IJAMC

International Journal of Applied Mathematics and Computing

E-ISSN: 3047-146X
p-ISSN :3048-1988

Using Mathematical Programming to Analyze and Improve Robust Queue Management in Healthcare Systems

Hasanain Hamed Ahmed*

Department of Administration Management, Faculty of Imam Alkadhum University College, Baghdad, Iraq.

*Corresponding Author : hasnin.hamid@iku.edu.iq

Abstract: Efficient management of patient queues is essential in healthcare systems to ensure timely care, optimize resource utilization, and enhance patient satisfaction. Mathematical programming, particularly when applied in conjunction with queuing theory and optimization models, provides a rigorous framework for analyzing and improving healthcare service delivery. This approach involves modeling arrivals and service processes, applying queuing models (such as single-server, multi-server, and priority queues), and formulating optimization objectives—often to minimize total costs, patient waiting times, or resource idling. Recent research demonstrates that combining queuing theory with mixed-integer programming and simulation techniques enables healthcare managers to allocate resources dynamically, set staffing levels, and assign priorities among different patient categories. For example, the use of mixed-integer programming can determine the optimal number of servers, beds, and service rates based on patient flow and priority needs, striking a balance between reducing waiting times for critical cases and controlling operational costs. These mathematical models also account for practical constraints and stochastic variability inherent in clinical settings. Applications span emergency departments, outpatient clinics, and even pharmacy and blood service centers—showing significant improvements in system efficiency, reduced patient wait times, and enhanced overall care quality. Thus, mathematical programming is a powerful decision-support tool for queue management, offering evidence-based strategies to address congestion and resource allocation challenges in complex healthcare environments.

Keywords: Analyzing for Queue; Improving for Robust Queue; Mathematical Programming; Queue Management; Robustness.

1. Introduction

Healthcare systems are growing increasingly complex, and the demand for healthcare services continues to increase. At the same time, health systems are under constant pressure to provide a higher level of service quality, reduce waiting times, and reduce sectoral costs. These demands must be addressed in a context of extreme uncertainty and randomness in patient arrivals, patient visit duration, and surgery durations.

Queues and waiting are inherent to nearly every healthcare system, where there is frequently a significant imbalance between supply and demand. Therefore, mathematical programming methods have been successfully applied to support decision-making in managing queues in healthcare systems. These methods can be used at the strategic/medical planning and operational levels.

Received: April, 10th 2025
Revised: April, 28th 2025
Accepted: May, 16th 2025
Online Available: August, 29th 2025
Curr. Ver.: August, 29th 2025

2. Literature Review

Queuing-inventory systems are critical components of the supply chain in many service organizations for efficient management of both inventory and customer service (K et al., 2023). Sebastian presents a finite-source inventory system and a service facility where demands are postponed. The service facility features a single server and a finite waiting hall of size K, which includes the customer receiving service. The analysis is carried out under assumptions of quasi-random arrival processes, phase-type service times and lead times, and predefined selection rules for choosing postponed customers. Numerical studies reveal that the squared coefficient of variation (SCV) of the service time and lead-time distributions significantly influences the optimal policy values. Consequently, SCV becomes essential in deciding ordering policies, selection strategies, and the appropriate waiting room size. Controlling the SCV enables minimizing costs or waiting times associated with lead times or service durations.

Mohammad Rahiminia pioneers a mathematical healthcare and waste-management model developed via a data-driven approach. Employing bulk service and two distinct multiserver queuing models, he presents a nonlinear multi-shift framework to manage congestion within the medical centre and the generation of medical waste during the pandemic. Patients are categorized according to their current health condition and analyzed using M/M/C queuing systems to control congestion levels. Integrating healthcare waste-queueing systems connected to the medical centre utilizes a single-server Markovian queue with bulk service M/M[y]/1. This approach combines a balking queuing system for inventory waste with two patient-service queuing systems.

Mathematical Programming Approaches

Mathematical programming is well-suited for modelling and optimizing queue management system design in a healthcare system (Rihm, 2017). Equations describing the system are formulated and solved as a mathematical programming problem, such as a mixed-integer linear or nonlinear optimization model. The objective function is based on reducing queue waiting times or lengths, increasing resource utilization rates, and other system performance indicators, to identify the best system configuration. Candidate solutions can be evaluated with analytical expressions and queueing theory.

Corresponding constraints ensure the system configuration is described correctly. The available system configurations are typically represented as alternatives or scenarios, encompassing the entire configuration space, so the identified optimal solution yields a highly robust design. Constraints also incorporate management policies required to be implemented. When analytical expressions describing quantitative performance indicators are unavailable, outputs from queueing or simulation models can be implemented iteratively within a mathematical programming framework to ensure optimal solution configurations (Kumar Yadav et al., 2021). Other iterative methods can also be combined with optimization techniques to efficiently determine robust queue management arrangements without fast analytical procedures.

