Admission, Discharge, and Transfer Control in Patient Flow Logistics: Overview and Future Research

Abstract

Patient flow logistics involves managing and coordinating the movement of patients within a healthcare system. It aims to optimize the patients' flow from their arrival to discharge or transfer, ensuring efficient and effective use of resources while minimizing delays and bottlenecks. Key components of patient flow logistics include capacity planning, resource allocation, appointment scheduling, and notably, admission, discharge, and transfer control. Our focus is on admission, discharge, and transfer control policies, which manage and regulate the flow of individual patients into and out of various facilities delivering care. We review a selection of 37 analytical, empirical, and experimental papers in this area published in leading operations research, operations management, and medical journals. These papers are categorized based on the specific department they target and the type of policies they consider. In particular, we identify: (i) four types of policies for intensive care units, (ii) three types of policies for emergency departments, and (iii) four types of policies for general wards. For each paper, we provide an overview of the research questions, formulations, solution methodologies, and results. This comprehensive review culminates in identifying future research directions for academics in this field.

Keywords: Admission, Discharge, and Transfer Control; Health Operations; Patient Flow Logistics, Literature Review.

1. Introduction

Admission control refers to process of deciding whether to accept or reject arriving customers into a service system or facility (Stidham, 1978). It is typically employed to reduce or prevent congestion, thereby improving system performance through measures such as limiting arrivals and closing gates (Stidham, 1985). Admission control is applicable to any system where demand for a resource exceeds its capacity.

Stidham (1985) provides a comprehensive review of studies addressing admission control in queueing systems. The review considers both static policies, where the probability of each arrival being accepted is predetermined regardless of the system state, and dynamic models, where admission decisions are made based on the current state of the system at the time of arrival. Static policies are typically formulated as static optimization models incorporating queueing relations and are solved either numerically or analytically using the Lagrangian method. In contrast, dynamic models are generally formulated as a Markov Decision Process (MDP), requiring more complex solution approaches, often involving heuristic methods.

Stidham (1985) reports that for a single-server system under relatively mild conditions, the (static) acceptance probability that maximizes the long-run average reward of admitting customers net of the cost of customers waiting in the queue can be obtained by solving a nonlinear equation. Moreover, for single-server queues with exponential service times, the dynamic policy that maximizes the total expected discounted net benefit over an infinite horizon is of a threshold type. This implies that customers will not be admitted when the total number of customers in the system exceeds a certain limit. Stidham (1985) also highlights the limited progress made at the time for more complex scenarios involving two-facility queues in series and parallel. In a subsequent update, Stidham (2002) provides further advancements on some of the problems discussed in his earlier work.

This review concentrates on studies that address admission, discharge, and transfer (ADT) control policies in the context of healthcare systems. These policies manage and regulate the flow of patients into and out of various facilities delivering care. Despite substantial research in this area, particularly in recent years, no systematic reviews have been conducted. Our study intends to fill this gap.

The papers we review aim to determine optimal or near-optimal policies for admitting, discharging, or transferring patients into/from different hospital departments. These papers' main questions usually revolve around several key issues: *What should patients' arrival and/or discharge rates be? Should a patient be admitted, rejected, discharged, or transferred? In which ward or facility should a patient be admitted? Which patient should be prioritized for room or clinician allocation? What are the human behavioral biases when making admission control decisions?*

We include analytical, empirical, and experimental articles specifically designed for healthcare applications and those that are generic but have direct applications in healthcare systems. We classify the papers based on their departmental focus and types of policies covered and provide a comparative analysis of their objectives, methodologies, assumptions, and distinctive features, thereby guiding future research directions.

The structure of this paper is as follows: Section 2 explains the scope and methodology of the review. Section 3 defines the proposed framework for classification of the papers. Sections 4 to 7 review and analyze the articles based on the proposed framework. Section 8 presents the general recommendations and directions for future research. The overall conclusion of the research is presented in Section 9.

2. Scope and Research Methodology

This section introduces the search scope and methodology in detail.

2.1. Scope

The current research systematically reviews the literature on ADT control in healthcare systems. This study does not include articles on capacity planning, resource allocation, and scheduling, as these problems typically focus on different aspects of healthcare logistics. Capacity planning literature centers on determining the optimal size and scale of resources required to meet anticipated demand. Resource allocation literature focuses on the efficient distribution and utilization of resources within a system. Scheduling articles aim to create schedules that balance conflicting objectives, such as resource efficiency and patient waiting times. In contrast, ADT control articles focus on developing appropriate policies for admitting, discharging, or transferring patients. ADT decisions are operational in nature and typically precede scheduling decisions. In comparison, resource allocation and capacity planning are more strategic decisions. By focusing specifically on ADT control, this review highlights the unique challenges and solutions associated with managing patient flow and optimizing healthcare delivery in real time.

2.2. Research Methodology

This research uses SCOPUS as the most prominent abstract database of peer-reviewed literature. While we did not limit the starting year of our search, we observed that the oldest relevant paper within the scope of this survey was published in 1987. Therefore, the selected papers include articles published in high-quality journals from January 1987 to December 2023. To enhance the quality and reliability of the reviewed papers, we exclude conference articles, book chapters, technical reports, and master’s and PhD theses. The selected journals are classified into three groups, described below.

***OR/MS journals*:** Given that the primary readership of this paper comprises researchers in Operations Research (OR) and Management Science (MS), we have selectively included journals from these fields that are highly ranked on a global scale. This includes top journals from the Academic Journal Guide (AJG), the Australian Business Deans Council (ABDC), the now-defunct Centre National de la Recherche Scientifique (CNRS; last published in 2021), and the Fondation Nationale de l'Enseignement de la Gestion des Entreprises (FNEGE) rankings. To ensure the highest quality, our selection criteria restrict inclusion to journals ranked in the top two tiers of these systems. For instance, from a scale of 1 to 4, such as that used by AJG, we consider only those journals ranked at levels 3 and 4, inclusive of those distinguished as 4*. From the ABDC list, we include journals rated 'A' and 'A*' as well as 'B'. Additionally, our selection includes journals recognized in the Financial Times (FT) rankings, which are noted for their rigorous standards.

The selected journals are as follows: *Annals of Operations Research*, *Computers and Operations Research*, *Decision Sciences*, *Decision Support Systems*, *European Journal of Operational Research*, *IISE*

Transactions, International Journal of Operations & Production Management, International Journal of Production Economics, International Journal of Production Research, Journal of Operations Management, Journal of Supply Chain Management, Journal of the Operational Research Society, Management Science, Manufacturing & Service Operations Management, Mathematical Programming, Mathematics of Operations Research, Naval Research Logistics, OMEGA: The International Journal of Management Science, Operations Research, OR Spectrum, Production and Operations Management, Production Planning and Control, Supply Chain Management: An International Journal, Transportation Research Part B: Methodological, Transportation Research Part E: Logistics and Transportation Review, and Transportation Science. However, no papers were found in some of the journals in this category.

In addition to the journals above, we consider the following journals because of their defined scope in the Scimago Journal and Country Rank (SJR) and citations to some of their papers in the journals we mentioned earlier: *IEEE Transactions on Automation Science and Engineering, IEEE Transactions on Communications, International Transactions in Operational Research.*

***Health-focused OR/MS journals*:** We include *Health Care Management Science*, *Operation Research for Healthcare*, and *Journal of Medical Systems*, ranked Q1 or Q2 in SJR ranking.

***Health journals*:** In addition to the two categories above, we reviewed several health journals that are not specifically focused on OR/MS but are recognized as top journals in the medical field. However, we found no papers on ADT control in these journals. The journals reviewed include *the New England Journal of Medicine*, the *Journal of the American Medical Association*, *Health Affairs*, *Health Services Research*, and *Annals of Internal Medicine*. These journals are recognized as top-tier in the health field, as evidenced by their rankings in studies conducted by Borkowski et al. (2018), Merigó and Núñez (2016), and Williams et al. (2002). Additionally, they are all classified as Q1 in the SJR system.

In the first step of the research, we used the following combination of keywords to search the title, abstract, and keywords of the selected journals: ("admission control" or "hospital admission" or "discharge policy" or "discharge control" or "transfer policy" or "transfer control") And ("healthcare" or "health" or "clinic" or "pandemic"). Then, after screening the abstract and text of the articles, the articles that were not related to the fields of patient ADT were excluded. We further excluded all articles on outpatient clinics or operating theatres. This is because these articles typically address admission control and appointment scheduling concurrently. Hence, reviews of these articles are already provided in the appointment scheduling literature; for example, refer to Marynissen and Demeulemeester (2019), and Youn et al. (2022). In the end, 37 papers are obtained from this screening.

2.3. Shortlisted Articles

Figure 1 displays the frequency of articles by their publishing outlet. Within the OR/MS journals, *Operations Research* and *Manufacturing & Service Operations Management* have the highest number of articles. Similarly, among health-focused OR/MS journals, *Health Care Management Science* has published the most.

Figure 2 illustrates the frequency of articles by publication year, showing an increasing trend in the number of published papers. Notably, a substantial number of articles have been published between 2020 and 2023. While the COVID-19 pandemic may have influenced this trend, the considerable time lag between the acceptance date and the publication year, especially for INFORMS journals complicates efforts to draw definitive conclusions based on publication dates alone.

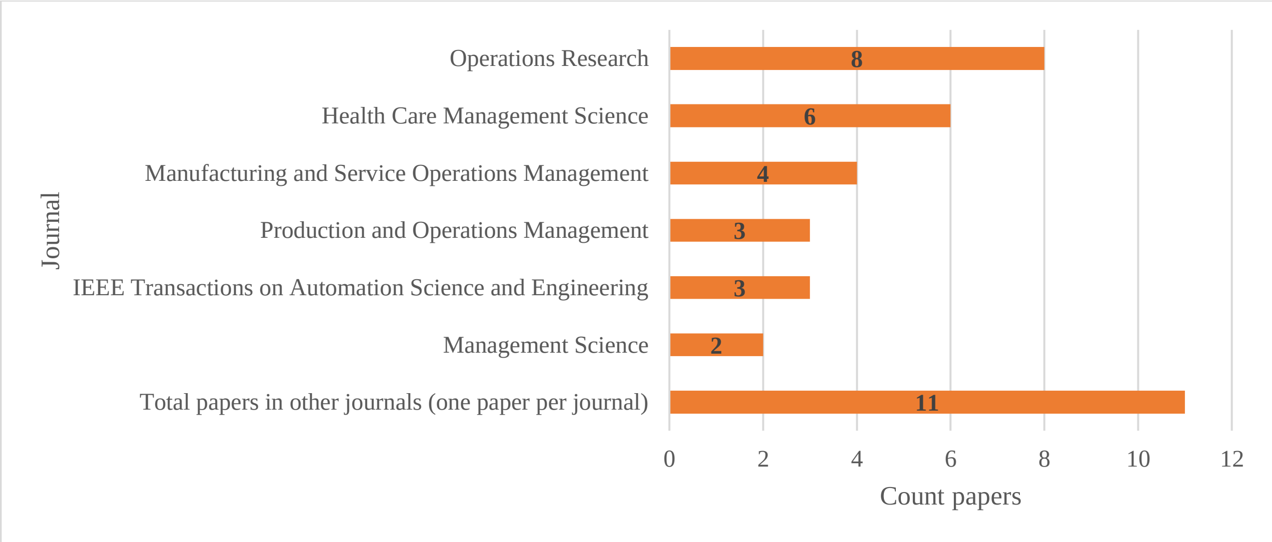

Figure 1. The frequency of articles by their publishing outlet.

