Dynamic Supervisory Allocation in Pooled-Intake Mental Health Systems

David Medina Rodríguez[*†]

Abstract

This paper studies pooled-intake mental health systems in which scarce supervisor time must be allocated across entrant screening, panel opening, and incumbent support. I develop a dynamic model of access expansion with a one-period opening lag: supervisor effort devoted to screening and onboarding is costly in the current period, while the resulting gains in treatment capacity, entrant composition, incumbent support capital, and backlog reduction materialize only in the following period. This timing generates an intertemporal trade-off between expanding access and preserving the quality of care for incumbent patients. Under pooled intake and protocolized panels, clinicians are ranked by posterior expected treatment value, and a static benchmark characterizes the marginal staffing rule as a function of screening precision, incumbent support, and supervisory scarcity. A dynamic planner then allocates a fixed supervisor-time budget across screening, opening, and maintenance, yielding comparative statics for panel opening, staged expansion, and responses to overload. These effects are strongest when support retention, opening-to-capacity delays, and backlog persistence are quantitatively important. The paper's contribution is organizational: in centralized behavioral-health networks, university counseling centers, and related protocolized access settings, expansion may be constrained not only by the value of marginal entrants, but also by deterioration in incumbent care when supervisory attention is reallocated too rapidly.

Keywords: mental health, provider networks, credentialing, capacity constraints, access and quality, welfare

JEL Codes: I11, I18, D82, L15

[*]Assistant Professor, Department of Economics, CUNEF Universidad. Email: david.medina@cunef.edu

[†]**Declaration.** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The author declares no competing interests. No empirical data were used in this study.

1 Introduction

In many mental-health delivery systems, access expansion is constrained less by reimbursement than by internal supervisory capacity. Behavioral-health insurers expand contracted panels through centralized credentialing teams, public outpatient systems open clinician panels through common intake cells, employee-assistance programs route referrals through pooled review teams, and university counseling centers use centralized screening and onboarding to respond to demand surges. In these settings, adding clinicians is not enough. Senior staff must review entrants, activate panels, and continue supporting incumbent clinicians through consultation, documentation review, escalation handling, and quality assurance. When those tasks compete for the same supervisory time, expanding access can come at the cost of incumbent care.

This paper studies that tradeoff in a dynamic model of pooled intake with standardized panels. The central organizational problem is not only whether to expand access, but how scarce supervisor time should be allocated across entrant screening, panel opening, and incumbent support. The key friction is temporal. Screening and onboarding absorb supervisor effort in the current period, but newly opened panels become usable only in the next period. Support effort is also intertemporal: it helps preserve future support capital and retained treatment quality rather than delivering a free contemporaneous gain. As a result, the organization faces an intertemporal tradeoff between rapid expansion and preservation of the installed stock of care.

The model combines three ingredients. First, a screening block characterizes staffing under pooled intake and protocolized assignment. Clinicians are ranked by posterior expected treatment value, so screening precision shapes the value of opening additional panels. Second, a static benchmark clarifies the marginal staffing condition: the value of additional treated load depends not only on the expected value of the marginal clinician, but also on the shadow cost of the supervisory time required to screen, onboard, and maintain that clinician. Third, a dynamic planner allocates a fixed supervisor-time budget across screening, panel opening, and incumbent maintenance when capacity creation is delayed, backlog evolves endogenously, and incumbent quality depends on support retention.

The main implication is that a system facing excess demand may rationally delay or stage expansion even when the next panel appears socially valuable in static terms. Opening more panels today diverts supervisory effort away from incumbent support, weakens future retained quality, and relieves backlog only with a lag. Expansion is therefore limited not only by the value of marginal entrants, but also by the deterioration in incumbent care that follows when

supervisory attention is reallocated too aggressively. The model delivers comparative statics for panel opening, staged expansion, and overload response, and these forces are strongest when support retention, opening delays, and backlog persistence are quantitatively important.

The analysis is intended for centralized or pooled-intake environments with standardized panel norms, compressed reimbursement, and meaningful onboarding bottlenecks. That description fits some behavioral-health networks, pooled tele-mental-health systems, employee-assistance intake models, public outpatient settings, and university counseling centers. It fits less well where therapist–patient matching is rich, clinician participation is decentralized, or service lines are highly specialized. Within that narrower domain, the emphasis is organizational: supervisory attention is a scarce input that jointly determines entrant selection, capacity creation, maintenance of incumbent quality, and the speed with which backlogs can be reduced.

This focus is consistent with several documented features of mental-health delivery. Reviews of centralized intake models emphasize pooled referral queues, standardized screening, and common routing protocols in non-acute services (Isaacs et al., 2023). Rapid-intake implementations in hospital-based and academic behavioral-health settings document waiting backlogs, referral-processing bottlenecks, and the use of dedicated intake teams to expand access once assessment and supervisory effort are reorganized (McMahan et al., 2022; Oneschuk et al., 2026). The supervision literature likewise treats senior review, fidelity monitoring, and case consultation as scarce organizational inputs that shape quality assurance rather than as costless background activities (Rabie et al., 2023; Center for Substance Abuse Treatment, 2009).

The paper relates to several strands of health economics. Work on quality and incentives emphasizes that delivery institutions matter when quality is difficult to observe and prices are compressed (Arrow, 1963; Chalkley and Malcomson, 1998; Gravelle and Masiero, 2000; Gravelle and Sivey, 2010; Brekke et al., 2010; Mak, 2018; Brosig-Koch et al., 2024). Work on provider networks and access shows that provider quality, network breadth, and contracting frictions shape welfare and effective capacity (Ho, 2009; Atwood and Lo Sasso, 2016; Drake, 2019; Kreider et al., 2024; Chandra et al., 2016). Work on congestion studies waiting and capacity pressure in regulated systems (Brekke et al., 2008; Chen et al., 2016; Brekke et al., 2021). This paper complements those literatures by focusing on a distinct internal margin of organization: the allocation of scarce supervisor time across entrant selection, panel creation, incumbent maintenance, and backlog management.

