. The functions $L_k^l(\cdot)$ depend on the specific application, as will be demonstrated in §4.2.1.

Similarly to the static case, demands are formalized by a matrix $V_{L\times K}(t)$, and capacities are formalized by the matrix $C_{L\times M}(t)$, for every time $t \ge 0$. Here, the matrix element $V_{l,k}(t)$ is the type-l demand associated with a *single* class-k customer who has been in the system for t units of time; and $C_{l,m}(t)$ represents the total capacity of server pool m available to accommodate type-l demand, at time t. We note that, in some circumstances, the capacity of interest is not point wise capacity but rather the threshold capacity of server pool m for type-l demand over a period of time (e.g., emotional capacity of nurses during a shift). In these cases, we denote this threshold capacity during time period i by $C(i)$.

A dynamic plan is a matrix-valued function $X = \{X(t), t \geq 0\}$: here $X(t) := [X_{k,m}(t)]$ is a $K \times M$ matrix that represents, for every $t \geq 0$, the cumulative number of services that started during the time interval $[0, t]$. Element-wise, $X_{k,m}(t)$ is the number of class k customers that started service, by servers from server pool m, up to time t inclusive (hence $X(\cdot)$ must and is assumed to be non-decreasing, right-continuous, with left-hand limits). We also assume that $X(0) \equiv 0$.

4.2.1. Example of Two-Dimensional Demand: Operational and Emotional. Consider two types of demand: first customer LOS and, second, emotional demand brought in by the customer. The first type of demand, $L_k^1\left(\bar{G}_k(t)\right)$, is the remaining proportion of service duration, after being in service for t units of time: that is, $L_k^1\left(\bar{G}_k(t)\right)$ is simply $\bar{G}_k(t)$. We therefore define the *operational offered load* derived from class k customers at time t to be

$$\begin{aligned}&\left(V_{1,k} * \Lambda_k\right)(t) = \left(L^1\left(\bar{G}\right) * \Lambda\right)_k(t) = \left(\bar{G} * \Lambda\right)_k(t)\\ &= \int_0^t \bar{G}_k(t-s)\, d\Lambda_k(s), \quad t \geq 0;\end{aligned} \tag{3}$$

this is the total number of class k customers in the system at time $t > 0$, assuming that there is ample server capacity; the symbol $*$ stands for the convolution operator between two functions.

Now assume that the emotional demand depends on the time a customer has spent in service, regardless of their remaining service duration. For example, in our case study, most tasks that nurses must perform depend on the time that a patient has spent in hospitalization (e.g., every x hours, starting from the hospitalization epoch, the patient must get medications). In this case, the total emotional offered load from class k customers at time $t > 0$ is

$$\begin{aligned}&\left(V_{2,k} * \Lambda_k\right)(t) = \left(L^2(\bar{G}) * \Lambda\right)_k(t)\\ &= \int_0^t L_k^2(t-s)\, \bar{G}_k(t-s) d\Lambda_k(s);\end{aligned} \tag{4}$$

here $L_k^2(t-s)$ is the emotional demand associated with a customer from class k that has been in service for $u = t - s$ units of time, while $\bar{G}_k(t-s)$ is the probability that a customer from class k

that arrived at time s, and started service also at time s (under ample capacity of servers), would still be in service at time t, that is, after $u = t - s$ units of service time.

As in the static case, a dynamic plan $X = \{X(t), t \geq 0\}$ is feasible if it satisfies the arrival rate constraints and the non-negativity constraints, namely for every $t \geq 0$,

$$(Xe)(t) \leq \Lambda(t), \quad X(t) \geq 0.$$

However, in the dynamic case, there are many ways to formulate capacity constraints. To mention a few:

1. *Pointwise constraint.* Bounding the load at every point in time t, that is

$$(V * X)(t) \leq C(t), \quad t \geq 0,$$

where $(V * X)_{l,m}(t) = \sum_{k=1}^{K} \left(V_{l,k} * X_{k,m}\right)(t)$. This type of constraint fits physical resources, such as beds in a ward: the number of hospitalized patients cannot exceed bed capacity, at all times.

2. *Cumulative constraint.* Here we accumulate the load over periods of length T (e.g., a shift). Then we bound the cumulative multi-featured load in every period $i = 1, 2, \ldots$ of length T:

$$\int_{T(i-1)}^{Ti} (V * X)(t) \leq C(i),$$

where $C_{l,m}(i)$ represents a threshold value for the capacity of server pool m to the cumulative load from type l, at the ith time period. This type of constraint fits resources that have some operational flexibility. For example, nurses can control when they measure fever of patients, though this activity needs to be done during a specific time window during the shift.

3. *Peak/end constraints.* Another option for capacity constraint is to bound the *peak* load and/or the *end* load that occurred during a given time interval before time t. For example, within every shift i of length T, peak load constraint is defined by

$$\max_{T(i-1) \leq t \leq Ti} (V * X)(t) \leq C(i),$$

and the end load constraint is defined by

$$\max_{t-1 \leq s \leq t} (V * X)(s) \leq C(t).$$

Theses constraints are inspired by papers showing that, on retrospective evaluations of certain experiences, people tend to associate high weights with peak and end events (Ariely and

Carmon 2000, Ashtar et al. 2023). For example, a nurse ending her shift is most likely to retain the most intense events (peak effect), and the most recent events (e.g., that occurred during the last hour of the shift). Therefore, when constraining behavioral dimensions such as emotional load, one may wish to bound these extremes.

Note that one may use, within the same model, different capacity constraints for different load types. For example, one may bound the operational offered load pointwise, while bounding the emotional offered load in a cumulative manner over a shift. In our case study, which we now turn to, we seek to balance emotional loads by minimizing the difference between the two maternity wards; while maintaining the operational load, that is the number of hospitalized patients, bounded above by the number of beds in each ward.

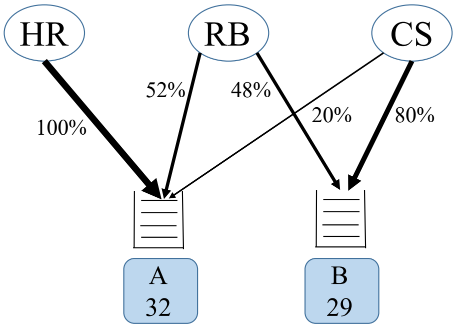
5. Case Study: Balancing Load in Patient Routing

In the following case study, we use hospitalization data from the wards of the Division of Obstetrics and Gynecology (OBGYN) in a large tertiary hospital in Israel. The OBGYN division has two maternity wards that provide care to various types of maternity patients, treating about 4,000 patients annually. Each patient is classified into one of three classes — Regular Birth (RB), C-Section Birth (CS), or High-Risk Pregnancy (HR) — each requiring a different level of care, associated with distinct lengths of stay (LOS) and varying degrees of emotional demand.