Formulating the Problem

The first essential step in mathematical programming for queue management is formulating the problem, which involves defining the objective function, decision variables, and constraints. Subsequently, one or more algorithms are applied to identify either an optimal or suboptimal solution. Analysis of queueing systems under both steady-state and transient conditions helps to determine whether waiting times will become unacceptably long or the queue is vulnerable to disruption: such an assessment tests the robustness of a queue management strategy. This approach is especially appropriate for evaluating proposals intended to mitigate queue congestion. As the queueing theory literature addresses approaches to assessing queue robustness in detail, the analysis concentrates on surveying various mathematical programming methods devised to model queueing situations. These methods then indicate ways in which the robustness of queue management strategies may be improved.

Queueing operations may be designed to be resilient to potential disruptions either by enhancing robustness—that is, making the system less sensitive to the disruption—or by enhancing recovery—enabling the system to restore its performance levels more quickly. For transient queues, appropriate objective functions include integral measures, such as minimizing the expected time-integrated cost of servicing the queue, which represents a balance between robustness and recovery. In addition, the transient nature of queues can also be incorporated into the constraints by limiting the queue size at all times rather than

in a steady state. Although similar solution techniques can be used, the discussion here addresses only the design of operations that are robust to disruptions.

Solution Techniques

Improving the queue management system's robustness involves effectively modeling and analyzing queues across diverse healthcare settings. Mathematical programming offers a rigorous approach to quantify the stochastic behavior of queues and assess the resilience of queue management procedures within healthcare systems (Ben Othman et al., 2018).

Many models for queue management rely on data and measurements from the healthcare system of interest. Careful consideration of the parameters to be utilized or estimated is essential since the degree of trustworthiness and the volume of historical data available can significantly impact the robustness of the analysis. Simulation techniques are typically used for validation (Jiang et al., 2019).

Case Studies

Mathematical programming and queuing models have been explored for various healthcare problems.

(Eugene Helm, 2012) develops mixed-integer-programming models for operating room resource allocation; the models optimize the surgical schedule, minimize costs, create fair and equitable procedure allocations, and reduce patient wait times.

(Jiang et al., 2019) Apply a queue-based approach to provider staffing configuration in hospital emergency departments; the models help administrators balance patient inflow, resource constraints, and provider availability in a dynamic environment. An M/M/R/N queuing model describes patient arrivals as a Poisson process feeding a single queue served by R identical providers with capacity N; performance measures such as staffing requirements and patient waiting times are estimated analytically.

(Ben Othman et al., 2018) Provide a comprehensive review of scheduling approaches considered for healthcare managers; a broad range of patient-scheduling problems, mathematical models, and solution procedures relevant to the emergency department are identified and categorized. Exact and heuristic methods have been proposed, including mathematical programming models, network programming, goal programming, and Varangian-based heuristics. Several studies address the scheduling of hospitals, nurses, and diagnostic resources. Approaches for staff scheduling problems, cyclic and non-cyclic nurse scheduling, multi-priority patient scheduling, and uncertainty-tolerant behaviors are also presented.

Emergency Departments

Emergency Departments (EDs) are the most familiar point of entry to hospitals and, consequently, have become an integral part of local communities. Those community hospitals that do not have an ED have become unpopular and eventually tend to close (E Hurwitz et al., 2014). Many challenges ED management and administration face are consequences of unpredictable demand and infeasible load. Patients tend to require resources and services in parallel and serial fashion while the available resources are limited (Choubey et al., 2020). The pressure on hospitals to minimize waiting times and to provide exemplary standards of patient care is considerable and is unlikely to diminish. Mathematical programming formulations can be employed to determine optimal allocation of human and material resources in EDs. Such models focus on minimizing patient waiting times by effectively scheduling and distributing available beds and staff.

Outpatient Clinics

Outpatient clinics face challenges related to variable and limited-capacity service due to constraints on the number of physician work hours and the start times of subsequent treatments (Laan et al., 2017). Appointment scheduling is traditionally used to organize patient flow and regulate access times. The outpatient service is subdivided into categories, and the allocation of service quantity across these categories is modeled as an integer linear programming problem. This approach enhances physician work efficiency, significantly reduces patient waiting times, and improves the overall quality of medical service (Hua et al., 2023). The model for outpatient appointment scheduling typically constitutes a stochastic mixed-integer program (SMIP) or a two-stage stochastic programming (TSSP) model. The SMIP is approximated by a deterministic mixed-integer program (MIP) to obtain unique and stable solutions, which assumes that a known fraction of appointment blocks may be canceled each week. The MIP determines a base weekly schedule of appointment slots and guides the numerical study of alternative strategies. A flexible schedule that adjusts to realized capacity variations can be derived from the MIP solution. Iterative model formulations for both queuing and SMIP frameworks facilitate the

incorporation of performance measures beyond simple averages, such as ensuring that the nominal access time upper bound is met for a specified percentage (e.g., 90%) of patients. The robust scheduling approach also applies to diagnostic clinics, where appointments are subject to dedicated capacity allocations and system capacity uncertainties (Kiani et al., 2019).