The rest of the paper is organized as follows. Section 2 sets up the institutional environment and develops the screening block. Section 3 presents the static marginal staffing benchmark.

Section 4 introduces the dynamic supervisory-allocation problem and derives the main comparative-statics results. Section 5 studies overload episodes and derives a backlog threshold for expansion response. Section 6 discusses robustness to limited panel heterogeneity. Section 8 concludes.

2 Institutional environment

2.1 Patients, clinicians, and protocolized panels

A managed mental-health system faces an exogenous treatment load $D > 0$, interpreted as the mass of arriving cases. Patient severity is indexed by

$$s \in \mathcal{S} \equiv [\underline{s}, \bar{s}],$$

with cumulative distribution function F and density $f > 0$ on $\mathcal{S}$.

There is also a unit mass of potential clinicians. Each clinician has latent quality

$$a \in A \equiv [\underline{a}, \bar{a}],$$

with cumulative distribution function G and density $g > 0$ on A.

If a patient of severity s is treated by clinician a and receives clinician time $t \geq 0$, treatment succeeds with probability

$$p(a, t, s) = \phi(a)u(t, s), \tag{1}$$

where $\phi(a) > 0$ captures clinician quality and $u(t, s)$ is the treatment technology. The multiplicative structure implies that clinician quality scales the productivity of treatment time uniformly across cases. This assumption is restrictive, but it is natural in pooled, protocolized environments and yields a scalar staffing object.

A successful treatment for severity s yields private benefit $B(s) > 0$ and social spillover $X(s) \geq 0$. Relative to no treatment, an unsuccessful treatment yields downside cost $C(s) \geq 0$. Define

$$\Psi(s) \equiv B(s) + \beta X(s) + C(s), \tag{2}$$

where $\beta \geq 0$ scales the social value of spillovers. Expected social value relative to no treatment

is therefore

$$v(a,t,s) = p(a,t,s)\big[B(s) + \beta X(s)\big] - \big(1 - p(a,t,s)\big)C(s) = \Psi(s)\phi(a)u(t,s) - C(s). \quad (3)$$

Each active clinician has one unit of time. Under the system's standard panel design, a fully staffed clinician treats $\kappa > 0$ patients. This panel size summarizes visit-frequency norms, reimbursement templates, supervision requirements, and internal caseload rules.

The system uses pooled intake and protocolized assignment, so active clinicians face the same severity distribution F. This assumption is intended for centralized-intake or stepped-care settings in which assignment margins are compressed. Section 6 relaxes it.

Fix a clinician of quality a facing panel size $n > 0$. The system chooses a severity-contingent time schedule $t(s) \geq 0$ to maximize expected treatment value:

$$m(a;n) \equiv \max_{t(\cdot)\geq 0} \left\{ \int_{\mathcal{S}} \big[\Psi(s)\phi(a)u(t(s),s) - C(s)\big]\, dF(s) : \; n \int_{\mathcal{S}} t(s)\, dF(s) \leq 1 \right\}. \quad (4)$$

The benchmark uses $n = \kappa$; Section 5 allows $n > \kappa$.

Assumption 1 (Technology and stakes)**.** *The function $\phi : A \to \mathbb{R}_{++}$ is continuously differentiable and strictly increasing. The functions B, X, and C are continuously differentiable on $\mathcal{S}$, and Ψ defined in (2) is strictly positive and weakly increasing. The treatment technology $u : \mathbb{R}_+ \times \mathcal{S} \to \mathbb{R}_+$ is twice continuously differentiable and satisfies*

$$u_t(t,s) > 0, \qquad u_{tt}(t,s) < 0, \qquad u_s(t,s) < 0, \qquad u_{ts}(t,s) > 0$$

for all $(t,s) \in \mathbb{R}_{++} \times \mathcal{S}$. For every s, $\lim_{t\downarrow 0} u_t(t,s) = \infty$ and $\lim_{t\uparrow\infty} u_t(t,s) = 0$. Finally, $0 < \phi(a)u(t,s) < 1$ on the feasible domain.

Proposition 1 (Common treatment protocol)**.** *Under Assumption 1, for every panel size $n > 0$ there exists a unique optimal schedule $t^*(\cdot;n)$ and a unique shadow value $\mu^*(n) > 0$ such that*

$$\Psi(s)u_t\big(t^*(s;n), s\big) = \mu^*(n) \qquad \text{for all } s \in \mathcal{S}, \quad (5)$$

and

$$n \int_{\mathcal{S}} t^*(s;n)\, dF(s) = 1. \quad (6)$$

Moreover,

$$\frac{\partial t^*(s;n)}{\partial s} > 0, \qquad \frac{\partial t^*(s;n)}{\partial n} < 0. \quad (7)$$

The optimal schedule is independent of clinician quality a.

Proposition 1 implies that more severe cases receive more time, while all active clinicians follow the same severity-contingent protocol. A larger caseload compresses time per case at every severity.

Define case-specific value and expected per-patient value under the optimal protocol by

$$\widetilde{m}(a, s; n) \equiv \Psi(s)\phi(a)u\big(t^*(s; n), s\big) - C(s), \tag{8}$$

and

$$m(a; n) \equiv \int_{\mathcal{S}} \widetilde{m}(a, s; n)\, dF(s). \tag{9}$$

Proposition 2 (Ranking and overload effects)**.** *Under Assumption 1, for every $n > 0$ and every $s \in \mathcal{S}$,*

$$\frac{\partial \widetilde{m}(a, s; n)}{\partial a} = \phi'(a)\Psi(s)u\big(t^*(s; n), s\big) > 0, \qquad \frac{\partial \widetilde{m}(a, s; n)}{\partial n} = \Psi(s)\phi(a)u_t\big(t^*(s; n), s\big)\frac{\partial t^*(s; n)}{\partial n} < 0. \tag{10}$$

Hence $m(a; n)$ is strictly increasing in clinician quality and strictly decreasing in caseload.