Function-wise, each maternity ward admits R patients and, in addition, has its own specialization: Ward A specializes in HR patients but admits *also* CS patients, whereas Ward B specializes in CS births and admits *no* HR patients. Accordingly, the *static routing policy* at the time of the study was as follows: all HR patients were routed to Ward A; most CS patients (about 80%) were routed to Ward B; and R patients were almost evenly split, with 52% assigned to Ward A and the rest to Ward B. Table 2 summarizes the distribution of patients by class and ward, and Figure 2 depicts the routing policy.

The hospital thus practices specialization between the two wards which, in general, is an important factor in healthcare: it can impact quality of care, LOS and even death rates (see Song et al. 2020, and references therein). Specialization becomes significant when medical conditions are challenging and/or rare (that has even led to creating speciality medical centers with superior outcomes; e.g., Dana-Farber Cancer Institute and Shouldice Hernia Hospital). Hence, speciality is of most significance to HR patients, who are the smallest patient group, and our mathematical articulation will account for it. Finally, we assume (for tractability) that patient LOS depends only

Figure 2 Original Routing Probabilities of the Three Patient Classes to the Two Maternity Wards



on the patient class, as opposed to also on the specific ward in which they are hospitalized or the specialization profile of that ward.

Ward A has 32 beds while Ward B has only 29 beds. Staffing levels in both wards are equal: 4, 3, and 2 nurses in morning, afternoon, and evening shifts, respectively. The aforementioned design results in bed occupancy of 78% for Ward A and 80% for Ward B; hence, a slightly higher ratio of patients-to-beds in Ward B.

The difference in emotional load is more pronounced than in operational load. Stationary emotional offered load on each ward is characterized by patient arrival rates, their static routing probabilities, and the total emotional demand each patient class brings, as presented in Table 1. According to the routing policy (Figure 2), the average emotional offered load on Ward A is **3.18** ($= 6.15 \cdot 0.52 \cdot 0.252 + 3.33 \cdot 0.2 \cdot 0.578 + 1.67 \cdot 1.19$), which is 39.5% more than the average emotional offered load on Ward B **2.28** ($= 6.15 \cdot 0.48 \cdot 0.252 + 3.33 \cdot 0.8 \cdot 0.578$); note that, the difference between (nurse level) emotional load is also 39.5% since the number of nurses in both wards is equal over the day. This result contrasts that of operational load (B higher than A), which highlights that load profiles can have conflicting facets that must be accounted for.

Next, we apply the balancing method developed in Section 4 while adding specialization goals. This yields an alternative routing policy, which we shall then examine for its impact on load balancing.

5.1. Optimal Static Design of Routing Policy

We start with the static routing policy that ignores load variability over the day. This approach fits the goal of long-term balance between wards, and it assumes that within day fluctuations are of less significance. A design (matrix) X defines the average number of patient hospitalizations, from each of the three patient classes, in each of the two wards, per day. That is, $X_{i,j}$, $i = 1\,(RB),\ 2\,(CS),\ 3\,(HR)$, $j = 1\,(A),\ 2\,(B)$, is the average number of hospitalizations of patients from type i in ward j, per day. Note that the routing probability of patients from type i to ward j is determined by $P_{i,j} = \frac{X_{i,j}}{X_{i,1}+X_{i,2}}$.

An optimal design seeks to achieve two goals: minimizing load differences jointly with maximizing specialization. We compare the current practice of balancing only the operational load with an alternative that balances multiple types of load simultaneously. To maximize specialization, we define a binary matrix $Z_{K\times M}$ and a large constant, $\mathcal{M}$, and add the constraints $\mathcal{M}Z \geq X$ and $X \geq Z$. Due to these constraints, if $Z_{k,m} = 0$, it forces $X_{k,m} = 0$ (since $X \geq 0$); conversely, if $X_{k,m} > 0$, it sets $Z_{k,m} = 1$. Then we add in the objective function the term $e_1^T Z e_2$, where e_1 is a cost vector and e_2 a unit vector of ones. Minimizing this term minimizes the cost of the active arcs on the routing graph, and therefore, the optimal solution avoids splitting a single patient class between the two wards, if possible. To ensure that the optimal solution prefers splitting RB patients over HR and CS patients, we assign higher cost to splitting HR and CS patients in the vector e_1.

We now present two versions of the design optimization problem: (a) minimizing the difference in operational load (i.e., bed occupancy levels); (b) minimizing multi-dimensional load (i.e., the sum of differences in both operational and emotional loads). Formally, (a) is given by

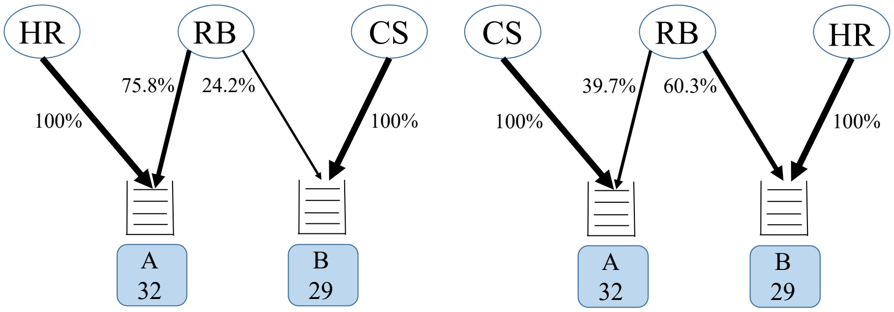
$$\begin{aligned} \min \quad & e_1^T Z e_2 + |\beta_1^1 - \beta_2^1| \\ s.t. \quad & Xe = \lambda, \\ & VX \leq C, \\ & X \geq 0, \\ & \mathcal{M}Z \geq X, \\ & X \geq Z, \\ & \beta_1^1 \leq 1.05\beta_2^1, \quad \beta_1^1 \geq \frac{1}{1.05}\beta_2^1. \end{aligned} \tag{5}$$

