healthcare

MDPI

Article

Designing Governance Boundaries for Hospital Bed Management: Balancing Centralized Pooling and Local Protection

Shao Wang [1], Feng Xiao [2,*], Kin Keung Lai [2,3] and Yang Xu [4]

[1] Competition Training Department, Xi'an Physical Education University, Xi'an 710068, China
[2] International Business School, Shaanxi Normal University, Xi'an 710119, China
[3] Department of Data and Systems Engineering, The University of Hong Kong, Hong Kong
[4] School of Economics Management, Xi'an Technological University, Xi'an 710021, China
* Correspondence: xiaofeng9727@snnu.edu.cn

Abstract

Background/Objectives: Hospital-wide bed pooling is widely used in bed-management reform to reduce ward-level mismatch and improve the use of scarce inpatient capacity. However, formally pooled beds are not automatically usable beds. Cross-ward reassignment may require coordination, placement judgment, and timely execution, especially when several wards are under pressure. This study examines hospital bed governance as a healthcare management problem of where to draw the boundary between local protection and centralized pooling. **Methods:** We develop an α-based governance framework that places full decentralization, bounded centralization, and full centralization within a common design space. The parameter α determines how much ward-level capacity is pooled and how much remains locally protected. The analysis first establishes a frictionless pooling benchmark, and then introduces coordination friction, stress-state pooled-allocation imperfection, and the protective value of local capacity. Scenario-based computational experiments examine the mechanism across operating conditions, alternative friction specifications, and multi-ward extensions. **Results:** In the frictionless benchmark, full centralization weakly dominates because it has the broadest feasible allocation set. When pooled allocation remains fully effective, full centralization is not substantively outperformed. Coordination friction can move the raw maximizing boundary inward, but this movement does not necessarily imply a meaningful improvement over full centralization. When reassignment becomes less reliable in high-pressure states, however, an interior governance boundary can become substantively attractive by preserving some locally accessible capacity. **Conclusions:** The findings do not reject centralized pooling. They suggest that bed reform should evaluate both the formal scope of pooling and the operational usability of pooled capacity under pressure.

Keywords: hospital bed management; healthcare management; capacity pooling; centralized bed allocation; local protection; operational resilience; governance boundary

Check for updates

Academic Editor: César Leal Costa

Received: 6 May 2026
Revised: 23 June 2026
Accepted: 28 June 2026
Published: 1 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. Hospital Bed Mismatch and Centralized Pooling

Inpatient bed pressure is often treated as a capacity and patient flow problem. In practice, however, the difficulty is not always a shortage of beds in the hospital as a whole. Some wards may have unused beds while others are unable to admit patients in time. Such mismatch can arise from uneven admissions, discharge timing, continuing occupancy, and

the organization of beds within clinical units rather than as a fully shared stock. Prior work on patient flow and hospital capacity planning has shown that congestion, delayed admissions, and bed blocking are closely tied to variability and coordination across hospital units, not only to the nominal number of beds available [1–4]. The issue, then, is not only how many beds a hospital owns. It also concerns how those beds are made operationally available when demand is uncertain and uneven across wards.

A bed listed on the board is not always a bed that can be used for a patient elsewhere. Before reassignment, staff still need to check whether the receiving ward can accept the patient, whether the team on duty has the required capacity, and whether the transfer can be completed before the demand state moves on. The frictionless benchmark sets these checks aside. Under that assumption, an observed bed is accepted and converted into a placement without delay, which gives full centralization the broadest allocation set after demand is realized. The benchmark is useful for defining that upper bound. It should not be read as a close account of daily bed work. On busy days, system visibility can run ahead of authority, staffing, and clinical fit. A bed may sit inside the shared pool while remaining practically unavailable, even as wards remain short and placement decisions wait for clearance.

1.2. Pooling Efficiency Versus Local Protection

Under routine conditions, the work needed to make a pooled bed usable is often hidden. A ward may show spare capacity, yet that capacity still has to be released, matched to the patient, accepted by the receiving ward, and absorbed into the ward's current workload. These checks may be quick. They become more consequential when pressure rises in several units at the same time.

Placement staff then have to compare competing requests, and a bed visible at the hospital level may not be released quickly enough for the shortage already forming. Patient flow studies describe congestion in similar terms. Variability, timing, and coordination across units shape access alongside average demand and nominal capacity [3–6].

The model represents this gap through the governance boundary and the reassignment work generated after that boundary is set. The parameter α determines how much ward capacity enters the shared pool and how much remains within immediate ward access. When pooled beds are used, Q records the coordination work attached to that use. The stress-state experiments add a further constraint by allowing pooled capacity to become less reliable when several wards are under pressure together.

Local protection has value only in this operational setting. Capacity kept under ward access can be used through familiar staffing arrangements, specialty routines, and local judgments about patient need. Its advantage lies in speed and reachability, before a centrally coordinated reassignment has been identified, approved, and completed.

1.3. From Regime Comparison to Governance Boundary Design

In practice, bed governance is rarely a clean split between ward control and a single hospital-wide pool. Many hospitals keep some beds within ward routines for immediate admission, while making other beds visible to a placement office for cross-service reassignment. A binary description of reform can obscure this mixed arrangement. The more relevant design question is how much access should remain close to the ward, and how much capacity should be placed under central coordination before demand is realized.

A possible match still has to become an actual placement. Operations management research makes the same point in broader terms. Pooling and flexibility create value when information, decision rights, and operating routines allow the match to be carried out in practice [7–9]. The frictionless case deliberately removes this difficulty by treating

every pooled bed as reliably available. When wards are already under pressure, however, reassignment becomes less automatic. A coordinator may see a bed, but authority has to be confirmed, clinical fit checked, placement work completed, and transfer timing arranged. If several wards are competing for the same capacity, these steps may slow down or break down. The practical issue for managers is therefore where to draw the boundary between capacity that wards can use at once and capacity that must move through shared allocation.

1.4. Research Questions

The preceding discussion leads to three research questions.

First, when pooled capacity can be reassigned without delay, coordination cost, or loss of clinical fit, does full centralization remain the natural benchmark?

Second, when reassignment becomes costly or less dependable under stress, how is the advantage of full centralization affected?

Third, under what operating conditions can an interior governance boundary become attractive, and does this mechanism remain visible beyond the two-ward canonical case?

These questions define the scope of the paper. We do not assume that bounded centralization is generally superior to full centralization. Instead, we first establish the frictionless pooling benchmark, and then examine how costly or less reliable reassignment, together with protected local access, can shift the preferred governance boundary.

1.5. Methodological Approach

We study these questions using a mechanism-driven analytical model that isolates the governance trade-off, complemented by scenario-based computational experiments. The aim is not to calibrate a predictive model for a particular hospital, nor to determine whether decentralization or centralization is always preferable. Instead, the stylized setting allows us to examine when the usual pooling argument holds, and when costly or unreliable reassignment may make a bounded governance design relevant.

We begin with a general multi-ward formulation because the boundary problem is not confined to two units. Ward capacity, demand variability, and cross-ward dependence need to be represented in a common setting before the governance regimes can be compared. For the analytical arguments, however, we deliberately reduce the setting to a two-ward canonical case. This simplification is useful because it shows the central trade-off in the smallest setting where cross-ward pooling can actually arise. It should not be read as a description of how real hospital bed systems are organized.

The analysis first gives pooling its strongest case. Every bed placed in the shared pool is assumed to be reachable, clinically suitable, and reassigned without delay. Under this idealization, full centralization has the widest feasible allocation set and weakly dominates bounded designs. This benchmark is useful precisely because it shows what pooling can achieve before hospital execution enters the problem. After the ideal case has done its work, the analysis relaxes it. Cross-ward reassignment may involve approval, matching, and transfer coordination. A ward may also need capacity it can use without waiting when pressure builds. The question then becomes practical. Can a bed in the shared pool still be made usable within the decision window?

Correspondingly, across the runs, the degree of centralization is varied first. The model then adds coordination work and reliability loss under stress to observe how performance changes. The purpose is to trace the mechanism. Other runs replace the coordination friction measure or expand the ward system. These checks test whether the interior-boundary pattern is observed only in the two-ward example or still appears under nearby modeling choices.

1.6. Contributions and Paper Organization

First, the paper reframes hospital bed management as a governance-boundary problem. Much of the existing discussion treats bed-system resilience as a matter of holding additional buffer capacity, protecting reserve beds, or deciding where overflow capacity can be shared. By contrast, the total bed stock is taken as fixed, and the managerial question is how access to that stock should be organized before demand is realized. In this sense, the α-boundary does not create redundant capacity; it determines how much capacity remains immediately accessible to wards and how much is opened to hospital-wide pooling.

Second, the paper develops a multi-ward framework that places full decentralization, bounded centralization, and full centralization on the same governance continuum. This framework is related to, but distinct from, clustered allocation approaches in which the sharing structure is usually specified through predefined service-line groups, overflow links, or specialty clusters. Here, the boundary of sharing is itself the object of design. The two-ward case is used for analytical clarity, and multi-ward computational extensions then examine whether the same mechanism remains visible beyond the simplest setting.

Third, the paper identifies conditions under which bounded centralization can become managerially attractive without dismissing the value of pooling. The analysis establishes a frictionless benchmark in which full centralization weakly dominates, and then examines how coordination friction, stress-state reassignment limits, and protected local access can jointly shift the preferred governance boundary. The same logic also reflects a practical feature of bed-management reform: a bed that is visible to a central team still has to be made available through information, authorization, placement judgment, and ward-level execution.

The remainder of the paper proceeds from literature positioning to model construction, analytical mechanisms, and computational evidence. Section 2 reviews the related literature; Sections 3 and 4 introduce the governance setting and model formulation; Sections 5 and 6 present the analytical and computational results; and Sections 7 and 8 discuss implications and conclude.

2. Literature Review and Research Positioning

2.1. Hospital Bed Allocation and Inpatient Capacity Management

For the boundary question studied in this paper, the bed-management literature is useful not primarily because it counts beds, but because it shows how difficult it is to make beds operationally available at the right time and in the right place. Admission, discharge, and transfer decisions shape patient flow across emergency departments, intensive care units, and inpatient wards [10]. Broader work on hospital capacity planning extends the same issue to a longer planning horizon, linking bed management to capacity planning, daily occupancy control, and hospitals' ability to respond to fluctuating demand [11]. Together, these studies provide the operational background for this paper: congestion and delayed admissions arise not only from the number of beds, but also from how capacity is accessed and coordinated across units.

A related stream of this literature has developed tools for managing beds under uncertain demand and occupancy. Data-driven admission control models use patient information and occupancy states to support placement decisions when reusable beds are limited [12]. In simulation studies, bed management is usually set within changing operating states, not treated as a fixed bed count. For short-term decisions, model inputs come on the one hand from real-time system status and length of stay information, and on the other hand from early warning signals [13]. Other studies examine resource balancing under severe pressure. Ring fencing during the COVID-19 pandemic is a typical case [14].

Taken together, these studies partly emphasize how changing occupancy conditions may affect timely response, demand anticipation, and patient flow control.

What remains less explicit in this literature is the governance structure within which allocation and response decisions are made. Much of the literature takes the capacity structure as given: beds are scarce resources to be allocated, forecasted, protected, or scheduled. A prior design choice concerns how much capacity remains under immediate ward-level access and how much is opened to hospital-wide pooling. The same physical bed stock can therefore have different operational consequences under full decentralization, bounded centralization, or full centralization. This is where the present study shifts from bed allocation alone to the governance boundary between local protection and centralized pooling.

This boundary is not only a modeling distinction; it is also visible in how hospitals manage beds in practice. Central teams, bed boards, and patient flow systems may improve visibility, but they do not by themselves resolve the operational questions of whether a patient can be placed in a particular ward, whether the receiving team has the required capability, and whether the transfer can be completed in time. The next subsection therefore turns to empirical and implementation evidence on centralized bed management before returning to pooling theory.

2.2. Empirical Evidence and Practical Applications of Centralized Bed Management

Studies of hospital command centers and patient flow coordination suggest that centralization is usually built through operational arrangements rather than through a bed-assignment rule alone. In the command center design described by Kane et al., real-time displays, predictive analytics, standard work, and rules-based protocols are combined as part of a systems-engineering approach to hospital flow [15]. Qualitative evidence from leading healthcare providers points in a similar direction, emphasizing central coordination, digital support, transfer structures, and discharge planning as practical supports for hospital-wide flow [16]. For this study, these findings give "centralized bed allocation" a more concrete operational meaning: central counting may improve visibility, but beds can be used across units only when information, authority, and coordination routines are organized to support that use.

In daily hospital operations, counts of available and occupied beds are often the easiest indicators to observe. What can be missed is that a patient placed in an available bed may still be outside the care setting and team support best suited to the case. Off-service placement can ease an immediate capacity shortage, but its clinical fit is not as visible as the bed number itself. A poor match may also carry costs, since Song et al. associate this form of hospital capacity pooling with longer remaining length of stay and higher readmission risk [17]. A similar gap appears in timing. Braaksma et al. show that early bed assignment may mark a bed as allocated before the patient is ready to move [18]. These findings do not reject pooling. They make the pooling claim more conditional. In bed management, bed numbers should not be the only consideration. Ward context, timing, and workable clinical arrangements also matter for decision makers.

Command center evidence points in the same direction. Central coordination may improve particular handoffs in patient flow, although current evaluations do not show that a command center automatically improves flow, data quality, or safety. Mebrahtu et al. report mixed NHS evidence, with some transition times improving and no consistent overall gain in patient flow indicators or data quality [19]. Another interrupted time-series study found limited evidence of patient-safety improvement after implementation [20]. Johnson et al. also report data-quality problems and tension between central coordination and local departmental practice in an artificial-intelligence command center study [21].

Centralization in practice depends on the hospital's ability to build reliable coordinating routines around the formal structure.

Taken together, these studies suggest that pool size is only part of the design problem. A larger common pool would be the obvious target if centralized bed management were only a question of how many beds to share. Hospital operations make the choice harder. A shared bed must be seen in the system, cleared under placement rules, accepted by a suitable ward, and worked into transfer and care routines that can safely handle the patient. The parameter α captures where this access boundary is drawn. It shows how much capacity is opened to central coordination, but the value of that capacity depends on whether the hospital can turn it into usable placements under its own operating conditions.

2.3. Resource Pooling and Centralization

When execution costs are set aside, pooling remains a strong starting point. Demand shocks are uneven across wards, and a common service base can absorb some mismatch that dedicated capacity would leave unresolved. Queueing studies give the formal version of this argument by comparing pooled and dedicated service structures. The same literature also cautions against treating the gain as automatic, since heterogeneity and operating conditions can change its size [22]. Studies of differentiated services add another qualification. Even after resources are pooled, allocation rules still shape who receives access and where the burden of sharing falls [23]. In this paper, these results are used for a limited purpose. They establish the clean case in which full centralization is the natural benchmark, but they do not close the hospital design question.

Hospital placement adds a further step between a nominal match and an actual admission or transfer. A bed may be listed, but the patient must be ready, the placement decision cleared, and the receiving ward able to take the case in time. Evidence on geographic virtual pooling speaks to this point. Wider matching can shorten waiting, yet it may also require movement across organizational or spatial boundaries [24]. When transfer timing, clinical fit, or ward-level approval makes reassignment less reliable, the issue is no longer just whether pooling helps in principle. It is how far pooling should extend before some capacity is better kept closer to local use.

2.4. Costly Coordination and Organizational Boundaries

A shared bed does not relieve pressure until the placement work has been done. Bed managers first have to find capacity that is usable in practice, confirm that the receiving ward can take the case, apply the relevant placement authority, and work the move into existing ward routines. On quieter days, these tasks may disappear into ordinary bed management. When several units are asking for capacity at the same time, they become part of the allocation constraint. The model locates coordination cost at this point in the process. The cost is attached to cross-ward reassignment work, rather than to central pooling as a formal arrangement.

Evidence from operations and organization studies supports this placement of the cost. Queue configuration can affect provider behavior because staff respond to queue ownership and to visible queue length. In healthcare delivery experiments, dedicated queues have sometimes processed patients faster than pooled queues without reducing quality when these behavioral channels are active [25]. Centralization studies point to a similar condition. More centralized structures improve adaptation only when communication patterns can carry the necessary information, including feedback from local units [26]. For bed governance, the implication is fairly direct. The burden is not the existence of a shared pool itself. It is the work required to turn a visible bed into an approved and timely

reassignment. The model therefore measures coordination cost through the amount of reassignment performed, instead of applying a fixed penalty to the governance regime.

2.5. Operational Resilience and Local Buffers

Local reserve is hardest to justify when pressure is low, or when ward-level shocks do not arrive together. In those conditions, a central pool can look more efficient. Beds held near one ward can be shifted toward another ward with a shortage, improving the use of capacity across the hospital. The argument changes when several wards are short at the same time. A ward with a small accessible reserve may be able to admit or hold patients through its own routines while the central team is still locating a bed, confirming that it can be used, and arranging the reassignment. What matters is not simply that extra capacity exists. The point is that the organization retains a usable option after disruption has already begun. This links the model to the resilience literature, where slack is valuable because it helps organizations absorb pressure, preserve options, and recover from stress [27].