2.2 Screening, posterior ranking, and the supervisor bottleneck

Clinician quality a is not directly observed. Instead, the system observes a credentialing signal

$$z \in Z \equiv [\underline{z}, \bar{z}],$$

generated by a screening technology indexed by precision parameter $\rho \in \mathcal{P}$. The conditional density of z given a is $k_\rho(z \mid a)$. A higher ρ corresponds to a more informative screening process, such as richer use of prior outcomes, more intensive audit review, or more detailed onboarding interviews.

For panel size n, define posterior expected per-patient value

$$M_\rho(z; n) \equiv \mathbb{E}_\rho\big[m(a; n) \mid z\big]. \tag{11}$$

The system staffs clinicians in descending order of posterior expected value.

Assumption 2 (Informative screening and screening time)**.** *For each $\rho \in \mathcal{P}$, the signal density $k_\rho(z \mid a)$ is continuous, has full support on Z, and satisfies the monotone likelihood ratio property in (z, a). The family $\{k_\rho\}_{\rho \in \mathcal{P}}$ is ordered by Blackwell informativeness. Implementing screening technology ρ requires supervisory time $s(\rho) \geq 0$ per clinician screened, where $s(\cdot)$ is weakly increasing and differentiable.*

Assumption 2 implies that better signals improve ranking but also consume more supervisory time.

Proposition 3 (Posterior ranking and gross screening gains)**.** *Under Assumptions 1 and 2, posterior expected per-patient value $M_\rho(z;n)$ is continuously differentiable and strictly increasing in z for every $n > 0$. Moreover, if $\rho' \succ \rho$ in the Blackwell order, then for any staffed load $D \in (0, \kappa)$,*

$$\widetilde{W}^I_{\rho'}(D) \geq \widetilde{W}^I_{\rho}(D), \tag{12}$$

where

$$\widetilde{W}^I_{\rho}(D) \equiv \kappa \int_{z^*_\rho(D)}^{\bar{z}} M_\rho(z;\kappa) h_\rho(z)\, dz$$

is gross staffed value before panel-opening cost, with strict inequality whenever the more informative technology changes the ranking of a positive-measure set of clinicians.

Opening a panel also uses organizational resources. Let $c \geq 0$ denote fixed administrative cost and let $s_0 \geq 0$ denote supervisory onboarding time per opened panel. Let $e \geq 0$ denote incumbent support intensity per active panel. This support includes consultation, chart review, documentation assistance, escalation review, and protocol monitoring. In reduced form, support scales effective per-patient value through a common quality-maintenance factor $\Gamma(e)$ satisfying

$$\Gamma(0) = 1, \qquad \Gamma'(e) > 0, \qquad \Gamma''(e) < 0. \tag{13}$$

Thus

$$m(a;n,e) \equiv \Gamma(e)\, m(a;n), \qquad M_\rho(z;n,e) \equiv \Gamma(e)\, M_\rho(z;n). \tag{14}$$

Because $\Gamma(e)$ scales value uniformly, it does not affect the staffing ranking, but it introduces a competing use of supervisor time.

Let $\bar{S}$ denote per-period supervisory capacity. If the system opens P panels in a period using screening precision ρ while maintaining support intensity e on those panels, the supervisory constraint is

$$[s(\rho) + s_0 + e]P \leq \bar{S}. \tag{15}$$

Let ω denote the shadow value of this constraint. The implied per-panel resource cost is

$$F(\rho, e) \equiv c + \omega[s(\rho) + s_0 + e]. \tag{16}$$

Better screening can improve selection while increasing the shadow cost of panel creation, and stronger incumbent support can raise the value of active care while tightening the same bottleneck.

2.3 Scope of the benchmark

The benchmark rests on a small set of assumptions. Active clinicians face the same severity distribution under pooled intake. The system chooses a common severity-contingent treatment protocol. Clinician quality scales the productivity of time uniformly across cases. Staffing is based on posterior expected per-patient value induced by the screening technology. Incumbent support scales value through a common factor $\Gamma(e)$. Each new panel contributes κ treatment slots and uses screening, onboarding, and support time additively. If these assumptions are relaxed, the formula below need not remain exact. Sections 5 and 6 show how overload and limited panel heterogeneity modify the benchmark.

3 Static accounting benchmark

Fix screening precision ρ, incumbent support intensity e, and the current shadow value of supervisory capacity ω. The system faces treatment load $D \in (0, \bar{D})$ and must open enough standard panels to serve it, where feasible capacity is

$$\bar{D} \equiv \kappa \int_{\underline{z}}^{\bar{z}} h_\rho(z)\, dz = \kappa,$$

and h_ρ is the density of the credentialing signal induced by screening technology ρ. Under the standard panel design, the system staffs all clinicians with signal above a cutoff $z_\rho^*(D)$ satisfying

$$D = \kappa \int_{z_\rho^*(D)}^{\bar{z}} h_\rho(z)\, dz. \tag{17}$$

This is a staffing rule rather than a decentralized equilibrium: the system screens and contracts clinicians until enough panels have been created to meet the service target.