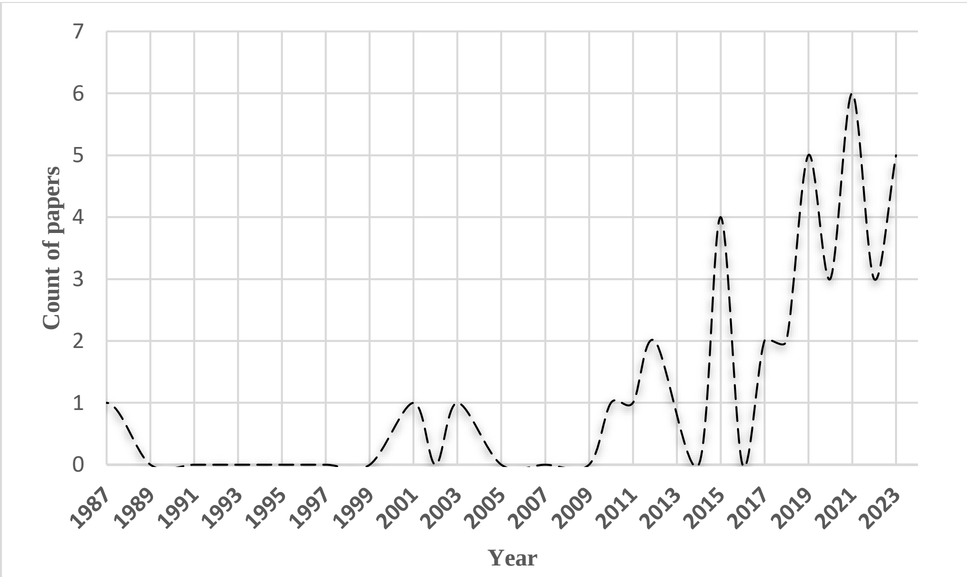

Figure 2. The frequency of articles by publication year.

3. Review Classification

Our literature classification is based on the departmental focus of reviewed papers. To have a better understanding of this classification, we first explain the main functions of seven major components of health systems: outpatient clinics (OCs), emergency departments (EDs), operating theatres (OTs), intensive care units (ICUs), general wards (GWs), step-down units (SDUs), and post-acute care units (PCUs).

The OCs are usually the initial point of contact for patients seeking medical evaluations, treatment, or routine check-ups without the need for hospital admission. Patients requiring further treatment that cannot be handled on an outpatient basis are referred either directly to the ED or scheduled for admission to GWs or

OTs, depending on the urgency and nature of their condition.

The ED is often the entry point for patients who require immediate medical attention due to acute illness or injury. The primary goal here is to stabilize patients. After initial assessment and stabilization, patients may be discharged home, admitted to a GW, transferred to an OT for urgent surgery, or moved to the ICU for intensive monitoring and treatment.

The OT is used for surgical procedures. Patients can be scheduled for surgery from OCs, admitted directly from the ED if urgent, or transferred from GWs when surgery is required as part of their treatment. Post-surgery, patients might return to GWs, be transferred to the ICU if they require closer observation and care due to their critical condition, or occasionally be sent directly to the SDU if their post-operative condition is stable but still requires more care than is typically provided in a GW. Some patients may also be discharged home immediately after surgery.

The ICU is designed for patients who require intensive treatment and monitoring, which cannot be provided in a GW. This includes patients with life-threatening conditions, severe illnesses, or those recovering from major surgeries. Once a patient's condition stabilizes and intensive monitoring or treatment is no longer necessary, they may be transferred to GWs or SDUs for further recovery before discharge. Some patients may also be discharged to PCUs if they require long-term care and rehabilitation following their spell in the ICU.

The GWs house patients who are admitted for a variety of medical reasons but do not require the intensive care provided in the ICU. These wards accommodate both patients admitted through the ED and those who have undergone surgery in the OT. Patients who recover can be discharged directly from GWs. Those requiring long-term care and rehabilitation are transferers to a PCU.

The SDUs serve as a transitional environment for patients who do not need intensive care but require more monitoring and medical attention than is typically available in a GW. From the SDU, patients are either discharged to their homes/PCUs, or sent back to the GW if further recovery is needed.

The PCUs include facilities such as nursing homes and long-term care centers. They cater to patients who require extended care after being discharged from acute care settings like ICUs, SDUs, or GWs, but are not ready to return home. In addition to providing some medical care, PCUs assist with patients' daily activities until they regain their independence.

Figure 3 provides a schematic illustration of the patient flow among the abovementioned components. It is crucial to recognize that patients may be transferred among different facilities/hospitals to receive the required care. Such transfers may occur not only due to a lack of available capacity but also due to the need for specific expertise or equipment essential for the patient's treatment. For instance, patients requiring complex cardiac surgery could be transferred from a general hospital where they were initially admitted to a specialized cardiac facility. Additionally, a patient might be transferred to another facility even when capacity or expertise is available at the initial hospital. This strategy can be employed to preserve resources at the current facility for handling more severe or urgent cases that may arise in the future or to reduce the overall cost of care.

We classify the shortlisted literature into three primary categories, each defined by its departmental focus. These categories include papers centered on ADT control policies within ICUs, EDs, and GWs. Furthermore, we introduce a fourth category for papers that do not clearly fit into the first three categories. As previously noted, we exclude papers focusing on OCs and OTs as reviews of ADT policies within these units are already provided in the scheduling literature; see, for example, Marynissen and Demeulemeester (2019) and Youn et al. (2022). Further, our review did not identify any studies addressing admission control in SDUs. Nevertheless, given the comparable nature of care in SDUs and ICUs, formulations and insights from the ICU-focused papers are potentially applicable to SDUs as well. Finally, we found only one article discussing ADT policies in PCUs; hence, its review is provided in the fourth category.

Within each category, we first identify the various types of ADT policies explored. This is followed by a comprehensive analysis of all papers addressing each policy within that category, including problem formulation, solution methodology, special features, and results. Potential areas for future research are presented at the end of each category's review.

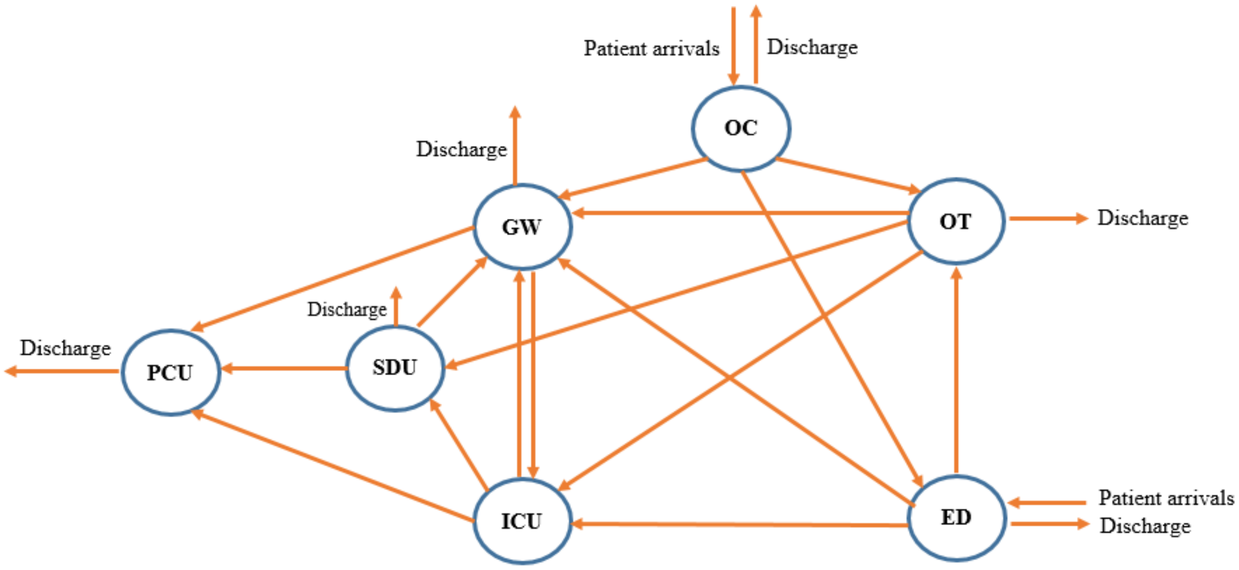

Figure 3. Patient flow within a hospital.

4. ICUs

ADT control decisions are critical in ICUs, primarily because they provide the most complex level of care utilizing highly expensive resources. This makes effective control vital for both operational efficiency and patient outcomes. Additionally, from a technical perspective, ICUs are often modeled as loss systems—where patients who cannot be admitted immediately are turned away (Bountourelis et al., 2013). This modeling approach simplifies analytical work by eliminating the need to account for queue dynamics. In this context, we identify five primary types of ADT policies: admission policies, early discharge policies, joint admission and early discharge policies, proactive transfer policies, and facility transfer policies.

Early or premature discharge policies outline the criteria for discharging existing ICU patients prematurely to reduce the cost of providing care or accommodate new arrivals who may derive greater benefit from ICU care. Patients discharged early are often relocated to an SDU or a GW.

Admission policies are designed to determine which patients, among those potentially benefiting from ICU care, should be admitted. These decisions typically depend on patient severity and operational considerations such as current ICU occupancy and anticipated future admission requests.

Joint admission and early discharge policies consider the two types of policies discussed above concurrently.

Proactive transfer policies focus on identifying which patients in GWs should be transferred to the ICU pre-emptively. The aim is to circumvent unplanned internal admissions when a patient's condition suddenly deteriorates within a GW, necessitating immediate ICU intervention. Studies such as Barnett et al. (2002) have found that these unplanned admissions often result in poorer patient outcomes than direct ED admissions.

Facility transfer policies determine whether patients needing intensive care within a multi-hospital network should be treated at the facility where they initially arrive or transferred to a different facility. These decisions aim to optimize care delivery by considering patient severity, transfer costs, and current and future occupancy levels of various facilities.

It is essential to recognize that admission policies might *reject* new arrivals even when beds are available. This strategy is intended to preserve ICU capacity for more critical future cases. Similarly, early discharge policies may be triggered regardless of bed availability or the arrival of new patients to reduce the overall cost of care delivery. In this context, Ouyang et al. (2020) define non-idling admission and non-idling early discharge policies, respectively, as those that consistently allocate available beds to new arrivals and avoid premature discharges from the ICU when beds are available. Proactive and facility transfers, on the other hand, only happen when ICU beds are available.