Hospital capacity studies place the same idea in a more concrete setting. Work on ICU infrastructure treats slack as one element of resilience management after the pandemic stress test [28]. Surge preparedness reviews describe emergency capacity as a bundle of staff, supplies, space, beds, systems, and the ability to expand or protect usable capacity as conditions worsen [29]. The meaning of local protection in this paper is narrower. It refers to capacity that remains reachable by the ward when shared reassignment slows or becomes unreliable. Full pooling may improve average matching, yet it can leave a ward with fewer local options when pressures become synchronized. Recent calls for models linking surge, demand, coordination, and system response are therefore relevant to this design choice [30]. Local protection is used here to capture the idea that exposure to severe-shortage states depends partly on where usable capacity is held.

2.6. Research Gap and Positioning

Off-service placement shows what pooling means once the decision is made inside a hospital rather than in an abstract capacity model. A patient can be moved to a ward with available space, and the immediate bed shortage may ease. At the same time, the patient is now cared for by a different team, under different ward routines, with extra coordination around handover, monitoring, and discharge planning. Empirical studies associate this form of pooling with longer remaining length of stay and higher readmission risk [17]. Related evidence also finds spillover effects for patients who stay on the original service, which indicates that the work consequences extend beyond the patients who are physically reassigned [31]. For this reason, pooling in hospitals should be read as a change in the organization of care, not only as a way to offset uneven demand.

Clustered overflow research has already developed intermediate designs between strict dedication and full pooling. These designs allow overflow capacity to be shared within specialty subsets [32]. Later studies optimize department clustering, dedicated bed allocation, and shared overflow beds under multiple operating costs [33]. Recent work on inpatient service reconfiguration also finds clustered overflow structures attractive under bed pressure [34]. In these models, however, the sharing structure is usually chosen in advance. Departments are assigned to clusters, overflow routes are specified, and dedicated or shared capacity is placed inside that architecture. Redundant bed and buffer capacity models address a neighboring question by deciding how much slack or backup capacity to hold against demand uncertainty.

The framework built around α takes a different design object. It fixes the total bed stock and lets α determine the split between ward access and hospital-wide coordination. As a result, the same model can examine the frictionless pooling benchmark, realized

coordination burden, stress-state limits on pooled allocation, the value of local protection, and robustness in multi-ward settings. This paper asks how far centralization should extend when pooled capacity is valuable, its use requires coordination, and local access may help during pressure states. Section 3 formalizes this setting through the bed stock, ward demand, and the centralization boundary.

3. Problem Setting and Governance Regimes

3.1. Hospital Bed Governance in an n-Ward System

We consider a hospital bed-management setting with n inpatient wards, indexed by $i \in I = \{1, \ldots, n\}$. Ward i has nominal bed capacity B_i, and the hospital has a fixed total bed stock

$$B = \sum_{i=1}^{n} B_i.$$

The focus is not short-term bed expansion, capacity investment, or the creation of redundant beds. Accordingly, we take the hospital bed stock as fixed and examine how access to that stock is governed across wards. The central question is therefore how much capacity should remain under immediate ward-level access and how much should be made available for hospital-wide coordination.

During a planning period, ward i faces random bed demand D_i. We use D_i to denote net bed demand, rather than new arrivals alone. This demand reflects admission pressure, continuing occupancy, and expected bed releases during the planning period. Patient-level length-of-stay (LOS) dynamics are therefore not modeled explicitly. LOS-related turnover is instead absorbed into the distribution of D_i. This treatment does not imply that LOS or discharge processes are unimportant. It allows the analysis to focus on how a realized net demand state is handled across ward boundaries under different governance arrangements. In this formulation, D_i is best read as the net pressure placed on ward i's bed capacity during the planning period. It is not intended to be a simple end-of-day census, although it may be summarized from data observed over a day or a shift. For empirical implementation, such a variable could be constructed from occupied beds at the beginning of the period, expected or realized admission requests, expected discharges, and transfer-related inflows and outflows. Depending on the available data, D_i can therefore be implemented as an admission-equivalent net demand or as a synthetic net-flow variable. This interpretation keeps LOS and discharge dynamics in the background as sources of the realized demand state, while allowing the model to focus on how that state is absorbed through local capacity and centrally coordinated pooled capacity. Let

$$D = (D_1, \ldots, D_n)$$

be the ward-level demand vector. We write

$$D \sim F(\mu, \Sigma), \tag{1}$$

where $\mu = (\mu_1, \ldots, \mu_n)$ is the mean demand vector and Σ is the covariance matrix. The covariance structure allows demand shocks to be correlated across wards. When ward demands are weakly correlated, spare capacity in one ward is more likely to offset pressure in another. When demands move together, several wards may compete for capacity at the same time, weakening the diversification benefit from pooling.

We describe the governance structure through a centralization degree $\alpha \in [0, 1]$. In managerial terms, α is a planning-level boundary between immediate ward-level access and hospital-wide coordination. When α is low, most capacity remains locally accessible to each ward. When α is high, more capacity is opened to hospital-wide reassignment.

Intermediate values represent designs that combine local protection with central pooling. Formally, α represents the share of each ward's nominal capacity placed in a hospital-wide shared pool. Given α, ward i retains a protected local reserve $r_i(\alpha)$, and the hospital-wide shared pool has capacity $S(\alpha)$

$$r_i(\alpha) = (1-\alpha)B_i, \quad S(\alpha) = \alpha B. \tag{2}$$

A larger α expands the shared pool but reduces protected ward-level capacity. Thus, α represents the governance boundary between local protection and centralized pooling. This n-ward formulation provides the general hospital setting for the paper. Later analytical sections use a two-ward canonical case to make the governance mechanism transparent, while the computational experiments return to multi-ward settings.

Table 1 summarizes the notation and model primitives for the governance framework. The objects especially relevant to the subsequent discussion are the centralization degree α, the protected local reserve $r_i(\alpha)$, the shared-pool capacity $S(\alpha)$, the realized allocation $x_i(D, \alpha)$, and the performance measure $\Pi(\alpha)$.

Table 1. Notation and model primitives for the governance framework.

Category	Symbol	Meaning
System	$i \in I = \{1, \ldots, n\}$	Ward index and ward set
Capacity	B_i	Nominal bed capacity of ward i
	$B = \sum_{i=1}^{n} B_i$	Total hospital bed stock
Demand	D_i	Net bed demand faced by ward i during the planning period
	$D = (D_1, \ldots, D_n)$	Ward-level demand vector
	μ	Mean vector of ward-level demand
	Σ	Covariance matrix of ward-level demand
	ρ_{ij}	Demand correlation between wards i and j
Governance	$\alpha \in [0, 1]$	Degree of centralization; governance boundary between local protection and shared pooling
	$r_i(\alpha) = (1-\alpha)B_i$	Protected local reserve of ward i
	$S(\alpha) = \alpha B$	Hospital-wide shared-pool capacity
Allocation	$x_i(D, \alpha)$	Effective capacity assigned to ward i after demand is realized
Outcomes	$A_i(D, \alpha)$	Admitted demand in ward i
	$U_i(D, \alpha)$	Idle capacity in ward i
	$L_i(D, \alpha)$	Unmet demand in ward i
Costs	p_i	Unit shortage penalty for ward i
	h_i	Unit idle-capacity cost for ward i
Coordination	c	Weight on coordination friction
	$Q(D, \alpha)$	Realized coordination burden under allocation $x(D, \alpha)$
	$Q_m(D, \alpha)$	Alternative coordination-friction specification, with $m = 1, \ldots, 4$

Table 1. *Cont.*

Category	Symbol	Meaning
Tail risk	$L_{sys}(D,\alpha) = \sum_i L_i(D,\alpha)$	Total system shortage
	η	Weight placed on tail-risk exposure
	γ	Convexity parameter for severe-shortage states
Performance	$C_{avg}(\alpha)$	Expected operating loss
	$C_{tail}(\alpha)$	Tail-risk exposure
	$\Pi(\alpha)$	Overall governance performance

3.2. Decision Timing and Information Structure

The decision sequence separates governance design from operational response. Before ward-level demand is realized, the hospital sets the centralization degree $\alpha \in [0,1]$. This choice fixes the governance boundary for the planning period and determines how much capacity remains under immediate ward-level access and how much is opened to hospital-wide coordination.

After demand is realized, the hospital observes the ward-level demand vector

$$D = (D1, \ldots, Dn)$$

and chooses an effective capacity allocation within the limits set by the chosen governance boundary α. We denote the resulting allocation by

$$x(D,\alpha) = (x1(D,\alpha), \ldots, xn(D,\alpha)),$$

where $xi(D,\alpha)$ is the effective bed capacity available to ward i after the demand state is observed. The local reserve remains available through immediate ward-level access, while capacity in the shared pool must be assigned through the central coordination process.

This timing sequence keeps governance design separate from operational response. The parameter α determines what allocations are feasible once demand is observed; it does not prescribe a fixed allocation for every demand realization. In this sense, α defines the scope of feasible coordination rather than a real-time dispatch rule. After the allocation $x(D,\ \alpha)$ is chosen, ward-level outcomes such as admitted demand, idle capacity, and unmet demand are determined. Coordination burden and tail-risk exposure are also evaluated at this stage. Section 4 formalizes the feasible allocation set, the outcome definitions, and the performance measures used in the analysis.

3.3. Governance Regimes Under the α-Framework

The centralization degree α places the three bed-governance arrangements in a common framework. As Figure 1 shows, full decentralization, bounded centralization, and full centralization are not treated as unrelated regimes. They occupy different positions on the same continuum between immediate ward-level access and hospital-wide pooling.

When $\alpha = 0$, all nominal bed capacity remains under immediate ward-level access. When $\alpha = 1$, the entire bed stock is opened to hospital-wide pooling. Values between these two endpoints describe bounded centralization, where each ward retains part of its nominal capacity as a protected local reserve and the remaining capacity is made available for central coordination.

When $\alpha = 0$, the system corresponds to full decentralization (FD). By Equation (2), each ward retains its full nominal capacity, $r_i(0) = B_i$, and no hospital-wide shared pool

is formed, $S(0) = 0$. FD gives wards the strongest immediate local access, but it does not allow hospital-wide rebalancing after demand is realized.

When $0 < \alpha < 1$, the system corresponds to bounded centralization (BC). By Equation (2), each ward retains a positive protected reserve, $r_i(\alpha) = (1 - \alpha)B_i > 0$, while the shared pool also has positive capacity, $S(\alpha) = \alpha B > 0$. BC therefore combines central coordination with protected local access. It keeps hospital-wide pooling in place, but it does not place all capacity under the shared pool.

When $\alpha = 1$, the system corresponds to full centralization (FC). By Equation (2), no ward-specific reserve is protected, $r_i(1) = 0$, and the full bed stock is placed in the shared pool, $S(1) = B$. FC provides the widest scope for hospital-wide pooling and serves as the natural benchmark when sharing is frictionless.

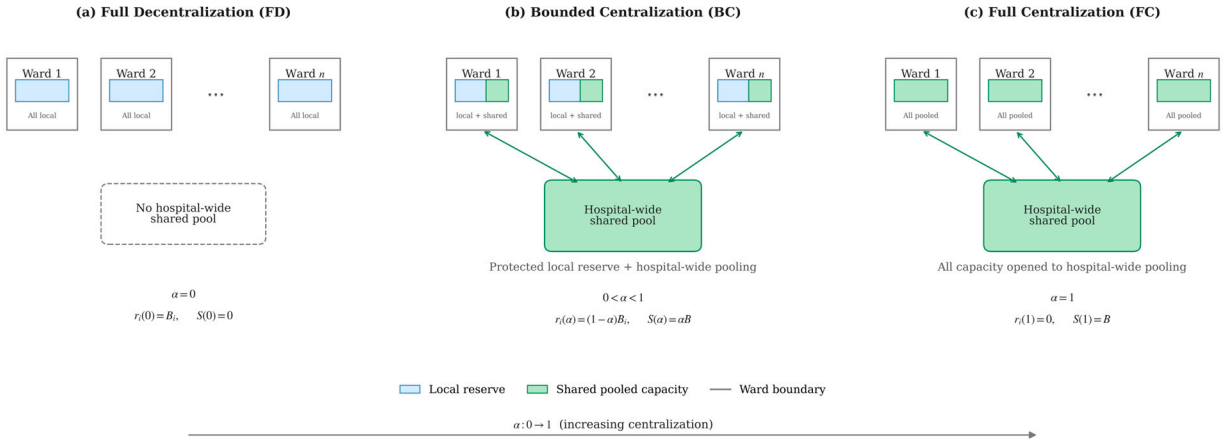


Figure 1. Hospital bed governance under (**a**) full decentralization, (**b**) bounded centralization, and (**c**) full centralization.

With this nesting, FD, BC, and FC are no longer separate regime labels to be compared one by one. They become different ways of drawing the governance boundary between capacity that is locally protected and capacity that is centrally pooled. The model leaves that boundary open. It does not build in a preferred regime; instead, the selected boundary is allowed to vary with demand conditions, coordination burden, and the value of local access under pressure.

3.4. The Two-Ward Canonical Case

Although the general formulation allows for n wards, the analytical discussion starts with two. The reason is not that hospital bed systems are this simple. More wards would introduce additional demand correlations, allocation combinations, and boundary cases before the basic governance tension is visible. With two wards, the first trade-off remains visible. The case shows how local access, pooled capacity, and the coordination work needed for reassignment can start to move in different directions, even before the setting becomes more complex.

Pooling across wards needs more than a single unit. In a case with two wards, the basic tension already appears when one ward comes under pressure while another still holds usable capacity. This structure is sufficient to show the pooling opportunity, but also sufficient to raise the governance question: how much capacity should remain locally accessible? The same setting can also represent synchronized demand shocks, protected local reserves, and the coordination burden created when capacity must be reassigned through a central process.

The two-ward case should therefore be read as a device for clarifying the mechanism, not as a claim that hospital bed systems are inherently two-ward. Its role is to keep the trade-

off visible before the analysis moves back to richer multi-ward settings. The computational experiments later examine systems with richer ward structures, heterogeneous capacities, and correlated demand patterns to examine whether the same boundary mechanism remains visible beyond the canonical case.

4. Model Formulation

4.1. Feasible Allocation Under a Governance Boundary

Section 3 introduced α as the governance boundary between protected ward capacity and centrally pooled capacity. We now translate that boundary into the set of effective capacity assignments available after ward-level demand is observed. The formulation is stated for the general n-ward system, although the analytical discussion later uses a two-ward canonical case for clarity. This keeps the regime comparison internally consistent and allows the computational experiments to return to larger ward systems.

Operationally, the feasible allocation set describes the assignments available to the bed-management team after demand is observed, given the governance boundary chosen in advance.

Protected local reserves remain immediately available to their own wards, whereas only the pooled portion can be redirected through hospital-wide coordination. In this formulation, $\mathcal{X}(\alpha)$ denotes the effective capacity assignments that the chosen governance design permits after demand is observed. It should not be interpreted as physical bed movement. The clinical usability of these assignments varies across scenarios. The frictionless benchmark assumes that the permitted assignments can be fully converted into usable capacity, while the later costly and stress-state settings allow that conversion to be limited by placement judgment, staffing, specialty requirements, and authorization. Let

$$x(D, \alpha) = (x_1(D, \alpha), \ldots, x_n(D, \alpha))$$

denote the effective capacity assigned across wards in demand state D. Using the reserve and shared-pool definitions in Equation (2), the feasible allocation set is

$$\mathcal{X}(\alpha) = \left\{ x \in \mathbb{R}^n_+ : x_i \geq r_i(\alpha), \sum_{i=1}^{n} [x_i - r_i(\alpha)] \leq S(\alpha) \right\}. \tag{3}$$

The lower-bound constraint $x_i \geq r_i(\alpha)$ reflects the protected reserve held by ward i. This reserve is a local claim on existing capacity, not a requirement that the ward keep beds idle. The shared-pool constraint limits how much capacity can be assigned above the protected reserves. Together, the two constraints translate the governance boundary α into an allocation set.

The endpoints of $\mathcal{X}(\alpha)$ recover the two polar regimes. When $\alpha = 0$, all nominal capacity remains locally protected. Since $r_i(0) = B_i$ and $S(0) = 0$,

$$\mathcal{X}(0) = \{x \in \mathbb{R}^n_+ : x_i = B_i, i = 1, \ldots, n\}. \tag{4}$$

This is the allocation set under full decentralization. When $\alpha = 1$, no ward-specific reserve is protected. Since $r_i(1) = 0$ and $S(1) = B$,

$$\mathcal{X}(1) = \left\{ x \in \mathbb{R}^n_+ : \sum_{i=1}^{n} x_i \leq B \right\}. \tag{5}$$

This is the allocation set under full centralization.

