Healthcare capacity management: Queuing theory and pooling approaches

(article starts on next page)

Healthcare Capacity Management: Queuing Theory and Pooling Approaches

Björn Lantz[1]

1. *Chalmers tekniska högskola*
412 96 Göteborg
031–7721381
bjorn.lantz@chalmers.se

ABSTRACT

This paper discusses the potential of patient pooling and staff pooling in healthcare capacity management. Through theoretical modeling within queuing theory, it is illuminated how these strategies can optimize resource allocation and patient flow in a system with variable supply and demand. Examples are presented in a two-unit system with separate arrival and service rates. Patient pooling demonstrates more efficient bed utilization and shorter waiting times through a shared queue for both units. Staff pooling is illustrated by flexible capacity allocation, enhancing the system's adaptability. The results underscore the theoretical potential of both strategies. For future research, emphasis is placed on empirical evaluations to validate these concepts in real healthcare contexts.

1. INTRODUCTION

Healthcare systems, particularly those providing acute care, are characterized by variation in both the supply and demand of healthcare services (Thorwarth et al., 2016; Winasti et al., 2018). While this variation is often perceived as random for practical purposes, it also stems from systematic sources. Factors such as sickness among staff, vacations, and educational interventions contribute to supply variation, while demand variation goes beyond randomness and includes seasonality (e.g., slipping accidents are naturally more common during the winter) and influenza outbreaks. When planning the capacity of a healthcare system, it is crucial not only to balance the expected supply and demand but also to account for the inherent variation within each dimension (Song et al., 2020).

Queuing theory is a mathematical discipline that focuses on analyzing and modeling the flow of entities, such as customers or patients, through a system of queues or waiting lines (Lantz, 2019). Originally developed to study the behavior of waiting lines in service systems, queuing theory has found widespread applications in other fields such as healthcare (Li et al., 2015; Proudlove, 2020). It provides a structured approach to understanding the dynamics of service processes, resource utilization, waiting times, and performance in production systems characterized by variation. At its core, queuing theory involves several key components (Lantz, 2019):

- Arrival Process: This refers to the pattern of entities entering the system, such as patients arriving at a healthcare facility.

- Service Process: This involves the way entities are processed or served within the system, reflecting the medical care and attention provided to patients.
- Queue Structure: Queuing theory models often consider the queue configuration, such as single or multiple queues, and how entities are selected for service.
- Service Capacity: This represents the resources available for providing service, including healthcare professionals, facilities, and equipment.

This paper delves into an in-depth analysis of two innovative capacity pooling approaches – patient pooling and staff pooling – within the analytical framework of queuing theory. These strategies, drawn from a growing body of literature, address the intricate dynamics of capacity optimization in healthcare settings. The integration of queuing theory into healthcare capacity analysis provides a fresh perspective that could potentially reshape resource allocation.

Historically, local fluctuations in supply and demand have been addressed by allocating surplus resources to individual units, leading to inefficient resource utilization and inflated costs (Fagefors et al., 2022). Patient pooling and staff pooling are two approaches that emerge as a promising strategy to transcend these limitations. Patient pooling optimizes bed utilization by centrally managing patient assignments, minimizing wait times, and enhancing overall patient flow. Studies have demonstrated the positive impact of patient pooling on reducing bottlenecks and streamlining resource allocation (Song et al., 2020; Winasti et al., 2018). Similarly, staff pooling offers another significant strategy. Staff pooling involves the dynamic allocation of healthcare professionals and support staff across multiple units, ensuring optimal coverage in response to fluctuating patient loads and variable capacity. By considering the dynamics of patient demand and the availability of skilled staff, this strategy enables a more efficient and flexible allocation of resources. The benefits of staff pooling in enhancing workforce flexibility and maintaining quality care delivery have been explored in previous research (Bouma, 2022; Fagefors et al., 2022).

In a similar vein, the literature has shown that the synergy between patient pooling and staff pooling has the potential for a powerful impact. The benefits of integrating both approaches have been highlighted, suggesting that their combined use could lead to enhanced capacity optimization and improved patient care outcomes (Winasti et al., 2018). However, investigations into the combined impact of patient pooling and staff pooling remain limited, warranting further exploration.