Institutional welfare net of panel-opening cost is

$$W_\rho^I(D; \omega, e) \equiv \kappa \int_{z_\rho^*(D)}^{\bar{z}} M_\rho(z; \kappa, e) h_\rho(z)\, dz - F(\rho, e) \int_{z_\rho^*(D)}^{\bar{z}} h_\rho(z)\, dz. \tag{18}$$

Define average net treatment value by

$$\bar{m}_\rho^I(D; \omega, e) \equiv \frac{W_\rho^I(D; \omega, e)}{D}.$$

Proposition 4 (Static marginal staffing formula)**.** *Under Assumptions 1 and 2, for every* (ρ, e)

and every $D \in (0, \bar{D})$, there exists a unique staffing cutoff $z_\rho^*(D)$, continuously differentiable with

$$\frac{dz_\rho^*(D)}{dD} = -\frac{1}{\kappa h_\rho(z_\rho^*(D))} < 0. \tag{19}$$

Moreover,

$$\frac{\partial W_\rho^I(D;\omega,e)}{\partial D} = M_\rho(z_\rho^*(D);\kappa,e) - \frac{F(\rho,e)}{\kappa}, \tag{20}$$

and

$$\frac{\partial \bar{m}_\rho^I(D;\omega,e)}{\partial D} = \frac{M_\rho(z_\rho^*(D);\kappa,e) - F(\rho,e)/\kappa - \bar{m}_\rho^I(D;\omega,e)}{D} < 0. \tag{21}$$

If there exists a unique $\hat{z}_{\rho,e} \in (\underline{z}, \bar{z})$ such that

$$M_\rho(\hat{z}_{\rho,e};\kappa,e) = \frac{F(\rho,e)}{\kappa}, \tag{22}$$

then institutional welfare is hump-shaped, with peak load

$$D_{\rho,e}^I(\omega) \equiv \kappa \int_{\hat{z}_{\rho,e}}^{\bar{z}} h_\rho(z)\, dz. \tag{23}$$

Proposition 4 provides a static accounting benchmark. Holding fixed screening precision, incumbent support, and supervisory scarcity, the marginal value of expanding access equals the posterior expected value of the marginal staffed panel net of the full per-panel resource cost. The dynamic problem in the next section endogenizes those objects by allowing supervisory time to be allocated jointly across screening, panel opening, and maintenance.

4 Dynamic supervisory allocation

The static benchmark treats screening precision, incumbent support, and the shadow value of supervisory capacity as fixed. This section endogenizes them in a dynamic organizational problem. The central friction is that the same supervisor-time budget must be allocated across entrant screening, panel opening, and incumbent maintenance, while the benefits of opening arrive only with a lag. The reduced-form structure is designed to capture three margins that are central in pooled-intake systems: support retention, the opening-to-capacity lag, and backlog persistence.

4.1 Mechanism at a glance

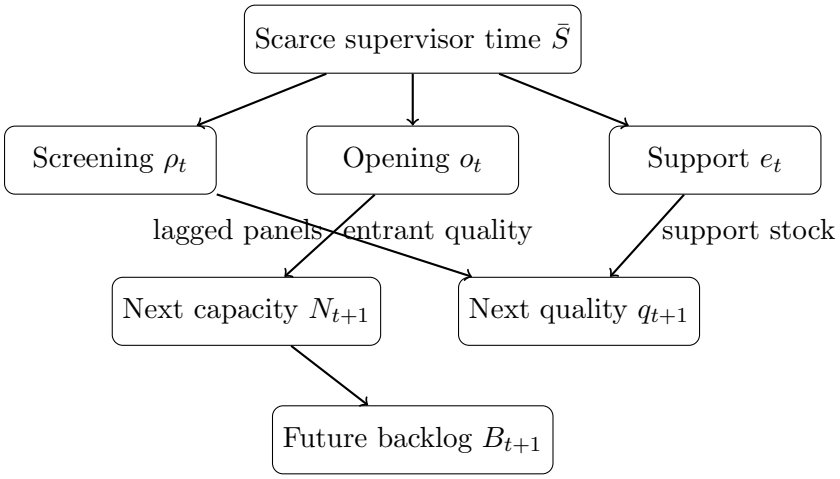


Figure 1. Supervisor time is allocated across screening, panel opening, and incumbent support. Opening raises capacity with a one-period lag; screening and support shape future quality; backlog relief arrives through future capacity rather than current opening.

The dynamic problem has four margins. Screening precision determines the informational content of entrant review and therefore the quality of newly opened panels. Opening uses scarce supervisory time and raises next-period capacity only with a lag. Incumbent support uses the same resource to preserve the quality of the installed stock of care. Unresolved demand becomes backlog, which generates waiting harm and attrition if expansion is delayed. The planner therefore chooses not only whether the next panel is valuable, but whether supervisory time should be reallocated today away from incumbents and toward entrants given its implications for tomorrow's quality and queue.

4.2 State variables, controls, and laws of motion

Time is discrete. At date t, the system is summarized by the state

$$x_t \equiv (N_t, q_t, g_t, B_t, \theta_t),$$

where $N_t > 0$ is the mass of active panels, $q_t > 0$ is average incumbent per-patient value under benchmark caseload κ, $g_t \geq 0$ is inherited support capital, $B_t \geq 0$ is backlog carried into the period, and $\theta_t \in \Theta$ is a persistent demand state. Flow demand is $D_t = D(\theta_t)$, where θ_t follows a Markov chain with transition kernel $P(\theta' \mid \theta)$.

Each period the planner chooses

$$o_t \geq 0 \quad \text{(new panels initiated)}, \qquad \rho_t \in \mathcal{P} \quad \text{(screening depth)}, \qquad e_t \geq 0 \quad \text{(support effort)}.$$

A panel initiated at date t becomes active at $t+1$. The common supervisor-time constraint is

$$[s(\rho_t) + s_0]o_t + e_t N_t \leq \bar{S}, \tag{24}$$

where $\bar{S}$ is effective supervisory capacity per period.