For intermediate values $0 < \alpha < 1$, the same feasible-set formulation describes bounded centralization: part of each ward's nominal capacity is protected, while the rest is pooled. Bounded centralization is therefore not an additional allocation rule, but an interior governance boundary within the same α-framework. Because protected reserves restrict reassignment, full centralization weakly expands the feasible set:

$$\mathcal{X}(\alpha) \subseteq \mathcal{X}(1), \forall \alpha \in [0,1]. \tag{6}$$

This inclusion anchors the frictionless pooling benchmark in Section 5. It is a feasibility statement made before coordination friction and tail-risk considerations are introduced. It should therefore be read as the strongest case for full centralization, not as a conclusion that full centralization is always preferred.

4.2. Ward-Level Outcomes and Realized Operating Loss

We next translate each feasible capacity assignment into ward-level outcomes. For a realized demand vector D and a feasible allocation $x \in \mathcal{X}(\alpha)$, the admitted demand, idle capacity, and unmet demand of ward i are

$$A_i(x,D) = \min\{D_i, x_i\}, \;\; U_i(x,D) = (x_i - D_i)^+, \;\; L_i(x,D) = (D_i - x_i)^+, \tag{7}$$

where $(z)^+ = max\{z, 0\}$. These definitions decompose the immediate consequences of an allocation into served demand, unused ward-assigned capacity, and demand that cannot be accommodated. Because the feasible set does not require the shared pool to be fully assigned in every demand state, we also record unused shared-pool capacity:

$$P(x,\alpha) = S(\alpha) - \sum_{i=1}^{n} [x_i - r_i(\alpha)].$$

By feasibility, $P(x,\alpha) \geq 0$. This term represents pooled capacity that remains operationally unassigned in the realized demand state. It is distinct from ward-level idle capacity $U_i(x,D)$, which measures capacity assigned to ward i but not used by demand in that ward. The realized operating loss under allocation x in demand state D is

$$\ell(x,D;\alpha) = \sum_{i=1}^{n} p_i L_i(x,D) + \sum_{i=1}^{n} h_i U_i(x,D) + \bar{h} P(x,\alpha) + cQ(x,D,\alpha). \tag{8}$$

where $\bar{h} = n^{-1}\sum_{i=1}^{n} h_i$. The first term captures shortage, the second term captures ward-level idle capacity, the third term captures pooled capacity that remains available but unassigned, and the fourth term captures coordination burden. We leave $Q(x,D,\alpha)$ general for the moment and define it in Section 4.3.

For a fixed governance boundary α, the second-stage allocation is defined by

$$x(D,\alpha) \in \arg\min_{x \in \mathcal{X}(\alpha)} \ell(x,D;\alpha). \tag{9}$$

This formulation makes the response to realized demand explicit. It represents the best feasible response under a given governance boundary, not a literal claim that managers solve a full optimization problem in real time for every demand realization. In the computational experiments, this response is operationalized by the scenario-level allocation procedure described in Algorithm A1, which uses the same feasible-boundary structure and cost components as the analytical model. The algorithm is therefore the numerical implementation of this best-response logic, not a separate modeling assumption. Once

$x(D, \alpha)$ is obtained from Equation (9), we write $A_i(D, \alpha)$, $U_i(D, \alpha)$, $L_i(D, \alpha)$, and $P(D, \alpha)$ as shorthand for the corresponding functions evaluated at $x(D, \alpha)$.

4.3. Coordination-Friction Specifications

The realized loss in Equation (8) includes $Q(x, D, \alpha)$, which represents the coordination burden that arises when capacity is used across ward boundaries. In practice, this burden may arise from communication, approval, placement adjustment, specialty compatibility, or delay. We do not model these processes separately. Instead, Q is a reduced-form measure of the coordination intensity implied by the realized allocation.

Using this reduced-form term does not imply that coordination burden arises from a single operational source. In a hospital bed-management setting, the difficulty of using pooled capacity may take several forms. It may increase when the final bed assignment departs substantially from the ward's usual capacity base. It may also increase when more shared-pool capacity must be activated. In other situations, the main burden may come from the number of wards involved in the reassignment process, or because some wards are clinically and organizationally harder to reassign to or from than others. We therefore use Q_1 as the main specification and retain Q_2–Q_4 as alternative readings of the same coordination problem.

The baseline specification measures coordination friction through realized reallocation intensity, rather than through the formal governance label:

$$Q_1(D, \alpha) = \frac{1}{2} \sum_{i=1}^{n} |x_i(D, \alpha) - B_i|. \tag{10}$$

Q_1 measures the total deviation of the realized allocation from the nominal capacity vector $B = (B_1, \ldots, B_n)$. The factor $1/2$ avoids counting the same reallocation twice. Under full decentralization, $x_i(D, 0) = B_i$ and therefore $Q_1(D, 0) = 0$. Under full centralization, Q_1 depends on the allocation chosen after demand is observed. FC therefore incurs coordination burden only when the realized allocation involves cross-ward adjustment, not simply because the beds are centrally governed.

Unless stated otherwise, Q_1 in Equation (10) is used as the baseline measure. This choice charges all regimes for realized departures from nominal ward capacity, not for their formal labels. To check that the analysis is not tied too closely to Q_1, we also define three additional specifications. Let

$$y_i(D, \alpha) = x_i(D, \alpha) - r_i(\alpha)$$

denote the amount of capacity assigned to ward i from the shared pool. By the definition of $\mathcal{X}(\alpha)$, $y_i(D, \alpha) \geq 0$ and $\sum_i y_i(D, \alpha) \leq S(\alpha)$.

If coordination burden is closer to the amount of pooled capacity actually used than to the distance from nominal ward capacity, the second specification is:

$$Q_2(D, \alpha) = \sum_{i=1}^{n} y_i(D, \alpha). \tag{11}$$

A different interpretation is that coordination becomes difficult when more wards have to be brought into the process. Even a moderate amount of shared capacity may require additional negotiation and handoff work when it is spread across several wards. This gives:

$$Q_3(D, \alpha) = \sum_{i=1}^{n} \mathbf{1}\{y_i(D, \alpha) > 0\}. \tag{12}$$

Finally, reassignment difficulty may vary across wards. For example, intensive care, surgical, infection control, or other highly specialized wards may impose staffing, clinical fit, and authorization requirements that differ from those of a general medical ward. To allow for this heterogeneity, we use:

$$Q_4(D,\alpha) = \frac{1}{2}\sum_{i=1}^{n} \omega_i |x_i(D,\alpha) - B_i|. \tag{13}$$

where $\omega_i \geq 0$ is a ward-specific reassignment weight. Table 2 summarizes the interpretation of the four coordination-friction specifications.

Table 2. Interpretation of coordination-friction specifications.

Measure	What the Measure Emphasizes	Interpretation in Bed Management
Q_1	Deviation from the ward's nominal capacity base	This is the main specification. It treats coordination work as increasing when the realized assignment moves capacity away from its usual ward base. This is useful because it charges FD, BC, and FC according to the realized reassignment, not according to the name of the regime.
Q_2	Amount of shared-pool capacity used	This specification reads coordination burden as the workload generated by activating the shared pool. It is relevant when the central bed-management team has to identify, release, and assign a larger amount of pooled capacity.
Q_3	Number of wards receiving pooled capacity	This specification focuses on the spread of coordination. The burden may increase when more wards are involved, even if the total amount reassigned is not large, because more communication, approval, and handoff decisions are needed.
Q_4	Ward-specific difficulty of reassignment	This specification allows some wards to be more difficult to reassign to or from. The weight may be higher for wards with stricter staffing, specialty fit, infection control, or clinical governance requirements.

These specifications should be read as simplified ways of representing coordination burden, not as calibrated estimates of the true cost of hospital bed reassignment. More modestly, they allow us to ask whether the model's boundary effect depends on one particular definition of coordination friction. In an empirical application, the relevant form of Q would need to be informed by hospital data, such as bed-assignment records, transfer timestamps, approval delays, staffing availability, and specialty compatibility rules.

4.4. Tail-Risk Exposure and Governance Performance

Once the allocation $x(D,\alpha)$ has been determined, we evaluate a governance boundary by the operating loss it generates across demand states. For brevity, write $Q(D,\alpha)$ for $Q(x(D,\alpha),D,\alpha)$. We define

$$C_{avg}(\alpha) = \mathbb{E}\left[\sum_{i=1}^{n} p_i L_i(D,\alpha) + \sum_{i=1}^{n} h_i U_i(D,\alpha) + \overline{h} P(D,\alpha) + cQ(D,\alpha)\right]. \tag{14}$$

This measure aggregates the average operating-loss components under a boundary: unmet demand, unused capacity, unused shared-pool capacity, and coordination effort.

Average loss alone does not distinguish mild shortages from severe system-wide shortages. This distinction matters for bed management because simultaneous shortages

across wards place greater pressure on hospital-wide coordination. We therefore define the total system shortage as

$$L_{\text{sys}}(D, \alpha) = \sum_{i=1}^{n} L_i(D, \alpha). \tag{15}$$

The tail-loss component is defined as

$$C_{\text{tail}}(\alpha) = \mathbb{E}\Big[\big(L_{\text{sys}}(D, \alpha)\big)^{\gamma}\Big], \gamma > 1. \tag{16}$$

The exponent γ gives disproportionate weight to larger shortage realizations. We use $C_{\text{tail}}(\alpha)$ as a compact proxy for exposure to adverse shortage states, not as a full measure of operational resilience. For the mechanism studied here, this proxy is sufficient because the model asks whether protected local capacity can reduce exposure to severe-shortage states. It does not attempt to measure broader organizational dimensions of resilience, such as recovery speed, staffing flexibility, or emergency expansion capacity.

The resulting governance performance measure is

$$\Pi(\alpha) = -C_{\text{avg}}(\alpha) - \eta C_{\text{tail}}(\alpha), \eta \geq 0. \tag{17}$$

Because C_{avg} and C_{tail} are loss terms, a larger value of $\Pi(\alpha)$ indicates better governance performance. The parameter η controls the weight placed on adverse shortage states. When $\eta = 0$, the criterion reduces to expected operating loss. When $\eta > 0$, severe system-wide shortages also enter the evaluation.

4.5. The Governance Boundary Problem

The preceding definitions allow us to state the boundary-design problem. The centralization degree α is a planning-level governance choice, not a real-time dispatching variable. Once the boundary is set, demand is realized, and bed capacity is allocated within the feasible set implied by that boundary. The boundary-design problem can be written as

$$\alpha^* \in \arg\max_{\alpha \in [0,1]} \Pi(\alpha). \tag{18}$$

Any selected optimizer α^* identifies where the governance boundary is drawn between locally accessible capacity and hospital-wide pooled capacity. If an optimizer lies at $\alpha = 0$, the preferred boundary corresponds to full decentralization. If an optimizer lies at $\alpha = 1$, the preferred boundary corresponds to full centralization. If an optimizer lies in the interior, $\alpha^* \in (0, 1)$, the selected design is bounded centralization: part of the ward-level capacity remains locally protected, while the rest is opened to hospital-wide pooling.

This formulation does not impose a preference for bounded centralization. The preferred boundary is determined by the demand process, shortage and idle-capacity weights, coordination friction, and the weight placed on adverse shortage states. The optimum may therefore occur at either endpoint or at an interior value. Equation (18) above defines the boundary problem in mathematical terms. Before turning to the analytical results, it is useful to make its managerial interpretation explicit. Section 4.6 first explains the main parameters in hospital-management terms. Section 4.7 then links the formal components of the model to clinical and operational bed-management processes. These two subsections provide the interpretation needed for the analytical comparison in Section 5.

4.6. Practical Interpretation of Key Parameters

The model uses a small set of parameters to represent several managerial features of hospital bed governance. Some of these parameters, such as the centralization degree α, describe how the bed stock is organized before demand is realized. Others, such as c and

δ, describe the operating conditions under which pooled capacity is actually used. The parameters η and γ describe how the evaluation weights severe-shortage states relative to average operating loss. Accordingly, these parameters should be read as reduced-form representations of governance design, coordination burden, and tail-risk concern. They are not calibrated estimates for a specific hospital in this study.

Table 3 summarizes the hospital-management interpretation of the main parameters.

Table 3. Practical interpretation of key governance parameters.

Parameter	Role in the Model	Practical Interpretation for Hospital Bed Management
α	Governance boundary between protected local reserve and hospital-wide pooling.	A higher α opens more ward capacity to central coordination; a lower α preserves more immediate ward-level access. It is a planning-level boundary, not a real-time dispatching rule.
c	Weight on realized coordination friction $Q(D, \alpha)$.	Captures how difficult, slow, or costly cross-ward reassignment is. The burden may come from authorization, patient-ward fit checks, transfer work, staffing limits, or receiving-ward coordination.
η	Weight placed on tail-risk exposure.	Indicates how much severe-shortage states matter beyond average operating loss. A higher η gives more importance to avoiding system-wide or clinically difficult shortages.
γ	Convexity of the severe-shortage penalty.	Determines how sharply the penalty increases when shortages become large or simultaneous across wards. A higher γ reflects disproportionate concern for escalation, ED boarding, or broader flow disruption.
δ	Stress-state pooled-usability loss.	Measures how much pooled capacity becomes less usable under high-pressure states. It represents loss of timely usability, not physical bed loss.

Read together, the parameters in Table 3 describe more than the size of a shared pool. The boundary α decides how much capacity is exposed to central coordination, but those beds still have to pass through placement work before they can help a ward in real time. The parameter c attaches weight to that work. The parameters η and γ make severe-shortage states matter in the objective, and δ represents the possibility that pooled capacity is harder to use when several wards are under stress. The model therefore treats pooling as valuable capacity only after it can be identified, coordinated, and turned into a clinically workable placement.

4.7. Clinical and Operational Mapping of the Model

The model has deliberately compressed many clinical details into a small number of terms. That compression is useful for analysis, but the terms still need to be read against the bed management work they stand for. Table 4 provides that link. It shows how demand states, pooled capacity, reassignment work, staffing limits, and stress related usability losses can be interpreted in hospital operations. The mapping is not meant to turn the model into a full clinical pathway description. Its purpose is to make clear which parts of bed management are being represented when the formal framework is used to study the boundary between central pooling and local protection.

Table 4. Clinical-operational mapping of model components and observable indicators.

Model Domain	Model Component	Operational Reading in Hospital Bed Management	Possible Empirical Traces
Demand	D_i	Net bed pressure on ward i during the planning period, combining admission pressure, continuing occupancy, and expected bed releases.	Beginning occupancy, admission requests, discharge records, transfer inflows and outflows.
Governance boundary	α	Planning-level degree to which ward capacity is made available for hospital-wide coordination rather than kept under immediate ward access.	Bed-pooling policies, central allocation rules, scope of beds managed by the bed-control team.
Local access	$r_i(\alpha)$	Capacity that ward i can use without first entering a central reassignment process.	Protected-bed rules, service-line reserves, ward-level bed-holding practices.
Shared capacity	$S(\alpha)$	Capacity visible to the central bed-management process and formally eligible for cross-ward assignment.	Bed-board entries, centrally managed bed lists, beds released to the hospital-wide pool.
Realized allocation	$x_i(D, \alpha)$	Final effective placement or capacity assignment to ward i after demand has been observed.	Final ward placement, off-service admissions, reassignment records.
Coordination burden	$Q(D, \alpha)$	Work needed to turn pooled capacity into usable capacity across wards.	Transfer delays, approval steps, handoffs, ward acceptance decisions.
Coordination intensity	c	Reduced-form weight on the difficulty, delay, or cost of cross-ward reassignment.	Staffing gaps, administrative bottlenecks, incomplete bed-status information, weak IT support.
Stress condition	Stress states	High-pressure demand states in which several wards compete for usable capacity at the same time.	High occupancy, ED boarding, admission waiting peaks, simultaneous ward shortages.
Pooled-capacity usability	λ_s, δ	Loss of effective pooled-capacity usability when the system is under stress.	Unstaffed beds, specialty mismatch, delayed placement, rejected placement requests.
Tail-risk concern	η, γ	Additional concern for severe or simultaneous shortages beyond average operating loss.	Escalation events, severe boarding, multi-ward shortage episodes, urgent capacity alerts.

Table 4 is intended to keep a simple operational distinction visible. A pooled bed does not become usable merely because it appears in the formal pool. Staff must be available, the patient's service needs must fit the receiving ward, placement approval must be secured, and the transfer still has to be carried out. When any of these steps slows, the bed may remain on the central list while the ward facing shortage receives little immediate relief. In the frictionless benchmark, capacity kept within ward access reduces the shared pool and can therefore look inefficient. Under pressure, the same capacity may work differently. Local routines can make it easier to act quickly, limit placement delay, or prevent admission saturation from worsening. The model does not follow the full clinical pathway behind each placement. The mapping instead clarifies how these operating frictions shape the governance choice between centralized pooling and local protection.