Here Z is the above-described binary matrix ($[Z]_{3\times2}$), $e_1 = \begin{bmatrix}1 & 2 & 2\end{bmatrix}$ is the cost vector, $e_2^T = \begin{bmatrix}1 & 1\end{bmatrix}$ a summing vector, and β_m^1 is as defined by Equation (2). According to Tables 1 and 2, the arrival rate vector λ, demand vector V and capacity vector C are given by

$$\lambda = \begin{bmatrix}6.15 \\ 3.33 \\ 1.67\end{bmatrix}, \quad V = \begin{bmatrix}2.75 & 5.67 & 7.5\end{bmatrix}, \quad C = \begin{bmatrix}32 & 29\end{bmatrix}.$$

The constraints $\beta_1^1 \leq 1.05\beta_2^1$, $\beta_1^1 \geq \frac{1}{1.05}\beta_2^1$, ensure that the difference in occupancy levels between the two wards does not exceed 5%. This constraint is added because $e_2^T Ze$ is of a higher magnitude than $|\beta_1^1 - \beta_2^1|$; therefore an optimal solution without it might overemphasize minimizing $e_2^T Ze$—that

Figure 3 Routing Probabilities of the Three Patient Classes to the Two Maternity Wards, while Balancing Load and Maximizing Specialty



(a) Balance operational load (b) Balance multi-dimensional load (option 1)

is, pushing toward specialization at the expense of load balancing. The inclusion of this balancing constraint circumvents this issue.

We solved the optimization problem numerically using the CVXPY Python solver. The optimal solution of (5) results in a W topology (which is a 2-level chain design), which is presented in Figure 4(a). This topology is indeed specialized, where all HR patients are sent to Ward A, and all CS patients are sent to Ward B. The operational load is balanced through the Regular-Birth patient class routing. The resulting performance, in terms of occupancy and emotional load for each ward, are presented in Table 4. We observe that this routing solution does not balance the emotional load — A exceeds B by 37.5%, which is ample.

Next, we add the emotional-load feature towards balancing both operational and emotional loads. In this case

$$V = \begin{bmatrix} 2.75 & 5.67 & 7.5 \\ 0.252 & 0.578 & 1.19 \end{bmatrix}, \quad C = \begin{bmatrix} 32 & 29 \\ 3 & 3 \end{bmatrix}.$$

To minimizing the differences of both operational and emotional loads, we change the objective function in (5) to

$$\min \quad e_1^T Z e_2 + |\beta_1^1 - \beta_2^1| + |\beta_1^2 - \beta_2^2|, \tag{6}$$

and add constraints regarding the maximal difference between β_1^2 and β_2^2. We note that an optimization problem in which the difference in both the operational load and the emotional load between the two wards is no more than 5% (i.e, $\frac{1}{1.05}\beta_2^1 \le \beta_1^1 \le 1.05\beta_2^1$ and $\frac{1}{1.05}\beta_2^2 \le \beta_1^2 \le 1.05\beta_2^2$) would inevitably split HR patients between the two wards. We, therefore, solve the optimization problem with varying values of maximal allowed operational load and emotional load differences, looking for an "efficient frontier". The results are presented in Table 3. In all cases, all CS patients are routed to Ward A while all HR patients are routed the Ward B. The only difference is the proportion of RB patients routed to the two wards.

Table 3 **Summary of the Operational and Emotional Loads Differences Given Different Constraints on Max Load Difference**

	Operational load (occupancy)			Emotional load		
Constraints	Ward A	Ward B	Diff (%)	Ward A	Ward B	Diff (%)
Option 1: Operational Diff ≤ 5%, Emotional Diff ≤ 15%	0.800	0.783	-2.1%	0.847	0.974	15.0%
Option 2: Operational Diff ≤ 10%, Emotional Diff ≤ 10%	0.821	0.761	-7.9%	0.867	0.954	10.0%
Option 3: Operational Diff ≤ 15%, Emotional Diff ≤ 5%	0.842	0.737	-14.0%	0.888	0.933	5.0%

As mentioned above, solving this optimization problem, with any of the values allowed for maximal offered load and emotional load differences, results again in a W specialized topology; see Figure 4(b). Note that this W topology is the exact opposite of the optimal solution of (5), where this time, all CS patients are treated in Ward A, while all HR patients are treated in Ward B. Figure 4(b) presents the routing probabilities given the optimal solution with Option 1 from Table 3. Note that, in this solution, the absolute value of the operational load difference is equivalent to the operational load difference in the current state; however, the emotional load difference is greatly reduced, from 39.5% to 15%. See Table 4 for a comparison between the performance measures of the current state, the optimal solution version (a)—balancing only operational load, and the optimal solution of version (b), option 1—balancing both operational load and emotional load.

Table 4 **Summary of the Operational and Emotional Loads for each Ward under Different Static Policies**

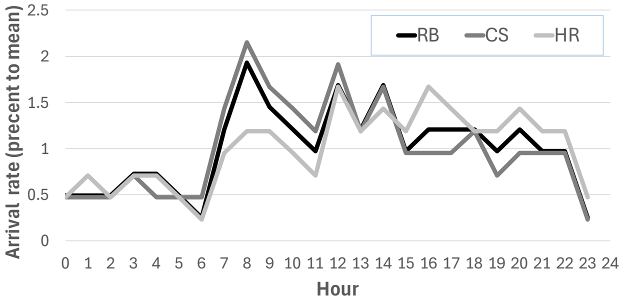
	Operational load (occupancy)			Emotional load		
Design	Ward A	Ward B	Diff (%)	Ward A	Ward B	Diff (%)
Current state (static)	0.780	0.800	2.1%	1.06	0.76	-39.5%
Balance operational load (version (a))	0.792	0.792	0.0%	1.05	0.77	-37.5%
Balance multi-dimensional load (version (b), option 1)	0.800	0.780	-2.1%	0.85	0.97	15.0%

5.2. Stochastic Dynamic Routing

We now continue with dynamic routing, aiming to optimally balance both types of load *at each point in time*. Variation over time can result from three sources: (a) time variability in the arrival rate process (see Figure 4) (b) LOS differences between patient classes (as presented in Table 1), and (c) time variability in the emotional demand profile during patient LOS (Figure 1).

Throughout the present section, we use the following LOS distributions: LOS of RB and CS patients follow a $\textit{Triangular}(48, 54, 96)$, and a $\textit{Triangular}(120, 120, 168)$ distribution, respectively

Figure 4 Time-varying Hourly Arrival Rate by Patient Class (RB, CS, HR), Percent to Mean



(parameters are in hours). The LOS distribution of HR patients is assumed $Lognormal(4.246, 1.415)$ (sample mean 190 hours, and its sample standard deviation 481 hours). (For more details see Appendix C.) We use the list of activities and their timing as described in Appendices A–B. Note that although patients are commonly discharged from the maternity ward between 10am and 3pm, we only use the above LOS distributions to determine the probability that a patient will be discharged during a given hour.