Inpatient Services

Hospital inpatient services require significant resources from hospital administrators, particularly regarding allocating capacity-demanding patient admissions. In bed-constrained environments, short-term strategies for reallocation must accompany long-term plans to support growth and expansion effectively. A burst-period model addresses this decision-making problem, allowing the temporary reassignment of existing patients to alternative floors or the cancellation of already-scheduled admissions (M. Thompson et al., 2009).

Using medical criteria, beds on each floor are identified; only appropriate floors may be used for each patient category. Single-sided bed blocks constrain reallocation options further. Although floor-hopping rules could allocate the smallest number of patients to their lowest-preference floors, this approach is often impractical. Decision-makers favor solutions that allow incremental flexibility, performing the least disruptive changes necessary. The model implements a three-stage sequence involving candidate solutions and employs an adjacency-pair model that favors continuity and incremental changes.

Optimal reallocation balances removing the minimum number of patients from their preferred floor while minimizing the number of patients assigned to lower-preference floors. A flexible approach also minimizes moves for lower-priority categories, ensuring fair and manageable patient transfers. Computational experiments based on data from a publicly funded hospital support the model and provide insights into its application.

Robustness in Queue Management

Robust optimization provides a mechanism for converting probabilistic inequalities into deterministic linear constraints, effectively characterizing sets of probability mass functions with specified characteristics. For discrete-time uncertainty sets, each probability mass function from the family generates a stationary distribution that resides within the defined polytope, yielding a polyhedral description of a family of stationary distributions. Consequently, performance measures such as average waiting times are guaranteed to remain below specified values when computed to any stationary distribution contained in this polytope (J. Bertsimas et al., 2009).

By applying robust optimization to multiclass single-station queueing networks, polyhedral uncertainty sets can be constructed for key performance metrics—including arrival counts, average service times, and routing probabilities—in a manner that aligns with available information. These uncertainty sets induce a family of associated stationary distributions, which, despite their accessibility, pose challenges for direct performance measure evaluation. Additional constraints, motivated by requirements like multi-stage service capacity and product-mix stationarity, can define reduced uncertainty sets for performance estimation. This framework is applicable across various priority structures—such as first-come-first-serve, fixed priority, and shortest-job-first—and permits the treatment of stationary marking and thinning policies as exogenous factors. It facilitates the derivation of analytically tractable yet probability-law-compliant estimates for maximum average waiting times for aggregate classes of heterogeneous customers within multiclass single-station, multi-server queueing systems.

Implementation of Solutions

Solving mathematical models to optimize hospital queuing systems yields prescription solutions for resource levels and patient admission sequences. Implementation can be either direct—immediately translating results into practice—or indirect—formulating abstract policies from which models are redeveloped or integrated within simulation platforms. Indirect implementation offers enhanced flexibility and confidence in field applications.

Developed software libraries compatible with C++, C#, and JavaScript, enabling the direct incorporation of development tools into information systems. While focused on the healthcare sector, these libraries can extend to other discrete event systems, such as manufacturing.

Software Tools

Modelling queuing systems is a broad subject where one of the objectives is to obtain quantitative performance indicators. Software packages that assist in modelling these systems can provide valuable decision support. Dimitrov (2012) developed software that uses VBA to determine queuing system indicators (Dimitri Dimitrov, 2012). Mathematical programming also supports operational decisions related to queue management. The sample software applications contain functionalities to optimize staff allocation: allocating staff to workstation stations; allocating a group of workers to queues; and a mixed model where both the queues and the workers are optimized. In-service scheduling problems are addressed to formulate and to solve with a mixed integer nonlinear programming model. This model is applied to manage the appointments for in-services purely based on duration with random service demands and random no-shows, where nonlinear cost functions occur (S. Schulz & Udwani, 2019). A queue-based approach proposes a decision support system to optimize the allocation of providers at the hospital emergency department. The staff requirements are estimated through a mathematical model based on the queueing theory (Jiang et al., 2019).