Incumbents can temporarily absorb higher caseloads up to overload cap $\bar{\kappa} > \kappa$. Treated load is therefore

$$Q_t = \min\{D_t + B_t, \bar{\kappa}N_t\}, \qquad n_t = Q_t/N_t \in [\kappa, \bar{\kappa}]. \tag{25}$$

Let $W_t \equiv [D_t + B_t - Q_t]_+$ denote unresolved backlog at the end of the period. Current flow value is

$$\pi(x_t, o_t, \rho_t, e_t) = \Gamma(g_t)\Lambda(n_t)Q_t q_t - h(W_t) - a(W_t) - co_t, \tag{26}$$

where $\Lambda(\kappa) = 1$, $\Lambda'(n) < 0$, and $\Lambda''(n) \leq 0$ capture overload losses from time compression; $h(\cdot)$ is waiting harm; and $a(\cdot)$ is attrition or deterioration loss.

The state evolves according to

$$N_{t+1} = (1 - \delta_N)N_t + o_t, \tag{27}$$

$$B_{t+1} = (1 - \lambda)W_t, \tag{28}$$

$$g_{t+1} = (1 - \delta_g)g_t + \psi e_t, \tag{29}$$

and

$$q_{t+1} = \frac{(1 - \delta_N)N_t\, \Xi(g_t, n_t)q_t + o_t\, m^E(\rho_t, o_t)}{N_{t+1}}. \tag{30}$$

Here $\Xi(g, n)$ is incumbent-quality retention, with $\Xi_g > 0$, $\Xi_n < 0$, and $0 < \Xi \leq 1$; and $m^E(\rho, o)$ is the average benchmark value of newly opened panels, increasing in screening precision and weakly decreasing in the scale of expansion.

The entrant-value block is inherited directly from the screening environment. Suppose each period brings a normalized applicant pool of unit mass, each successful opening uses one clinician, and the organization staffs the upper tail of posterior expected benchmark value.

For any feasible opening scale $o_t \in (0, 1]$, let $z^o_{\rho_t}(o_t)$ solve

$$o_t = \int_{z^o_{\rho_t}(o_t)}^{\bar{z}} h_{\rho_t}(z)\, dz. \tag{31}$$

Then define the average benchmark value of newly opened panels by

$$m^E(\rho_t, o_t) \equiv \frac{1}{o_t} \int_{z^o_{\rho_t}(o_t)}^{\bar{z}} M_{\rho_t}(z; \kappa)\, h_{\rho_t}(z)\, dz. \tag{32}$$

Under Assumptions 1 and 2, m^E is increasing in ρ_t and weakly decreasing in o_t: more informative screening raises the mean value of the staffed upper tail, while larger opening lowers the cutoff.

The planner's Bellman problem is

$$V(x_t) = \max_{o_t, \rho_t, e_t} \Big\{ \pi(x_t, o_t, \rho_t, e_t) + \beta \mathbb{E}\big[V(x_{t+1}) \mid x_t\big] \Big\} \quad \text{s.t.} \quad (24)\text{–}(30). \tag{33}$$

4.3 Interpretation of the dynamic backbone

The dynamic structure separates current support effort from inherited support capital. In the static benchmark, support is a contemporaneous flow variable. Here, e_t is current support effort, while g_t is the stock inherited from prior periods. Current flow value depends on g_t through $\Gamma(g_t)$, and current support effort affects the future through $g_{t+1} = (1 - \delta_g)g_t + \psi e_t$. This timing makes the intertemporal support margin explicit.

The remaining objects play distinct roles. Screening precision ρ_t affects the informational quality of entrant review and therefore the value of newly opened panels through $m^E(\rho_t, o_t)$. Opening o_t uses onboarding time and expands future capacity through (27). Incumbent support e_t preserves future quality through support capital. The retention term $\Xi(g_t, n_t)$ governs how much of current incumbent quality survives into the next period once support and overload are realized. These margins are analytically distinct even though they all draw on the same supervisory budget.

The results should be read in that spirit. The primitive frictions are the shared supervisory constraint, the opening lag, and the sensitivity of incumbent quality to overload and weak support. Linear support accumulation, scalar backlog, and average-quality retention are reduced forms used to represent those frictions. The analysis therefore emphasizes conditional comparative statics rather than a full structural queueing model.

4.4 Dynamic allocation conditions

Let ν_t denote the multiplier on (24). The first-order conditions characterize how supervisor time is allocated once current and continuation values are both relevant.

Proposition 5 (Dynamic supervisory allocation)**.** *Suppose an interior optimum exists and V is differentiable. Then the dynamic planner equalizes the marginal value of supervisor time across entrant screening, panel creation, and incumbent support:*

$$\nu_t s'(\rho_t) o_t = \beta \mathbb{E}_t \left[V_q(x_{t+1}) \frac{\partial q_{t+1}}{\partial \rho_t} \right], \tag{34}$$

$$\nu_t N_t = \beta \mathbb{E}_t \left[V_g(x_{t+1}) \psi + V_q(x_{t+1}) \frac{\partial q_{t+1}}{\partial e_t} \right], \tag{35}$$

and

$$c + \nu_t [s(\rho_t) + s_0] = \beta \mathbb{E}_t \left[V_N(x_{t+1}) + V_q(x_{t+1}) \frac{\partial q_{t+1}}{\partial o_t} \right]. \tag{36}$$

Equations (34)–(36) are the core allocation conditions. Screening is increased until its continuation-value benefit through entrant quality equals its use of supervisor time. Support is valued because it builds future support capital and protects future clinical quality. Opening is valuable only through the continuation problem: it changes future capacity and the composition of future care. Because new panels become usable only next period, opening does not reduce current backlog directly.

Proposition 6 (Conditional monotone comparative statics of opening)**.** *Suppose h and a are increasing and convex, m^E has increasing differences in (ρ, o), and the continuation value is supermodular in $(N, -B, g, \theta)$. Then the optimal opening rule $o_t^*(x_t)$ is weakly increasing in current backlog B_t, in the persistence of high-demand states, and in the slope of waiting harm and attrition; and it is weakly decreasing in incumbent-quality fragility, as measured by $|\Xi_n|$ and δ_g.*

Proposition 6 is a comparative-statics result on reallocation. When waiting is more costly or demand shocks are more persistent, the system opens sooner; when incumbent quality is more fragile to overload and weak support, it opens more cautiously.