According to Section 3, nurses provide discharge guidance in the evening prior to the discharge day. We assume that this guidance takes place between 7pm and 8pm; alternative time windows were examined but yielded negligibly different results.

Current State: Figures 6(a) and 6(b) present the operational and emotional loads at the two wards under the hospital's current routing probabilities, as stated in Figure 2. We used Equations (3) and (4) to numerically calculate the time-varying dynamics of the operational and emotional offered loads, respectively, and then divided by the time-varying capacity. Specifically, let $EL_j(t)$ be the emotional offered load imposed on ward j at time t:

$$EL_j(t) = \left(V_{2,RB} * X_{RB,j}\right)(t) + \left(V_{2,CS} * X_{CS,j}\right)(t) + \left(V_{2,HR} * X_{HR,j}\right)(t),$$

where $\left(V_{2,k} * X_{k,j}\right)(t)$ is a discrete formulation of Equation (4), that is:

$$\left(V_{2,k} * X_{k,j}\right)(t) = \sum_{s=1}^{t} L_k^2(t-s)\bar{G}_k(t-s)X_{k,j}(s).$$

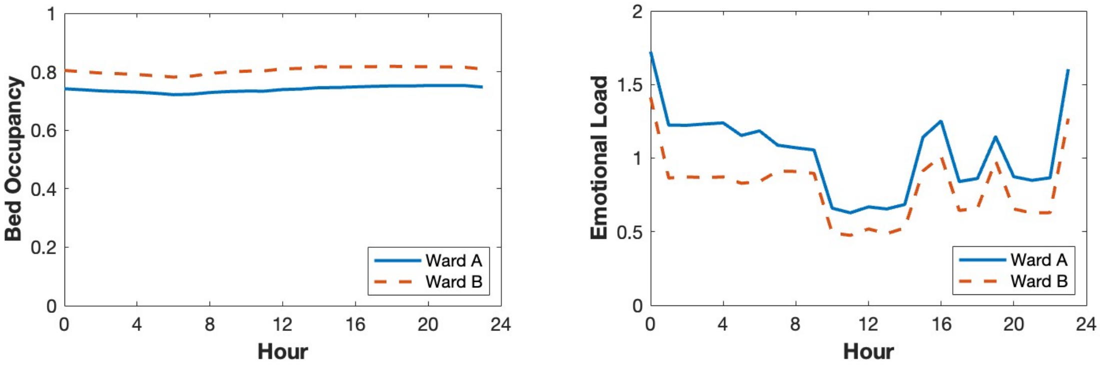
Here $X_{k,j}(s)$ denotes the arrival rate of patients from type k at time s ($\lambda_k(s)$) multiplied by $P_{k,j}(s)$ — the probability that those patients will be routed to ward j at time s (see Figure 2). The emotional load at ward j during hour t is given by $EL_j(t)$ divided by the number of nurses in the ward at time t: as indicated already, 4 nurses in the morning shift, 3 in the afternoon, and 2 at night at both wards. Let $OL_j(t)$ be the operational offered load on ward j at time t. The

$$OL_j(t) = \left(V_{1,RB} * X_{RB,j}\right)(t) + \left(V_{1,CS} * X_{CS,j}\right)(t) + \left(V_{1,HR} * X_{HR,j}\right)(t),$$

where $\left(V_{1,k} * X_{k,j}\right)(t)$ is a discrete formulation of Equation (3), and $X_{k,j}(s)$ is as defined above. The operational load (i.e., bed occupancy) at ward j during hour t is given by $OL_j(t)$ divided by the total number of beds in the ward, which is constant over time. Note that the offered loads formulas are time-dependent due to the time-varying emotional demand profile of each patient class, as well as the time-varying arrival rate of each class.

Figure 5 was calculated over a period of 40 days, with day 40 presented; this is to avoid the influence of the initial state (empty system) on load calculations. As in the static case, we observe that, although the operational load in Ward B is higher, Ward A nurses suffer from a higher emotional load throughout the day. Specifically, the proportional difference between Ward B occupancy and Ward A occupancy is on average 8.8% and stable over the day, while the average proportional difference in emotional load between Ward A and Ward B is 30.9% and changes over the day between 17% and 42%.

Figure 5 Average Operational and Emotional Loads for each Ward (Ward A - Blue, Solid; Ward B - Red, Dashed), Given the Original Static Routing Policy, During a Single Day in an Hourly Resolution



(a) Operational load (occupancy)

(b) Emotional load

Speciality Dynamic Routing Design: Next, we optimize patient routing dynamically. Here, we fix the system design to the W topology found in Section §5.1. This W design routes all CS patients to Ward A, all HR patients to Ward B, and it balances load by dividing the RB patient class between wards. While forcing this W topology, we allow the routing probabilities of RB patients to change every hour. Therefore, for every $1 \le s \le t$, we fix $X_{CS,A}(s) = \lambda_{CS}(s)$, and $X_{HR,B}(s) = \lambda_{HR}(s)$, while allowing $X_{RB,j}(s)$ to depend on $P_{RB,j}(s)$ such that $X_{RB,j}(s) = \lambda_{RB}(s) * P_{RB,j}(s)$, $j \in \{A, B\}$.

To determine the optimal routing for Regular-Birth patient class, we solve the following dynamic load balancing problem:

$$\begin{aligned}
\min \quad & \sum_{t=1}^{T} \left| \frac{EL_A(t)}{C_A^2(t)} - \frac{EL_B(t)}{C_B^1(t)} \right| && (7)\\
s.t. \quad & P_{RB,A}(t) \geq 0, \quad P_{RB,B}(t) \geq 0, && \forall 1 \leq t \leq T;\\
& P_{RB,A}(t) \leq 1, \quad P_{RB,B}(t) \leq 1, && \forall 1 \leq t \leq T;\\
& OL_{t,A}(t) \leq C_A^1(t), && \forall 1 \leq t \leq T;\\
& OL_{t,b}(t) \leq C_B^1(t), && \forall 1 \leq t \leq T;\\
& \frac{OL_{t,A}(t)}{C_A^1(t)} \leq 1.1 \cdot \frac{OL_{t,B}(t)}{C_B^1(t)}, && \forall 1 \leq t \leq T;\\
& \frac{OL_{t,A}(t)}{C_A^1(t)} \geq \frac{1}{1.1} \cdot \frac{OL_{t,B}(t)}{C_B^1(t)}, && \forall 1 \leq t \leq T,
\end{aligned}$$