Amidst these challenges and opportunities, this paper navigates through theoretical insights and practical implications. By drawing inspiration from queuing theory, we critically examine the theoretical underpinnings of patient pooling and staff pooling. By merging queuing theory's analytical precision with the pragmatic context of healthcare management, this paper aims to offer a comprehensive understanding of the potential of patient pooling and staff pooling as instrumental tools for redefining capacity optimization in healthcare systems.

We show that patient pooling and staff pooling emerge as innovative paths to reshape healthcare capacity management, as seen through the lens of queuing theory. Theoretical promises are bright, yet their real-world application requires keen attention to the practical barriers within each healthcare system. By merging the theoretical potential with practical adaptability, these pooling strategies offer a transformative framework that may optimize resource distribution and enhance patient care delivery across diverse healthcare settings.

2. PATIENT POOLING IN THEORY AND PRACTICE

Patient pooling, a strategic approach in healthcare capacity optimization, revolves around the centralization of patient assignment and bed allocation (e.g., Li et al., 2015). This strategy is also known in the literature as off-service placement (e.g., Song et al, 2020) or inpatient flow management (e.g., Winasti, 2018). From a queuing theory perspective, this approach introduces a significant departure from the traditional decentralized model, where each service unit operates independently with its own queue. To comprehend the implications of patient pooling, it is essential to explore its theoretical foundation within a queuing framework.

In the absence of pooling, each service unit functions as an independent server with its distinct queue. This scenario translates into a series of isolated queuing systems, where patients are confined to the queue of the specific unit they initially enter. Patient movement between units is precluded, leading to potential imbalances in resource utilization and waiting times. Any variations in demand across units remain isolated, resulting in suboptimal capacity utilization and potentially prolonged patient waiting times.

Conversely, patient pooling introduces a consolidated approach, where multiple service units share a common queue. This configuration allows patients to be dynamically allocated to available beds across different units based on real-time demand and resource availability. From a queuing theory standpoint, this shared queue mirrors a single integrated queuing system with multiple servers. As patient demand fluctuates, the centralized pool of patients can be intelligently routed to available beds, effectively mitigating imbalances in resource usage and minimizing waiting times.

The shift from isolated queuing systems to a unified, pooled queue fundamentally alters the dynamics of patient flow. Patient pooling serves as a mechanism for load balancing, where units experiencing high demand can tap into the excess capacity of units with lower demand. Queuing theory models offer insights into how this load balancing mechanism operates, providing an analytical foundation to predict how patient pooling impacts waiting times, resource utilization, and overall system performance.

In theory, patient pooling improves resource allocation in healthcare settings by centralizing patient assignments and bed utilization through a shared queue. This approach can enhance system efficiency in the following ways (Song et al., 2020; Winasti et al., 2018):

- Optimal Bed Utilization: In a traditional model without patient pooling, each unit operates independently with its own queue. This can lead to uneven patient distribution among units, resulting in underutilization of beds in some units and overcrowding in others. Patient pooling allows for a unified view of available beds across units, ensuring that patients are assigned to the most suitable and accessible beds in real-time. This reduces the likelihood of beds being left vacant while patients wait for admission.
- Dynamic Patient Routing: Patient pooling enables dynamic allocation of patients to available beds based on factors such as urgency, medical needs, and resource availability. This contrasts with the traditional model, where patients are confined to the queue of a specific unit. By centrally managing patient assignments, patient pooling optimizes patient flow, preventing bottlenecks and ensuring a smoother transition between care units.

- Load Balancing: Patient pooling facilitates load balancing across units by redistributing patient demand. Units experiencing high patient volume can benefit from excess capacity in units with lower demand. This redistribution minimizes the risk of resource overutilization in some areas and underutilization in others, leading to more consistent and efficient resource allocation.
- Reduced Wait Times: In a non-pooled system, patients might experience longer wait times due to imbalances in resource allocation. Patient pooling reduces wait times by enabling patients to be allocated to available beds across units, reducing congestion and streamlining patient flow. This results in quicker access to care and improved patient satisfaction.
- Adaptive Response to Demand: Patient pooling allows healthcare systems to adapt to fluctuating patient demand. As demand increases, the shared queue can prioritize and efficiently allocate patients to available beds, preventing delays in care provision. During periods of low demand, resources can be allocated to units with higher demand, optimizing capacity utilization.
- Flexibility in Resource Management: The flexibility inherent in patient pooling allows for resources, such as staff and equipment, to be allocated more dynamically across units. This responsiveness ensures that units have the appropriate level of resources to match patient needs, improving overall service quality.