Real-world Applications

Queueing management is an essential topic in healthcare systems, where optimization models provide suitable decision-making tools to evaluate the impact of different policies. In this work, the authors formulate a multi-objective mathematical programming model as an optimization tool and apply it to a real-life healthcare service provider (Jiang et al., 2019). A production-cost-optimization model appears to be the preferred method in healthcare systems to estimate costs. Given the number of products and services offered, this approach often involves large-scale mathematical programs. Nonetheless, the interest in using such models within healthcare services has gained importance in the last decade. Because public health centers feature particular characteristics not present in generic systems, adopting these models is often complicated, leading to estimates that often do not reflect the real situation. The problem is relevant for public providers whose income depends on the number of patients treated, e.g., the earnings of the Spanish public health system are directly related to the number of patients treated/attended to.

Policy makers who seek to manage long queues and apply efficient appointment systems require a set of methodologies and techniques that provide information, data, and reliable solutions to estimate waiting times, assignment policies, and the impact on economic figures. The specific context of healthcare systems is challenging, as it involves mixed clientele and services related to private and/or public provider systems. Under these circumstances, standard problem formulations used in queuing models and systems of appointments do not reflect the real situation completely (Bandi et al., 2019). The SC model can incorporate realistic constraints and handle multi-bodied problems effectively.

Evaluation of Queue Management Strategies

Hospital queue management strategies constitute an active research domain, attracting scientific interest throughout the decades. Initial investigations have focused on waiting list management. (Eugene Helm, 2012) examines models from the healthcare literature, highlighting relevant strategies. A mixed integer programming method has been adopted to support operating room capacity allocation, thereby illustrating an optimization technique for queue management. (Jiang et al., 2019) Consider the Emergency Department scenario, where service is provided continuously, but the number of doctors available per shift fluctuates. The authors develop a mixed-integer linear programming model to determine the optimal number of doctors to employ in each shift. The objective is to minimize the system's total costs, encompassing the fixed expense of assigning doctors to operate during each shift, the idle-time costs incurred when doctors wait for patients, and the overtime costs arising when doctors exceed their scheduled shift duration. Provides further discussion on the topic.

Cost-Benefit Analysis

Cost-benefit analysis is used to develop an estimator for the impact of design and service modifications in a healthcare setting. One approach optimizes the number of providers needed to maximize expected monetary benefits given system parameters (Jiang et al., 2019). Expected costs of employing a given number of providers are analyzed for benchmark parameters to assist decision makers in examining the sensitivity of staffing decisions to low patient-arrival rates or high provider-consumption rates. Reducing overcrowding in an emergency department and increasing patient flow to subsequent treatment areas motivated the reforming of the staffing process at the Flemish Red Cross

in Genk, Belgium. A combination of queuing modeling, simulation, and optimization was used to identify the optimal number of nurses and caregivers needed in a given location for each time block during the day. Analytical modeling procedures determine whether current provider numbers are sufficient to ensure adequate service levels and forecast the bottleneck location within the system. When policy problems become too complex for purely analytical solutions, preliminary information may be obtained from the model that can be used as input to a simulation. The simulation approach determines the optimal workforce to be scheduled in each subarea for every time block. A mathematical-programming formulation identifies the shift combination that covers the demand with the fewest personnel while observing constraints on workforce size and shift requirements.

Mathematical Models for Queue Management

Mathematical programming methods for managing queues have recently been formulated within an optimization-based framework that considers both queue and service dynamics (Jiang et al., 2019). Such models objectively determine the optimal number of servers, the allocation of staffing resources, and the maximum acceptable number of waiting patients, or queue length, to ensure a desired service level.

The primary objective of mathematical programming models is to determine the minimum required number of servers and the optimal allocation of resources that yield the desired steady-state probabilities of queue metrics, as dictated by the constraints. Because allocation decisions involve both the number of servers and the maximum allowed queue length, the objective function can be defined to minimize (or maximize) the total number of servers and patients in the queue. This approach allows decision makers to ascertain a solution that delivers a satisfactory service level at minimum cost. The use of mathematical programming for queue management is illustrated by two models—steady-state and transient—which select the number of servers and queue length necessary to satisfy given limits on a specific queue length probability or expected waiting time.

In healthcare settings, analytical queuing models offer powerful and realistic tools for decision support, facilitating the design of effective strategies for epidemic containment whenever necessary resources are insufficient. They generate direct insights into the relationships among service time, utilization, waiting time, queue length, and resource allocation, permitting systematic investigations into the effects of changing each parameter. Critical performance metrics obtained from the models are used in optimization procedures to identify the best operating configurations. Consequently, these models provide a better understanding of the processes involved in care delivery and enable the identification of efficient service-level strategies without resorting to expensive simulation, which requires significant calibration and computational efforts. Analytical models are quick to develop and straightforward to implement using optimization software packages.