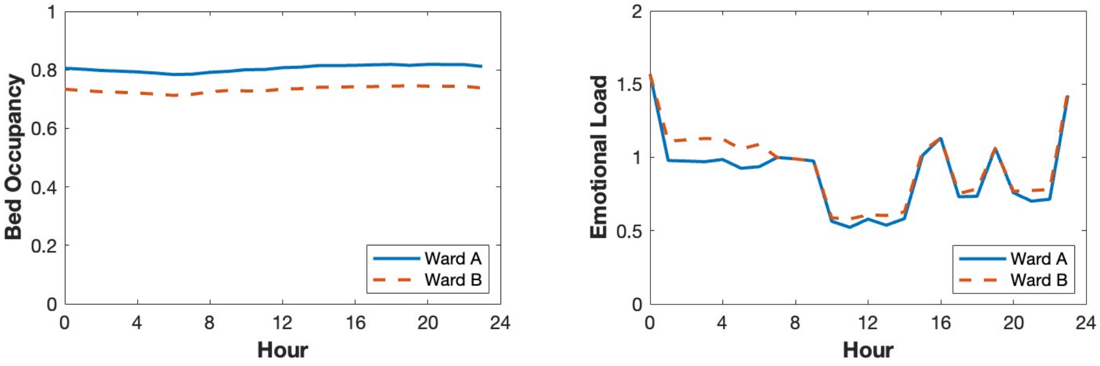
where the operational and emotional capacity are as in the current state (i.e, $C_A^1(t) = 32$ beds $\forall t$, $C_B^1(t) = 29$ beds $\forall t$, while $C_A^2(t) = C_B^2(t)$ nurses $\forall t$, but both change over time — 4 nurses in the morning shift, 3 in afternoon shift, and 2 at night shift). This hourly discrete formulation minimizes the $L1$ norm of the difference of emotional offered loads over the planning horizon T, while maintaining the difference between operational loads bounded by 10% (which turned out the minimal difference with feasible solution). Note that minimizing the emotional offered load is equivalent to minimizing the emotional load (at the nurse level) , because the number of nurses in both wards is equal throughout the day.

Remark 2. To solve time-varying formulations, one must determine the initial condition over which the optimization starts. For that we use the static routing probabilities found in §5.1. They were calculated for 34 days, to initialize the system with some form of "steady-state", while days 35 to 40 were optimized. We solved this optimization problem numerically, using the MATLAB CVX tool. △

Figures 7(a) and 7(b) present the hourly operational load (i.e., bed occupancy) and emotional load per ward, given the W topology, with the time-varying routing probabilities of RB patients as found by the above optmization problem, during day 40.

We note that enabling time-varying routing probabilities greatly improves the difference between emotional loads, while maintaining similar difference in the operational load compared to the current state, as seen in Figure 5. To be concrete, the proportional difference between Ward A

Figure 6 Hourly Operational and Emotional Loads, Given a W Topology, with Time-Varying Routing Probabilities of the RB Patient Class



(a) Occupancy levels

(b) Emotional load

occupancy and Ward B occupancy is on average 9.9%, while the average proportional difference in emotional load between Ward B and Ward A reduced to 6.9% and varies over the day between 0% and 16%.

Note also that when considering the full LOS distribution of each patient class and not just their mean LOS, the occupancy differences between the wards become more pronounced, both in the original hospital routing and in the suggested dynamic routing under the W topology specialization constraint. The dynamic routing allows us to maintain the occupancy differences fairly small (within 10% difference) while greatly reducing the the emotional load differences from 30.9% to less than 7%. A solution summary of the dynamic routing problem, given as averages over the 24 hours of the 40th day, are presented in Table 5.

Table 5 Summary of Operational and Emotional Loads Within a Day for each Ward under Different Dynamic Policies

	Operational load (occupancy)			Emotional load		
Design	Ward A	Ward B	Diff (%)	Ward A	Ward B	Diff (%)
Current state (dynamic)	0.740	0.805	8.8%	1.034	0.798	-30.9%
Balanced emotional load (dynamic)	0.800	0.728	-9.9%	0.885	0.935	6.9%

6. Concluding Remarks

Developing Measurement of Emotional Load: Existing literature emphasizes that healthcare workload affects the performance of healthcare systems. However, most studies on workload measure only operational load, overlooking emotional load — a significant component that extends beyond operational demands. This gap stems primarily from the lack of suitable methods for measuring emotional load. Existing measures often focus on abstract situations or rely on proxies (for example, text analyses). Yet, to be incorporated into operational decision-making, the measurement of emotional load must be less abstract and linked to the physical aspects of healthcare job requirements.

In the present research, we address this gap by proposing a measurement of emotional load that refers directly to the activities performed by healthcare employees. We applied this approach in a hospital case study, collecting task-level emotional load data from two maternity wards. This enabled us to construct emotional load measures for different patient classes; this reveals, among other things, that the emotional load generated by each patient class varies along the patient's length of stay (LOS), reflecting the differing emotional demands of admission, care, and discharge phases.

Advantages and Limitations: Our approach can be easily implemented for other medical wards or work environments. It is scalable, and applicable to both small and large organizations.

Yet, we acknowledge that our measure does not capture all aspects of the emotional burden that occur in the wards. Indeed, two types of events that may cause emotional load are excluded from this study. First, rare but impactful events — our analysis focused on predictable events that appear in the healthcare worker's everyday tasks; however, extreme events such as harmful medical errors or patient deaths may occur randomly and are likely to result in high emotional stress. Second, other types of events not related to patient–personnel tasks may also create emotional load — for example, a manager–nurse disagreement. Such random or non–task-related events were outside the scope of our model. We note that while these "rare events" are less important for long-term planning decisions such as load balancing, they are important for real-time management of emotional load, and are therefore worthy of future research.

Balancing Operational and Emotional Loads: We utilized the emotional load measurement to balance multi-dimensional loads between agent groups. First, we developed a conceptual measure of emotional offered load, by analogy to the operational measure of offered load. Then, we proposed a RAN time-varying optimization model that balances both operational and emotional loads between agent groups, and tested it on data from two maternity wards. The model is based on a dynamic fluid formulation for balancing long-term averages. Future research may extend this approach to real-time balancing.

One of the challenges in multi-type load balancing is that different load measures may have different scales, making it difficult to determine which types of load are most important to balance.