Despite the apparent theoretical advantages, patient pooling may also be characterized by negative practical consequences, for example (Song et al., 2020; Winasti et al., 2018):

- Unequal Patient Prioritization: Patient pooling might lead to certain patient groups being prioritized over others. Those with more urgent medical needs could face delays if they are pooled with patients requiring less critical care, potentially compromising patient outcomes.
- Resource Imbalance: If not managed effectively, patient pooling could result in imbalances in resource utilization across units. Units with higher patient demand might become overloaded, leading to increased wait times and reduced quality of care.
- Complex Workflow Management: Centralized patient assignments can complicate workflow management, making it challenging to coordinate patient transfers between units, especially when specialized equipment or staff are needed. Implementing patient pooling requires changes to administrative processes, such as appointment scheduling and patient tracking, which could introduce logistical complexities.
- Patient Preferences: Some patients may have preferences for specific units or medical teams, and patient pooling could potentially limit their choice and disrupt the patient-provider relationship.
- Communication Challenges: Patient pooling may require improved communication and coordination between healthcare professionals in different units, which could introduce communication challenges and possible errors.
- Privacy and Confidentiality Concerns: Centralized patient pooling may raise concerns about patient privacy and confidentiality, as medical information may be shared across different units.
- Resistance to Change: Healthcare professionals and staff members might resist the shift to patient pooling due to concerns about changes in work routines, job roles, and unfamiliar processes.

3. STAFF POOLING IN THEORY AND PRACTICE

Staff pooling stands as a pivotal approach within healthcare capacity optimization, focusing on the dynamic allocation of healthcare professionals and support staff across multiple service units (e.g., Fagefors et al., 2022). This strategy is also known in the literature as flexible resource allocation (e.g., Thorwarth et al, 2016). A fundamental departure from the conventional model, staff pooling redefines the allocation of workforce capacity within a queuing theoretical framework.

In the absence of staff pooling, each service unit operates as an individual server, possessing its own dedicated capacity. This configuration translates into a scenario where units function independently with limited capacity interaction. The capacity of each unit remains isolated, leading to potential inefficiencies when fluctuations in patient demand occur. Units with excess capacity may underutilize resources, while units with high demand may struggle to meet patient needs due to insufficient staffing.

Staff pooling introduces a paradigm shift by enabling the sharing of workforce capacity across units. In the extreme case of full staff pooling, all service units share a common workforce pool, allowing for the aggregation of staff resources to meet the overall system's demands. Queuing theory models provide insights into how staff pooling redistributes capacity allocation, optimizing workforce utilization and enhancing patient care delivery.

The core principle of staff pooling is its ability to adaptively allocate staff resources based on real-time patient demand. As patient loads vary across units, staff can be dynamically allocated to ensure that each unit has adequate coverage. This dynamic allocation addresses imbalances in staffing levels that might occur under the non-pooled model. Staff pooling's queuing theoretical model illuminates the capacity reallocation process, showcasing how units with excess capacity can assist those with higher demand and/or capacity deficits, fostering an equilibrium in workforce utilization.

Furthermore, staff pooling bolsters the system's flexibility in responding to sudden changes in patient volume. The shared workforce capacity allows for a more efficient allocation of staff during peak demand periods, ensuring that patient care needs are met promptly. During quieter periods, staff can be redistributed to units experiencing higher demand, thereby maximizing workforce efficiency.

In summary, the queuing theoretical model of staff pooling highlights its transformative impact on workforce allocation. By introducing a dynamic approach to staff sharing, staff pooling optimizes workforce utilization, mitigates inefficiencies caused by uneven patient demand distribution, and enhances the overall quality of patient care.