Linear Programming

Various mathematical programming methods have been developed to optimize queue management in healthcare settings, including linear programming, goal programming, and mixed-integer linear programming. Each technique aims to trade off competing variables such as cost, waiting time, and patient throughput. The optimization problem involves determining the correct number of service counters to minimize the total cost and achieve operational efficiency. Cost components comprise the remuneration paid to service personnel and the waiting costs incurred by patients in the queue. Waiting costs reflect potential customer loss and the decline in satisfaction resulting from extended delays.

In modeling the process as a queue system, an M/M/R/N rationale is applied, where 'R' represents the number of service agents and 'N' the maximum system capacity. Patient arrivals follow a Poisson distribution, while service durations are exponentially distributed. Performance metrics derived from this model include the cost profile, the expected number of patients in the system, and the average waiting time. The goal is to identify the configuration of servers and system capacity that balances these factors effectively, ensuring timely access to care while controlling operational expenses. Queue-length reduction thereby enhances patient satisfaction, making strategic deployment of personnel a crucial managerial decision. Optimization techniques thus provide a systematic basis for designing patient-flow strategies and reallocating resources in response to dynamic needs (Jiang et al., 2019).

Integer Programming

Central Decision Models (Jiang et al., 2019). Given these system-building blocks, we consider the optimal strategy to decide how many agents to add to each station. An exact Integer Programming model of the problem provides the optimal solution. The interdependence of the queueing network performance and the discrete staffing decision can be incorporated in a Mixed Integer Nonlinear Programming formulation solved to optimality.

Stochastic Models

Stochastic models allow exploration of system behaviour influenced by random parameters. This produces more reliable insights than deterministic models and facilitates the reuse of existing solution techniques (R. Chakravarthy, 2021). They also accommodate multiple service phases, abandonment, retrials, and vacation time for staff. Demonstrates the matrix-geometric method, which originates from Neuts' work in the late 1970s. An algorithm investigates both steady state and transient behaviour. The matrix-analytic method is also suited to continuous-time Markov chains with neutrophil arrival processes and incorporates phase-type architecture in the service pattern. (van de Klundert et al., 2023) Study the relationship between waiting time and hospital choice. A very general constraint system permits the investigation of many queuing regimes where waiting times increase with choice probabilities. Several random utility-based choice models are accommodated; for instance, calms-in, calms-out, and calms-out-in. Opting out is modelled, but only under the assumption that it does not affect waiting time. Multiple extensions concern hospital utility as a nonlinear function of expected waiting time: evidence from clinics in China suggests the relationship is not linear. Variability in waiting time depends on demand and capacity. Randomised service rules can moderate fluctuations in waiting time. Uncertainty about waiting times would therefore be a valuable addition to the analysis, along with nonlinear utility relations such as prospect theory. Other extensions relate to visits beyond the first. Service tasks and time differ depending on severity. Modelling learning dynamics is also essential; while expected waiting times are difficult to observe, patients learn about waiting times from only a few visits to a hospital, and this is the primary channel through which policies can reduce waiting times.

Healthcare Systems Overview

Underlying healthcare delivery are several interacting components; patients (customers) arrive and wait for access to facilities staffed by healthcare providers (servers) and equipped with diagnostic equipment. Numerous individuals, from clinicians to nurses, administrators to billing clerks, each have a role in effectively providing patient care and successfully operating healthcare systems. Healthcare provision depends on the value of several key organizational resources or variables (such as beds and pharmaceutical supplies). Indeed, the effective deployment and utilisation of these critical resources is of utmost importance. When they are physically constrained, queues for services will inevitably develop (Hu, 2017). Queueing theory and models are therefore powerful tools for describing, analysing, and supporting related operational decisions (Jiang et al., 2019).

Types of Healthcare Systems

Healthcare systems have evolved rapidly in recent years, driven by medical technology and medicine advances. This transformation has created opportunities to enhance hospital services and presents new challenges for managers and decision makers. Hospitals must respond to diverse requirements while utilizing sophisticated equipment and highly qualified personnel, leading to elevated operational costs. Optimizing patient flow and alleviating bottlenecks can reduce expenses and improve the quality of care.

Healthcare systems for patient flow can be broadly classified into two types: (1) pooled or integrated systems, where healthcare systems are combined into one, and (2) isolated or separated systems, in which each healthcare system manages its patient flow independently. In a pooled system, patient arrival distributions determine access rules, which may prioritize certain patient groups such as those with scheduled appointments or random walk-ins without prior appointments (Jiang et al., 2019). Model_P describes a pooled queue with two patient types, where scheduled-appointment patients receive higher priority. Walk-in patients are served only when no scheduled appointment patients are waiting. Model_S considers separate queues and distinct resources for each patient type (Kyung Kwak, 2023). Statistical distributions commonly used to describe patient arrivals include the normal distribution for scheduled patients and the exponential distribution for

walk-in patients. We can model the inter-arrival time for walk-in patients, an exponential distribution with rate λ_w, denoted $X_w \sim Exp(\lambda_w)$.