5. Analytical Results and Mechanism Development

5.1. Frictionless Pooling Benchmark

We first consider an idealized pooling benchmark in which the shared bed stock can be used without operational loss. This benchmark gives full centralization its strongest case: beds are perfectly substitutable across wards, reassignment is immediate after demand is observed, and centralized allocation creates no delay, mismatch, or coordination burden. Because the benchmark is intentionally strong, the standard pooling logic should apply: if all beds can be allocated centrally, the hospital has a larger feasible allocation set than it would have when part of the bed stock remains locally protected.

The benchmark should therefore be read narrowly. It does not assume that all hospital beds are interchangeable in routine clinical practice. Rather, it describes an upper-bound case in which the beds considered for pooling are sufficiently standardized, clinically compatible, and supported by placement rules that allow demand to be absorbed across ward boundaries. Such an approximation may be reasonable for some general medical beds or low-acuity overflow arrangements, but it is far less appropriate for beds tied to specialized equipment, infection-control requirements, dedicated nursing skills, or specialty-specific clinical responsibility, such as ICU, isolation, postoperative specialty, pediatric, maternity, or psychiatric beds. Beginning with this strong case gives full centralization its clearest possible advantage. The later introduction of coordination friction and stress-state pooled-allocation imperfection then examines how the conclusion changes once the conversion from formal pooled capacity to usable ward-level capacity becomes less immediate.

Formally, consider a frictionless pooling environment. The central planner observes the realized demand vector D before making the second-stage allocation decision. Beds are fully substitutable across wards, and assigning capacity away from its nominal ward does not create clinical mismatch or service-quality loss. Reassignment is immediate and costless, so coordination friction is absent, $c = 0$. The realized loss therefore contains only shortage and idle-bed costs:

$$\ell_0(x, D) = \sum_{i=1}^{n} p_i L_i(x, D) + \sum_{i=1}^{n} h_i U_i(x, D), \tag{19}$$

where $L_i(x, D)$ and $U_i(x, D)$ denote, respectively, unmet demand and idle capacity for ward i under allocation x. For a given centralization degree α, define the realized benchmark loss as

$$\ell_0^{\alpha}(D) = \min_{x \in \mathcal{X}(\alpha)} \ell_0(x, D).$$

Let $\Pi_0(\alpha) = -\mathbb{E}\left[\ell_0^{\alpha}(D)\right]$ denote the corresponding benchmark performance under centralization degree α. The benchmark suppresses coordination and tail-risk terms to isolate the classical pooling effect. Let $\mathcal{X}^{FC}$ denote the feasible allocation set under full centralization, and let $\mathcal{X}(\alpha)$ denote the feasible allocation set under centralization degree $\alpha \in [0, 1]$.

Theorem 1. *Frictionless FC dominance.*

Under the frictionless pooling benchmark,

$$\Pi_0(1) \geq \Pi_0(\alpha), \forall \alpha \in [0, 1]. \tag{20}$$

That is, full centralization weakly dominates any bounded-centralization regime under the benchmark loss ℓ_0.

The proof follows from feasible-set inclusion. Under the formulation in Section 4, bounded centralization restricts how the fixed bed stock can be assigned; it does not create allocations that FC could not implement. For any $\alpha \in [0,1]$, every allocation feasible under $\mathcal{X}(\alpha)$ is also feasible under full centralization by (6). Hence, for every realized demand vector D,

$$\min_{x \in \mathcal{X}^{FC}} \ell_0(x, D) \leq \min_{x \in \mathcal{X}(\alpha)} \ell_0(x, D). \tag{21}$$

Taking expectations over D and using $\Pi_0(\alpha) = -\mathbb{E}\left[\ell_0^{\alpha}(D)\right]$, gives Equation (20).

Theorem 1 provides the reference point for the rest of the analysis. When pooling is costless, timely, and operationally neutral, local protection does not improve on full centralization in the benchmark shortage–idle loss. In hospital-management terms, this is the case in which pooled beds can be converted into usable ward-level capacity without delay, mismatch, or coordination burden. Figure 2a illustrates this benchmark mechanism, under frictionless pooling, benchmark performance reaches its highest value at $\alpha = 1$. The next question is how this conclusion changes when the use of pooled capacity carries operational burden.

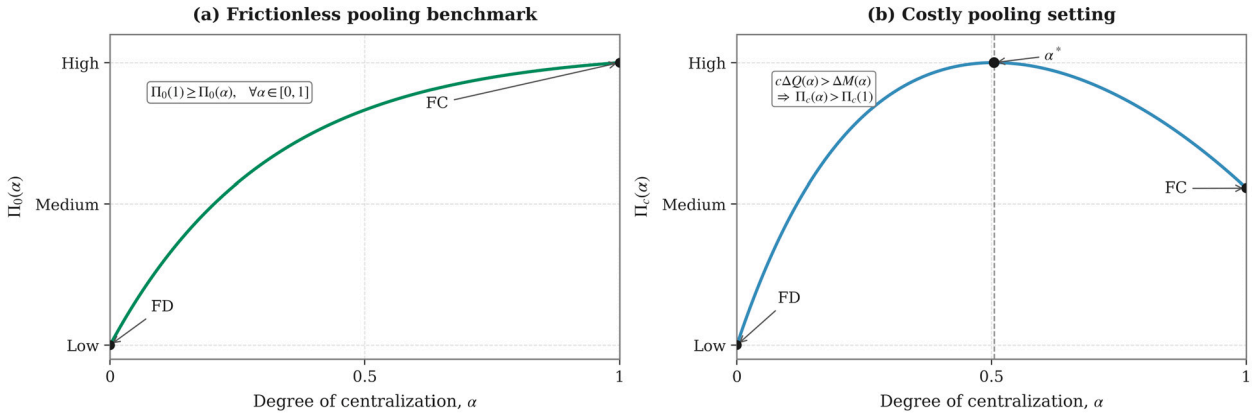


Figure 2. Performance over the governance boundary under frictionless and costly pooling. Panel (**a**) shows the frictionless reference case. When pooled beds can be reassigned without delay, mismatch, or coordination burden, the performance curve reaches its highest value at full centralization. Panel (**b**) incorporates the cost of using the shared pool. The asterisk denotes the performance-maximizing governance boundary α^* In this case, the curve may bend back before α reaches 1, because a larger formal pool can also create more work in turning pooled capacity into usable ward-level capacity. The curves are normalized to highlight this change in mechanism, rather than to report calibrated performance levels for a particular hospital.

5.2. Costly Pooling and the Erosion of FC Dominance

Theorem 1 rests on a demanding premise, the flexibility created by full centralization can be used at no cost. In hospital bed governance, central reassignment may involve communication, capacity search, approval, patient-placement decisions, and coordination across ward boundaries. The issue is not only whether capacity is formally pooled, but also whether pooled capacity can be converted into usable ward-level capacity without additional burden. Once these activities carry operational burden, feasible-set inclusion alone no longer determines the comparison. Let

$$M(D, \alpha) = \sum_{i=1}^{n} p_i L_i(D, \alpha) + \sum_{i=1}^{n} h_i U_i(D, \alpha)$$

denote the shortage–idle component of realized loss before coordination burden is added. With positive coordination friction,

$$\ell_c(D,\alpha) = M(D,\alpha) + cQ(D,\alpha), c > 0, \tag{22}$$

where $Q(D,\alpha)$ captures realized coordination burden. The corresponding performance measure is

$$\Pi_c(\alpha) = -\mathbb{E}[M(D,\alpha) + cQ(D,\alpha)]. \tag{23}$$

The key modeling point is that $Q(D,\alpha)$ is not a fixed surcharge imposed on FC. It is a reduced-form measure of realized coordination intensity. Under the baseline specification Q_1 in Equation (10), FC incurs coordination burden only when the implemented allocation involves cross-ward adjustment. BC is treated under the same rule: it is not rewarded for being an interior design and improves on FC only in cases where the reduction in realized coordination effort is large enough to outweigh the additional shortage–idle loss.

Proposition 1. *Costly pooling can erode FC dominance.*

For a given $\alpha \in (0,1)$, define

$$\Delta M(\alpha) = \mathbb{E}[M(D,\alpha) - M(D,1)],\ \Delta Q(\alpha) = \mathbb{E}[Q(D,1) - Q(D,\alpha)]. \tag{24}$$

The first term $\Delta M(\alpha)$ records how much additional shortage–idle loss the boundary α carries relative to FC. The second, $\Delta Q(\alpha)$, records the expected reduction in coordination burden relative to FC. If that reduction is positive and large enough to cover the additional shortage–idle loss,

$$\Delta Q(\alpha) > 0,\ \ c\Delta Q(\alpha) > \Delta M(\alpha), \tag{25}$$

then

$$\Pi_c(\alpha) > \Pi_c(1). \tag{26}$$

In that case, the feasible-set advantage of FC no longer settles the comparison once realized coordination burden is included.

The proof follows directly from the definition of Π_c. Subtracting $\Pi_c(1)$ from $\Pi_c(\alpha)$ gives

$$\Pi_c(\alpha) - \Pi_c(1) = -\Delta M(\alpha) + c\Delta Q(\alpha). \tag{27}$$

Hence, if $c\Delta Q(\alpha) > \Delta M(\alpha)$, then $\Pi_c(\alpha) - \Pi_c(1) > 0$, and the result follows.

This result is deliberately limited. It does not rank BC above FC in general. It shows only that a bounded regime can outperform FC when the coordination effort saved from using less pooled reassignment is larger than the shortage–idle loss created by restricting the feasible allocation set. For hospital managers, the point is not that a smaller pool is better, but is that pooled capacity may have to be coordinated before it becomes usable, and that this cost can affect where the governance boundary is drawn. Figure 2b illustrates this departure from the frictionless benchmark: with positive coordination friction, the performance curve need not peak at full centralization.

5.3. Local Protection as Resilience Capacity

Coordination friction explains why the use of the shared pool may be costly. However, cost is not the only reason why the degree of pooling matters. Full pooling also reduces immediate ward-level access. These are separate channels: one concerns the burden of using pooled capacity; the other concerns the value of retaining capacity that can be reached locally under pressure. Under bounded centralization, ward i retains $r_i(\alpha)$ as protected

local capacity. This should not be read as a preference for keeping capacity idle, but as a local claim on capacity that can be used before the ward depends on central reassignment.

Under the frictionless benchmark in Theorem 1, this channel is intentionally absent. If reassignment is immediate, reliable, and costless, FC can replicate any bounded arrangement. There is then no separate advantage to keeping capacity locally protected. Local protection begins to matter only when access to pooled capacity is imperfect, time-sensitive, or costly under adverse states.

This mechanism becomes most visible when ward demands move together. Pooling is most valuable when high demand in one ward can be offset by slack in another. As ward-level demand shocks become more synchronized, this offsetting logic weakens. The shared pool may face simultaneous claims, while protected local capacity can provide immediate ward-level absorption. Using Equations (15) and (16), this channel enters the model through total system shortage $L_{\text{sys}}(D, \alpha)$ and the tail-risk component $C_{\text{tail}}(\alpha)$, where $\gamma > 1$ places disproportionate weight on severe-shortage realizations. This adverse-state channel is reflected in the tail-risk component shown later in Figure 3d.

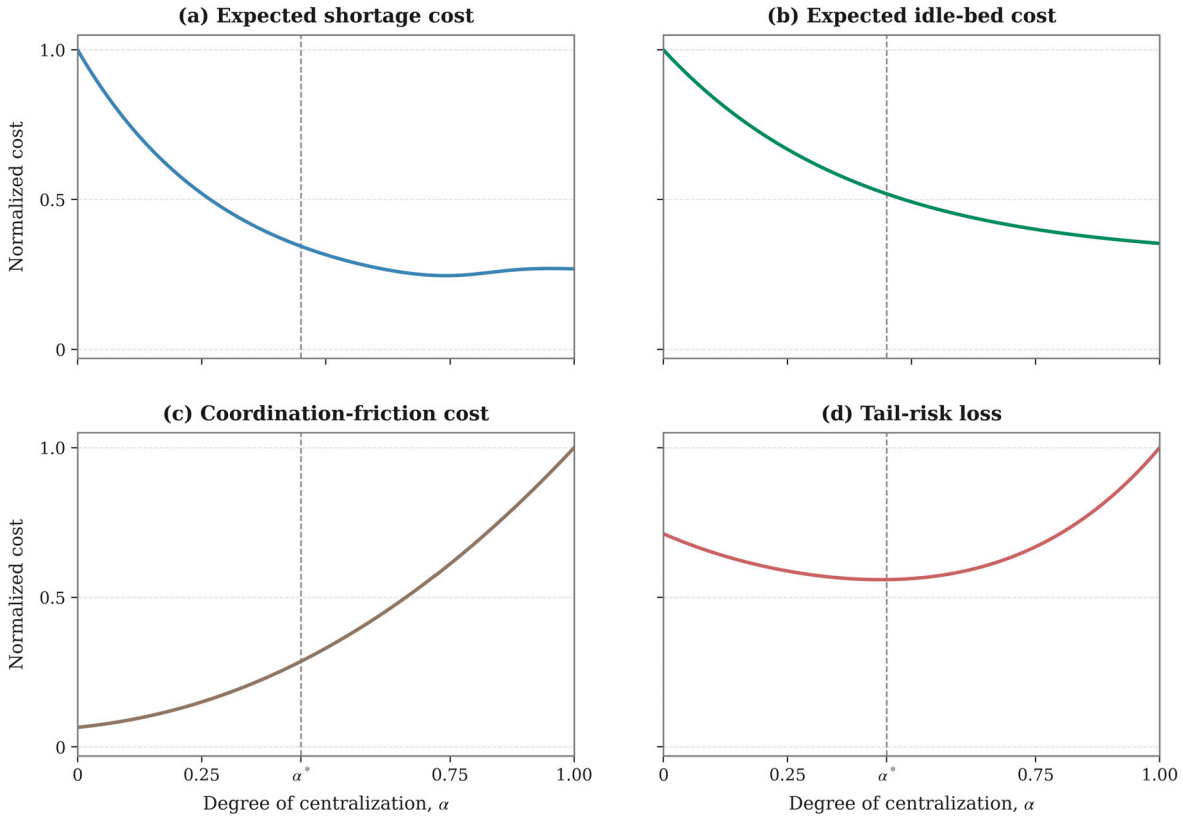


Figure 3. Cost-component decomposition over the centralization boundary. The figure decomposes the normalized performance curve into expected shortage cost, expected idle-bed cost, coordination-friction cost, and tail-risk loss. The vertical dashed line marks the maximizing interior boundary α^*. The decomposition is included because an interior boundary is not produced by a single cost component. At lower values of α, expanding the shared pool can reduce mismatch between wards. At higher values, the marginal pooling gain may become smaller, while the system relies more heavily on coordinated reassignment and retains less immediately accessible local capacity. Components are normalized to show their relative movement, rather than calibrated hospital-level magnitudes.

Mechanism Proposition 2. *Local protection and tail exposure.*

Suppose ward demands are positively correlated, shortage losses are convex, and centralized reassignment is costly or delayed in adverse demand states. Protected local

capacity may reduce exposure to severe simultaneous shortage states. In such parameter regions, bounded centralization can provide resilience value that is not captured by average shortage–idle loss alone.

The mechanism is conditional. Local protection helps only when the value of immediate ward-level absorption outweighs the loss of some pooling flexibility. If reassignment is frictionless or demand shocks are weakly synchronized, the pooling benchmark in Theorem 1 remains the natural reference point.

Proposition 1 and Mechanism Proposition 2 identify two distinct departures from the frictionless benchmark. The first concerns the cost of using pooled capacity, the second concerns the loss of protected local access in adverse states. Together, they imply that the effect of α on overall performance may be non-monotone.

5.4. Interior Governance Boundary

We now return to the full performance measure in Equation (17). Increasing α expands the shared pool, but it also changes coordination burden and the amount of protected local capacity. Accordingly, the effect of centralization depends on the balance among shortage–idle loss, coordination friction, and tail-risk exposure.

This balance can vary across the centralization range. Near $\alpha = 0$, increasing α enlarges the shared pool and may sharply reduce cross-ward mismatch. Near $\alpha = 1$, additional pooling may bring smaller marginal gains while further reducing protected local reserves and increasing reliance on central reassignment. Together, these forces can make the performance curve non-monotone.

Proposition 3. *Interior governance boundary.*

When coordination friction is positive and protected local capacity has adverse-state value, $\Pi(\alpha)$ need not be monotone on $[0, 1]$. There may exist parameter regions in which

$$\alpha^* \in (0, 1). \tag{28}$$

A sufficient condition for such an interior boundary is that there exists $\overline{\alpha} \in (0, 1)$ such that

$$\Pi(\overline{\alpha}) > \Pi(0), \;\; \Pi(\overline{\alpha}) > \Pi(1), \tag{29}$$

Then the preferred governance boundary cannot be represented by either full decentralization or full centralization alone; it corresponds to a bounded design that combines hospital-wide pooling with protected local access.