Staff pooling enhances resource allocation in healthcare settings by promoting the dynamic sharing of healthcare professionals and support staff across different service units. This approach can improve system efficiency in the following ways (Bouma, 2022; Fagefors et al., 2022; Thorwarth et al., 2016):

- Optimal Workforce Utilization: In a traditional model without staff pooling, each service unit operates with its own dedicated staff capacity. This can result in imbalances, where some units are overstaffed while others are understaffed. Staff pooling allows for a more flexible allocation of staff resources, ensuring that staff are efficiently distributed to units with higher patient demand. This minimizes underutilization of staff capacity and reduces instances of staff shortages.

- Adaptive Response to Patient Demand: Staff pooling enables healthcare systems to dynamically respond to changes in patient volume. As patient demand increases, staff resources can be shifted to units where they are most needed. This prevents understaffed units from becoming overwhelmed and ensures that patients receive timely and high-quality care during peak periods.
- Equitable Resource Allocation: Staff pooling helps in achieving a more equitable distribution of resources. Units with lower patient loads can contribute their surplus staff to units facing higher demand, promoting a balanced distribution of workload. This approach prevents the burden of patient care from falling disproportionately on a few units, leading to a more uniform distribution of staff resources.
- Enhanced Staff Flexibility: Staff pooling encourages a more flexible and adaptable workforce. Staff members can be allocated to different units based on patient needs, ensuring that specialized skills are efficiently utilized across the healthcare facility. This flexibility enables staff to gain diverse experience and contribute to patient care where their expertise is most valuable.
- Cost Efficiency: By optimizing staff allocation, staff pooling can lead to cost savings. Rather than maintaining excess staff in each unit to handle peak demand, staff can be allocated to units as needed, reducing the need for overstaffing. This can lead to reduced labor costs while maintaining high-quality patient care.
- Improved Patient Care: Staff pooling enhances patient care by ensuring that units have adequate staffing levels to provide optimal care. Balanced staffing levels reduce the likelihood of delays in patient care and help maintain a high standard of care delivery. Patients benefit from shorter wait times, increased attention from healthcare professionals, and improved overall satisfaction.
- Reduction of Staff Burnout: Staff pooling can mitigate the strain on individual units and prevent staff burnout. By sharing the workload across units, staff are less likely to be overwhelmed during high-demand periods. This can contribute to higher morale, lower turnover rates, and a more sustainable work environment.

Despite the theoretical advantages, staff pooling may also be characterized by negative practical consequences, for example (Bouma, 2022; Fagefors et al., 2022; Song et al., 2020; Thorwarth et al., 2016):

- Skill Mismatch: Reallocating staff to different units might lead to a mismatch between the required skills in each unit and the skills possessed by the staff members, potentially impacting patient care quality.
- Reduced Continuity of Care: Frequent staff rotations between units due to pooling could result in reduced continuity of care for patients, as they may interact with different healthcare professionals during their treatment. Patients may develop relationships with specific healthcare providers, and staff pooling might disrupt these relationships, affecting patient satisfaction and engagement.
- Staff Resistance: Healthcare professionals might resist staff pooling due to concerns about unfamiliar workflows, changes in team dynamics, and potentially diminished job satisfaction. Constantly shifting staff members to different units to meet fluctuating demands might also increase stress and burnout due to the need for rapid adaptation and variable workloads.
- Loss of Unit Identity: Staff pooling might dilute the unique identity and culture of individual units, affecting teamwork, morale, and unit-specific processes.

- Communication Challenges: Frequent movement of staff members between units could lead to communication gaps and misunderstandings, impacting effective information sharing and patient handoffs.

4. ILLUSTRATIVE NUMERICAL EXAMPLE

In this section, we present a number of illustrative and fictitious numerical examples to showcase the potential benefits of patient pooling and staff pooling within a queuing theoretical framework. The foundation of these examples is a small healthcare system with two separate units. Each unit has an arrival intensity of 1 patient per hour and a service capacity of 1.5 patients per hour. Clearly, if there was no variation in neither the arrival process nor the service process, there would arrive exactly 1 patient per hour to each unit, and each patient would be served in exactly 40 minutes. There would be no queues anywhere, as long as the service capacity in each unit exceeds the arrival intensity. In such situations, pooling approaches would not create any value.

Now, let us model both the arrival and the service processes in each unit as Poisson processes. This implies that the arrival intensities and the service capacities are treated as expected values rather than constants, around which there is random variation.