Challenges in Queue Management

Managing queues is crucial in healthcare management literature because operations research models help create efficient systems. The main objectives are minimizing waiting times, improving satisfaction levels, and reducing healthcare expenditures.

In the hospital emergency department (HED) setting, the mismatch between the daily workflow and the critical requirement for emergency care leads to the inefficient use of resources and excessive stress on patients and staff. Long queues can form when the appropriate number of doctors to deal with the expected number of patients is not determined. The HED must operate 24 hours a day and respond to multiple demands. It requires sophisticated and expensive equipment, laboratories, and highly trained medical staff. The operating cost of the emergency department is higher than that of other departments. Maintaining the appropriate number of doctors or healthcare professionals for the number of patients is a crucial problem (Jiang et al., 2019). The patients may be prescribed outpatient services or hospitalization after the initial triage and examination of the first and stable cases. Patient approaches other services, such as the Pharmacy Department or X-ray Department, for different purposes. Hence, a patient's entire chain of activities must be considered.

A general problem in healthcare delivery is allocating scarce resources to keep waiting times as short as possible. The resource distribution problem arises when the timescale of the decisions does not match the demand variability. Staff and medical equipment, among others, cannot always be varied on short notice, depending on the type of resources (Ben Othman et al., 2018). There is also significant daily and weekly fluctuation in arrival rates when medical and administrative resources remain constant, as the work schedules of healthcare professionals are fixed for several weeks ahead. Moreover, the exact number of patients for each specialty remains unknown, and healthcare management has to ensure that queues are as short as possible during periods of maximal demand.

Data Collection and Analysis

Queueing theory provides a framework for the effective management of healthcare systems. Empirical data are used to develop filtering techniques that identify leading measures for system efficiency and modify mathematical models of the system accordingly. Examples, location planning and schedule coordination, are shown to improve the overall efficiency of healthcare systems (Jiang et al., 2019).

Data on arrival rates, service times, and the evolution of queues for a range of service systems are collected, with clustering and event-based approaches to data analysis that are tailored to service systems. Techniques for parameter estimation and data classification that correspond with all phases of the service process are developed, providing real-time support and data-driven decision methods for systems such as call centres, hospitals, and supply chains (Kim, 2014).

Data Sources

In recent decades, increased mobility and population shifts have increased attendance rates at Healthcare Centres and hospital emergency departments (HEDs). Excessive attendance often exceeds current resource levels during peak periods, diverting incoming patients to other hospitals. Widespread saturation negatively impacts the quality of treatments provided to patients and heightens the workload on medical staff and pressure on existing healthcare resources. Planning strategies for ambulatory care require engineers and healthcare planners to accurately analyse current and future operation policies and resource allocation plans. A large amount of healthcare data containing information on patient profiles and health records, along with diagnostic procedures and prescriptions, is generated daily by national and international healthcare authorities. Demonstrating the utility of a Decision Support System (DSS), a queue-based approach has been presented for optimizing staffing under different patient arrival patterns and service rates within HEDs. Investigating the performance of a healthcare provider at different staffing levels, this DSS supports managing the flow of patients, allocating healthcare resources, and selecting effective and efficient operation plans. Managers can analyse previous arrangements of staffing and scheduling and access this information to identify sources of problematic delays, allocate existing resources effectively, and plan future arrangements accordingly. (Jiang et al., 2019)

Agent-based modeling (ABM) considers a system as comprising a collection of autonomous agents that pursue their interests while interacting with each other and their environment, designating home locations and moving stochastically from location to location. Using ABM, systems are more accurately modeled with realistic, complex, time-varying behaviours that are non-linear and stochastic. Open-source technologies such as NetLogo and Repast make ABM accessible for detailed simulation because they support rapid development and execution for tens or hundreds of thousands of agents. This approach has been used for assessing system-level performance, patient safety, medical staff workload, changes in policy or procedures, and investigating patient flows through a hospital during everyday and critical incident situations. Little work, however, has applied ABM to healthcare policy development. Queuing-based modeling is a complementary technique. It is well established within the operations research community and has a history of application to health care, where it has been widely used. In particular, queuing-based methods have been applied to study patient flows, delays and bottlenecks, admissions strategies and scheduling policies, and various resource allocation problems. Specialist technologies such as intra-hospital tracking and internetworking can improve patient care, facilitate distributed management of emergency departments, and provide flexible communication between ambulances and providers. Applications exist in ambulance redirection policies and patient diversion schemes. (Laskowski et al., 2009)

Statistical Methods

Statistical models linking data elements to random variables enable the prediction of key performance measures for queueing systems. For example, econometric approaches can examine the relationship between primary interest measures—such as length of stay (LOS), transfer time, and crude mortality rate—and the explanatory variables derived from data. Count data regression techniques, like negative binomial models, can estimate LOS and the number of ICU admissions, while Poisson models may assess counts. Queueing methods provide an alternative approach to fitting models and generating predictive insights (Kim, 2014).