Proposition 7 (Staged opening)**.** *Suppose h and a are strictly convex, $m^E(\rho, o)$ is concave in o, and the current marginal cost of diverting supervisor time toward entrants rises with o because incumbent support is sufficiently fragile. Then there exists a nonempty region of the*

state space in which the dynamic optimum satisfies

$$0 < o_t^*(x_t) < \bar{o},$$

so that the system opens panels gradually rather than fully responding in one period. This staged-opening region expands when shocks are less persistent and when incumbent support is more valuable at the margin.

The proposition is qualitative rather than universal. The point is not that gradual expansion is always optimal, but that it can arise for a structural reason in pooled-intake systems: opening many panels at once consumes the same supervisory input needed to maintain the installed stock of care.

5 Dynamic overload and backlog response

This section specializes the dynamic program to an overload episode and makes the expansion rule explicit.

5.1 A persistent-shock specialization

Consider two demand states, $\theta \in \{L, H\}$, with $D(H) > \kappa N_t > D(L)$ and persistence parameter

$$p \equiv P(H \mid H) \in (0, 1).$$

When the system enters state H, incumbents operate at overload caseload $n_t > \kappa$, backlog accumulates according to (28), and the planner chooses whether to open panels immediately or wait. Newly opened panels become active next period.

Let

$$\Delta_t(o) \equiv \beta \mathbb{E}_t\big[V(x_{t+1} \mid o) - V(x_{t+1} \mid 0)\big] - co - \nu_t[s(\rho_t) + s_0]o \tag{37}$$

denote the net dynamic gain from opening o panels during the overload spell, holding screening policy fixed.

Proposition 8 (Dynamic overload threshold)**.** *In the two-state specialization, there exists a threshold rule*

$$B_t \geq B^*(g_t, N_t, p) \quad \Longrightarrow \quad o_t^*(x_t) > 0,$$

where $B^*(g, N, p)$ is weakly decreasing in shock persistence p, waiting harm, and attrition, and weakly increasing in incumbent-quality fragility and supervisor scarcity. Moreover, if p is sufficiently low and g_t is sufficiently low, then $B^*(g_t, N_t, p)$ exceeds the current backlog, so the planner optimally tolerates an additional period of overload even when the static marginal value of expansion is positive.

Proposition 8 shows that a positive static marginal value of opening need not imply immediate expansion. The planner compares the future benefit of additional capacity with the intertemporal cost of diverting supervisory effort away from incumbent support and with the option value of waiting for the demand shock to dissipate.

In the mental-health settings motivating the paper, $h(W_t)$ captures expected harm among unresolved cases, including deterioration while untreated, escalation into urgent care, or delayed diagnosis after intake. The term $a(W_t)$ captures permanent loss from patients who abandon the queue before the first visit. The model summarizes these channels through a scalar backlog state because the relevant operational choice is whether the pending queue is large enough to justify the current supervisory cost of rapid expansion.

6 Robustness to limited panel heterogeneity

The benchmark assumes common panels under pooled intake. This section shows that the main staffing formulas remain approximately valid when panel heterogeneity is limited rather than absent.

Suppose that under a more general assignment rule, clinicians with signal z face severity distribution F_z rather than common distribution F. Let

$$\Delta \equiv \sup_{z \in Z} \sup_{s \in \mathcal{S}} \left| F_z(s) - F(s) \right| \tag{38}$$

measure the maximal dispersion in panel composition.

Assumption 3 (Bounded panel heterogeneity)**.** *The primitives of Assumption 1 hold, $\mathcal{S}$ is compact, and the indirect panel value is Lipschitz in the panel distribution under the sup metric. In particular, there exists $L < \infty$ such that for every clinician type a, panel size n, and panel distributions F_1, F_2,*

$$\left| m(a; n; F_1) - m(a; n; F_2) \right| \leq L \left\| F_1 - F_2 \right\|_\infty.$$

Proposition 9 (Approximate staffing formula)**.** *Under Assumptions 1, 2, and 3, there exists a constant $L < \infty$ such that*

$$\sup_{z,n} \left| M_\rho^\Delta(z;n) - M_\rho(z;n) \right| \leq L\Delta, \tag{39}$$

where M_ρ^Δ is posterior expected per-patient value under heterogeneous panels. Consequently,

$$\left| \partial_D W_\rho^{I,\Delta}(D;\omega,e) - [M_\rho(z_\rho^*(D);\kappa,e) - F(\rho,e)/\kappa] \right| \leq L\Delta. \tag{40}$$

A useful example is diagnosis-share heterogeneity. Suppose the pooled-intake system serves two diagnostic groups with pooled share $\bar{\alpha}$, but clinicians with different scores receive shares $\alpha_z \in [\bar{\alpha} - \varepsilon, \bar{\alpha} + \varepsilon]$ because some specialize modestly in trauma, substance use, youth services, bilingual care, or telehealth. If treatment value is Lipschitz in diagnosis share, then $\Delta = O(\varepsilon)$, and Proposition 9 implies that the benchmark staffing formula remains accurate up to the same order.

The same logic applies to modest heterogeneity in support sensitivity. If panels differ slightly in how supervisory support affects retention and the retention block is Lipschitz in that parameter, then the induced laws of motion for quality remain locally stable, and the signs of the staged-opening and overload-threshold comparative statics are preserved. The paper's qualitative logic is therefore robust to limited within-cell heterogeneity, but not to environments in which specialization is a first-order margin.

7 Policy implications

The policy implications are necessarily conditional on the institutional setting. In pooled-intake mental-health systems with standardized panels and genuinely scarce senior review, access expansion should be evaluated jointly with supervisory capacity rather than clinician headcount alone. Within that domain, hiring or contracting additional clinicians need not raise welfare one-for-one if rapid panel opening diverts supervisory effort away from incumbent support and lowers the quality of existing care.