We identify and review the following papers: five papers addressing early discharge policies, three papers focusing on admission policies, four papers considering joint admission and early discharge policies, three papers discussing proactive admissions, and one paper examining facility transfer policies. The details of these papers are discussed below.

4.1. Early Discharge Policies

Dobson et al. (2010) propose a model for evaluating an early discharge policy that discharges patients with the fewest remaining days in their clinical Length of Stay (LOS) when the pool of existing and newly arriving patients exceeds the number of available beds. The model categorizes patients into scheduled and unscheduled groups, with scheduled patients arriving on specific days and unscheduled patients arriving on any day. The numbers of arrivals and LOSs for both groups are random variables. The system is modeled as a discrete-time Markov chain, with the state of the system in each day defined by the remaining LOS of patients in each bed. The authors introduce a method for reducing the state space size, allowing for the efficient computation of stationary probabilities. The findings indicate that increasing the number of beds reduces the percentage of early discharges, albeit with diminishing returns. Additionally, increasing the proportion of unscheduled admissions results in a higher discharge rate.

Chan et al. (2012) examine early discharge policies with the consideration that patients discharged prematurely may deteriorate over time and require readmission. They assume that newly arrived patients must always be admitted, necessitating deciding which patient to discharge when the ICU is full. Patients are categorized into multiple classes, each with a specific LOS and associated early discharge cost. The authors propose an MDP formulation to minimize the expected total discharge costs over a finite planning horizon.

They establish performance guarantees for an index policy, which prioritizes discharging a patient from the group with the lowest discharge cost. Using real data, the authors demonstrate that employing such index policies, in conjunction with a predictive model of readmission risk, can significantly improve ICU throughput without increasing mortality.

Mallor et al. (2015) address early discharge policies by formulating a static optimization model that identifies the ICU discharge rate as a function of the number of current patients such that the probability of rejecting new admissions falls below a given threshold at steady state. They derive optimal state-dependent discharge rates using the theoretical results for loss queues, assuming that LOSs are exponentially distributed. A range of discharge policies are then crafted to reproduce the optimal discharge rates obtained from the optimization model. These policies are compared through a simulation, allowing slight deviations from the Exponential assumption for LOSs. The authors ultimately identify a policy that discharges the longest-staying patient upon the arrival of a new one with a state-dependent probability, not only aligning with their optimization model's recommended discharge rates but also being clinically viable.

Mahmoudian-Dehkordi and Sadat (2017) compare the impact of premature discharge of ICU patients with two other ICU management policies during a natural disaster. They develop a system dynamic simulation that captures the flow of emergency patients among ED, ICU, and GWs. This model, populated by data from the literature, compares all combinations of three specific policies: (i) the premature discharge of ICU patients, (ii) the temporary placement of ED patients awaiting ICU admission in GWs, and (iii) prioritizing ward admission for ICU patients ready to be discharged over ED patients seeking a general bed. Their finding indicates that implementing the third policy alone results in the lowest total hospital mortality. The authors argue that this somewhat counterintuitive outcome arises because this policy avoids compromising patient health by ensuring that patients are not downgraded to a lower level of care than necessary, unlike the other two policies.

Azcarate et al. (2020) develop a detailed simulation model capturing emergency and elective (scheduled) arrivals to get insight into the performance of early discharge policies. The progress of patients' severity levels is modeled as a phase-type distribution, with early discharge permitted in some phases. The discharge process is captured through a probability function, representing the discharge probability of an existing patient as a function of the ICU occupancy, the severity level of the patient, and the schedule of future elective arrivals. The authors utilize a simulation-optimization approach to estimate the discharge probability function that minimizes the expected loss in LOS due to early discharges subject to admission rejection probability falling below a threshold. They provide an example of the application of their model in a real-world scenario.

4.2. Admission Policies

Shmueli et al. (2003) conduct one of the earliest studies on ICU admission control policies with the primary objective of identifying an admission policy that maximizes the expected incremental number of lives saved by operating an ICU. The study evaluates three distinct policies: (1) the First-Come, First-Served (FCFS) policy; (2) a threshold-based FCFS policy, which admits patients only if their expected incremental survival benefit from ICU admission exceeds a fixed threshold; and (3) a bed-dependent threshold FCFS policy, where the admission threshold varies depending on bed availability. Note that the first policy is non-idling, whereas the second and third policies are idling. Assuming the increment in survival probability is a random variable following a certain distribution, the authors integrate loss queueing results into a numerical optimization framework to derive the overall expected value of operating the ICU under each policy. Using data from a hospital in Jerusalem, the study demonstrates that the threshold-based FCFS policy improves survival benefits by 18% compared to the standard FCFS policy, with a further 1.2% improvement achieved through the bed-dependent threshold policy.

Kim et al. (2015) integrate empirical research and analytical modeling to examine the effects of different ICU admission policies on patient outcomes. Their empirical study evaluates the influence of ICU admissions on three key patient outcomes: hospital LOS, hospital readmission, and transfer up to a higher level of care (for those admitted to GWs). Their analytical study includes a simulation model for predicting the impacts of various admission policies as well as an MDP model aimed at identifying the optimal admission policy. The MDP model specifically determines a state-dependent threshold policy minimizing the total costs associated with ICU admission denials over a finite horizon, where costs are defined as increases in hospital LOS, hospital readmission, or transfer up, as estimated from their empirical analysis. The state-dependent threshold policy advises admitting a patient when a bed is available, and the cost of denying ICU care exceeds a certain threshold, which varies depending on the ICU occupancy and rejecting them otherwise. A unique feature of the MDP model is that it incorporates observable patient characteristics, such as demographics, unobservable severity, and operational factors that physicians may consider when making admission decisions but are not explicitly available in the data. This integration into the MDP model, enabled by the empirical analysis, allows for a more nuanced approach to admission control. Simulation results demonstrate that implementing the optimal policy, which utilizes both observable and unobservable information, significantly enhances performance compared to policies that rely solely on observable data. Specifically, these improvements are orders of magnitude larger than those that can be achieved by adding a new ICU bed.

Yang et al. (2015) examine a scenario faced by ICU managers concerning the allocation of limited ICU beds for planned postoperative care, requiring them to choose which elective surgeries to admit and which to defer until the next day. They model this using an MDP approach, aiming to minimize the p-norm of the total postponement costs over the planning horizon. These costs vary based on the type of surgery and the surgeon performing it. Using the p-norm in the objective function is intended to balance efficiency with equity, where equity represents the fair distribution of postponed surgeries among surgeons. The model accommodates the randomness of request batch sizes and adds postponed requests to the next day's demand; it allows rejection only when beds are insufficient. Due to the complexity of the original model, a myopic policy is proposed as a pragmatic alternative, and its performance boundaries are rigorously assessed. When applied to actual data from a surgical ICU, this myopic policy demonstrates superior performance compared to two other alternative policies.

4.3. Joint Admission and Early Discharge Policies

Focusing on a short planning horizon, e.g., a single day, Li et al. (2015) consider admission and early discharge policies by proposing an MDP model that maximizes the benefit of admission to the ICU net the cost of early discharges. They divide patients into two classes and assume the benefit of admission, as measured by the difference between survival probability in and out of the ICU, is higher for class 1 patients than class 2. Their formulation allows class 2 patients to be declined admission or early discharged. In contrast, class 1 patients can never be discharged early and must always be admitted unless class 1 patients occupy all beds. The cost associated with early discharge of a class 2 patient is assumed to be the corresponding survival benefit plus the negative impact of patient transfer. Assuming that patients' classes do not change within the planning horizon, the authors solve their model using value iteration and illustrate its benefits compared with FCFS and a variation of FCFS, which always admits class 1 patients by discharging a class 2 patient when all beds are occupied. Li et al. (2019) extend their earlier work by considering a multiday planning horizon in which class 1 patients can also be denied admission when beds are available. Assuming that only one arrival or discharge occurs within each time interval, they derive several structural properties for the optimal policy. They also design a heuristic threshold-based policy utilizing these properties and show that it yields satisfactory performance in terms of total survival benefits, acceptance rates, and the percentage of time the ICU is full.

Ouyang et al. (2020) develop a stylized model featuring a GW with unlimited capacity alongside an ICU with limited capacity to explore admission and early discharge policies. They model the progression of patient severity levels as a discrete-time Markov chain with death and discharge as the absorbing states. The model distinguishes between ICU and ward patients by using different transition probabilities to capture the enhanced survival benefits of ICU care. Assigning a unit cost to every death, the authors formulate an MDP model to minimize the expected long-term average cost. They derive an optimal policy for a single-bed scenario and extend their analysis to demonstrate that, for a multi-bed setup, an optimal non-idling admission and early discharge policy can be established under certain conditions. Restricting their attention to non-idling policies, the authors propose a heuristic that prioritizes the early discharge of the patient whose reduction in survival probability relative to the mean LOS is minimal when a new arrival occurs. Conversely, upon a bed becoming available, the heuristic readmits the patient who shows the greatest improvement in survival probability per unit of mean LOS. Using a simulation model that includes a more complex severity progression dynamic, they demonstrate that their heuristic significantly outperforms other common strategies, such as the FCFS policy and a greedy policy that solely considers survival probabilities.

Bai et al. (2021) consider a setting and decision framework akin to Ouyang et al. (2020) review above. However, their MDP model has a distinct objective: to minimize the long-term weighted average of monetary and medical costs. In this context, monetary costs are associated with financial losses due to patients' rejection or early discharge, whereas medical costs represent the resulting increase in mortality rates. The authors provide an approximate solution to their MDP formulation by restricting the number of patients transitioning to other states at each time interval. Their findings indicate that under certain conditions, the policy derived from this approximation is an idling admission policy, where new arrivals may not be admitted even when beds are available. The application of their model on data from a German hospital shows that their approximate policy outperforms a myopic policy as evaluated by a simulation model.

4.4. Proactive Admission Policies

Hu et al. (2018) combine an empirical study with analytical modeling to explore the impact of proactively admitting GW patients to the ICU at the individual-patient level and system-wide level. Their empirical analysis indicates that ICU transfer of patients in the ward can reduce both their hospital LOS and mortality risk, with the extent of these benefits varying according to patient severity, as measured by a specific scoring mechanism. The analytical component of their investigation deploys a simulation model calibrated by empirical results to explore the impact of several proactive admission policies. The simulation comprises a GW with unlimited capacity and an ICU with limited capacity. It utilizes a discrete-time Markov chain to capture the evolution of patients' severity levels within the ward. In addition to proactive admissions, the ICU accommodates unplanned internal admissions and direct admissions from the ED, which are given the highest priority. When the ICU is at capacity, the patient with the shortest remaining ICU LOS is prematurely discharged to make room for these high-priority admissions. The authors analyze both static threshold policies, wherein patients above a given severity level are proactively admitted to the ICU if any beds are available, and state-dependent threshold policies, where severity thresholds reduce with the number of beds available. Their results show that while proactive admissions can reduce mortality and hospital LOS, they may also lead to higher rates of readmissions and premature discharges if the policy is applied too aggressively. However, applying a state-dependent threshold policy can mitigate some of these negative outcomes.