4.1. Example 1: Patient pooling

In the absence of patient pooling, the two units operate independently with their own queues. In technical terms, the healthcare system consists of two parallel and independent M/M/1 queuing systems (Lantz, 2019). Patients arriving at each unit are served according to the available capacity. However, this may lead to imbalanced resource utilization and varying waiting times between the units. It can be shown that a patient arriving to any unit should expect a queuing time of on average 1 hour and 20 minutes and a throughput time (i.e., queuing time plus service time) of on average 2 hours.

By implementing patient pooling and creating a single common queue for both units, patients are assigned to available service regardless of the unit they entered. In this situation, where the healthcare system has been transformed to one M/M/2 queuing system (Lantz, 2019), it can be shown that a patient arriving to any unit should expect a queuing time of on average 32 minutes and a throughput time of on average 1 hour and 12 minutes. Hence, the introduction of patient pooling creates a substantial reduction in the expected queuing time without adding any additional capacity, as patients can be efficiently allocated to available staff across the combined capacity of both units. Patient pooling thus ensures a more balanced utilization of resources and a smoother patient flow.

4.2. Example 2: Staff pooling

Let us revert to the original situation with two units operate independently with their own queues. Assume that some of the total staff capacity is pooled, so that service capacity corresponding with 0.25 or 0.5 patients per hour can be transferred between the units. As long as the arrival intensities to the units are equal, the optimal configuration of the system as a whole is clearly to keep the service capacities equal too. However, assume that some fundamental change occurs so that 60% of all patients arrives at unit 1 and 40% of all patients arrives at unit 2. By keeping the service capacity at 1.5 patients per hour in both units (i.e., by not transferring any staff in the pool from unit 2 to unit 1), it can be shown that a patient arriving to unit 1 should expect a queuing time of on average 2 hours and 40 minutes, while a patient arriving to unit 2 should expect a queuing time of on average 46 minutes. The average throughput times are 40 minutes longer than the average queuing times. However, if staff corresponding with a service capacity of 0.25 patient per hour were transferred from unit 2 to unit 1, the expected average queuing times would become 1 hour and 15 minutes, and 1 hour and 25 minutes, respectively.

Now, assume that some fundamental change occurs so that 70% of all patients arrives at unit 1 and 30% of all patients arrives at unit 2. By keeping the service capacity at 1.5 patients per hour in both units (i.e., by not transferring any staff in the pool from unit 2 to unit 1), it can be shown that a patient arriving to unit 1 should expect a queuing time of on average 9 hours and 20 minutes, while a patient arriving to unit 2 should expect a queuing time of on average 27 minutes. However, if staff corresponding with a service capacity of 0.25 patient per hour were transferred from unit 2 to unit 1, the expected average queuing times would become 2 hour and 17 minutes, and 44 minutes, respectively. And if staff corresponding with a service capacity of 0.5 patient per hour were transferred from unit 2 to unit 1, the expected average queuing times would become 1 hour and 10 minutes, and 1 hour and 30 minutes, respectively. Thus, the ability to redistribute capacity optimally across units enhances the system's responsiveness to sudden demand changes. Staff that can be flexibly shifted between units thus ensures a more balanced workload and utilization.

4.3. Discussion of results

The presented numerical examples highlight the potential of patient pooling and staff pooling in healthcare capacity management. Patient pooling demonstrates how a single queue for multiple units can lead to improved resource utilization and reduced waiting times. Staff pooling showcases the adaptability of capacity distribution, allowing the system to effectively respond to changing demand scenarios. These examples, based on simple yet foundational parameters, underscore the theoretical insights of these strategies within a queuing theoretical framework. While these examples provide a conceptual understanding of the power in pooling approaches, practical implementation would require substantial consideration (Fagefors et al., 2022).

5. CONCLUSION

The exploration of patient pooling and staff pooling within a queuing theoretical framework has illuminated their potential to reform healthcare capacity management. Theoretical modellings have unveiled the underlying mechanisms through which these strategies optimize resource allocation, patient flow, and ultimately enhance the quality-of-care delivery. In this concluding discussion, we interpret the theoretical insights gained from patient pooling and staff pooling, examine their practical implications, address challenges in implementation, and propose directions for future research.