Two implications are especially robust. First, policies that relax the senior-review bottleneck—for example by reducing onboarding time per panel, temporarily expanding reviewer hours, or delegating parts of the credentialing process—are most valuable when waiting harm and backlog persistence are high and incumbent quality is not excessively fragile to reduced support. Second, delayed or staged expansion need not indicate administrative inefficiency.

In pooled-intake systems, it can arise as an efficient response to the intertemporal tradeoff between expanding access and preserving the installed stock of care, especially when demand shocks may dissipate and current supervisory diversion is costly.

These implications do not extend mechanically to settings with rich clinician–patient matching, decentralized participation, or highly specialized case assignment. In those environments, matching, specialization, and queue composition may be more important than the supervisory-allocation channel emphasized here.

8 Conclusion

This paper studies access expansion in pooled-intake mental-health systems in which supervisory time is the scarce organizational input. The static benchmark provides a useful accounting object: holding fixed screening, support, and the current bottleneck, the local welfare effect of additional load equals the value of the marginal staffed panel net of its full shadow cost. The paper's main contribution, however, is dynamic. The same scarce supervisory resource must be allocated across entrant screening, panel opening, and incumbent maintenance, and the benefits of opening arrive only with a lag.

That dynamic structure generates a tradeoff between expansion speed and preservation of incumbent care. Opening more panels today raises future capacity and can improve backlog dynamics, but it also diverts supervisory effort away from incumbent support and can reduce the future quality of the installed stock of care. Within that environment, the paper derives conditional results for monotone opening, staged expansion, and overload delay. These forces are strongest when support retention, opening delays, and backlog persistence are quantitatively important.

The paper is best understood as a conceptual health-economics model of managed pooled-intake systems with centralized review, standardized panels, and scarce supervisory attention. Its central implication is organizational. Rapid access expansion may be constrained not only by the value of marginal entrants, but also by the deterioration in incumbent care that follows when supervisory attention is reallocated too quickly.

References

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, 53(5):941–973.

Atwood, A. and Lo Sasso, A. T. (2016). The effect of narrow provider networks on health care use. *Journal of Health Economics*, 50:86–98.

Brekke, K. R., Canta, C., Siciliani, L., and Straume, O. R. (2021). Hospital competition in a national health service: Evidence from a patient choice reform. *Journal of Health Economics*, 79:102509.

Brekke, K. R., Cellini, R., Siciliani, L., and Straume, O. R. (2010). Competition and quality in health care markets: A differential-game approach. *Journal of Health Economics*, 29(4):508–523.

Brekke, K. R., Siciliani, L., and Straume, O. R. (2008). Competition and waiting times in hospital markets. *Journal of Public Economics*, 92(7-8):1607–1628.

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., Kokot, J., and Wiesen, D. (2024). A new look at physicians' responses to financial incentives: Quality of care, practice characteristics, and motivations. *Journal of Health Economics*, 94:102862.

Center for Substance Abuse Treatment (2009). *Clinical Supervision and Professional Development of the Substance Abuse Counselor*. Treatment Improvement Protocol (TIP) Series 52. Substance Abuse and Mental Health Services Administration, Rockville, MD.

Chalkley, M. and Malcomson, J. M. (1998). Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics*, 17(1):1–19.

Chandra, A., Finkelstein, A., Sacarny, A., and Syverson, C. (2016). Healthcare exceptionalism? performance and allocation in the U.S. healthcare sector. *American Economic Review*, 106(8):2110–2144.

Chen, Y., Meinecke, J., and Sivey, P. (2016). A theory of waiting time reporting and quality signaling. *Health Economics*, 25(11):1355–1371.

Drake, C. (2019). What are consumers willing to pay for a broad network health plan? evidence from Covered California. *Journal of Health Economics*, 65:63–77.

Gravelle, H. and Masiero, G. (2000). Quality incentives in a regulated market with imperfect information and switching costs: Capitation in general practice. *Journal of Health Economics*, 19(6):1067–1088.

Gravelle, H. and Sivey, P. (2010). Imperfect information in a quality-competitive hospital market. *Journal of Health Economics*, 29(4):524–535.

Ho, K. (2009). Insurer-provider networks in the medical care market. *American Economic Review*, 99(1):393–430.

Isaacs, A., Bonsey, A., and Couch, D. (2023). Centralized intake models and recommendations for their use in non-acute mental health services: A scoping review. *International Journal of Environmental Research and Public Health*, 20(9):5747.

Kreider, A. R., Layton, T. J., Shepard, M., and Wallace, J. (2024). Adverse selection and network design under regulated plan prices: Evidence from Medicaid. *Journal of Health Economics*, 97:102901.

Mak, H. Y. (2018). Managing imperfect competition by pay for performance and reference pricing. *Journal of Health Economics*, 57:131–146.

McMahan, K., Martin, K. M., Greenfield, M. J., Hay, P., Bates Redwine, M., Fargason, R., and Lokken, K. (2022). Using a tele-behavioral health rapid intake model to address high demand for psychotherapy at an academic medical center during COVID-19. *Frontiers in Psychiatry*, 13:989838.

Oneschuk, H., Tuckerman, A., Archer, C., and Vasudev, K. (2026). Reducing wait times for hospital-based outpatient mental healthcare: What works? *BMJ Open Quality*, 15(1):e003031.

Rabie, S., Poudyal, A., King, A., et al. (2023). A supervision framework for task-shared mental health workers: Implications for clinical trials and beyond. *Global Health: Science and Practice*, 11(6):e2300092.