This result is conditional and does not require $\Pi(\alpha)$ to be globally concave or the interior maximizer to be unique. It states only that, under some operating environments, selective pooling combined with protected local capacity can outperform both extremes. Figure 3 decomposes this pattern into shortage, idle-bed, coordination-friction, and tail-risk components.

5.5. Computational Implications

The analytical results above do not provide a universal ranking of FD, BC, and FC. Instead, they specify the conditions under which the frictionless dominance of FC may be weakened: realized coordination burden, adverse-state exposure, and the loss of immediate local access. These mechanisms define the role of the computational experiments.

Accordingly, the computational experiments in Section 6 are used as mechanism checks rather than calibrated hospital-level predictions. They examine whether the interior-boundary pattern remains visible as the operating environment changes.

The experiments serve three roles. First, they verify the frictionless benchmark by checking whether FC remains performance-consistent when pooled allocation is fully effective and costless. Second, they examine whether coordination friction alone can produce a substantive improvement over FC. Third, they test whether an interior boundary becomes more meaningful when pooled allocation becomes less reliable in high-pressure states.

Finally, an interior numerical maximizer should not automatically be interpreted as evidence against FC. Section 6 therefore distinguishes between a raw interior maximizer and a substantively meaningful improvement over FC. This distinction is important for reading the computational results because an inward shift in the boundary is not necessarily managerially meaningful. Its relevance depends on whether the performance gain over FC is large enough to be substantive.

6. Computational Experiments

6.1. Experimental Design and Parameterization

The computational experiments are used to check the proposed mechanisms under controlled assumptions, not to forecast bed-management outcomes for a specific hospital. They compare how the governance boundary α behaves as the ideal pooling benchmark is gradually relaxed, moving from frictionless pooling to costly pooling and then to stress states in which pooled reassignment becomes less reliable. This structure makes the scenarios transparent and keeps the interpretation limited. The simulations illustrate when an interior boundary can emerge; they do not identify the optimal α for any particular hospital.

We begin with the two-ward canonical setting introduced in Section 3. A one-ward system cannot generate a cross-ward pooling problem, while a larger multi-ward system would introduce heterogeneity before the baseline mechanism is visible. For this baseline, each ward is assigned the same nominal capacity, $B_i = 50$, so that $B = (50, 50)$. The scale is not meant to reproduce the bed stock of a particular hospital department. It gives a transparent setting in which changes in the governance boundary can be examined before capacity heterogeneity is introduced.

Demand pressure is expressed as $\kappa_i = \mu_i / B_i$. In the baseline setting, $\kappa = (0.96, 0.92)$, giving $\mu = (48, 46)$. The baseline demand levels are chosen so that both wards operate close to nominal capacity, but not under chronic overload. This leaves room for governance design to show an effect. If demand were much lower, neither pooling nor local protection would have much operational consequence in the experiment. With demand well above capacity, every governance boundary would perform poorly. The chosen baseline therefore leaves room for pooling to reduce mismatch while still allowing coordination burden and local protection to matter. Demand variability is set at $cv = 0.17$, and the baseline correlation is $\rho = 0.4$. This gives a moderate level of uncertainty and demand synchronization: ward pressures are related enough for pooled capacity to be contested in some states, but not so strongly synchronized that the pooling benchmark is undermined by construction. Demand scenarios are then generated from a nonnegative transformation of a multivariate normal vector:

$$\widetilde{D} \sim N(\mu, \Sigma), D_i = \max\left\{0, \widetilde{D}_i\right\}, \tag{30}$$

Thus, μ should be interpreted as the pre-truncation demand level used in scenario generation rather than the exact mean of the transformed nonnegative demand distribution. The cost parameters are chosen to keep the comparison interpretable in clinical-operational terms rather than as monetary estimates. Shortage costs are $p = (10, 12)$, while idle-capacity costs are $h_i = 1$. This assigns a substantially higher penalty to unmet bed demand than to unused assigned capacity, consistent with the governance focus of the paper: the main concern is not merely whether a bed is unused, but whether admission pressure

cannot be absorbed in time. The small difference between the two ward-level shortage penalties introduces limited heterogeneity, without making the experiment depend on a dominant ward.

The baseline coordination-friction weight is $c = 2.0$. This value is used as a moderate reference level, not as a calibrated estimate of transfer delay or administrative cost. Its role is to make the coordination burden visible while allowing the robustness checks to show how the classification changes when this burden is weaker or stronger. The same logic applies to the tail-risk terms. We set $\eta = 0.05$ and $\gamma = 2$ so that severe-shortage states are represented in the performance measure, but do not overwhelm the average shortage, idle-capacity, and coordination components. The main simulations use N = 12,000 demand scenarios. This sample size is large enough to stabilize the reported baseline comparisons, and the larger-sample checks reported in Table 5 are used to verify that the qualitative classifications are not driven by a particular simulation draw.

Table 5. Threshold and simulation robustness of benchmark classifications.

Robustness Check	Scenario	Test Setting	Raw α^*	Δ_{BC-FC}	Classification	Interpretation
Threshold sensitivity	Frictionless benchmark	$\epsilon_\Delta = 0.25, 0.5, 1.0$	1.00	0.0000	FC-consistent for all thresholds	FC remains on the performance plateau.
Threshold sensitivity	Costly pooling only	$\epsilon_\Delta = 0.25, 0.5, 1.0$	0.26	0.0453	FC-comparable for all thresholds	The raw interior gain is too small to be substantive.
Threshold sensitivity	Costly pooling + mild stress imperfection	$\epsilon_\Delta = 0.25, 0.5, 1.0$	0.20	4.7138	BC for all thresholds	The interior boundary remains substantively better than FC.
Monte Carlo stability	Frictionless benchmark	N = 20,000	1.00	0.0000	FC-consistent	The frictionless benchmark is unchanged.
Monte Carlo stability	Costly pooling only	N = 20,000	0.26	0.0366	FC-comparable	The small interior gain remains practically negligible.
Monte Carlo stability	Mild stress-imperfection baseline	N = 20,000	0.20	4.7190	BC	The substantive interior pattern is stable.

$\Delta_{BC-FC} = \max_{\alpha \in (0,1)} \hat{\Pi}(\alpha) - \hat{\Pi}(1)$. The threshold ϵ_Δ is used only for substantive classification; it does not affect simulated performance values or the raw maximizing α. The Monte Carlo stability checks use N = 20,000 demand scenarios.

Stress-state pooled-allocation imperfection is introduced after the baseline environment has been specified. We define stress states as the upper 15% of total demand realizations, because the concern is not ordinary variation around the mean but the upper-tail situations in which several wards may be trying to draw on the shared pool at the same time. In those states, a centrally pooled bed may still be visible in the system, but the work of converting it into usable ward-level capacity can become more difficult. We therefore set

$$\lambda_s = \max\{\lambda_{\min}, 1 - \alpha\delta(1+\rho)\mathbf{1}_{s \in \mathcal{S}^{\text{stress}}}\}. \quad (31)$$

Here, $\lambda_{min} = 0.94$ sets a lower bound on the usable share of pooled capacity, and the baseline stress-imperfection case uses $\delta = 0.02$. In these experiments, the scenario-level allocation set is obtained from $\mathcal{X}(\alpha)$ by replacing the nominal shared-pool capacity

$S(\alpha)$ with the usable shared-pool capacity $\lambda_s S(\alpha)$. The adjustment therefore changes the operational usability of pooled capacity in stress states, not the physical bed stock. When $\delta = 0$, pooled allocation remains fully effective. When $\delta > 0$, pooled capacity becomes less readily converted into usable ward-level capacity under synchronized high-pressure states, for example because approval, placement judgment, patient-ward fit, and competing ward claims become more difficult to resolve in time. The baseline value is deliberately mild, so that the experiment tests whether an interior boundary can appear without assuming a large failure of pooling.

For each α, simulated governance performance is computed as

$$\hat{\Pi}(\alpha) = -\frac{1}{N}\sum_{s=1}^{N}\left[\sum_i p_i L_{is}(\alpha) + \sum_i h_i U_{is}(\alpha) + \bar{h} P_s(\alpha) + c Q_s(\alpha) + \eta (L_s^{sys}(\alpha))^{\gamma}\right]. \quad (32)$$

We evaluate performance over the grid

$$\mathcal{A} = \{0, 0.02, \ldots, 1\}. \quad (33)$$

To distinguish a raw interior maximizer from a substantively meaningful improvement over FC, we define

$$\Delta_{BC-FC} = \max_{\alpha \in (0,1)} \hat{\Pi}(\alpha) - \hat{\Pi}(1). \quad (34)$$

An interior boundary is classified as substantive BC only if

$$\Delta_{BC-FC} > \epsilon_{\Delta}. \quad (35)$$

The baseline analyses use $\epsilon_{\Delta} = 0.5$, and alternative thresholds are examined in the robustness checks. If the raw maximizer is interior but $\Delta_{BC-FC} < \epsilon_{\Delta}$, the case is reported as FC-comparable. The threshold is used only for substantive classification. It does not change the simulated performance values or the raw maximizing α. This distinction matters because a numerical maximum can occur at an interior value of α even when the performance difference from FC is practically negligible. The classification rule prevents such cases from being interpreted as substantive evidence against full centralization.

6.2. Benchmark Experiments

We compare three benchmark cases that progressively relax the ideal conditions underpinning full centralization. The first is frictionless pooling, in which reassignment is fully effective and carries no coordination cost. The second adds coordination cost while keeping pooled allocation fully effective: central reassignment may be costly, but pooled capacity remains fully usable once assigned. The third adds mild stress-state pooled-allocation imperfection to costly pooling. The sequence is intended to separate the classical pooling effect from the additional question of whether pooled capacity remains usable when reassignment becomes costly or pressure-sensitive.

In the third benchmark, a pooled bed can remain visible in the central system but not be usable by another ward in time. As noted earlier, before a transfer can occur, placement staff may still need authorization and a sequence of follow up arrangements. Under system pressure in particular, these processes may slow down or fail to be completed. This gives Equation (31) its operational interpretation. The experiment still keeps the physical pool in place. Pooled beds do not disappear, and they are not treated as wholly unusable. The adjustment is intended to stay mild. It represents delayed conversion rather than the loss of pooled beds themselves. Only the share of nominal pooled capacity that can become effective ward-level capacity under stress is reduced. Figure 4 reports the resulting performance curves.

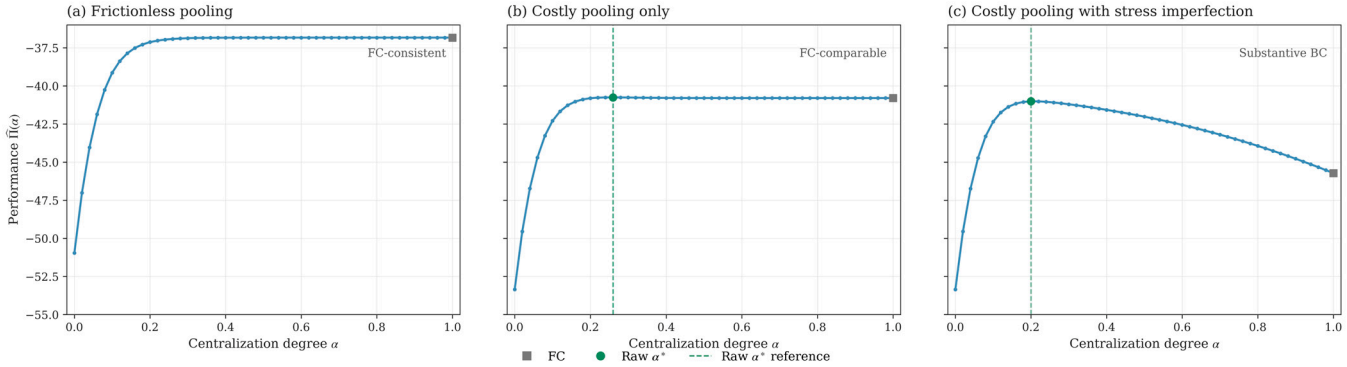


Figure 4. Benchmark performance under progressively relaxed pooling assumptions. Panel (**a**) starts from the frictionless benchmark. Panel (**b**) keeps pooled allocation fully effective but adds coordination friction. Panel (**c**) further allows for a mild loss in pooled-allocation effectiveness during high-pressure states. The panels should therefore be read as a sequence, not as three unrelated examples. The comparison separates a numerical movement of the raw maximizer from a substantively meaningful improvement over full centralization, which is reported using $\Delta_{\mathrm{BC}-\mathrm{FC}}$ and $\epsilon_{\Delta} = 0.5$.

This sequence leads to three different interpretations. In the frictionless case, FC is performance-consistent: the best interior boundary ties FC but does not improve on it. Adding coordination friction changes the location of the raw maximizer, but not the substantive conclusion. The best raw value is now interior, yet its gain over FC remains below the reporting threshold, so the case is classified as FC-comparable. The interpretation changes only in the third benchmark, where pooled capacity is both costly to use and mildly less reliable in stress states. In that case, the best interior boundary produces a gain above the substantive threshold and is classified as BC. This result should therefore be read as conditional. The experiment does not rank BC above FC in general; it shows that the governance boundary can become meaningful when the formal size of the pool and the clinical-time usability of that pool begin to diverge.

6.3. Mechanism Decomposition

The benchmark results show when an interior boundary becomes substantive, but the total performance curve does not identify which objective component drives the shift. We therefore decompose the baseline stress-imperfection case into its main cost components. The purpose is not to introduce another performance criterion, but to show how pooling gains, coordination burden, and adverse-state exposure jointly shape the governance boundary. For each value of α, the simulated loss corresponding to (32) can be written as

$$-\hat{\Pi}(\alpha) = \hat{C}^{short}(\alpha) + \hat{C}^{idle}(\alpha) + \hat{C}^{Q}(\alpha) + \hat{C}^{tail}(\alpha),$$

where $\hat{C}^{short}(\alpha)$ is the expected shortage cost, $\hat{C}^{idle}(\alpha)$ is the expected idle-capacity cost, $\hat{C}^{Q}(\alpha)$ is the coordination-friction cost, and $\hat{C}^{tail}(\alpha)$ is the tail-risk loss. The idle-capacity component includes both unused ward-assigned capacity and unused shared-pool capacity, consistent with the operating-loss definition in Section 4.2. Figure 5 reports these four components for the baseline stress-imperfection case.

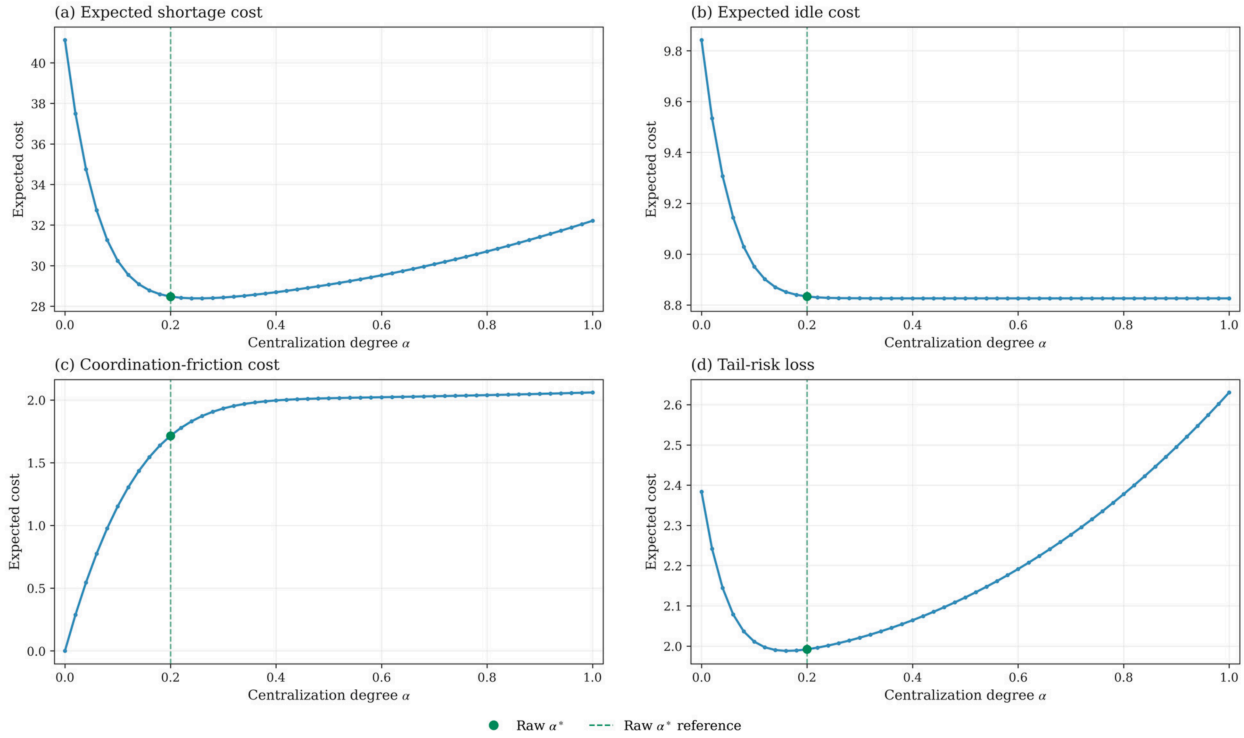


Figure 5. Cost-component decomposition in the baseline stress-imperfection case. The figure separates total simulated loss into expected shortage cost, expected idle-capacity cost, coordination-friction cost, and tail-risk loss. The vertical dashed line marks the raw interior maximizer, $\alpha^* = 0.20$. This decomposition explains why the interior result in Figure 4c arises. Pooling is beneficial over the first part of the range because it reduces mismatch between wards. The advantage weakens when α becomes high, as the system depends more heavily on centrally reassigned capacity that is costly to coordinate and less reliable under stress. The component values are used to interpret the mechanism, not to provide calibrated cost estimates for a particular hospital.