Patient pooling, as seen through queuing theory, offers a structured approach to mitigating resource imbalances and reducing patient waiting times. Centralized patient assignments and shared queues allow for efficient utilization of available beds, smoothing patient flow, and improving overall resource allocation. Staff pooling, on the other hand, promotes dynamic workforce allocation. By sharing some staff resources, this strategy can optimize service levels and maintains patient care quality while increasing staffing efficiency.

While the theoretical underpinnings of patient pooling and staff pooling are promising, translating these concepts into real-world healthcare systems requires careful consideration. Practical implementation hinges on factors such as existing organizational structures, professional competence, technological capabilities, and stakeholder engagement. Challenges may arise in ensuring smooth transitions from traditional operational models to pooling approaches, addressing concerns related to staff flexibility, patient preferences, and potential disruptions during implementation.

At their core, patient pooling and staff pooling are both strategies that can be used to increase resource utilization. When considered as mutually exclusive alternatives, patient pooling is generally the better strategy in contexts where staff and/or other resources required to provide care are significantly less mobile than patients. Conversely, when staff and other required resources can be transferred between units more easily than patients, staff pooling is typically a good strategy. However, these strategies can also be used in combination. The synergy between patient pooling and staff pooling presents an enticing avenue for capacity optimization. The combination of these strategies holds the potential to offer enhanced patient flow, reduced waiting times, and improved resource allocation. However, implementing combined pooling approaches necessitates a delicate balance between patient needs, staff distribution, and systemic adaptability. Strategies for effectively integrating these approaches warrant further investigation to maximize their collective benefits.

To conclude, patient pooling and staff pooling, when viewed through the lens of queuing theory, offer innovative pathways to reform healthcare capacity management. While the theoretical insights are promising, their successful implementation demands attention to practical challenges and considerations specific to each healthcare system. By fostering a dialogue between theoretical insights and real-world applicability, this study contributes to the ongoing discourse on reshaping capacity optimization in healthcare. As researchers, practitioners, and policymakers collaborate, we can navigate the intricacies of implementation, steer through challenges, and unlock the full potential of combined pooling approaches in enhancing patient care and resource utilization.

The theoretical foundation laid by this study paves the way for future research endeavors aimed at validating the practical applicability and effectiveness of patient pooling and staff pooling in diverse healthcare contexts. Future research could delve deeper into real-world case studies based on operational data and empirical validations based on experimental approaches to fully unlock the potential benefits of patient pooling, staff pooling, and their combined implementation in diverse healthcare contexts.

REFERENCES

Bouma, S. (2022). *On the Role of Procurement in Inter-Organisational Staff Pooling in Healthcare*. Master thesis. Department of Technology Management and Supply, Faculty of Behavioural, Management and Social Sciences. https://purl.utwente.nl/essays/93647

Fagefors, C., Lantz, B., Rosén, P., & Siljemyr, L. (2022). Staff pooling in healthcare systems – results from a mixed-methods study. *Health Systems*, in press. https://doi.org/10.1080/20476965.2022.2108729

Lantz, B. (2019). *Kömodeller – teori och övningar*. Studentlitteratur, Lund. https://www.studentlitteratur.se/9789144129846

Li, S., Geng, N. & Xie, X. (2015). Radiation Queue: Meeting Patient Waiting Time Targets. *IEEE Robotics & Automation Magazine*, 22(2), 51-63. https://dx.doi.org/10.1109/MRA.2015.2401232

Proudlove, N. C. (2020). The 85% bed occupancy fallacy: The use, misuse and insights of queuing theory. *Health Services Management Research*, 33(3), 1-12. https://doi.org/10.1177/0951484819870936

Song, H., Tucker, H., Graue, R., Moravick, S. & Yang, J. J. (2020). Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science*, 66(9), 3825–3842. https://doi.org/10.1287/mnsc.2019.3395

Thorwarth, M., Rashwan, W. & Arisha, A. (2016). An analytical representation of flexible resource allocation in hospitals. *Flexible Services and Manufacturing Journal*, 28, 148–165. https://doi.org/10.1007/s10696-015-9216-4

Winasti, W., Elkhuizen, S., Berrevoets, L., van Merode, G. and Berden, H. (2018). Inpatient flow management: a systematic review. *International Journal of Health Care Quality Assurance*, 31(7), 718-734. https://doi.org/10.1108/IJHCQA-03-2017-0054