As we saw, there may also be other important considerations, such as specialization, that can interfere with load balancing. The advantage of our approach is that it provides such considerations with a clear mathematical articulation. In this way, we demonstrate that the psychological aspect of emotional load can—and should—be incorporated into the operational design of service systems.

References

Adams JS (1965) Inequity in social exchange. *Advances in experimental social psychology*, volume 2, 267–299 (Elsevier). 4

Afzali M, Nouri JM, Ebadi A, Khademolhoseyni SM, Rejeh N (2017) Perceived distributive injustice, the key factor in nurse's disruptive behaviors: a qualitative study. *Journal of caring sciences* 6(3):237. 4

Altman D, Yom-Tov GB, Olivares M, Ashtar S, Rafaeli A (2021) Do customer emotions affect agent speed? An empirical study of emotional load in online customer contact centers. *Manufacturing & Service Operations Management* 23(4):854–875. 2, 3, 5, 8

Ariely D, Carmon Z (2000) Gestalt characteristics of experiences: The defining features of summarized events. *Journal of Behavioral Decision Making* 13(2):191–201. 17

Armony M, Ward A (2010) Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research* 58(3):624–637. 6

Ashtar S, Rafaeli A, Yom-Tov GB, Akiva N (2021) When do service employees smile? response-dependent emotion regulation in emotional labor. *Journal of Organizational Behavior* 42:1202–1227. 2

Ashtar S, Yom-Tov GB, Rafaeli A, Wirtz J (2023) Affect-as-information: Customer and employee affective displays as expeditious predictors of customer satisfaction. *Journal of Service Research* 0(0):1–18. 18

Atar R (2008) Central limit theorem for a many-server queue with random service rates. *Ann. Appl. Probab.* 18(4):1548–1568. 6

Atar R, Shaki YY, Shwartz A (2011) A blind policy for equalizing cumulative idleness. *Queueing Systems* 67(4):275–293. 6

Berry Jaeker JA, Tucker AL (2017) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* 63(4):1042–1062. 2

Campello F, Ingolfsson A, Shumsky RA (2017) Queueing models of case managers. *Management Science* 63(3):882–900. 7

Daw A, Castellanosa A, Yom-Tov GB, Pender J, Gruendlinger L (2023) The co-production of service: Modeling service times in contact centers using hawkes processes, working paper. 2, 3, 6, 7

Delasay M, Ingolfsson A, Kolfal B, Schultz K (2019) Load effect on service times. *European Journal of Operational Research* 279(3):673–686. 2

Do HT, Shunko M (2020) Constrained load-balancing policies for parallel single-server queue systems. *Management Science* 66(8):3501–3527. 6

Dolan P, Gudex C, Kind P, Williams A (1996) The time trade-off method: results from a general population study. *Health economics* 5(2):141–154. 9

Field JM, Victorino L, Buell RW, Dixon MJ, Meyer Goldstein S, Menor LJ, Pullman ME, Roth AV, Secchi E, Zhang JJ (2018) Service operations: what's next? *Journal of Service Management* 29(1):55–97. 2

Fishburn PC (1967) Methods of estimating additive utilities. *Management science* 13(7):435–453. 8

Johnson RM (1974) Trade-off analysis of consumer values. *Journal of marketing research* 11(2):121–127. 8

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management science* 55(9):1486–1498. 2

Mandelbaum A, Momcilovic P, Tseytlin Y (2012) On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Science* 58(7):1273–1291. 6

Meese KA, Colón-López A, Singh JA, Burkholder GA, Rogers DA (2021) Healthcare is a team sport: stress, resilience, and correlates of well-being among health system employees in a crisis. *Journal of Healthcare Management* 66(4):304. 2

Momčilović P, Mandelbaum A, Carmeli N, Armony M, Yom-Tov G (2022) Resource-driven activity-networks (RANs): A modelling framework for complex operations, working paper, Technion—Israel Institute of Technology. 4, 7, 12

Rafaeli A, Erez A, Ravid S, Derfler-Rozin R, Treister DE, Scheyer R (2012) When customers exhibit verbal aggression, employees pay cognitive costs. *Journal of applied psychology* 97(5):931–950. 2

Ripp J (2021) Cognitive load as a mediator of the relationship between workplace efficiency and well-being. *Joint Commission Journal on Quality and Patient Safety* 47(2):74–75. 2

Song H, Tucker AL, Graue R, Moravick S, Yang JJ (2020) Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science* 69(9):3825–3842. 18

Topa G, Guglielmi D, Depolo M (2016) Effort–reward imbalance and organisational injustice among aged nurses: a moderated mediation model. *Journal of Nursing Management* 24(6):834–842. 4

Tversky A, Sattath S, Slovic P (1988) Contingent weighting in judgment and choice. *Psychological review* 95(3):371. 8

Van Kleef GA, De Dreu CK, Manstead AS (2010) An interpersonal approach to emotion in social decision making: The emotions as social information model. *Advances in experimental social psychology*, volume 42, 45–96 (Elsevier). 2

Wakker P, Deneffe D (1996) Eliciting von neumann-morgenstern utilities when probabilities are distorted or unknown. *Management science* 42(8):1131–1150. 3, 8, ec2

Ward AR, Armony M (2013) Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research* 61(1):228–243. 6

Weiss HM, Cropanzano R (1996) Affective events theory. *Research in organizational behavior* 18(1):1–74. 3, 7

Whitt W (2007) What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics* 54(5):476–484. 2

Whitt W (2013) Om forum—offered load analysis for staffing. *Manufacturing & Service Operations Management* 15(2):166–169. 5, 11

Wittels P, Johannes B, Enne R, Kirsch K, Gunga HC (2002) Voice monitoring to measure emotional load during short-term stress. *European journal of applied physiology* 87:278–282. 3

Wittenbrink B (2007) Measuring attitudes through priming. *Implicit measures of attitudes* 17–58. ec1

Yom-Tov GB, Ashtar S, Altman D, Natapov M, Barkay N, Westphal M, Rafaeli A (2018) Customer sentiment in web-based service interactions: Automated analyses and new insights. *In WWW '18 Companion: The 2018 Web Conference Companion, April 23–27*, 8 pages (New York, NY, USA: ACM). 3, 15

Yom-Tov GB, Mandelbaum A (2014) Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* 16(2):283–299. 7

Yom-Tov GB, Rafaeli A (2025) The impact of procedural and distributive justice on patient flow in hospitals, working paper. 4, 6

Appendix A: Computing the Emotional Demand per Task - Details

Our method for determining emotional task-level demand consists of the following steps:

A.1. Identifying Employee Tasks

Nurses working at hospitals perform many different tasks during a day. Using observations and motion studies, we identified over 50 nurse activities that nurses regularly perform. We then clustered these activities into 12 task categories (13 in the case of high-risk patients), which represent the range of work preformed with all types of maternity patients. Task categories were double-checked with the head nurses of the wards and, to simplify terminology, we refer to them also as tasks — see Table EC.1 for their list. The motion-studies also yield T_i — the average time required for task i to be preformed (see Tables EC.2–EC.4). These times may vary across patient types.