The component curves show the same pattern. At low-to-moderate α, pooling reduces shortage and idle capacity, and the interior result should therefore not be interpreted as evidence that pooling is ineffective. The pooling benefit is visible in the first part of the range. As α increases, coordination-friction cost rises, and the system becomes more exposed to reassignment loss in high-demand states. The interior optimum therefore reflects a balance: moderate pooling reduces mismatch, while excessive reliance on the shared pool raises coordination burden and tail-risk exposure.

6.4. Boundary Analysis over Coordination Friction and Pooled-Allocation Imperfection

The benchmark value $\delta = 0.02$ represents only one way to operationalize the usability adjustment in Equation (31). To check whether the benchmark classification depends on this specific value, we vary both the coordination-friction weight c and the stress-state pooled-allocation imperfection intensity δ. For each pair (c, δ), we evaluate $\hat{\Pi}(\alpha)$ over the grid in Equation (33), compute Δ_{BC-FC} using Equation (34), and apply the classification rule in Equation (35).

Because the transition near $\delta = 0$ is substantively important, we use a refined low-δ grid, $\delta \in \{0, 0.0025, 0.005, 0.0075, 0.01, 0.015, 0.02\}$, together with $c \in \{0, 1, \ldots, 10\}$. This refinement allows us to distinguish the fully effective pooling case from mild departures from fully usable pooled capacity under stress.

Figure 6 shows that the BC classification is not produced simply by choosing an interior governance boundary or by adding coordination friction. When pooled allocation remains fully effective, the cases remain largely FC-comparable, which is consistent with

the costly-pooling-only benchmark. The BC region emerges only after the usability loss in stress states is introduced, and only for part of the c–δ space. This distinction matters for interpreting Equation (31). The formula is not being used to make any positive δ automatically favor BC; instead, it identifies the operating region in which the pooling benefit is offset by the combined effect of coordination burden and stress-state conversion loss. The result remains conditional on the simulated high-pressure environment, the severe-shortage penalty, and the substantive-classification threshold.

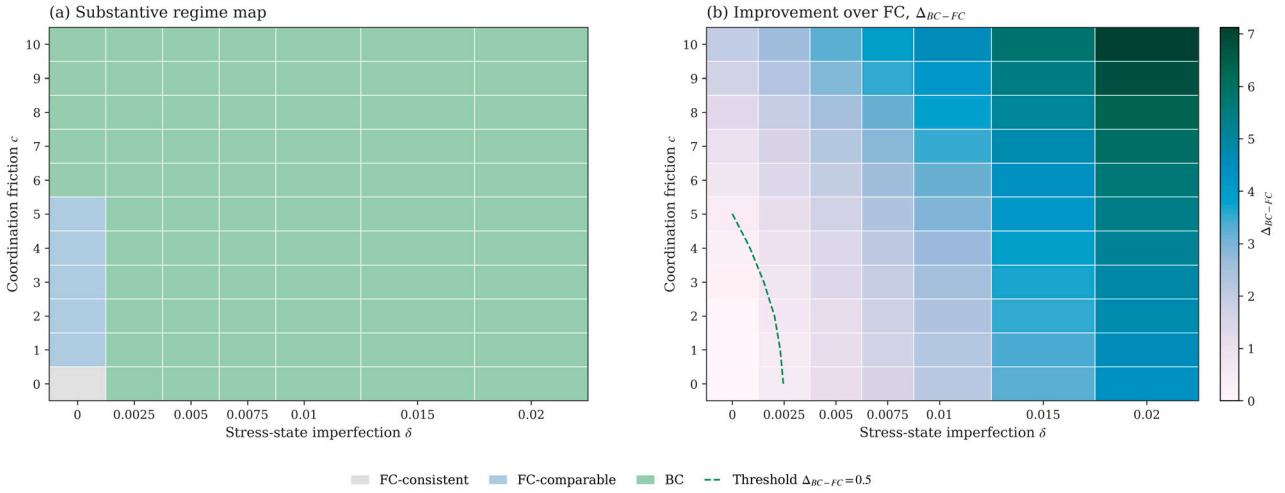


Figure 6. Boundary analysis across coordination friction and stress-state pooled-allocation. Panel (**a**) reports the substantive regime classification over c and δ, using the threshold $\epsilon_\Delta = 0.5$. Panel (**b**) reports Δ_{BC-FC}, with the dashed contour marking $\Delta_{BC-FC} = 0.5$. The figure shows how the benchmark classification changes across operating conditions, not to identify calibrated cutoff values for a particular hospital.

This pattern raises a more specific robustness question. The result would be less persuasive if it depended mainly on the reporting threshold, a single simulation draw, or a particular measure of coordination burden. We therefore check these issues next before returning to the broader question of whether the same mechanism remains visible in larger ward systems.

6.5. Sensitivity and Robustness Checks

The preceding results depend on a classification rule, simulated demand scenarios, a baseline measure of coordination burden. We therefore treat the robustness checks as a way to address a practical question: whether the main interpretation would materially change if the reporting threshold were tightened or relaxed, if the Monte Carlo sample were enlarged, or if coordination friction were measured from a different operational angle. Table 5 reports the threshold and simulation-size checks.

The classifications remain stable under both checks. Varying $\varepsilon\Delta$ changes only the reporting threshold for what counts as a substantive improvement; it does not alter the simulated performance curves or the raw maximizing boundary. Enlarging the Monte Carlo sample leads to the same qualitative reading. The frictionless benchmark remains FC-consistent, the costly-pooling-only case remains FC-comparable, and the mild stress-imperfection case remains stable under both checks.

We also consider whether the interior pattern depends on the baseline definition of coordination friction. The main experiments use Q_1, which measures realized deviation from nominal ward capacity. This is a natural baseline because it links coordination burden to actual reassignment rather than to the formal name of the governance regime. Coordina-

tion work, however, can also be measured from other operational perspectives. Appendix Table A2 repeats the baseline stress-imperfection experiment using the normalized alternatives Q_2, Q_3, and Q_4, which respectively emphasize shared-pool use, the number of wards involved, and ward-specific reassignment difficulty. The raw best boundary remains interior and the substantive classification remains BC under these alternatives. This does not prove robustness to every possible coordination-cost formulation, but it reduces the concern that the finding is an artifact of measuring coordination burden only through Q_1. The next subsection examines whether the same mechanism remains visible beyond the two-ward setting.

6.6. Multi-Ward Extension

The previous experiments use a two-ward system to keep the mechanism transparent. We now examine whether the classification pattern persists in larger systems. The extension considers n = 5 and n = 10 wards with moderate heterogeneity in ward capacity, demand pressure, demand variability, and shortage penalties, while keeping the same allocation logic and α-grid. The extension is used as a mechanism check, not as a calibrated comparison of hospitals with different ward-system sizes.

For each system size, we compare two pooling conditions. The first assumes fully effective pooled allocation. The second adds the same mild stress-state pooled-allocation imperfection used in the benchmark experiments. This comparison separates the effect of reduced pooled-reassignment reliability under pressure. Table 6 reports the results.

Table 6. Multi-ward extension under fully effective and stress-imperfect pooling.

System Size	Pooling Condition	Raw α^*	Δ_{BC-FC}	Classification	Interpretation
$n = 2$	No imperfection	0.26	0.0453	FC-comparable	The raw optimizer is interior, but the gain over FC is practically small.
$n = 2$	Mild stress imperfection	0.20	4.7138	BC	The canonical case shows a substantive interior boundary.
$n = 5$	No imperfection	0.46	0.0236	FC-comparable	The raw interior gain remains practically negligible.
$n = 5$	Mild stress imperfection	0.24	12.5639	BC	The interior-boundary pattern appears in a larger ward system.
$n = 10$	No imperfection	0.70	0.0097	FC-comparable	FC remains performance-comparable despite an interior raw maximizer.
$n = 10$	Mild stress imperfection	0.26	29.8877	BC	The interior-boundary pattern persists in the larger system.

Δ_{BC-FC} is defined in (34). "No imperfection" refers to fully effective pooled allocation. "Mild stress imperfection" uses $\delta = 0.02$. Substantive classification follows (35) with $\epsilon_\Delta = 0.5$. Larger Δ_{BC-FC} values in larger systems partly reflect aggregate system-level cost accounting rather than calibrated hospital-level policy gains.

Table 6 shows that the multi-ward extension does not automatically favor BC. When pooled allocation is fully effective, the raw maximizer can still lie in the interior, but the gain over FC is negligible. For $n = 2$, 5 and 10, all no-imperfection cases are classified as FC-comparable. This is useful evidence for interpreting the model, adding more wards does not, by itself, make bounded centralization substantively better than FC. When the shared pool works as intended, FC remains a strong benchmark.

The pattern changes once mild stress-state imperfection is introduced. In all three systems, the best-performing boundary is interior, and the improvement over FC exceeds

the substantive threshold. The exact value of α^* varies with system size, as expected, but the classification does not, these stress-imperfection cases are classified as BC. Appendix Figure A1 reports the corresponding full α-grid curves. The visual pattern is consistent with Table 6. Under fully effective pooling, the curves rise quickly and then flatten near FC. Under stress-state imperfection, the curves peak in the interior and then decline as the system approaches full centralization.

These multi-ward results should be read as mechanism checks rather than calibrated policy comparisons across hospital sizes. The larger Δ_{BC-FC} values for $n = 5$ and $n = 10$ partly reflect aggregate system-level cost accounting. The important point is therefore not the numerical value of α^* or Δ_{BC-FC} in any one system. Rather, it is that the boundary mechanism is not confined to the two-ward case. With fully effective pooling, FC remains performance-comparable, whereas stress-state reassignment loss restores the value of retaining some local protection. This supports the broader view of bounded centralization as a conditional governance design rather than a general substitute for full centralization.

The preceding analyses examine the governance mechanism from several complementary angles, including benchmark scenarios, cost decomposition, boundary analysis, robustness checks, and multi-ward extensions. Table 7 synthesizes these results to show how the qualitative interpretation changes across computational scenarios and how it is shaped by the main robustness checks.

Table 7. Summary of computational scenarios and robustness analyses.

Evidence Covered	Analysis/Source	Key Finding	Interpretation
Benchmark scenarios	Figure 4a–c	FC is performance-consistent in the frictionless case; costly pooling alone is FC-comparable; mild stress imperfection yields substantive BC.	The main shift toward BC occurs only when pooled capacity remains useful but becomes less reliable under pressure.
Mechanism and boundary	Figure 5; Figure 6 (c-δ grid)	The interior pattern reflects several cost components and appears only in part of the c-δ space.	BC is conditional on operating regions with coordination burden and stress-state usability loss.
Threshold and simulation robustness	Table 5	Classifications are preserved when ε_Δ is varied and when N is increased.	The FC-comparable versus BC distinction is not driven by one cutoff or one simulation sample.
Alternative friction specifications	Appendix Table A2	The stress-imperfection case remains BC under Q_1–Q_4 specifications.	The mechanism is robust to alternative reduced-form measures of coordination burden.
Multi-ward extension	Table 6; Appendix Figure A1	For larger n, fully effective pooling remains FC-comparable, while stress imperfection restores BC.	The mechanism is not confined to the two-ward case, though results remain mechanism checks rather than calibrated hospital-size comparisons.

Δ_{BC-FC} is defined in Equation (34), and ε_Δ denotes the substantive-improvement threshold. Robustness refers to the qualitative classification and mechanism pattern, not to identical numerical values of α^* or Δ_{BC-FC} across all settings.

6.7. Summary of Computational Findings

The computational results should be read as a sequence of mechanism checks rather than a simple ranking of the three governance regimes. The frictionless benchmark preserves the usual pooling argument: when pooled capacity can be reassigned without loss, FC remains the natural reference point. When coordination friction is added, the perfor-

mance curve changes and the raw maximizer can move away from $\alpha = 1$. That numerical shift should not be read, on its own, as evidence for a substantively meaningful BC regime. In the costly-pooling-only case, the gain relative to FC remains too small to carry much managerial weight.

When pooled capacity becomes less reliable in stress states, the interior solution should not be judged only by where the maximizer occurs. The cost decomposition gives a clearer reading of what is happening. At low and moderate values of α, pooling still helps: it shifts capacity toward wards with unmet demand and reduces cross-ward mismatch. The trade-off changes as α increases. More of the system's performance then depends on beds that must pass through central coordination before they can be used, and these beds are exposed to the stress-state usability loss. The interior BC pattern reflects this tension. It emerges when c and δ are high enough to make heavy reliance on the shared pool costly, while pooling still retains enough matching value to be useful.

The robustness checks give little reason to revise this reading. The BC classification in the stress imperfection case is not created by a loose substantive improvement threshold, a small Monte Carlo sample, or the particular baseline choice of Q_1. Tightening or relaxing the reporting threshold leaves the qualitative classification unchanged, and the larger simulation sample gives the same pattern. Replacing Q_1 with alternative normalized measures of coordination friction also preserves the interior result. The multi-ward extension points in the same direction. A larger ward system does not make BC substantively better than FC when pooled allocation is still fully effective. The interior pattern reappears only after stress-state reassignment loss is introduced, which keeps the mechanism tied to pooled capacity that becomes harder to use under pressure.

The results therefore call for a conditional reading. They do not show that BC generally dominates FC. They point instead to a more limited operating logic. Bounded centralization becomes substantively attractive when pooling still helps reduce mismatch, while the conversion of pooled capacity into usable capacity becomes costly or less reliable under pressure. For managers, the issue is therefore not simply the size of the shared pool. It is also whether that pooled capacity can be made usable in the demand states where it is most needed.

7. Managerial and Policy Implications

7.1. Pooling as Usable Capacity, Not Only Formal Capacity

Ward demand seldom rises in the same pattern across the hospital. One unit may have an unused bed while another is trying to keep admissions moving. That unevenness is the basic reason pooling remains important in bed reform. If patients can be placed safely across wards, decisions are made without much delay, and coordination adds little work, a hospital-wide pool gives bed teams the widest room to respond. In that setting, full centralization remains the benchmark.

The difficulty starts once a bed has been counted as part of the pool. It still has to be found, cleared for use, checked against the patient's needs and the receiving ward's capacity, and turned into a placement before the demand situation changes. On a quiet day, this work may barely be noticed. When several wards need help at once, it can become the step that slows the whole response. The bed board may still show a large shared pool, but managers may have less usable flexibility than the formal number implies. The practical test is whether shared capacity can be converted into ward-level placements quickly enough when pressure is concentrated across wards.

7.2. Coordination Burden and Local Protection as Governance Variables

After demand is observed, Q reflects the work needed to turn an allocation into an actual placement. Full centralization may add little burden when patients can still be placed close to each ward's usual capacity. A bounded design is not automatically lighter. If several wards become short at the same time, the hospital may have to use pooled beds repeatedly, with each use bringing approval, movement, and placement work. The relevant issue is the demand state that appears, rather than the governance label selected in advance.

The protected reserve $r_i(\alpha)$ in Equation (2) should be read in the same operational way, but from the access side. It does not create new beds, and it should not be interpreted as capacity deliberately left empty. It is a priority claim on beds already inside the hospital. Its value lies in reachability. A ward can draw on this capacity through its own admission routines, staffing arrangements, and clinical judgment before waiting for central reassignment. When shared placements are quick and reliable, this advantage may be limited. It matters more when pressure rises across several wards at once, or when central staff cannot identify, authorize, and complete transfers quickly enough.

For bounded centralization to be workable, hospitals need explicit rules around these claims. The shared pool remains available, but some usable capacity is not placed entirely behind the full reassignment process. It stays close enough for wards to use directly as pressure rises. In practice, hospitals would need to define when local claims are protected, when they give way to the shared pool, and who has authority to override them during hospital-wide pressure.

