Chapter 1

Literature Review

1.1 Introduction and Definition of Key Terms

This literature review situates the current research within the existing body of academic work on the mathematical modeling of hospital bed occupancy. As established in the problem statement, the efficient management of patient flow is a critical challenge in healthcare operations, where the goal is to mitigate both costly underutilization of resources and dangerous overcrowding that can compromise patient care [**?**]. To address this complex optimization problem, queuing theory provides a robust and well-established analytical framework. The following key terms, central to this research, are defined to ensure clarity.

1.2 Definition of Key Terms:

Queuing Theory: A branch of mathematics dedicated to the study of waiting lines, or queues. It uses mathematical models to represent, analyze, and predict the behavior of systems where entities (e.g. patients) arrive for a service, wait if all servers are busy, and then leave after receiving service.

Markovian Property (Memorylessness): A core assumption in many basic queuing models. It stipulates that the future state of a process depends only on its present state, not on its past history.

- **Poisson Process (for Arrivals, λ):** An arrival process is Poisson if the number of arrivals in any time interval follows a Poisson distribution, and the times between successive arrivals are independent and exponentially distributed. This is a direct consequence of the Markovian property and is used to model random, independent patient arrivals.

- **Exponential Distribution (for Service, μ):** A service process is exponential if the time taken to serve a single entity follows an exponential distribution. This implies that shorter service times are more probable than longer ones and that the remaining service time for a patient is independent of how long they have already been in service.

M/M/c/K Notation: A standard shorthand, known as Kendall's notation, for describing a queuing model:

- **M (Arrivals):** Denotes a Markovian (Poisson) arrival process.

- **M (Service):** Denotes a Markovian (Exponential) service time distribution.

- **c (Servers):** Represents the number of parallel servers. In the context of this study, the servers are explicitly defined as the **healthcare staff** who provide the direct care that facilitates patient discharge.

- **K (Capacity):** Represents the total system capacity (patients being served + those waiting). For this research, K corresponds to the **total number of physical beds** in the ward.

Key Performance Indicators (KPIs): These are metrics used to evaluate the efficiency of the queuing system. Common KPIs include the probability of having n patients in the system (P_n), the average waiting time (W_q), system utilization (ρ), and the Blocking Probability (P_K), which is the probability that an arriving patient is turned away because the system is full (i.e., all beds are occupied).
This review will now examine the literature as it pertains to each of the study's three specific objectives, culminating in a synthesis that identifies the precise research gap this study intends to fill.

1.3 Literature Review by Specific Objectives

1.3.1 Modeling Bed Occupancy with the M/M/c/K Framework (Objective 1)

The first objective of this research is to model hospital bed occupancy using the M/M/c/K queuing framework. The literature has firmly established this model as a foundational tool for healthcare systems analysis due to its ability to represent systems with finite capacity. While early studies used infinite-capacity models (e.g., M/M/c), the work of researchers like [**?**] highlights that the physical constraint of a fixed number of beds makes finite-capacity models essential for realistic analysis.

The M/M/c/K model's primary strength lies in its ability to calculate the probability of the system being full, known as the blocking or rejection probability. This is of paramount importance to hospital administrators who must manage patient diversions. The probability that an arriving patient is rejected is the probability that the system is in state K, which can be calculated using the steady-state probability formula [**?**]:

$$P_{\text{reject}} = P_K = \frac{\frac{(\lambda/\mu)^K}{c!c^{K-c}}}{\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!} \sum_{n=c}^{K} \left(\frac{\lambda}{c\mu}\right)^{n-c}} P_0 \tag{1.1}$$

where P_0 is the probability of the system being empty. Almomani2023 directly validate the use of such "Finite Queuing Models," demonstrating their accuracy in predicting real-world bed occupancy and related KPIs. Their work confirms that M/M/c/K provides a reliable structural baseline.

The practical utility of this framework was further proven during the COVID-19 pandemic. [**?**] used the M/M/c/K model to develop a dynamic bed allocation strategy, using the model's predictive power to manage extreme patient surges and minimize patient rejection rates. Similarly, [**?**] analyze "finite buffer queues" in healthcare, providing a thorough mathematical treatment of the model's properties and reinforcing its suitability for systems with hard physical constraints. These foundational works establish the M/M/c/K model as the appropriate starting point for analyzing hospital bed occupancy, providing a robust mathematical structure upon which more complex, nuanced analyses can be built.