Table EC.1 List of Tasks Performed by Nurses in Maternity Wards

Category no. (i)	Task Name	Activity Example
1	External Examinations	Taking vital signs
2	Invasive Examinations	Taking a blood test
3	Performing an intimate Treatment	Intimate wash
4	Assisting a Patient	Transporting newborn to nursery
5	Conversations with a Patient	Providing nursing guidance
6	Tasks Accompanying Treatment	Transporting medical equipment
7	Shift Handoff	Receiving shift briefing
8	Administrative Actions	Reviewing patient file
9	Assisting Another Professional	Escorting patient rounds
10	Conversation with Family Members	Conversing with visitors
11	Admitting a Patient	
12	Discharging a Patient	Guiding patient for discharge
13	Monitor checkup (high risk only)	

A.2. Assessing Perceived Emotional Difficulty for Each Task

We designed a survey, distributed to all 32 maternity nurses. The survey included a questionnaire aimed at measuring the emotional difficulty of each task. However, to help nurses focus on the tasks' emotional aspect alone when answering this questionnaire, we first asked them to respond to two other questionnaires.

In the first questionnaire, nurses were asked to rank, for each patient class, the 12 task categories from the easiest (ranked as 1) to the most difficult (ranked as 12), without explicit mention of either emotional or operational difficulty. This forced-ranking step was meant to instill in the nurses the idea that not all tasks are of similar difficulty and prepare them for the ratings step of the next questionnaires (cf. Wittenbrink 2007).

In the second questionnaire, we asked the nurses to score, for each task category and for each patient class, the operational difficulty in conducting the task (i.e. how much *time* a task demands), on a scale of 1 (least time consuming) to 7 (most time consuming). Specifically, the instructions stated: ”We now try to understand, using numbers, how difficult each of the tasks is. We understand that there are two types of difficulties: difficulty resulting from the duration of time the task takes, and difficulty resulting from the emotional burden it carries for the nurse. We would first like to understand the difficulty resulting from the time load. Use a scale of 1 to 7 to rate each task's difficulty under this definition. Tasks that consume the most amount of time will get the score 7, whereas tasks that consume the least amount of time will get the score 1.” These explicit ratings of the operational difficulty were meant only to allow nurses to mentally distinguish between the two types of difficulties a task carries, and thus increase the validity of the emotional difficulty ratings that followed in the third and last questionnaire provided.

In this third questionnaire, the focus of the survey, we asked nurses to rate, for each task and for each patient class, the emotional difficulty in conducting the task (i.e., how much a task is emotionally difficult), on a scale of 1 (least difficult) to 7 (most difficult). Here, we explicitly defined emotional difficulty as the mental or emotional hardship a task inflicts, reflecting how difficult, irritating, or annoying a task is. Specifically, the instructions stated: ”Now, putting the time question aside, we would like to focus on how difficult the task is for you emotionally. Use the same 1 to 7 scale to rate how much the task is difficult, burdening or upsetting for the nurse. The difficulties can result from the nature of the task or from feelings that you carry with you after the task is completed”. Based on these responses, we calculated the average rating of the perceived emotional difficulty for each task across all nurses (see Tables EC.2–EC.4). From these data, we identified the easiest and most difficult tasks.

A.3. Assessing the Emotional Factor and Emotional Demand for Each Task

As explained in the main text, measures of emotional demand must be additive since we aim to aggregate the emotional demands of tasks to the patient level and then to the ward level. The raw ratings derived from the questionnaire in the previous step do not have this property, But they do provide *relative* differences between emotional difficulties of the tasks, and are helpful by identifying the tasks that the nurses on average perceive to be the least and the most emotionally demanding.

To translate the raw emotional difficulty ratings to a quantitative scale with interpretable, additive units, we elicited a mapping from these ratings to an additive scale measured in ”easiest-emotional-task minutes per minute”, using a trade-off method (e.g., Wakker and Deneffe 1996). Specifically, we interviewed nurses individually, and asked them what they preferred: performing the least emotionally difficult task or performing the most emotionally difficult task, for the same exact duration of time (e.g., x minutes). As expected, all nurses preferred performing the least emotionally difficult task. We then gradually increased the duration of the least emotionally difficult task and again asked the nurses to *choose* between performing the least emotionally difficult task, now for a longer period of time, and the most emotionally difficult task for the same original duration x. Through this interview process, we identified the point in which nurses were indifferent between the two options. Hence, this process allows us to identify the amount of additional time nurses prefer to spend performing the least emotionally difficult task so as to avoid performing the most emotionally difficult task. This allowed us to calculate the emotionally hardest tasks’ emotional factor, as explained in the main text.

Next, for each patient class, the other tasks were assigned F_i scores according to their relative ratings from the survey. These new scores were calculated using linear transformations of the emotional difficulty ratings of the questionnaire from the previous step to the emotional factor scale. For example, had the emotional difficulty ratings been in the range of 3–5 (for all classes), and Task 1 raw rating had been 3.5, then the emotional factor of Task 1 would have been 25% $(= (3.5 - 3)/(5 - 3))$ of the range of the emotional factor, namely $1 + 0.25 \cdot (1.64 - 1) = 1.16$. Accordingly, every task is assigned an emotional factor, which is computed in terms of the time units required to perform the easiest emotional task, in “easiest-emotional-task minutes per minute”.

Finally, the "emotional demand" of a task i, ED_i, is computed as the product of the task's emotional factor with its duration in minutes: $ED_i = T_i \cdot F_i$. The emotional demand is thus measured in "easiest-emotional-task minutes". For the easiest task, it is equal to the task's actual duration, because the easiest task's emotional factor equals 1 (i.e, $ED_{shift-handoff} = T_{shift-handoff} \cdot 1 = T_{shift-handoff}$). For the other tasks, however, the emotional demand is larger than their duration and reflects their perceived duration if their emotional difficulty had been equal to the easiest task. Appendix B provides a detailed list of the emotional factors and demands, computed for each of the tasks per patient class.