Meisami et al. (2019) examine proactive admission policies within a network of a single GW, multiple SDUs, and multiple ICUs. They introduce a static optimization model to determine patient severity thresholds (driven by a mortality risk score), which dictate proactive admission to either SDUs or ICUs. Instead of using a simulation model, the authors propose a stochastic location process to capture the flow of patients in the network. This process accurately reflects patient movement among wards across different days, conditional upon their initial admission unit, and is integrated within the optimization framework. The optimization objective is to maximize the time-averaged sum of risk scores for patients admitted to the ICUs and SDUs, aiming to optimally allocate the limited resources of these units to the patients most in need. Utilizing actual hospital data, the authors illustrate how the risk-based threshold policies derived from their model could enhance admission rates while keeping blockings in ICUs and SDUs to a minimum, offering improvements over current practices.

Grand-Clément et al. (2023) use the same problem structure as in the simulation model of Hu et al. (2018) but follow an MDP modeling approach to identify the optimal proactive admission policy to the ICU. Given bed availability, the optimal policy identifies whether and when a patient must be proactively transferred to the ICU. By focusing on a single patient and making assumptions about the ordering of rewards associated with different patient outcomes, the authors show that a threshold proactive admission policy exists maximizing the expected discounted rewards. This policy recommends admitting patients whose severity exceeds a threshold when ICU beds are available. To account for the potential error in estimating transition probabilities among different severity states, they also propose a robust formulation that maximizes the worst-case reward across all plausible values of transition probabilities. They show that this policy also has a threshold structure and transfers more patients than the policy that does not consider parameter uncertainty.

Their simulation analysis based on realistic data indicates that not accounting for parameter uncertainty could lead to an optimistic estimation of the hospital performance and, consequently, unreliable insights about the impact of threshold policies.

4.5. Facility Transfer Policies

Marquinez et al. (2021) explore facility transfer policies in a public-private partnership where patients needing critical care from a public hospital within the network can be transferred to another public hospital or a private clinic within the partnership. The patients are categorized into various classes based on their diagnosis related group and transfer costs between facilities are calculated to include transportation expenses as well as disparities in the quality and costs of care. The authors develop an MDP model to minimize total transfer costs over an infinite planning horizon. They employ an approximate approach to solve the model. Utilizing illustrative data, they demonstrate that their policy may sometimes opt to transfer patients to a different hospital even when there are available beds at the current location, suggesting that an idling policy might be optimal under certain conditions. A case study using realistic data indicates that policies derived from their approximate model yield significant cost savings compared to a myopic policy.

4.6. Observations and Future Research

Table 1 summarizes the main features of the papers reviewed in Sections 4.1-4.5. We observe that 9 out of the 16 papers utilize MDP formulations, while the remaining papers employ alternative approaches such as static optimization, simulation-optimization, or system-dynamics simulation. The stochastic location model proposed by Meisami et al. (2019) offers a suitable alternative to complex dynamic or simulation models, though its application has been limited. Several recent studies have addressed proactive admission policies, and newer studies on admission policies also incorporate early discharge policies. There is an increasing emphasis on using empirical models to estimate the impact of interventions on patient outcomes. Finally, only one study has been conducted on facility transfer policies.

Table 1. Literature review of ADT problem for ICUs.

Article	Admission control policies	Methodology	Readmission	Patient Classification	Patient severity progression	Objective function/ Performance measures
Shmueli et al. (2003)	Admission	Queueing theory embedded in an optimization model	-	Multiple groups based on the APACHE II score	-	The expected survival benefit
Dobson et al. (2010)	Early discharge	Stochastic Markov chain	-	Two groups (scheduled patients and unscheduled patients)	-	Probability of bumping
Chan et al. (2012)	Early discharge	MDP	✓	Multiple groups based on the ailment/health condition	Based on a random progression of the state of health condition	Expected discharge costs over a finite horizon.
Kim et al. (2015)	Admission	Empirical, MDP	✓	Multiple groups based on the Laboratory Acute Physiology Score (LAPS)	-	Total costs of ICU admission denials over a finite horizon in terms of increases in hospital LOS, hospital readmission, or transfer up)
Li et al. (2015)	Admission & early discharge	MDP	-	Two groups based on the APACHE II score	-	The benefit of admission to the ICU net the cost of early discharges
Mallor et al. (2015)	Early discharge	Static optimization	-	Single group	-	The probability of patient rejection and the shortening of patients' LoS
Yang et al. (2015)	Admission	MDP	-	Multiple groups based on the required surgery and surgeon	-	p-norm of the total postponement costs (These costs vary based on the type of surgery and the surgeon performing it)
Mahmoudian-Dehkordi & Sadat (2017)	Early discharge	System dynamic simulation	✓	Two groups (critical and non-critical)	-	Accumulated total hospital mortality
Hu et al. (2018)	Proactive transfer	Empirical model and simulation study	✓	Multiple groups based on the Early Detection of Impending Physiologic deterioration score, version 2 (EDIP2)	Based on the discrete-time Markov chain	LOS, morality, and readmission rate
Li et al. (2019)	Admission & early discharge	MDP	-	Multiple groups based on the APACHE II score	-	The benefit of admission to the ICU net the cost of early discharges
Meisami et al. (2019)	Proactive transfer	Static optimization	-	Multiple groups based on the mortality risk score	-	The time-averaged sum of risk scores
Azcarate et al. (2020)	Early discharge	Simulation-optimization	-	Two groups (scheduled patients and unscheduled patients)	Based on the Phase-type distribution	Expected loss in LOS due to early discharges
Ouyang et al. (2020)	Admission & early discharge	MDP	✓	Two groups (patients with highly critical conditions, and patients with critical conditions)	Based on the discrete-time Markov chain	Expected long-term average cost
Bai et al. (2021)	Admission & early discharge	MDP	-	Two groups (low-severity and high-severity patients)	Based on the a probability transition matrix of health conditions	The long-term weighted average of monetary (due to the rejection or early discharge of patients) and medical costs (increase in mortality rates)
Marquinez et al. (2021)	Facility transfer	MDP	-	Multiple groups based on the Diagnosis Related Group (DRG) and transfer costs	-	Total transfer costs over an infinite horizon
Grand-Clément et al. (2023)	Proactive transfer	MDP	✓	Multiple groups based on the Advance Alert Monitor (AAM) early warning score	Based on the discrete-time Markov chain	The expected discounted rewards

Below, we propose four new avenues for research within the context of ICU.

When is prioritizing ward admission for ICU patients ready to be discharged effective in reducing mortality? As reported earlier, using a simulation approach, Mahmoudian-Dehkordi and Sadat (2017) found that prioritizing ward admission for ICU patients ready to be discharged over ED patients seeking a general bed is more effective in reducing hospital mortality than the premature discharge of ICU patients or the temporary placement of ED patients awaiting ICU admission in GWs. Future research should analytically identify the conditions under which this policy is most effective and how it can best be integrated with the other two policies. This requires developing a model that captures the flow between the ED, ICU, and GWs, along with an empirical model to estimate the impact of delayed ICU admissions and premature ICU discharges on mortality or other patient outcomes.

When might idling admission and discharge policies be optimal? As outlined earlier, Ouyang et al. (2020) prove that an optimal non-idling policy exists under certain conditions, implying that beds should not be left empty when demand exists. In contrast, Bai et al. (2021) show that under certain conditions, an optimal admission policy could be idling, meaning beds might be intentionally left vacant. This discrepancy indicates the need for a thorough investigation to determine when idling admission or discharge policies could be optimal. Understanding the specific conditions and factors that favor idling versus non-idling policies is crucial for developing effective ICU admission strategies.

What if GWs have finite capacity? Many studies, such as those by Ouyang et al. (2020) and Bai et al. (2021), assume that GWs have infinite capacity, allowing them to admit ICU-rejected arrivals or ICU early discharged patients without restriction. However, exploring how the derived policies would be affected when capacity constraints for GWs are explicitly considered would be interesting.

When can permitting early discharge of patients eliminate the need for idling admission policies? It is reasonable to assume that if early discharges are allowed and costless, there would be no need to reject new arrivals when beds are available. This is because new arrivals can be admitted and subsequently discharged early if necessary, such as when the system is full and another new arrival occurs. However, early discharge may impose additional costs on the health system or the patient. For instance, discontinuity in care for early discharged patients from the ICU may increase the likelihood of adverse outcomes compared to admitting them to a lower care unit and keeping them there for the duration of their stay. Therefore, an empirical analysis is needed to investigate the impact of changing patients' levels of care during their stay, along with an analytical model to explore the implications of this impact on early discharge and admission policies.

5. EDs

Similar to ICUs, patients cannot wait long before admission to EDs. This makes them operate similarly to a loss system. We categorize ADT policies in EDs into three major types: admission, assignment, and early transfer and discharge. *Admission* policies focus on whether a new patient arriving at the ED should be accepted or rejected. In contrast, *assignment* policies determine the specific ED facility to which a patient should be admitted or if they should be rejected outright. *Early transfer and discharge* policies consider moving some patients outside the ED earlier than usual by transferring them to inpatient wards or discharging them entirely, aiming to create additional capacity in the ED. We identified only one paper for each type of policy, as reviewed below.

Lee and Lee (2018) study the admission policy for patients arriving at an ED following a mass casualty incident. They classify casualties into different classes based on severity, with admission reward for each class defined as the improvement in survival probability following admission to the ED. Given a short planning horizon (e.g., half a day), the problem is formulated as an MDP model aiming to maximize the total expected reward. Under certain conditions, the authors prove that the optimal policy is a state-dependent threshold policy, which admits a new arrival if the number of available beds exceeds a certain threshold determined by the arrival time and severity of the patient. Focusing on a scenario with only two severity classes, they demonstrate the advantages of this policy over the FCFS policy and a policy that only admits high-severity patients through numerical experiments.

Pehlivan and Xie (2021) tackle the general problem of assigning a customer, who may belong to one of several classes, to one of multiple multi-server stations or rejecting them altogether upon their arrival. They assume that LOSs are exponentially distributed across all classes and stations with the same mean. However, the reward for admitting each class varies depending on the station to which they are admitted. This scenario has applications in deciding which ED facility an emergency patient must be transported to, where the reward for assignment decreases as patients are taken to facilities farther from their homes. The authors formulate the problem as an MDP model to maximize the expected reward over an infinite planning horizon. They demonstrate that the optimal policy for the two-station setup is a switching curve policy, meaning that the assignment policy switches from assignment to one station to the other, then to rejection, as the number of busy servers in each station increases. For networks with more than two stations, the authors propose an approximate method that is shown to produce near-optimal solutions through numerical experiments. Additionally, they propose a static optimization model for calculating a tight upper bound on the optimal average reward, which can serve as a benchmark for assessing the efficiency of different heuristic policies.