7.3. Designing a Data-Informed Boundary for Bed Reform

Hospitals do not have to treat centralization as the inevitable endpoint of bed reform. Before moving more capacity into a shared pool, they can ask what the change is likely to do under their own conditions. Occupancy histories, bed-assignment records, stress scenarios, and simulation runs can all help trace how performance changes as the access boundary moves.

The α framework is useful because it puts full decentralization, bounded centralization, and full centralization on the same scale. They are not three unrelated policy labels, but different positions along a common boundary between ward access and central coordination. This framing shifts the reform question from which regime sounds preferable to where access should actually sit.

The evidence will not always point cleanly to one answer. Full centralization may remain the best design in some hospitals. In other cases, several designs close to $\alpha = 1$ may perform so similarly that the operational difference is small, even if one value ranks first in a simulation. An interior value also needs to be interpreted rather than copied into policy. Its apparent advantage may come from fewer unmet bed requests, less unused capacity, lower coordination work, or reduced exposure to severe-shortage states. Those sources of gain have different managerial meanings.

The useful α is therefore likely to vary across hospitals. Service mix, staffing cover, transfer routines, and placement authority all affect how much shared capacity can be used in practice. Managers need to know which beds a central team can identify and release quickly, which services require stronger local access because reassignment is difficult, and when central authority should expand because protected local access is no longer the safest use of capacity. A defensible boundary should fit observed demand, staffing conditions, and coordination capacity. It should also be revisited as those conditions change.

7.4. Limited Centralization in Hospital Practice

In many hospitals, centralization begins with shared visibility rather than a full transfer of placement control. Patient flow centers, command centers, electronic dashboards, daily bed meetings, and placement coordination teams can give staff a view of capacity beyond their own wards and help them manage cases across service boundaries [19,35,36].

That shared view is valuable, but it is still only the start of a placement decision. The receiving service must decide whether the patient fits its specialty scope, whether staffing is sufficient, whether infection-control requirements can be met, and whether the ward's routines can absorb the case. A visible bed therefore remains only a candidate bed until these local conditions are satisfied.

Limited centralization is most plausible when the shared pool is built around clinically adjacent services. General medical units, geriatric medicine, and low-acuity overflow areas may be able to share capacity because their staffing patterns and care routines overlap. In such cases, reassignment is less likely to require a full renegotiation of clinical responsibility.

Other beds are harder to move. Specialist teams, equipment needs, isolation rules, and specialty accountability may make local control more appropriate. The practical task is to distinguish beds that can be shared routinely from beds that require additional clinical approval, and from beds where ward access should remain protected. Evidence on off-service and outlying placements is a useful warning here. Overflow beds can relieve immediate pressure, but patients placed outside their primary specialty context may face added risks in quality, length of stay, readmission, and coordination [17,37,38].

The boundary between local access and central authority may also need to change as pressure rises. In ordinary periods, wards may keep priority over protected local capacity. Once emergency department boarding, admission waits, or simultaneous ward pressure passes agreed trigger points, the central bed management team may need stronger authority to release capacity across units.

The difficult part is making those triggers work while the system is under pressure. Command center evaluations show that better visibility does not guarantee smoother flow when data quality is weak, processes are poorly integrated, or units are not engaged [19]. Bed-assignment research adds a timing problem: a bed may be marked as allocated before the intended patient is ready to move, while other ready patients continue to wait [18]. Evidence on medical outliers and off-service placement raises the same concern from the clinical side [17,39]. Limited centralization is therefore workable only when shared visibility is supported by timely identification, authorization, assignment, and a placement that the receiving ward can actually use.

7.5. Implementation Challenges

An empty bed shown in the central system is still only a possible placement. The receiving ward must be able to take the patient safely, and that judgment depends on nursing cover, skill mix, and local support at the time of transfer. General wards often allow more sharing because their routines overlap. Beds in intensive care, isolation, postoperative specialties, pediatrics, maternity, or psychiatry are different. Equipment, staffing, infection-control requirements, and specialty accountability make these beds less flexible. Overflow or off-service placement can relieve immediate pressure, but the evidence also points to risks in quality, length of stay, readmission, and coordination work [17,37,38].

These clinical limits need to be settled in rules before congestion makes each placement harder. Hospitals need to know which unit accepts the patient, which team carries clinical responsibility, how escalation is triggered, and when central coordinators can override local claims on capacity. If those decisions are left until the system is already under pressure, the shared pool may add delay rather than flexibility. Protocols, accountability arrangements,

and the authority given to the bed management team therefore shape whether shared capacity can be used in practice [35].

Information systems have to support these rules at the moment decisions are made. Bed teams need current information on bed status, room constraints, expected discharges, transfer progress, and pending placement decisions. Dashboards and bed-management platforms can help, but command-center studies show that data quality and departmental engagement still affect performance [19,36,40]. A listed bed is not enough. Staff must be able to identify it, clear its use, and complete the placement while the demand state still matters.

8. Conclusions

8.1. Summary and Main Findings

The same stock of beds can be governed by giving wards and the central system different degrees of access. In the α framework, full decentralization, bounded centralization, and full centralization are positions on a single continuum between immediate ward access and hospital-wide coordination. When reassignment is reliable and almost frictionless, full centralization remains the benchmark because it gives the hospital the widest feasible allocation set after demand is observed.

The later results do not overturn that benchmark. They show where it begins to need qualification. Coordination friction can move the numerical maximizer away from full centralization, but this shift may have limited managerial meaning when the performance difference is small. The interior boundary becomes more important when high pressure reduces the reliability or usable share of pooled reassignment, and when capacity close to the ward can absorb pressure before delays spread. Local protection has value in this narrower sense. It is useful when it can be reached at the right time and used with sufficient reliability, not merely because it is locally held.

These findings still support centralized pooling as a strong reference point, but they change the practical test for reform. Managers need to ask how much pooled capacity the hospital can identify, authorize, staff, and convert into placements while the demand state remains active. Because the evidence comes from a stylized, mechanism-oriented simulation, the framework should be used for hospital-specific assessment rather than as a universal prescription for α.

8.2. Limitations and Future Research

The simulations are intended to make one governance mechanism easier to see, rather than to recommend a policy for any specific hospital. The framework has not been calibrated with hospital-level operational data, and its results have not been validated against observed placement records. This controlled setting is useful because it separates the formal scope of pooling from the capacity that can actually be used. It does not, however, produce an α value that is ready to be implemented.

Net bed demand is the main simplification that makes the model tractable. It combines admission pressure, continuing occupancy, and expected bed releases into a single variable, so the comparison across governance regimes remains manageable. That choice also leaves some important dynamics outside the main framework. Length of stay, discharge timing, and changing occupancy are not modeled directly. Future work could bring these elements into the analysis through a rolling-horizon allocation model or a Markov occupancy model.

For empirical validation, the first task would be to rebuild the net demand variable from hospital records. Relevant sources would include admission requests, current occupancy, expected and actual discharges, transfer requests, and time stamped placement decisions. Several operating constraints are also simplified in the present framework.

Staffing availability, skill mix, patient acuity, specialty compatibility, infection-control requirements, clinical authorization, and transfer execution enter only through reduced form coordination and stress-state imperfection terms. Command center records, bed management platforms, transfer logs, approval timestamps, and placement histories could make these mechanisms more observable.

A hospital-level test could then follow the boundary as it is used in practice. Ward-level net demand and the realized split between locally accessible and centrally coordinated capacity would need to be reconstructed first. Coordination burden could be approximated from placement timestamps, approval delays, transfer logs, and the number of wards involved in a reassignment. High-pressure states could be identified from emergency department boarding, high occupancy, simultaneous ward shortages, or admission waiting peaks. The gap between beds visible in a central pool and beds converted into timely placements would provide a trace of stress-state usability loss. These boundary patterns could then be compared with delayed admission, off-service placement, length of stay, readmission, transfer to higher acuity care, and waiting time [17,19,40].

One further limitation concerns the scale of the boundary. The present model uses a hospital-wide boundary, while many hospitals operate with service-line clusters, specialty overflow rules, or hybrid command center structures. Extending the α framework in that direction would allow future work to distinguish routine sharing within compatible services from less routine reassignment across specialty boundaries.

Author Contributions: Conceptualization, F.X. and Y.X.; methodology, Y.X., F.X. and K.K.L.; software, F.X.; validation, F.X. and K.K.L.; formal analysis, F.X.; investigation, S.W. and F.X.; resources, S.W.; writing—original draft preparation, F.X.; writing—review and editing, F.X. and K.K.L.; visualization, S.W.; supervision, K.K.L.; project administration, Y.X.; funding acquisition, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Xi'an Technological University, grant number H202512531, under the project "Research on Green Supply Chain Optimization of Medical Logistics Enterprises".

Institutional Review Board Statement: Not applicable. This study did not involve human subjects, human material, human tissues, identifiable patient records, or individual-level human data.

Informed Consent Statement: Not applicable. This study did not involve human participants, human material, human tissues, identifiable patient records, or individual-level human data.

Data Availability Statement: The original contributions presented in this study are included in the article. The computational analysis is based on simulated scenarios and parameter settings described in the manuscript. No patient-level, identifiable human, clinical-record, survey, or third-party data were used. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors gratefully acknowledge the support provided by their institutions and sincerely thank the editors and reviewers.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Additional Computational Details

This appendix provides additional details on the computational experiments, including the parameter settings, allocation procedure, robustness checks, and multi-ward extensions. Table A1 reports the baseline values used in the main experiments and the robustness dimensions along which those values are varied. The experiments are not calibrated to a specific hospital. The baseline values define the reference high-pressure setting studied in the main text, while the robustness checks examine whether the qualitative classifications change when key inputs such as demand correlation, coordination friction, pooled-allocation imperfection, improvement thresholds, and simulation size are varied.

Table A1. Baseline experimental parameter settings and robustness dimensions.

Category	Parameter	Baseline Setting	Robustness/Sensitivity
System size	n	2	5, 10
Capacity	B	$(50, 50)$	Heterogeneous B_i in multi-ward tests
Demand pressure	$\kappa_i = \mu_i / B_i$	$(0.96, 0.92)$	Varied in multi-ward settings
Demand mean	μ	$(48, 46)$	Implied by $\kappa_i B_i$
Demand variability	cv	0.17	$cv_i \in [0.12, 0.20]$ in multi-ward tests
Demand correlation	ρ	0.4	0 to 0.9
Idle cost	h_i	1	Fixed baseline
Shortage cost	p_i	$(10, 12)$	Heterogeneous penalties in multi-ward tests
Coordination friction	c	2.0	0 to 10
Tail-risk weight	η	0.05	Fixed in baseline checks
Tail convexity	γ	2	Fixed in baseline checks
Pooled imperfection intensity	δ	0.02	0 to 0.08; refined low-δ grid
Effectiveness lower bound	$\lambda_{\min}$	0.94	Fixed
Stress-state definition	—	Upper 15% of total demand	Fixed
Governance grid	α	$0, 0.02, \ldots, 1$	Fixed
Baseline coordination measure	Q	Q_1	Q_2, Q_3, Q_4
Practical improvement threshold	ϵ_Δ	0.5	0.25, 1.0
Monte Carlo sample size	N	12,000	20,000 for stability checks

The baseline coordination measure is Q_1, defined in (10). Alternative definitions Q_2, Q_3, Q_4 are normalized by their reference expected value under $\alpha = 1$ and $c = 0$, so that different friction specifications can be compared on a common scale. The practical threshold ϵ_Δ is used only for reporting substantive regime classifications; it does not affect the performance calculation or the choice of the raw maximizing α.

Appendix A.1. Allocation Procedure

Algorithm A1 describes the scenario-level allocation and performance-evaluation procedure used in the computational experiments. The procedure is applied for each candidate governance boundary $\alpha \in \mathcal{A}$, where $\mathcal{A} = \{0, 0.02, \ldots, 1\}$. The algorithm operationalizes the definitions in Sections 4 and 6 and computes local reserves, shared-pool capacity, stress-state pooled effectiveness, realized allocations, coordination burden, cost components, and the estimated performance value $\hat{\Pi}(\alpha)$.

Algorithm A1. Scenario-level shared-pool allocation and performance-evaluationprocedure.

Input:

Ward capacities B_i; demand scenarios $D_s = (D_{s1}, \ldots, D_{sn})$, $s = 1, \ldots, N$; centralization grid $\mathcal{A}$; shortage costs p_i; idle-capacity costs h_i; coordination-friction weight c; tail-risk weight η; tail convexity γ; demand correlation ρ; pooled-imperfection parameters $(\delta, \lambda_{\min})$; stress-state quantile τ; coordination-friction measure Q_m; and normalization scale q_m^{scale}.

Output:

Estimated governance performance $\hat{\Pi}(\alpha)$ and average cost components for each $\alpha \in \mathcal{A}$.

For each $\alpha \in \mathcal{A}$:

1. Compute protected local reserve and shared-pool capacity:

$$r_i(\alpha) = (1 - \alpha)B_i,$$

$$S(\alpha) = \alpha \sum_{i=1}^{n} B_i.$$

2. Identify stress-state scenarios. Let

$$T_s = \sum_{i=1}^{n} D_{si}.$$

Scenario s is classified as a stress-state if T_s is no smaller than the τ-quantile of $\{T_s\}_{s=1}^{N}$. Denote the set of stress-state scenarios by $\mathcal{S}^{stress}$.

3. Compute the pooled-allocation effectiveness factor for each scenario:

$$\lambda_s = \begin{cases} 1, & \text{if pooled imperfection is disabled or } \delta = 0, \\ \max\{\lambda_{\min},\, 1 - \alpha\delta(1+\rho)\mathbf{1}_{s \in \mathcal{S}^{stress}}\}, & \text{otherwise.} \end{cases}$$

4. Compute the remaining service need after protected local capacity:

$$R_{si} = \max\{D_{si} - r_i(\alpha), 0\}.$$

Convert this to the physical shared-pool requirement:

$$\widetilde{R}_{si} = \frac{R_{si}}{\lambda_s}.$$

5. Initialize shared-pool allocation and remaining shared-pool capacity:

$$y_{si} = 0, \qquad P_s = S(\alpha).$$

6. Compute ward-level allocation scores. Let

$$\bar{\lambda} = \frac{1}{N} \sum_{s=1}^{N} \lambda_s.$$

For each ward i, define

$$a_i = \bar{\lambda} p_i - \frac{c\, \xi_i}{q_m^{scale}},$$

where

$$\xi_i = \begin{cases} \omega_i, & \text{if } Q_m = Q_4, \\ 1, & \text{otherwise.} \end{cases}$$

Order wards by decreasing a_i, and retain only wards with $a_i > 0$.

7. Allocate the shared pool greedily according to this order. For each retained ward i,

$$y_{si} = \min\left\{\widetilde{R}_{si}, P_s\right\},$$

$$P_s = P_s - y_{si}.$$

8. Compute physical and effective allocations:

$$x_{si}^{phys} = r_i(\alpha) + y_{si},$$
$$x_{si}^{eff} = r_i(\alpha) + \lambda_s y_{si}.$$

9. Compute ward-level outcomes:

$$L_{si} = \max\left\{D_{si} - x_{si}^{eff}, 0\right\},$$
$$U_{si} = \max\left\{x_{si}^{eff} - D_{si}, 0\right\}.$$

Here L_{si} denotes unmet demand and U_{si} denotes idle ward capacity. The unused shared-pool capacity in scenario s is P_s.

10. Compute the coordination-friction measure. The baseline specification is

$$Q_1(D_s, \alpha) = \frac{1}{2}\sum_{i=1}^{n} |x_{si}^{phys} - B_i|.$$

The alternative specifications used in robustness checks are

$$Q_2(D_s, \alpha) = \sum_{i=1}^{n} y_{si},$$
$$Q_3(D_s, \alpha) = \sum_{i=1}^{n} \mathbf{1}\{y_{si} > 0\},$$
$$Q_4(D_s, \alpha) = \frac{1}{2}\sum_{i=1}^{n} \omega_i |x_{si}^{phys} - B_i|.$$

The normalized coordination burden is

$$Q_s^{norm} = \frac{Q_m(D_s, \alpha)}{q_m^{scale}}.$$

11. Compute scenario-level cost components:

$$C_s^{short} = \sum_{i=1}^{n} p_i L_{si},$$
$$C_s^{idle} = \sum_{i=1}^{n} h_i U_{si} + \bar{h} P_s, \qquad \bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i,$$
$$C_s^{coord} = c Q_s^{norm},$$
$$L_s^{sys} = \sum_{i=1}^{n} L_{si},$$
$$C_s^{tail} = \eta (L_s^{sys})^{\gamma}.$$

12. Compute total scenario loss:

$$C_s^{total} = C_s^{short} + C_s^{idle} + C_s^{coord} + C_s^{tail}.$$

13. Estimate performance under α:

$$\hat{\Pi}(\alpha) = -\frac{1}{N}\sum_{s=1}^{N} C_s^{total}.$$

Also record average shortage cost, idle-capacity cost, coordination-friction cost, tail-risk cost, normalized coordination burden, unused shared-pool capacity, and average pooled-allocation effectiveness.