Appendix B: Data of Each Task

Tables EC.2–EC.4 present, for each task, information regarding its duration, average emotional difficulty score as received by the questionnaires described in Appendix A.2, and the calculated emotional factor and emotional demand described in Appendix A.3. Table EC.2 presents protocol tasks—tasks that depend on the time that passed since a patient was admitted to the ward. Table EC.3 presents ward work profile tasks—tasks that are carried out as part of the ward schedule at specific ties during the day. The duration and the emotional factor of these tasks does not depend on the patient class. Lastly, Table EC.4 presents continuous treatment tasks—ongoing tasks that are done throughout the day. The load of these tasks is given in hours of work per hour. The range of emotional factors for RB, HR, and CS patients is $1-1.365$, $1-1.453$, and $1-1.64$, respectively.

Table EC.2 Emotional Data for each Task: Protocol Tasks

Patient Type	Task No.	Description	Duration (Hours)	Difficulty (1–7)	Emotional Factor	Emotional Demand
RB	1	Admission	0.619	4	1.365	0.845
	2	Hospitalization guidance	0.167	3.45	1.257	0.209
	3	Assisting patient with basic activities	0.043	3	1.169	0.051
	4	Measuring vital signs	0.033	3.09	1.186	0.039
	5	Intimate examination	0.048	2.72	1.114	0.054
	6	Discharge guidance	0.156	2.63	1.096	0.171
	7	Discharge	0.023	2.63	1.096	0.025
HR	1	Admission	0.488	5.2	1.6	0.78
	2	Distribution of medication	0.019	4.6	1.482	0.028
	3	Measuring vital signs	0.034	3.2	1.208	0.041
	4	Monitor Check	0.037	2.4	1.051	0.039
	5	Booking a counselor	0.066	3.2	1.208	0.08
	6	Blood test	0.041	4.6	1.482	0.061
	7	Handling blood test	0.065	2.6	1.09	0.071
	8	Urine test	0.031	3.2	1.208	0.038
	9	Handling urine test	0.065	2.6	1.09	0.071
	10	Escorting a counselor	0.032	4.6	1.482	0.047
	11	Discharge guidance	0.114	3.2	1.208	0.137
	12	Discharge	0.028	3.2	1.208	0.033
CS	1	Admission	0.621	4.45	1.453	0.902
	2	Measuring vital signs	0.04	3.54	1.275	0.051
	3	Intimate examination	0.051	3.09	1.186	0.061
	4	Distribution of medication	0.025	3.9	1.345	0.034
	5	Removing catheter	0.038	3.9	1.345	0.05
	6	Assisting patient with basic activities	0.048	3.27	1.222	0.059
	7	Discharge guidance	0.177	3.27	1.222	0.216
	8	Discharge	0.018	3.27	1.222	0.021

Table EC.3 Emotional Data for each Task: Ward Work Profile Tasks

Task No.	Description	Duration (Hours)	Difficulty (1–7)	Emotional Factor	Emotional Demand
8	Reviewing patient data in the system	0.044	2.14	1	0.044
9	Shift briefing	0.035	2.14	1	0.035
10	Escorting patient rounds	0.053	3.67	1.3	0.069

Table EC.4 Emotional Data for each Task: Continuous Treatment Tasks

Patient Type	Task No.	Description	Hours of Work per Hour	Difficulty (1–7)	Emotional Factor	Emotional Demand
RB	1	External examination	0.0007	3.09	1.186	0.0008
	2	Invasive examination	0.0007	3.45	1.257	0.0008
	3	Intimate treatment	0.0008	2.72	1.114	0.0009
	4	Assisting a patient	0.0011	3	1.169	0.0012
	5	Conversation with a patient	0.0025	3.45	1.257	0.0032
	6	Tasks accompanying treatment	0.0053	2.63	1.096	0.0059
	7	Administrative actions	0.0198	3.9	1.345	0.0266
	8	Conversation with family members	0.0024	4.81	1.524	0.0037
	9	Assisting another professional	0.0074	3.54	1.275	0.0094
HR	1	External examination	0.0011	3.2	1.208	0.0013
	2	Invasive examination	0.0003	4.6	1.482	0.0005
	3	Intimate treatment	0.0014	2.8	1.129	0.0017
	4	Assisting a patient	0.0019	2.8	1.129	0.0021
	5	Conversation with a patient	0.0094	3.6	1.286	0.0121
	6	Tasks accompanying treatment	0.0096	2.6	1.09	0.0105
	7	Administrative actions	0.0355	3.2	1.208	0.0429
	8	Conversation with family members	0.0049	5.4	1.639	0.008
	9	Assisting another professional	0.0119	4.6	1.482	0.0176
CS	1	External examination	0.0011	3.54	1.275	0.0014
	2	Invasive examination	0.0023	3.9	1.345	0.0032
	3	Intimate treatment	0.0003	3.09	1.186	0.0004
	4	Assisting a patient	0.0008	3.27	1.222	0.001
	5	Conversation with a patient	0.0055	3.63	1.292	0.0071
	6	Tasks accompanying treatment	0.0082	2.72	1.114	0.0091
	7	Administrative actions	0.0204	3.72	1.31	0.0267
	8	Conversation with family members	0.0024	4.72	1.1.506	0.0037
	9	Assisting another professional	0.0072	3.45	1.257	0.0091

Appendix C: LOS Distribution

The sojourn times of Regular-Birth and C-Section patients have strict lower bounds placed by the Ministry of Health. Moreover, only in extremely rare cases do the sojourn times exceed a certain upper bound and they also have a palpable mode. Therefore, we decided to estimate the sojourn times for these patients to be of Triangular distribution. The lower limit parameter was taken to be the lower bound stated in clinical regulations. The other two parameters were determined following experts' estimations. The sojourn times of High Risk patients are distributed very differently as there is no obvious mode and the tail of the distribution could be very long. Since the nature of treatment is very similar to that of patients admitted in internal care units, and since empirical results show that the sojourn times of patients in internal care are distributed Lognormal, we estimated that the sojourn time distribution of High Risk patients would also be Lognormal. We further verified this hypothesis by speaking with experts. We used MLE as the estimated distribution’s parameters based on a sample of 34 High Risk patients. The derived distributions were (parameters in hours):

- Regular-Birth patients: $\mathit{Triangular}(48, 54, 96)$
- C-Section patients: $\mathit{Triangular}(120, 120, 168)$
- High-Risk patients: $\mathit{Lognormal}(4.182, 1.196)$