Mills et al. (2021) consider early transfer and discharge policies from a hospital's ED and inpatient wards to respond to the immediate need for surge capacity and compare it with workload smoothing in inpatient wards to mitigate the need for surge capacity. Specifically, they formulate static optimization models to determine the number of high-acuity (low-acuity) ED patients that must be early transferred to inpatient wards (early discharged home) as well as the number of inpatients that must be early transferred to a lower complexity unit given a certain amount of excess ED demand. These models minimize the cost of rejecting additional demand plus the cost associated with early discharges or transfers under two different scenarios: a coordinated decision-making scenario in which early discharge and transfer decisions are made centrally at the hospital level, and an independent decision-making scenario, wherein ED and inpatient wards make these decisions following a negotiation process. The optimal early discharge and transfer policies are compared with workload smoothing policies in a simulation model, where the latter is implemented by imposing a cap on the number of elective patients admitted each day. The results suggest that virtually every hospital can increase its surge capacity by coordinating early transfer and discharge. However, hospitals with certain utilization and elective case mixes can substantially reduce the need to surge capacity by implementing workload-smoothing policies.

Table 2 illustrates the characteristics and technical aspects of articles focused on admission control in EDs. Only three papers are identified in this section, which is not surprising given the acute nature of ED treatments. This leaves limited scope for controlling patient arrivals or discharges. However, there appears to be more potential in deciding which ED facility a patient should be transferred to based on the severity of their condition and the occupancy levels of different facilities.

Table 2. Literature review of ADT problem for EDs.

Article	Admission control policies	Methodology	Readmission	Patient Severity	Severity progression	Objective function/ Performance measures
Lee and Lee (2018)	Admission	MDP	-	Multiple groups based on severity	-	Total expected reward in finite horizon
Mills et al. (2021)	Early transfer & discharge	Static optimization	-	Four groups (High-acuity emergency, low-acuity emergency, emergency hospitalization, and elective patients)	-	The cost of rejecting additional demand and the cost associated with early discharges or transfers
Pehlivan and Xie (2021)	Assignment	MDP	-	Multiple groups based on the severity	-	Total expected reward in infinite horizon

Below, we propose four new avenues for research within the context of ED.

To which multi-server station should a customer be sent, given differences in service time and desired service levels? As outlined earlier, Pehlivan and Xie (2021) study the problem of deciding which multi-server station a customer should be assigned to if all customer classes have the same service time. It would be interesting to generalize this problem to settings where service times or desired service levels vary between different customer classes. This scenario would naturally occur when assigning patients to EDs, where patients are divided based on the criticality of their condition, with the most critical ones having longer LOS and being less tolerant of delays.

How do we model the negotiation process between the ED and other wards? Mills et al. (2021) propose a model for capturing the negotiation process between EDs and inpatient wards in response to the immediate need for surge capacity. It would be interesting to investigate such negotiations more generally, including other wards such as SDU and ICU. Additionally, comparing the outcomes of such negotiations with scenarios where decisions to admit, discharge, or transfer patients are made centrally would provide valuable insights into hospital management.

6. GWs

ADT control in GWs can manifest in various forms, including ward selection policies, admission policies with and without queuing, and discharge policies. *Ward selection* policies are designed to allocate the most suitable ward for a patient based on their medical condition and the availability of beds. Typically, one or more wards may be deemed appropriate as their resources and nursing team skill sets align with the patient's needs. These wards are referred to as the patient's primary wards. When a primary ward is at capacity, the patient may be admitted to a non-primary ward, a practice known as patient outlying (Izady et al., 2024a) or off-service placement (Song et al., 2020). Admission policies allowing *queuing* enable patients to be placed in a waiting queue if immediate admission is impossible. In contrast, policies *without queuing* stipulate that patients must either be admitted immediately or rejected, often resulting in the patient being transferred to another facility. *Discharge* policies focus on determining which patients are ready to be discharged home, ensuring that beds are available for incoming patients and that those discharged have sufficiently recovered to avoid readmission.

6.1. Ward Selection Policies

Chalgham et al. (2019) address the challenge of identifying an appropriate non-primary ward for ED patients whose primary ward is full when the decision to admit is made. They employ multi-criteria decision-making (MCDM) methods to rank alternative wards based on several factors, including the patient's pathology, the availability of empty beds, the number of multi-skilled nurses in each ward, and the ward's capacity to manage critically ill patients. Their numerical results demonstrate a reduced boarding time for patients in the ED.

Dai and Shai (2019) consider the problem of deciding whether a patient awaiting admission to their primary ward must be admitted to a non-primary ward. They model the hospital inpatient flow as a multi-pool parallel queueing system with multiple customer classes, where each pool represents the primary ward for a specific class of patients. The patients waiting for a primary bed to become available can be admitted to a non-primary ward at a cost. This is formulated as an MDP model, requiring the decision-maker to determine at each decision epoch the number of waiting patients from each class to allocate to each non-primary pool. The objective is to minimize the long-term average cost of these non-primary assignments, plus the costs associated with patients waiting in the queue. A simulation-based approximate solution approach is designed to overcome the curse of dimensionality of the MDP formulation. The numerical results demonstrate that this approach effectively reduces congestion while keeping non-primary assignments manageable.

Heydar et al. (2022) address the assignment of patients to different inpatient wards within a hospital, acknowledging that a patient's LOS can increase if placed in non-primary wards. They classify patients into several classes, assuming that each class can only be admitted to a certain subset of wards and that the mean LOS varies depending on the ward to which they are assigned. The ward with the shortest LOS for a given class is considered the primary ward for that class. The authors formulate the problem as an MDP model to determine where to admit new patients and whether to relocate existing patients to other wards. The objective is to minimize the weighted average of accumulated LOS, relocation costs, and penalties for admitting patients to non-primary wards over a finite planning horizon. Due to the high dimensionality of the problem, an approximate approach is used, with its efficiency demonstrated through numerical examples.

6.2. Admission Policies with Queueing Allowed

Helm et al. (2011) address admission control in GWs for scheduled elective patients and a subset of unscheduled patients referred to as expedited patients. These expedited patients generally have lower acuity than most emergency patients admitted through the ED. However, they often seek hospital admission via the ED due to lengthy waiting times associated with elective admission. The authors suggest separating expedited patients from emergency cases and holding them in a call-in queue managed by an admission policy. This policy not only determines whether an expedited patient should be admitted or held in the queue but also whether a scheduled elective patient should be canceled. The problem is formulated as an MDP model that aims to minimize the holding costs for expedited patients in the queue, the opportunity costs of unoccupied beds, the cost of overcapacity when more patients than beds are present, and the costs associated with elective cancellations. A double-threshold policy, featuring thresholds for both call-in and cancellation actions, is proven optimal for a special case of the problem. Based on this finding, the authors propose a practical interval-based admission control policy that divides hospital occupancy into three intervals, with specific actions linked to each interval. Applying this policy to real hospital data shows a reduction in emergency blockages and elective cancellations.

Samiedaluie et al. (2017) investigate admission control policies for ED patients requiring admission to a neurology ward with the possibility of patient queueing prior to admission. Specifically, when an admission request is made for an ED patient, the ward must decide whether to admit them to the ward, place them in a queue with limited capacity, or transfer them to another hospital. Similarly, when a patient is discharged from the ward, the ward must decide whether to admit a patient from the queue or reserve the bed for future cases. Patients are categorized into multiple types based on their severity, each having distinct mean LOSs, waiting costs, and transfer costs. The authors adopt a continuous-time MDP model to minimize the average cost of waiting and transfer per unit time, assuming exponential LOS for all patients. A notable aspect of their model is using a regression model to estimate waiting costs based on the expected loss in Health-Related Quality of Life (HRQoL) measure for each additional day of ED boarding. Due to the high dimensionality of the MDP formulation, an approximate algorithm is proposed. Computational results indicate that the policies derived from the approximation effectively mitigate overall patient health deterioration compared to several alternative strategies. Specifically, the results show a substantial reduction in the average HRQoL loss per day under their approximate dynamic programming policy relative to the existing policy at their partner hospital, which earmarks beds for different patient types and transfers patients whose waiting time in the ED exceeds 48 hours.

6.3. Admission Policies with no Queueing

Dwyer-Matzky et al. (2021) examine the admission policy to observational units for ED patients who require a brief period (less than 48 hours) of surveillance and treatment before discharge. They categorize patients into multiple types based on their characteristics and propose a straightforward admission criterion. This criterion involves assigning each patient to the unit—a GW or an observational unit—where the expected reward, multiplied by an increasing function of the unit's available capacity, is maximized. Utilizing a static optimization model, they derive an upper bound for the effectiveness of their admission policy and explore various methods for calculating the admission rewards to different units. Their numerical experiments, informed by real data, demonstrate that their policy can improve the total hospital LOS and reduce its variability.

Liu et al. (2022) address the admission control problem for buffer zones established in some hospitals during the COVID-19 pandemic to prevent the spread of the virus. ED patients requiring hospital admission are placed in these zones for COVID-19 testing before moving to a GW. The authors categorize patients into high- and low-severity groups, each with distinct arrival rates, exponential LOSs, admission rewards, and rejection penalties. They stipulate that only low-severity patients may be denied admission to these zones when beds are available. The study considers two scenarios: one with constant patient arrival rates and another with time-varying rates. For the scenario with constant rates, an infinite-horizon MDP formulation is proposed, while a finite-horizon MDP formulation is developed for the time-varying scenario. The authors propose several iteration algorithms to find optimal and near-optimal policies for constant and time-varying scenarios, respectively. Numerical experiments demonstrate the effectiveness of these policies compared to the FCFS policy. Additionally, the results highlight that accounting for the non-stationarity of the arrival process (through the finite-horizon model) leads to higher admission rates, highlighting the value of such models in environments with fluctuating demand.

Jiang et al. (2023) explore the admission problem in inpatient departments such as urology, where mixed-gender rooms are avoided. The study categorizes patients into two severity types, each associated with a distinct reward, requiring the decision-maker to decide whether a new arrival must be admitted. Admission is denied if no beds are available or if admitting the patient would result in a mixed-gender room. However, the model does permit the rearrangement of existing patients to different rooms to free up space. Given that each room contains two beds, the problem is formulated as an MDP model to maximize the long-term discounted reward. The study identifies the structural properties of the optimal solution and proposes a value iteration method to determine the optimal policy. Numerical analysis reveals that the optimal policy maintains the threshold-type structure typically observed in scenarios that do not consider gender constraints.

6.4. Discharge Policies

Shi et al. (2021) investigate discharge policies for patients in GWs, considering their readmission risk. They develop a predictive model to estimate each patient's probability of readmission for each day following their (potential) discharge based on their characteristics and the number of days they have already been in the hospital. This prediction model is incorporated into an MDP formulation, which determines the optimal number of patients of each class (representing their characteristics) to discharge each day to minimize the long-term average costs of waiting and readmissions. Analyzing the MDP model's structural properties reveals a ranking of patients to discharge at each decision point and a threshold policy that aligns with this ranking. These insights are combined to create a heuristic algorithm, which is shown to produce near-optimal solutions for several small-scale problem instances. Implementing the heuristic-based policy using realistic hospital data shows improved performance compared to an empirical policy based on historical discharge behavior and a static threshold policy. The authors further detail the implementation process for their discharge risk prediction and optimization tools in their partner hospital, indicating their approach's practical applications and benefits.