1.3.2 Influence of Staff Experience on Treatment Times (Objective 2)

The second objective is to evaluate how different levels of staff experience affect treatment times and bed turnover. This stream of literature moves beyond static structural modeling to consider the qualitative nature of healthcare staff, arguing that not all "servers" are equal. While less commonly integrated into analytical queuing models, the impact of staff experience is well-documented in healthcare management and operations research. The work of [**?**] using a data-driven "P Model" for inpatient bed assignment showed that service times are significantly influenced by "provider factors," lending strong empirical support to the premise that staff characteristics are a critical variable.

To integrate this concept into a queuing model, researchers have proposed modifying the service rate parameter, μ. One approach is to define an effective service rate, μ_{eff}, as a weighted average of the service rates of different experience cohorts within a team [**?**]. If a team consists of m experience categories (e.g., junior, mid-level, senior), the effective service rate can be modeled as:

$$\mu_{\text{eff}} = \sum_{i=1}^{m} w_i \cdot \mu_i \tag{1.2}$$

where w_i is the proportion of staff in experience category i, and μ_i is the service rate for a staff member in that category. Another approach models the efficiency of an individual staff member as a function of their years of experience, y, often using an exponential learning curve. This captures the rapid initial skill acquisition followed by diminishing returns:

$$\mu(y) = \mu_{\max}(1 - e^{-\beta y}) \tag{1.3}$$

where $\mu_{\max}$ is the maximum service rate achieved by a fully experienced staff member, and β is a parameter controlling the learning rate. These formulations allow for a more granular analysis, enabling administrators to assess the operational impact of different hiring and staffing strategies, such as determining an optimal mix of senior and junior staff to balance cost and efficiency.

1.3.3 Impact of Staff-to-Patient Ratios on Performance (Objective 3)

The third objective is to assess the impact of varying staff-to-patient ratios on key performance indicators. This extensive stream of literature addresses the quantitative aspect of staffing. The research here explicitly links the number of staff members to system efficiency, moving beyond the simple assumption of a fixed number of servers, c. Jiang2024, for example, provide a comparative analysis of traditional and new methods for "calculating optimal patient to nursing capacity," framing the staff-to-patient ratio as a central determinant of operational performance and service quality.

To model this relationship mathematically, the service rate can be made dependent on the number of staff, s, and the number of patients currently in the system, n. One powerful formulation proposed in the literature treats the service rate as load-dependent Asaduzzaman2020:

$$\mu(s, n) = \mu_0 \cdot \min\left(1, \alpha \cdot \frac{s}{n}\right) \quad \text{for } n > 0 \tag{1.4}$$

where μ_0 is the baseline service rate under ideal (non-overloaded) conditions, s is the number of staff members on duty, n is the number of patients, and α is a scaling factor representing the maximum number of patients a single staff member can effectively manage. This formula elegantly captures the reality of service degradation under pressure: as the patient load per staff member (n/s) increases beyond a certain threshold ($1/\alpha$), the effective service rate per patient decreases proportionally. This reflects the real-world effects of staff being spread too thin during periods of high demand, leading to longer treatment times and increased bottlenecks. This approach directly connects staffing levels to core performance metrics like waiting time and throughput, allowing for data-driven decisions on nurse staffing mandates and resource allocation.

1.4 Gap Analysis and Summary

A synthesis of the literature across these three distinct but interconnected streams reveals a clear and compelling research gap. The academic inquiry has progressed logically through three stages. The first stream of literature successfully established the M/M/c/K model as a structurally valid framework for analyzing hospital bed

management, yet it primarily treated the service process as static. The second stream provided crucial insights into the qualitative impact of staff experience on efficiency, but these findings were rarely integrated back into tractable, analytical queuing models. The third stream effectively modeled the quantitative impact of staff-to-patient ratios, but in doing so, it often treated staff as homogenous, interchangeable units, overlooking the significant performance variations due to experience.

The critical gap, therefore, lies at the intersection of these three streams: there is a lack of a unified, analytical model that integrates the structural M/M/c/K framework with a service rate parameter that is a dynamic function of *both* the quantity (staff-to-patient ratio) and the quality (experience mix) of the healthcare staff.

This research aims to fill this precise gap. By developing and analyzing an enhanced M/M/c/K model that incorporates these multifaceted staffing variables, this study will provide a more comprehensive and actionable tool for hospital administrators. Acknowledging the study's own limitations, such as the assumption of Markovian processes and the focus on a single department, this research serves as a critical step toward a more holistic and realistic understanding of hospital operations. It provides a robust foundation for evidence-based decision-making and paves the way for future research into more complex, non-Markovian systems with interconnected departments and additional human factors like staff fatigue.

.