These techniques have critical applications in healthcare, where improved flow for ICU patients—both in terms of care and discharge—can increase access to critical care and reduce costly emergency room stays. External factors such as policies, protocols, and internal unit factors influence LOS and can be analyzed using these models. Understanding the impact of transfer delays on vulnerable ICU patients may shed light on the use of clinical treatment or health information technology. The ability of transfers to provide this information largely depends on LOS data because potential changes to care must occur before discharge. Counting models, like the negative binomial regression, quantify and identify how operational factors affect LOS and ICU admissions, assisting policy and system design decision-making.

Hospital admissions from emergency departments or other units generate count data that can be modeled using negative binomial or Poisson frameworks to capture admission patterns. Patient readmissions often serve as key quality indicators, and the length of the readmission window critically influences conclusions about frequency. Risk-adjusted hospital mortality is another important measure that can benefit from refined methodologies for mortality rate computation by various indicator categories. Queueing theory, regarded as the "physics of queues," is essential for explaining service system behavior and provides approaches for staffing in time-varying systems. The unit's ICU bed flow problem can be improved through analyses based on queueing systems and optimization models.

3. Results & Discussion

- **Patient Waiting Time Reduction**: Applying mathematical programming, especially queueing models, significantly reduces patient waiting times. For example, studies found that adopting queueing theory and integer programming models to outpatient services noticeably improved patient satisfaction by decreasing wait times and enhancing the quality of care.
- **Resource Allocation Optimization**: Mathematical models, including Markov decision models and mixed-integer programming, enable healthcare managers to determine the optimal number of service staff (doctors, nurses) and other resources. For instance, robust models allowed for balancing operational costs with waiting time costs, minimizing overutilization of staff while preventing excessive patient delays.

- **Handling Uncertainty and Fairness**: Robust programming and dynamic queue models accommodate uncertainty in patient arrivals, service rates, and priority levels, thereby enhancing system flexibility. Some research emphasizes fairness, ensuring early arrivals are not bypassed, while others prioritize urgent medical needs, often using preemptive priority strategies for emergency services.
- **Improvement in Staff Skills and Attitudes**: Training staff in queue management systems based on mathematical programming led to improved knowledge and a more positive attitude among nurses and other staff, further enhancing the overall quality of care delivery.
- **Selection of Queue Models**: The effectiveness of queue management depends on the choice of the mathematical model. Simple M/M/1 models provide a basic framework. Still, more complex models (e.g., M/Erlang/1, multi-server queues, or network models) better represent real-world variations and can achieve lower average response times in practice.
- **Integration with Machine Learning and Simulation**: Some recent research integrates deep learning and simulation with mathematical programming to predict waiting times and compare optimization strategies. For example, deep learning models reduced the mean absolute error in waiting time prediction, complementing traditional mathematical approaches.
- **Practical Applications**: Case studies highlight tangible benefits, such as reducing hospital bed allocation delays by up to 50%, optimizing emergency department (ED) staff levels, and improving operational efficiency across various hospital departments. These improvements can translate into both patient satisfaction and overall cost savings.
- **Robustness and Scalability**: Mathematical programming models provide robust solutions that are adaptable to changing scenarios (e.g., sudden patient surges, staff shortages) and scalable from small clinics to large hospitals.
- **Limitations and Future Research**: Despite these benefits, healthcare settings often encounter unpredictable variables (patient behavior, emergencies), which require ongoing refinement of models. Future research explores combining artificial intelligence with mathematical programming for adaptive real-time queue management.
- In summary, applying mathematical programming for robust queue management in healthcare systems offers measurable improvements in service delivery, resource utilization, and patient satisfaction, especially when models are carefully selected and combined with modern data analytics and simulation tools.