Chuang et al. (2023) address the decision-making process for discharging GW patients into long-term care centers (LTCs) with limited bed availability. Patients are categorized into multiple classes based on their characteristics, with each class's severity represented through several states that evolve according to a discrete-time Markov Chain. Given a fixed number of patients ready for discharge and a random number of available LTC beds in each period, they develop an MDP formulation that determines whether each patient should be discharged to LTC or remain in the hospital. The model's objective is to minimize the total cost over an infinite horizon, where costs are calculated based on patient resource usage, depending on where patients stay in the hospital, pass away, are discharged home without readmission, or are discharged but readmitted within 30 days. The authors also introduce a prediction-driven formulation incorporating predictions of the patient's next state using machine learning algorithms. They propose index policies to overcome the curse of dimensionality inherent in both formulations. The results demonstrate that these policies outperform both FCFS and myopic policies, offering more effective and efficient discharge decision-making in healthcare settings.

6.5. Observations and Future Research

GW beds are often used as a last resort for keeping patients in the hospital due to their larger availability and lower cost of care than ED, SDU, and ICU beds. Consequently, overflows from these units frequently need to be accommodated in GWs. Further, their ability to discharge patients is often constrained by external factors such as delays in insurance authorization, lack of adequate home healthcare services, nonavailability of beds in PCUs, and insufficient family or caregiver support. These challenges contribute to high occupancy rates in GWs, along with a significant amount of patient outlying and frequent ward changes. This highlights the importance of designing and implementing effective admission, discharge, and assignment policies. Our literature review identified at least two papers addressing each of these policies. The dominant formulation methodology remains the MDP, with only one study employing a non-MDP approach. Table 3 presents the characteristics and technical aspects of the articles reviewed.

Table 3. Literature review of ADT problem for GWs.

Article	Admission and control policies	Methodology	Readmission	Patient Severity	Severity progression	Objective function/ Performance measures
Helm et a. (2011)	Admission policies with queueing allowed	MDP	-	Three groups (elective, emergency, and expedited patients)	-	Long-term average holding costs, the opportunity costs of unoccupied beds, the cost of overcapacity, and the elective cancellations
Samiedaluie et al. (2017)	Admission policies with queueing allowed	MDP	-	Multiple groups based on the HRQoL	-	Long-term average cost of waiting and transfer per unit time
Chalgham et al. (2019)	Ward selection	MCDM	-	Multiple groups based on the pathologies	-	LOS
Dai and Shai (2019)	Ward selection	MDP	-	Multiple generic groups	-	Long-term average cost of non-primary assignments plus the costs associated with patients waiting
Dwyer-Matzky et al. (2021)	Admission policies with no queueing	Static optimization	-	Multiple groups based on the medical characteristics	-	Expected reward from admitting patients to inpatient units
Shi et al. (2021)	Discharge	MDP	✓	Multiple groups based on the readmission probability	Based on the two stages of disease progression: critical and stable	Long-term average costs of waiting and readmissions
Heydar et al. (2022)	Ward selection	MDP	-	Multiple groups based on a subset of possible wards for hospitalization	-	Weighted average of accumulated LOS, relocation costs, and penalties for admitting patients to non-primary wards
Liu et al. (2022)	Admission policies with no queueing	MDP	-	Two groups (high- and low-severity groups)	-	Long-term average rewards for accepting or rejecting patients
Chuang et al. (2023)	Discharge	MDP	✓	Multiple groups based on the Charlson-Deyo Comorbidity Index (CDCI) and the Elixhauser Comorbidity Index (ECI)	Based on the machine learning algorithms	Long-term discounted total discharge and congestion cost
Jiang et al. (2023)	Admission policies with no queueing	MDP	-	Two groups (high- and low-severity groups)	-	Long-term discounted reward

Below, we propose four new avenues for research within the context of GW.

How do we capture the impact of outlying patients on non-outlying patients in assignment policies? A recent empirical study by Lim et al. (2021) shows that the mean LOS of non-outlying patients is longer in wards that receive a larger number of outlying patients. Capturing this trend in assignment policies would be an interesting area for future research. This would require extending the formulation proposed by Heydar et al. (2022) to account for the number of outlying and non-outlying patients in each ward. By incorporating these dynamics, it would be possible to develop assignment policies that mitigate the negative impact of outlying patients while using the available capacity effectively.

How do we capture the impact of ward changes in assignment policies? Heydar et al. (2022) consider the possibility of moving patients to different wards during their stay in the hospital in their investigation of optimal assignment policies. However, empirical results presented by Arabzadeh et al. (2022) show that an increased number of ward changes correlates with a longer mean LOS for patients, with this impact being more significant for elderly patients. Capturing this impact in assignment policies would be a valuable area for future research, ultimately enhancing patient outcomes and the operational efficiency of assignment policies.

How do we capture the impact of focus and workload in admission and assignment policies? Empirical results from Best et al. (2015) show that as more specialties are assigned to an inpatient ward, the mean LOS increases due to a loss of focused care. This effect is modest for smaller workloads but becomes more significant as the workload grows, eventually reaching an asymptotic value. Capturing this empirical trend in admission and assignment policies would help develop policies that balance specialization and generalization under different levels of workload.

How do we devise admission policies in the presence of ICU and ED demand? Two streams of demand typically compete for the limited resources of inpatient beds, ICU patients ready to be discharged and ED patients ready to be admitted. The literature has only focused on only one of these streams. For example, Samiedaluie et al. (2017) consider only ED demand, whereas Dobson et al. (2010) focus on ICU demand. An interesting area for future research is investigating admission policies that account for both demand streams simultaneously. Incorporating elective surgeries seeking inpatient admission would provide further insight into devising effective admission policies.

How do we devise admission, discharge, and transfer policies for virtual wards? As healthcare systems evolve, virtual wards have emerged as a promising solution to manage patient care remotely, particularly for those with chronic conditions or post-acute care needs. Virtual wards provide continuous monitoring and medical support to patients in their homes, thereby reducing the burden on physical healthcare facilities. However, it is important to devise effective admission policies, identifying which patients are most likely to benefit from these wards, discharge policies, clearly defining when patients should be discharged from these wards, and transfer policies, deciding when patients should be transferred from virtual wards to a physical ward.

7. Other papers

Some papers do not fall within either of the categories of ICU, ED, or inpatient wards. One article compares two readmission policies for nursing home patients, one study examines admission policies for a generic multi-class loss queueing model, one paper evaluates prioritizing patient access to examination rooms and physicians in EDs, one paper examines early discharge policies in ICUs, SDUs, and GWs concurrently, two papers conduct controlled experiments to study behavioral biases in clinical admissions, one study compares different admission systems during a pandemic, and finally, one study decides whether patients must be admitted to intensive, intermediate, or general care units. These studies are reviewed below.

Hannan and Gimbrone (1987) utilize a simulation model to evaluate two policies for nursing home patients who temporarily require hospital care: a bed reservation policy and a priority readmission policy. The bed reservation policy holds the patient's bed in the nursing home until their hospital discharge. In contrast, the priority readmission policy prioritizes the returning patient over new admissions but does not guarantee immediate readmission. This simulation assesses the trade-off between increased nursing home capacity under the priority readmission policy and the potential increase in waiting times for hospital admissions. Using cost data from the Medicaid program, the authors demonstrate that the priority readmission policy could lead to significant losses due to a substantial increase in waiting times for nursing home patients seeking readmission.

Altman et al. (2001) study admission policies for a generic loss queueing model, where multiple customer classes arrive at the system with finite resources according to a Poisson process with different rates. Each customer class requires a specific number of resources and occupies those resources for a duration that follows an exponential distribution with a specific rate. If a customer is admitted, the system receives a class-dependent reward. The authors propose an MDP approach to maximize the average discounted rewards and prove several structural properties of the optimal policy. Specifically, the optimal policy has a threshold form for the special case where service rates and resource requirements are equal across different classes. This means that for each customer class, there is a critical level for the total number of customers present in the system above, for which no additional customers are admitted.

Saghafian et al. (2012) examine patient prioritization in EDs in two phases. Phase 1 concerns prioritizing the allocation of an examination room to patients who have completed triage. Phase 2 involves prioritizing physician allocation to patients ready for their first or subsequent consultation. It is assumed that, following an initial consultation, each patient undergoes a series of testing/waiting and further consultation rounds until they are eventually either discharged or admitted to the hospital. During the testing/waiting periods, the patient occupies the examination room but is unavailable for consultation. The authors evaluate the two phases of prioritization within two distinct configurations: a pooled configuration, where resources are shared among patients eventually admitted or discharged, and a streaming configuration, where dedicated pathways are used for patients likely to be admitted or discharged, as anticipated by the triage nurse. Using a combination of analytical and simulation results, they demonstrate that the streaming configuration, which prioritizes high-severity patients in both admission and discharge pathways for phase one and prioritizes the latest (earliest) arrivals in admission (discharge) pathways for phase 2, is likely to be beneficial for busy EDs which experience a high percentage of admitted patients with substantial day-to-day variations.

González et al. (2019) focus on early discharge policies for patients whose conditions have stabilized sufficiently to no longer require intensive care yet are not stable enough to clearly warrant transfer to a less complex care unit. Specifically, they model the flow of patients among an ED, an ICU, an SDU, and a GW. ED patients are assumed to require admission to either ICU, SDW, or GW, and once admitted, their condition progresses from 'serious' to 'fair' and from 'fair' to 'good.' In addition to ED demand, elective patients must be admitted to ICU, SDU, or GW following their procedure. Patients reaching good condition in each unit move to the unit providing a lower care complexity (and are discharged if they are in the GW), where their condition follows the same path from 'serious' to 'good.' A decision must be made in each time period regarding how many patients in 'fair' condition in each unit must be discharged to a lower complexity unit and how many procedures must be canceled. When the destination unit of a patient is full, they wait in a queue until a bed becomes available. The problem is formulated as an MDP model whose objective is maximizing the total net income minus the total cost associated with cancellations and waiting time for ED patients. A linear programming model approximates the problem, the solution of which can be obtained numerically. The simulation results, calibrated by data from a real hospital, show that the proposed policy could potentially improve waiting time in the ED as compared to the current situation and two alternative policies.

Kim et al. (2020) and Kim and Tong (2023) study the impact of human behavioral biases on admission control in high-complexity care units. Both studies focus on a decision-maker who must accept or reject a patient, assuming that new patients are turned away when the system is full. Patients are categorized into two groups, high-reward, and low-reward, where the "reward" represents the benefits patients derive from admission. The goal is to maximize the long-term average reward. According to analytical results derived from sources like Altman (2001), the optimal policy is to admit all high-reward patients and only low-reward patients if the total occupancy is below a certain threshold. In this context, Kim et al. (2020) examine biases assuming perfect diagnostic information, whereas Kim and Tong (2023) consider uncertainty in diagnosis at the time of admission. Both studies involve controlled experiments where physicians manage a simulated hospital unit. We review these two articles below.