End for.

Return $\hat{\Pi}(\alpha)$ and all recorded cost components for each $\alpha \in \mathcal{A}$.

After Algorithm A1 has been applied over the full grid $\mathcal{A}$, the raw maximizing boundary is obtained as $\arg\max_{\alpha\in\mathcal{A}} \hat{\Pi}(\alpha)$. The improvement of the best bounded-centralization design over FC is computed as $\Delta_{BC-FC} = \max_{\alpha\in(0,1)} \hat{\Pi}(\alpha) - \hat{\Pi}(1)$. The substantive classification then follows the threshold rule in Equation (35). Thus, the threshold affects only the reporting classification; it does not affect the estimated performance values or the raw maximizing value of α.

Appendix A.2. Additional Robustness Results

Table A2 reports the same stress-state imperfection experiment under alternative normalized coordination-friction measures. The baseline measure Q_1 follows the reallocation-intensity interpretation used in the main text. The alternatives capture different aspects of the same coordination problem: Q_2 emphasizes the amount of shared-pool capacity used, Q_3 emphasizes the number of wards brought into pooled allocation, and Q_4 allows reassignment difficulty to vary across wards. This check examines whether the interior-boundary pattern is tied to treating coordination burden only as deviation from nominal ward capacity.

Table A2. Robustness to alternative coordination-friction measures.

Measure	Definition	Raw α^*	Δ_{BC-FC}	Classification
Q_1	$Q_1 = \frac{1}{2}\sum_i \lvert x_i - B_i \rvert$	0.20	4.7138	BC
Q_2	$Q_2 = \sum_i y_i$	0.22	6.1097	BC
Q_3	$Q_3 = \sum_i \mathbf{1}\{y_i > 0\}$	0.20	4.7864	BC
Q_4	$Q_4 = \frac{1}{2}\sum_i \omega_i \lvert x_i - B_i \rvert$	0.20	4.7160	BC

The definitions of Q_1–Q_4 follow (10)–(13). In Table A2, y_i denotes the realized shared-pool allocation to ward i, consistent with Algorithm A1. Δ_{BC-FC} is defined in Equation (34). All Q measures are normalized by their reference expected value under $\alpha = 1$ and $c = 0$, so the reported comparisons are on a common scale. The table reports the baseline stress-imperfection setting. These checks do not establish robustness under every possible coordination-cost specification; they show that the interior-boundary pattern is not eliminated by the alternative normalized Q definitions considered here.

Appendix A.3. Additional Multi-Ward Results

Figure A1 reports the full α-grid performance curves for the multi-ward extensions. These curves supplement the summary statistics reported in the main text. They are included to assess whether the qualitative pattern observed in the two-ward canonical case persists when the number of wards increases.

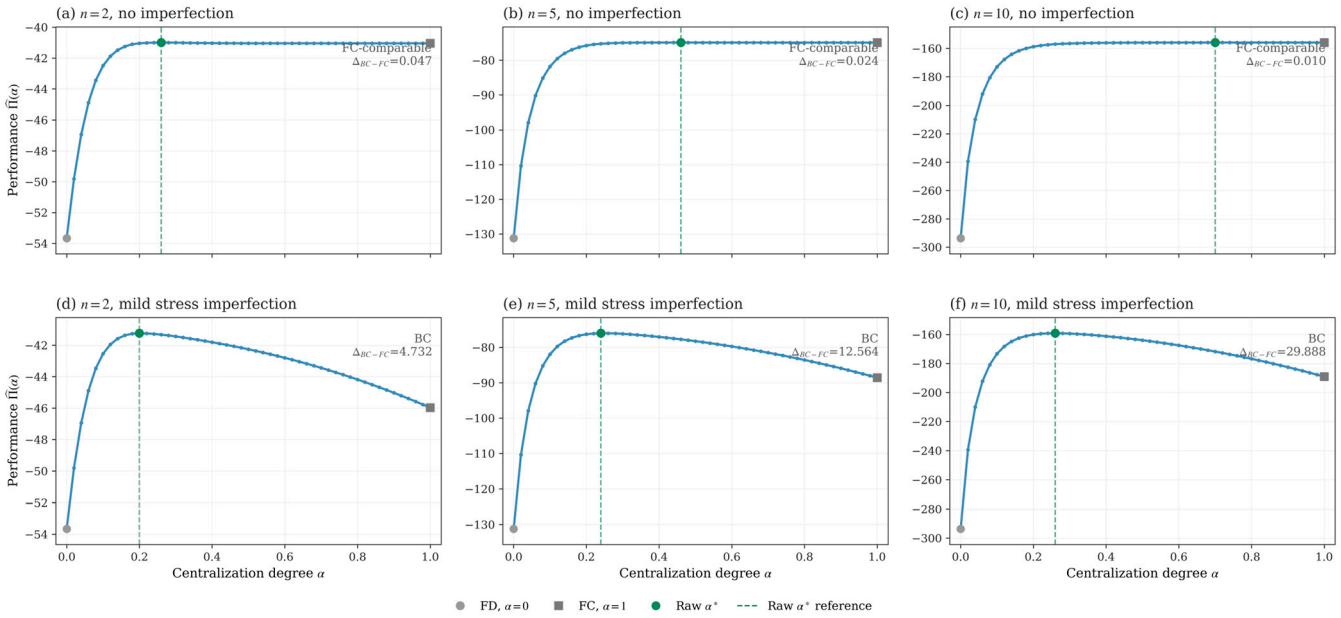


Figure A1. Multi-ward performance curves under fully effective and stress-imperfect pooling. The figure reports $\hat{\Pi}(\alpha)$ over the full centralization grid for multi-ward systems with $n = 2, 5, 10$. Panels (**a**–**c**) assume fully effective pooled allocation, while panels (**d**–**f**) introduce mild stress-state pooled-allocation imperfection. The green marker and dashed vertical line indicate the raw maximizing boundary α^*, and the gray square marks full centralization ($\alpha = 1$). In the fully effective pooling cases, the raw optimizer may lie in the interior, but the improvement over FC remains small and is classified as FC-comparable. Under mild stress imperfection, the interior boundary becomes substantively attractive across all three system sizes. Larger Δ_{BC-FC} values in larger systems partly reflect aggregate system-level cost accounting and should not be interpreted as calibrated hospital-level policy gains.

References

1. Bekker, R.E.; Koeleman, P.M. Scheduling admissions and reducing variability in bed demand. *Health Care Manag. Sci.* **2011**, *14*, 237–249. [CrossRef] [PubMed]
2. De Bruin, A.M.; Bekker, R.E.; Van Zanten, L.; Koole, G.M. Dimensioning hospital wards using the Erlang loss model. *Ann. Oper. Res.* **2010**, *178*, 23–43.
3. Green, L. Queueing analysis in healthcare. In *Patient Flow: Reducing Delay in Healthcare Delivery*; Springer: Boston, MA, USA, 2006; pp. 281–307.
4. Hulshof, P.J.; Kortbeek, N.; Boucherie, R.J.; Hans, E.W.; Bakker, P.J. Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Syst.* **2012**, *1*, 129–175. [CrossRef]
5. Litvak, E.; Buerhaus, P.I.; Davidoff, F.; Long, M.C.; McManus, M.L.; Berwick, D.M. Managing unnecessary variability in patient demand to reduce nursing stress and improve patient safety. *Jt. Comm. J. Qual. Patient Saf.* **2005**, *31*, 330–338. [CrossRef] [PubMed]
6. Rechel, B.; Wright, S.; Barlow, J.; McKee, M. Hospital capacity planning: From measuring stocks to modelling flows. *Bull. World Health Organ.* **2010**, *88*, 632–636. [CrossRef] [PubMed]
7. Cachon, G.E.R.P.; Lariviere, M.A. Capacity choice and allocation: Strategic behavior and supply chain performance. *Manag. Sci.* **1999**, *45*, 1091–1108. [CrossRef]
8. Chen, M.; Lin, C. Effects of centralization on expected costs in a multi-location newsboy problem. *J. Oper. Res. Soc.* **1989**, *40*, 597–602. [CrossRef]
9. Van Mieghem, J.A. Investment strategies for flexible resources. *Manag. Sci.* **1998**, *44*, 1071–1078. [CrossRef]
10. Zamani, H.; Parvaresh, F.; Izady, N.; Farahani, R.Z. Admission, discharge, and transfer control in patient flow logistics: Overview and future research. *Transp. Res. Part E Logist. Transp. Rev.* **2024**, *191*, 103722. [CrossRef]
11. Humphreys, P.; Spratt, B.; Tariverdi, M.; Burdett, R.L.; Cook, D.; Yarlagadda, P.K.; Corry, P. An overview of hospital capacity planning and optimisation. *Healthcare* **2022**, *10*, 826. [CrossRef] [PubMed]
12. Zhalechian, M.; Keyvanshokooh, E.; Shi, C.; Van Oyen, M.P. Data-driven hospital admission control: A learning approach. *Oper. Res.* **2023**, *71*, 2111–2129. [CrossRef]

13. Oakley, D.; Onggo, B.S.; Worthington, D. Symbiotic simulation for the operational management of inpatient beds: Model development and validation using Delta-method: D. Oakley et al. *Health Care Manag. Sci.* **2020**, *23*, 153–169. [PubMed]
14. Melman, G.J.; Parlikad, A.K.; Cameron, E. Balancing scarce hospital resources during the COVID-19 pandemic using discrete-event simulation. *Health Care Manag. Sci.* **2021**, *24*, 356–374. [CrossRef] [PubMed]
15. Kane, E.M.; Scheulen, J.J.; Püttgen, A.; Martinez, D.; Levin, S.; Bush, B.A.; Huffman, L.; Jacobs, M.M.; Rupani, H.; Efron, D.T. Use of systems engineering to design a hospital command center. *Jt. Comm. J. Qual. Patient Saf.* **2019**, *45*, 370–379. [CrossRef] [PubMed]
16. Åhlin, P.; Almström, P.; Wänström, C. Solutions for improved hospital-wide patient flows–a qualitative interview study of leading healthcare providers. *BMC Health Serv. Res.* **2023**, *23*, 17. [PubMed]
17. Song, H.; Tucker, A.L.; Graue, R.; Moravick, S.; Yang, J.J. Capacity pooling in hospitals: The hidden consequences of off-service placement. *Manag. Sci.* **2020**, *66*, 3825–3842. [CrossRef]
18. Braaksma, A.; Copenhaver, M.S.; Zenteno, A.C.; Ugarph, E.; Levi, R.; Daily, B.J.; Orcutt, B.; Turcotte, K.M.; Dunn, P.F. Evaluation and implementation of a Just-In-Time bed-assignment strategy to reduce wait times for surgical inpatients. *Health Care Manag. Sci.* **2023**, *26*, 501–515. [PubMed]
19. Mebrahtu, T.F.; McInerney, C.D.; Benn, J.; McCrorie, C.; Granger, J.; Lawton, T.; Sheikh, N.; Habli, I.; Randell, R.; Johnson, O. The impact of hospital command centre on patient flow and data quality: Findings from the UK National Health Service. *Int. J. Qual. Health Care* **2023**, *35*, mzad072. [CrossRef] [PubMed]
20. Mebrahtu, T.F.; McInerney, C.D.; Benn, J.; McCrorie, C.; Granger, J.; Lawton, T.; Sheikh, N.; Randell, R.; Habli, I.; Johnson, O.A. Effect of a hospital command centre on patient safety: An interrupted time series study. *BMJ Health Care Inform.* **2023**, *30*, e100653. [CrossRef] [PubMed]
21. Johnson, O.A.; McCrorie, C.; McInerney, C.; Mebrahtu, T.F.; Granger, J.; Sheikh, N.; Lawton, T.; Habli, I.; Randell, R.; Benn, J. Implementing an artificial intelligence command centre in the NHS: A mixed-methods study. *Health Soc. Care Deliv. Res.* **2024**, *12*, 1–108. [CrossRef] [PubMed]
22. Cao, P.; He, S.; Huang, J.; Liu, Y. To pool or not to pool: Queueing design for large-scale service systems. *Oper. Res.* **2021**, *69*, 1866–1885. [CrossRef]
23. Zhong, Y.; Zheng, Z.; Chou, M.C.; Teo, C. Resource pooling and allocation policies to deliver differentiated service. *Manag. Sci.* **2018**, *64*, 1555–1573. [CrossRef]
24. Jiang, Y.; Abouee Mehrizi, H.; Van Mieghem, J.A. Geographic virtual pooling of hospital resources: Data-driven trade-off between waiting and traveling. *Manuf. Serv. Oper. Manag.* **2023**, *25*, 1527–1544. [CrossRef]
25. Song, H.; Armony, M.; Roels, G. Queue configurations and operational performance: An interplay between customer ownership and queue length awareness. *Manuf. Serv. Oper. Manag.* **2024**, *26*, 2284–2304. [CrossRef]
26. Bernstein, E.S.; Shore, J.C.; Jang, A.J. Network centralization and collective adaptability to a shifting environment. *Organ. Sci.* **2023**, *34*, 2064–2096. [CrossRef]
27. Essuman, D.; Bruce, P.A.; Ataburo, H.; Asiedu-Appiah, F.; Boso, N. Linking resource slack to operational resilience: Integration of resource-based and attention-based perspectives. *Int. J. Prod. Econ.* **2022**, *254*, 108652.
28. Marczyk, C.E.S.; Saurin, T.A.; Bulhões, I.R.; Patriarca, R.; Bilotta, F. Slack in the infrastructure of intensive care units: Resilience management in the post-pandemic era. *BMC Health Serv. Res.* **2023**, *23*, 579. [CrossRef] [PubMed]
29. Hasan, M.K.; Nasrullah, S.M.; Quattrocchi, A.; González, P.A.; Castro-Delgado, R. Hospital surge capacity preparedness in disasters and emergencies: A systematic review. *Public Health* **2023**, *225*, 12–21. [CrossRef] [PubMed]
30. Mahmoud, H.; Kirsch, T.; O'Neil, D.; Anderson, S. The resilience of health care systems following major disruptive events: Current practice and a path forward. *Reliab. Eng. Syst. Safe* **2023**, *235*, 109264. [CrossRef]
31. Lim, J.M.; Song, H.; Yang, J.J. The spillover effects of capacity pooling in hospitals. *Manag. Sci.* **2024**, *70*, 7692–7711. [CrossRef]
32. Izady, N.; Mohamed, I. A clustered overflow configuration of inpatient beds in hospitals. *Manuf. Serv. Oper. Manag.* **2021**, *23*, 139–154. [CrossRef]
33. Gong, X.; Wang, X.; Zhou, L.; Geng, N. Managing hospital inpatient beds under clustered overflow configuration. *Comput. Oper. Res.* **2022**, *148*, 106021. [CrossRef]
34. Izady, N.; Arabzadeh, B.; Sands, N.; Adams, J. Reconfiguration of inpatient services to reduce bed pressure in hospitals. *Eur. J. Oper. Res.* **2024**, *316*, 680–693. [CrossRef]
35. Lovett, P.B.; Illg, M.L.; Sweeney, B.E. A Successful Model for a Comprehensive Patient Flow Management Center at an Academic Health System. *Am. J. Med. Qual.* **2016**, *31*, 246–255. [PubMed]
36. Martinez, D.A.; Kane, E.M.; Jalalpour, M.; Scheulen, J.; Rupani, H.; Toteja, R.; Barbara, C.; Bush, B.; Levin, S.R. An electronic dashboard to monitor patient flow at the Johns Hopkins Hospital: Communication of key performance indicators using the Donabedian model. *J. Med. Syst.* **2018**, *42*, 133. [CrossRef] [PubMed]
37. Stowell, A.; Claret, P.; Sebbane, M.; Bobbia, X.; Boyard, C.; Genre Grandpierre, R.; Moreau, A.; de La Coussaye, J. Hospital out-lying through lack of beds and its impact on care and patient outcome. *Scand. J. Trauma Resusc. Emerg. Med.* **2013**, *21*, 17. [CrossRef] [PubMed]

38. Liu, J.; Griesman, J.; Nisenbaum, R.; Bell, C.M. Quality of care of hospitalized internal medicine patients bedspaced to non-internal medicine inpatient units. *PLoS ONE* **2014**, *9*, e106763. [CrossRef] [PubMed]
39. Stylianou, N.; Fackrell, R.; Vasilakis, C. Are medical outliers associated with worse patient outcomes? A retrospective study within a regional NHS hospital using routine data. *BMJ Open* **2017**, *7*, e015676. [CrossRef] [PubMed]
40. Seo, H.; Ahn, I.; Gwon, H.; Kang, H.; Kim, Y.; Choi, H.; Kim, M.; Han, J.; Kee, G.; Park, S. Forecasting Hospital Room and Ward Occupancy Using Static and Dynamic Information Concurrently: Retrospective Single-Center Cohort Study. *JMIR Med. Inform.* **2024**, *12*, e53400. [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.