4. Conclusions

Mathematical programming enables healthcare providers to design solutions that assign patients to multiple priority classes across clusters of geographically distributed service facilities. Such models require queues with batch arrivals and service processes beyond the standard assumptions of Poisson arrivals and exponential service times. Prompt scheduling of patients and assignment to the best facilities is crucial in areas such as elective surgery, diagnostic testing, and home healthcare. The interaction between fixed costs, quality of care, and customer convenience is captured using models with multiple objectives. Distinguishing between demand generated by referrals and demand originating from walk-ins is essential. Realistic modeling of patient referrals requires knowledge of joint probabilities involving the referring physician, receiving physician, timing, and type of referral. Parameter estimation methodologies guide healthcare organizations in minimizing the demotivation of family physicians through suboptimal referral decisions (Kumar Yadav et al., 2021).

Effective queue management plays a pivotal role in enhancing customer relationships. Hospitals must carefully balance the number of patients served and the resources employed. Well-designed mathematical programming models facilitate prompt and patient-friendly service delivery. Accurate estimations of future demand enable organizations to allocate the workforce efficiently and enhance flexibility. Queuing theory assists hospitals in analyzing demand, optimizing staffing deployment, and efficiently utilizing facilities to minimize patient waiting times. Models that track the pathways of referred patients help maximize staff motivation and improve service efficiency (Jiang et al., 2019).

References

Bandi, C., Trichakis, N., & Vayanos, P. (2019). Robust multiclass queuing theory for wait time estimation in resource allocation systems.

Ben Othman, S., Ajmi, F., Zgaya, H., & Hammadi, S. (2018). A cubic chromosome representation for patient scheduling in the emergency department.

Bush, N. (2019). Impact of queueing theory on capacity management in the emergency department.

Chakravarthy, S. R. (2021). Queueing models with MAP arrivals are useful in service sectors.

Chouba, I., Amodeo, L., Yalaoui, F., Arbaoui, T., & Laplanche, D. (2020). A mixed-integer linear program for human and material resources optimization in the emergency department.

Ciuiu, D. (2008). Solving nonlinear systems of equations and nonlinear systems of differential equations by the Monte Carlo method using queueing networks and game theory.

Dimitrov, S. D. (2012). Program for modelling queuing systems in transport.

Helm, J. E. (2012). Stochastic and deterministic methods for patient flow optimization in care service networks.

Hu, X. (2017). Application of mathematical and computational models to mitigate the overutilization of healthcare systems.

Hua, L., Dongmei, M., Xinyu, Y., Xinyue, Z., Shutong, W., Dongxuan, W., Hao, P., & Ying, W. (2023). Research on outpatient capacity planning combining lean thinking and integer linear programming. National Center for Biotechnology Information (NCBI).

Hurwitz, J. E., Lee, J. A., Lopiano, K. K., McKinley, S. A., Keesling, J., & Tyndall, J. A. (2014). A flexible simulation platform to quantify and manage emergency department crowding. National Center for Biotechnology Information (NCBI).

Jiang, F. C., Shih, C. M., Wang, Y. M., Yang, C. T., Chiang, Y. J., & Lee, C. H. (2019). Decision support for the optimization of provider staffing for hospital emergency departments with a queue-based approach. National Center for Biotechnology Information (NCBI).

K., S., P. S., A., & Rangaswamy, M. (2023). Queueing-inventory systems: A survey.

Kiani, M., Eksioglu, B., Isik, T., Thomas, A., & Gilpin, J. (2019). Evaluating appointment postponement in scheduling patients at a diagnostic clinic.

Kim, S. H. (2014). Data-driven decisions in service systems.

Kwak, J. K. (2023). Analysis of the waiting time in clinic registration of patients with appointments and random walk-ins. National Center for Biotechnology Information (NCBI).

Laan, C., van de Vrugt, M., Olsman, J., & Boucherie, R. J. (2017). Static and dynamic appointment scheduling to improve patient access time. National Center for Biotechnology Information (NCBI).

Laskowski, M., McLeod, R. D., Friesen, M. R., Podaima, B. W., & Alfa, A. S. (2009). Models of emergency departments for reducing patient waiting times. National Center for Biotechnology Information (NCBI).

Rihm, T. (2017). Applications of mathematical programming in personnel scheduling.

Schulz, A. S., & Udwani, R. (2019). Robust appointment scheduling with heterogeneous costs.

Thompson, S. M., Nunez, M., Garfinkel, R., & Dean, M. D. (2009). Efficient short-term allocation and reallocation of patients to hospital floors during demand surges.

van de Klundert, J., Cominetti, R., Liu, Y., & Kong, Q. (2023). The interdependence between hospital choice and waiting time: A case study in urban China.

Yadav, S. K., Singh, G., Sarin, N., Singh, S., & Gupta, R. (2021). Optimization of manpower deployment for COVID-19 screening in a tertiary care hospital: A study of the utility of queuing analysis. National Center for Biotechnology Information (NCBI).