Kim et al. (2020) explore the impact of “occupancy information hurdles” and “decision noise” on unit utilization. Occupancy information hurdles are tasks that physicians must perform to observe the occupancy of the unit, such as calling the admission controller or entering a password on a computer. Decision noise refers to operation errors in admission decisions, such as admitting a patient when it would be optimal to reject or vice versa. The results of experiments indicate that occupancy information hurdles cause physicians to overestimate the utilization of the unit, possibly because discharge events are less noticeable than admissions, resulting in more patients being rejected. The study also finds that while physicians' admission decisions generally align with optimal choices without information hurdles, decision noise can result in higher- or lower-than-optimal utilization depending on system parameters. The researchers conclude that even removing minor information hurdles can significantly improve occupancy. Moreover, it may be beneficial to decompose admission decisions into clinical and operational components, with the operational component managed by an automated algorithm.

While Kim et al. (2020) assume perfect diagnosis, Kim and Tong (2023) study how uncertainty in patient diagnosis affects admission control performance. Their experiments show that while physicians adjust their perceived reward for admitting a patient to account for diagnostic errors, these adjustments often need to be revised. As a result, physicians tend to overestimate the difference in rewards between those diagnosed as high-reward and low-reward patients, leading to an over-rationing bias. This implies that they reject more low-diagnosis patients than a rational decision-maker might. The study shows this bias has a more significant impact in hospitals with less accurate diagnoses, a higher influx of high-reward patients, and less decision support on conditional probabilities. In such settings, mitigating this bias by providing feedback to admission decision-makers about the outcomes of rejected patients becomes critically important. The authors suggest several approaches to delivering this feedback.

Chen and Kong (2022) examine the necessary scope of hospital admissions during a pandemic such as COVID-19. They analyze three distinct admission systems: a mixed system where both mildly and severely infected patients are admitted, a hierarchy system that admits only severely infected patients, and the Fangcang system, where asymptomatic and mildly symptomatic individuals are treated in temporary hospitals. They propose a modified Susceptible–Exposed–Infected–Recovered (SEIR) model to capture the impact of individual access to limited medical resources under each admission system. Using data from Wuhan hospitals, the researchers demonstrate that the Fangcang system leads to the lowest rates of infections, deaths, and hospital bed occupancy. A survey of healthcare workers in Wuhan, China corroborates these findings.

Zhalechian et al. (2023) examine the selection of a ward with the appropriate level of care— intensive, intermediate, or general—for patients requiring hospital admission. This choice requires balancing the improved patient outcomes typically associated with higher levels of care against the opportunity cost of not having high-care beds available for more complex future cases. The authors propose a batch online learning algorithm that adjusts the readmission probability based on feedback from previous ward assignments. This updated information is integrated into an MDP model, which determines the proportion of patients from each class to be assigned to each ward during each time interval. The objective of the MDP model is to minimize the expected number of readmissions over a finite horizon. To overcome the curse of dimensionality, the authors propose a deterministic linear programming model that identifies the optimal state-independent policy, updated in each time interval based on the most recent readmission probabilities. Using hospital data, they demonstrate that this approach outperforms both a policy approximating the hospital's current practices and a greedy policy that assigns patients to the ward with the highest expected reward. Table 4 presents the characteristics and technical aspects of the articles reviewed.

Table 4. Literature review of ADT problem for miscellaneous papers.

Article	**Admission control policies**	**Methodology**	**Readmission**	**Patient Severity**	**Severity progression**	**Objective function/ Performance measures**
Hannan and Gimbrone (1987)	The bed reservation and priority readmission	Simulation	✓	Single group	-	Waiting time, LOS, and other performance measures
Altman et al. (2001)	Admission	MDP	-	Multiple generic groups	-	Average rewards
Saghafian et al. (2012)	Streaming and pooling	MDP	-	Multiple groups based on the Emergency Severity Index (ESI)	-	Time to First Treatment (TFT) and LOS
González et al. (2019)	Early discharge	MDP	-	Three groups (serious, fair, and good)	Based on transition probabilities	The total net income and cost associated with cancellations and ED patients' waiting time
Kim et al. (2020)	Admission	Control experiment	-	Two groups (high- and low-severity groups)	-	Utilization, total reward
Chen and Kong (2022)	Admission	Simulation	-	Two groups (mildly and severely infected patients)	Based on the SEIR model	Morality and occupancy rate
Kim and Tong (2023)	Admission	Control experiment	-	Two groups (high- and low-severity groups).	-	Utilization, total reward
Zhalechian et al. (2023)	Ward selection	Data-Driven algorithm integrated into an MDP model	✓	Multiple groups based on the desirability of the unit in which they will be hospitalized	-	Expected number of readmissions over a finite horizon

8. Generic Research Areas

We have already proposed areas for future research concerning ADT policies within ICU, ED, and GW domains. Here, we suggest some broader areas of research that extend beyond these specific categories.

How do we devise readmission policies for nursing homes? The only study that considers readmission policies for nursing homes is that of Hannan and Gimbrone (1987), which compare two policies using simulation. Given that patients in these units and other long-term care centers may require frequent visits to the hospital, it is imperative to design appropriate readmission policies that, while reducing the pressure on inpatient beds, provide efficient use of nursing care facilities. Following an analytical approach using an MDP formulation will likely provide new insight into the structure and performance of such policies.

How should ADT policies be revised in the presence of pandemic patients? The recent COVID-19 pandemic put unprecedented pressure on health systems worldwide. In preparation for future pandemics, planning how ADT policies for different departments should adapt in response to pandemic demand is crucial. Key features that must be accounted for include the impact of nosocomial transmission and associated mitigation costs (Cooper et al., 2023), uncertain and time-varying patient arrival patterns (Izady, 2010), and potentially longer and less variable LOSs (Izady et al., 2024b). Developing models incorporating these factors will enable healthcare systems to create robust policies that can effectively manage the surge in patient numbers while minimizing the risk of infection spread within healthcare facilities.

How do we decide the appropriate level of care? The only study that considers designing admission policies for selecting the appropriate level of care, i.e., intensive, intermediate, or general care, is that of Zhalechian et al. (2023). Due to the curse of dimensionality, however, they work with state-independent policies, which may not be optimal. It is, therefore, important to consider other approximations or heuristic approaches that allow state-dependent policies. Extension to non-exponential LOSs common in healthcare settings is another area of future research.

How do we identify and mitigate behavioral biases in early discharge and ward selection policies? We outlined two studies that apply controlled experiments to investigate human biases in admission decision-making in a two-class patient environment. Applying the same methodology to evaluate behavioral biases in early discharge and ward selection policies is important. These studies would provide valuable insight into how close human decision-making is to optimal policies and what interventions are required to mitigate the impact of biases.

9. Conclusion

We reviewed 37 shortlisted articles focusing on ADT control policies in healthcare settings. As illustrated in Figure 4, ICU has the largest concentration of papers with 16 shortlisted papers. This is not surprising given the complexity and resource-intensive nature of ICU services. Within the ICU category, admission and early discharge policies have garnered the most attention, accounting for three-quarters of the papers in this area. GWs have the second-highest concentration, with 10 reviewed articles, followed by the ED with only 3 articles.

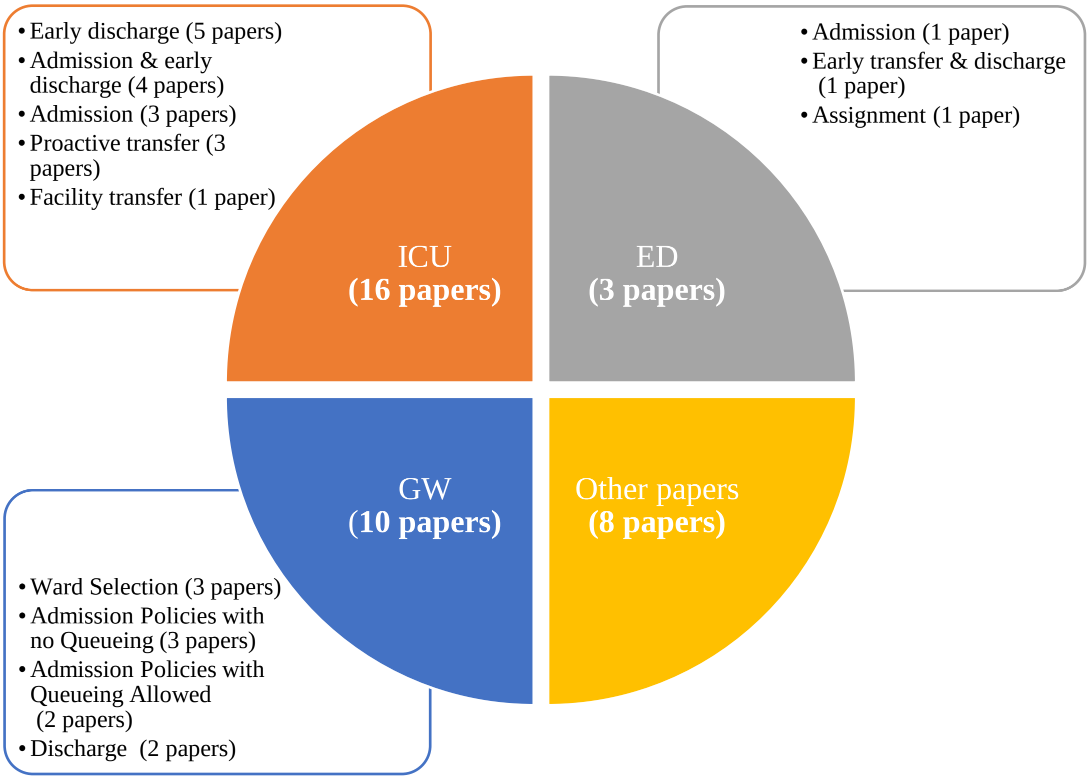

Figure 4. A summary of the reviewed articles based on their departmental focus and type of policies.

In addition to traditional admission policies, which are the primary focus of the admission control literature, our review identified several other critical aspects of control decisions within healthcare settings. For ICUs, we found that early discharge and facility transfer policies are significant considerations. In the context of EDs, early transfer and discharge policies play a crucial role in reducing congestion levels. For GWs, ward selection policies are essential to balance care efficiency and care quality. Additionally, the proactive transfer of patients from GWs to ICUs highlights the importance of pre-emptive action to improve patient outcomes with a minimum level of resources.

Moreover, sporadic attention is given to selecting the appropriate level of care, identifying and mitigating human biases, and developing effective readmission policies for nursing homes. These insights emphasize the multifaceted and unique nature of ADT control policies within healthcare and the necessity for comprehensive strategies that encompass various policies to enhance the efficiency and effectiveness of healthcare delivery.

We proposed a comprehensive range of directions for future research. Analytically, this involves developing admission and prioritization policies for ICU and ED patients who are competing for the limited resources of GWs as well as devising models which inform decision making on the appropriate level of care. We also recommend investigating the optimality of idling admissions and early discharge policies, and analysing their interactions. Additionally, there is a need to model negotiation processes among different departments and comparing the outcomes with centralized decision-making approaches. Further, we suggest exploring the impact of factors such as patient outlying, workload, focus, and ward changes on ward assignment policies.

On the experimental side, there is a critical need to identify and mitigate behavioural biases across the full spectrum of ADT control policies. Empirically, the effects of varying levels of care during a patient's stay on their health outcomes merit deeper investigation.

Finally, the development of ADT policies for virtual wards emerges as a critical yet underexplored area. Similarly, revising existing ADT policies in the context of a pandemic and formulating readmission policies for PCUs are both areas where significant gaps in research exist. These present unique challenges for OR and MS scholars, requiring pioneering efforts to adapt to the rapidly evolving healthcare landscape.

REFERENCES

Altman, E., Jiménez, T., & Koole, G. (2001) On optimal call admission control in resource-sharing system. *IEEE Transactions on Communications* 49(9) 1659-1668.

Arabzadeh, B. (2022). Reconfiguration of inpatient services to reduce bed pressure in hospitals (Ph.D. thesis), City, University of London, URL https://openaccess.city. ac.uk/id/eprint/28994/1/.

Azcarate, C., Esparza, L., & Mallor, F. (2020) The problem of the last bed: Contextualization and a new simulation framework for analyzing physician decisions. *Omega* 96(1) 102120.

Bai, J., Fügener, A., Gönsch, J., Brunner, J. O., & Blobner, M. (2021) Managing admission and discharge processes in intensive care units. *Health Care Management Science 24*(4) 666-685.

Barnett, M. J., Kaboli, P. J., Sirio, C. A., & Rosenthal, G. E. (2002) Day of the week of intensive care admission and patient outcomes: a multisite regional evaluation. *Medical Care* 40(6) 530-539.

Best, T. J., Sandıkçı, B., Eisenstein, D. D., & Meltzer, D. O. (2015) Managing hospital inpatient bed capacity through partitioning care into focused wings. *Manufacturing & Service Operations Management* 17(2) 157-176.

Borkowski, N., Williams, E.S., O'Connor, S. J., & Qu, H. (2018) Outlets for health care management research: an updated assessment of journal ratings. *Journal of Health Administration Education* 35(1) 47-63.

Bountourelis, T., Ulukus, M. Y., Kharoufeh, J. P., & Nabors, S. G. (2013) The modeling, analysis, and management of intensive care units. *Handbook of Healthcare Operations Management: Methods and Applications* 153-182.

Chalgham, M., Khatrouch, I., Masmoudi, M., Walha, O. C., & Dammak, A. (2019) Inpatient admission management using multiple criteria decision-making methods. *Operations Research for Health Care 23* 100173.

Chan, C. W., Farias, V. F., Bambos, N., & Escobar, G. J. (2012) Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research* 60(6) 1323-1341.

Chen, Z., & Kong, G. (2022) Hospital admission, facility-based isolation, and social distancing: An SEIR model with constrained medical resources. *Production and Operations Management 32*(5) 1397-1414.

Chuang, Y. T., Zargoush, M., Ghazalbash, S., Samiedaluie, S., Kuluski, K., & Guilcher, S. (2023) From prediction to decision: Optimizing long-term care placements among older delayed discharge patients. *Production and Operations Management* 32(4) 1041-1058.

Cooper, B. S., Evans, S., Jafari, Y., Pham, T. M., Mo, Y., Lim, C., ... & Knight, G. M. (2023). The burden and dynamics of hospital-acquired SARS-CoV-2 in England. *Nature* 623(7985) 132-138.

Dai, J. G., & Shi, P. (2019) Inpatient overflow: An approximate dynamic programming approach. *Manufacturing & Service Operations Management* 21(4) 894-911.

Dobson, G., Lee, H. H., & Pinker, E. (2010) A model of ICU bumping. *Operations Research* 58(6) 1564-1576.

Dwyer-Matzky, K., Pachamanova, D., & Tilson, V. (2021) Accounting for capacity: A real-time optimization approach to managing observation unit utilization. *Naval Research Logistics (NRL) 68*(5) 534-555.

González, J., Ferrer, J. C., Cataldo, A., & Rojas, L. (2019) A proactive transfer policy for critical patient flow management. *Health Care Management Science 22*(2) 287-303.

Grand-Clément, J., Chan, C. W., Goyal, V., & Escobar, G. (2023) Robustness of proactive intensive care unit transfer policies. *Operations Research* 71(5) 1653-1688.

Hannan, E. L., & Gimbrone, C. J. (1987) Predicting the impact of instituting a priority readmission policy in nursing homes. *Computers & Operations Research*. 14(6) 493-505.

Helm, J. E., AhmadBeygi, S., & Van Oyen, M. P. (2011) Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management 20*(3) 359-374.

Heydar, M., O'Reilly, M. M., Trainer, E., Fackrell, M., Taylor, P. G., & Tirdad, A. (2022). A stochastic model for the patient-bed assignment problem with random arrivals and departures. *Annals of Operations Research* 1-33.

Hu, W., Chan, C. W., Zubizarreta, J. R., & Escobar, G. J. (2018) An examination of early transfers to the ICU based on a physiologic risk score. *Manufacturing & Service Operations Management 20*(3) 531-549.

Izady, N. (2010) On Queues with Time-varying Demand. PhD dissertation, Lancaster University Management School.

Izady, N., Arabzadeh, B., Sands, N., & Adams, J. (2024a) Reconfiguration of inpatient services to reduce bed pressure in hospitals. *European Journal of Operational Research* 316(2) 680-693.

Izady, N., Savin, S., & Farahani, R. Z. (2024b) Preparing for the next health emergency: Concurrent management of pandemic and non-pandemic demand, Working paper, Available online on https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4785352.

Jiang, Y., Yang, F., Tang, Z., & Li, Q. L. (2023) Admission control of hospitalization with patient gender by using Markov decision process. *International Transactions in Operational Research* 30(1) 70-98.

Kim, S. H., & Tong, J. (2023) Admission control bias and path-dependent feedback under diagnosis uncertainty. *Manufacturing & Service Operations Management 26*(1) 117-136.

Kim, S. H., Chan, C. W., Olivares, M., & Escobar, G. (2015) ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science 61*(1) 19-38.

Kim, S. H., Tong, J., & Peden, C. (2020) Admission control biases in hospital unit capacity management: How occupancy information hurdles and decision noise impact utilization. *Management Science 66*(11) 5151-5170.

Lee, H. R., & Lee, T. (2018) Markov decision process model for patient admission decision at an emergency department under a surge demand. *Flexible Services and Manufacturing Journal 30* 98-122.

Li, J., Dong, M., & Zhao, W. (2015) Admissions optimisation and premature discharge decisions in intensive care units. *International Journal of Production Research 53*(24) 7329-7342.

Li, X., Liu, D., Geng, N., & Xie, X. (2019) Optimal ICU admission control with premature discharge. *IEEE Transactions on Automation Science and Engineering 16*(1) 148-164.

Lim, JM, H Song, JJ Yang (2024) The spillover effects of capacity pooling in hospitals. *Management Science* Accepted, Available online at https://doi.org/10.1287/mnsc.2022.02202

Liu, R., Xu, J., & Liu, Y. (2022) Dynamic patient admission control with time-varying and uncertain demands in Covid-19 pandemic. *IEEE Transactions on Automation Science and Engineering* 19(2) 620-631.

Mahmoudian-Dehkordi, A., & Sadat, S. (2017) Sustaining critical care: using evidence-based simulation to evaluate ICU management policies. *Health Care Management Science 20* 532-547.

Mallor, F., Azcárate, C., & Barado, J. (2015) Optimal control of ICU patient discharge: from theory to implementation. *Health Care Management Science 18*(3) 234-250.

Marquinez, J. T., Sauré, A., Cataldo, A., & Ferrer, J. C. (2021) Identifying proactive ICU patient admission, transfer and diversion policies in a public-private hospital network. *European Journal of Operational Research 295*(1) 306-320.

Marynissen, J., & Demeulemeester, E. (2019) Literature review on multi-appointment scheduling problems in hospitals. *European Journal of Operational Research 272*(2) 407-419.

Meisami, A., Deglise-Hawkinson, J., Cowen, M. E., & Van Oyen, M. P. (2019) Data-driven optimization methodology for admission control in critical care units. *Health Care Management Science 22*(2) 318-335.

Merigó, J.M. and Núñez, A. (2016) Influential journals in health research: a bibliometric study, Globalization and Health 12(46) 1-12.

MHA (2021) Most Widely Read Journals in Healthcare, MHADegree.org, https://mhadegree.org/25-most-widely-read-journals-in-healthcare/ visited on 5 July 2024.

Mills, A. F., Helm, J. E., & Wang, Y. (2021) Surge capacity deployment in hospitals: effectiveness of response and mitigation strategies. *Manufacturing & Service Operations Management 23*(2) 367-387.

Ouyang, H., Argon, N. T., & Ziya, S. (2020) Allocation of intensive care unit beds in periods of high demand. *Operations Research 68*(2) 591-608.

Pehlivan, C., & Xie, X. (2021) Admission control policies in loss networks. *IEEE Transactions on Automation Science and Engineering 19*(4) 3152-3171.

Saghafian, S., Hopp, W. J., Van Oyen, M. P., Desmond, J. S., & Kronick, S. L. (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research 60*(5) 1080-1097.

Samiedaluie, S., Kucukyazici, B., Verter, V., & Zhang, D. (2017) Managing patient admissions in a neurology ward. *Operations Research 65*(3) 635-656.

Shi, P., Helm, J. E., Deglise-Hawkinson, J., & Pan, J. (2021) Timing it right: Balancing inpatient congestion vs. readmission risk at discharge. *Operations Research* 69(6) 1842-1865.

Shmueli, A., Sprung, C. L., & Kaplan, E. H. (2003) Optimizing admissions to an intensive care unit. *Health Care Management Science* 6 131-136.

Song, H., Tucker, A. L., Graue, R., Moravick, S., & Yang, J. J. (2020) Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science* 66(9) 3799-4358.

Stidham Jr, S. (1978) Socially and individually optimal control of arrivals to a GI/M/1 queue. *Management Science* 24(15) 1598-1610.

Stidham Jr, S. (2002) Analysis, design, and control of queueing systems. *Operations Research* 50(1) 197-216.

Stidham, S. (1985) Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control* 30(8) 705-713.

Williams, E.S., Stewart, R.T., O'Connor, S., Savage, G.T. & Shewchuk, R. (2002). Rating outlets for health care management research: An update and extension. *Medical Care Research and Review* 59(3), 337-352.

Yang, M., Fry, M. J., & Scurlock, C. (2015) The ICU will see you now: Efficient–equitable admission control policies for a surgical ICU with batch arrivals. *IIE Transactions 47*(6) 586-599.

Youn, S., Geismar, H. N., & Pinedo, M. (2022) Planning and scheduling in healthcare for better care coordination: Current understanding, trending topics, and future opportunities. *Production and Operations Management 31*(12) 4407-4423.

Zhalechian, M., Keyvanshokooh, E., Shi, C., & Van Oyen, M. P. (2023) Data-driven hospital admission control: A learning approach. *Operations Research 71*(6) 2111-2129.