Appendix: Proofs

Proof of Proposition 1

Fix clinician quality a and panel size $n > 0$. Since $\phi(a) > 0$ and the term $-C(s)$ does not depend on the choice of $t(\cdot)$, problem (4) is equivalent to

$$\max_{t(\cdot)\geq 0} \left\{ \int_{\mathcal{S}} \Psi(s) u(t(s), s)\, dF(s) : \; n \int_{\mathcal{S}} t(s)\, dF(s) \leq 1 \right\}.$$

Because $u(\cdot, s)$ is strictly concave for every s, the objective is strictly concave in the schedule $t(\cdot)$. Hence any optimum is unique.

The Lagrangian is

$$\mathcal{L} = \int_{\mathcal{S}} \Psi(s) u(t(s), s)\, dF(s) + \lambda \left(1 - n \int_{\mathcal{S}} t(s)\, dF(s) \right).$$

The first-order condition is

$$\Psi(s) u_t(t^*(s; n), s) = \lambda n \equiv \mu^*(n) \qquad \text{for all } s \in \mathcal{S}.$$

By the Inada conditions, for each (s, μ) there is a unique strictly positive solution $t(s; \mu)$ to

$$\Psi(s) u_t(t, s) = \mu.$$

Moreover, $t(s; \mu)$ is continuous and strictly decreasing in μ. Since $u_t > 0$ everywhere, the time constraint binds at the optimum. Thus $\mu^*(n)$ is uniquely determined by

$$n \int_{\mathcal{S}} t(s; \mu^*(n))\, dF(s) = 1,$$

which establishes existence and uniqueness of $t^*(\cdot; n)$.

For the severity comparative static, totally differentiate

$$\Psi(s) u_t(t^*(s; n), s) = \mu^*(n)$$

with respect to s, holding n fixed. Since $\mu^*(n)$ does not vary with s,

$$\Psi'(s) u_t + \Psi(s) u_{tt} \frac{\partial t^*(s; n)}{\partial s} + \Psi(s) u_{ts} = 0.$$

Because $\Psi'(s) \geq 0$, $u_t > 0$, $u_{ts} > 0$, and $u_{tt} < 0$, it follows that

$$\frac{\partial t^*(s;n)}{\partial s} > 0.$$

For the caseload comparative static, an increase in n tightens the time constraint and therefore raises the shadow value $\mu^*(n)$. Since $t(s;\mu)$ is strictly decreasing in μ, this implies

$$\frac{\partial t^*(s;n)}{\partial n} < 0.$$

Finally, the optimal schedule does not depend on a because the factor $\phi(a)$ scales the objective multiplicatively and therefore does not affect the choice of $t(\cdot)$.

Proof of Proposition 2

Differentiate (8) with respect to clinician quality:

$$\frac{\partial \widetilde{m}(a,s;n)}{\partial a} = \phi'(a)\Psi(s)u(t^*(s;n),s) > 0,$$

since $\phi'(a) > 0$, $\Psi(s) > 0$, and $u > 0$.

Differentiating with respect to caseload gives

$$\frac{\partial \widetilde{m}(a,s;n)}{\partial n} = \Psi(s)\phi(a)u_t(t^*(s;n),s)\frac{\partial t^*(s;n)}{\partial n} < 0,$$

where the sign follows from Proposition 1.

Integrating both expressions over $s \in \mathcal{S}$ yields that $m(a;n)$ is strictly increasing in a and strictly decreasing in n.

Proof of Proposition 3

By Proposition 2, $m(a;n)$ is strictly increasing in latent quality a for every fixed n. Under the monotone likelihood ratio property in Assumption 2, the posterior distribution of a conditional on a higher signal z' first-order stochastically dominates the posterior distribution conditional on a lower signal $z < z'$. Since $m(a;n)$ is increasing in a, it follows that

$$M_\rho(z;n) = \mathbb{E}_\rho[m(a;n) \mid z]$$

is strictly increasing in z. Continuity follows from continuity of the signal density and dominated convergence.

Now let $\rho' \succ \rho$ in the Blackwell order. Any staffing rule feasible under the less informative experiment remains feasible under the more informative experiment, since the planner can always garble the finer signal back to the coarser one. Because the institution staffs the upper tail of posterior expected value, the more informative experiment weakly raises gross staffed value at every target load D. This proves (12). Strict inequality obtains whenever the finer experiment changes the posterior ranking on a set of positive measure.

Proof of Proposition 4

Fix (ρ, e) and define

$$H_\rho(z) \equiv \kappa \int_z^{\bar{z}} h_\rho(u)\, du.$$

Since $h_\rho > 0$ on its support, H_ρ is continuous and strictly decreasing, with

$$H_\rho(\underline{z}) = \kappa \qquad \text{and} \qquad H_\rho(\bar{z}) = 0.$$

Hence for every $D \in (0, \bar{D})$ there exists a unique cutoff $z_\rho^*(D)$ satisfying (17). Differentiating that identity yields

$$\frac{dz_\rho^*(D)}{dD} = -\frac{1}{\kappa h_\rho(z_\rho^*(D))} < 0,$$

which is (19).

Using (17), welfare can be written as

$$W_\rho^I(D; \omega, e) = \kappa \int_{z_\rho^*(D)}^{\bar{z}} M_\rho(z; \kappa, e) h_\rho(z)\, dz - \frac{F(\rho, e)}{\kappa} D.$$

Applying Leibniz's rule gives

$$\frac{\partial W_\rho^I}{\partial D} = -\kappa M_\rho(z_\rho^*(D); \kappa, e) h_\rho(z_\rho^*(D)) \frac{dz_\rho^*(D)}{dD} - \frac{F(\rho, e)}{\kappa},$$

which simplifies, using (19), to

$$\frac{\partial W_\rho^I(D; \omega, e)}{\partial D} = M_\rho(z_\rho^*(D); \kappa, e) - \frac{F(\rho, e)}{\kappa}.$$

This proves (20).

For average net value,

$$\bar{m}_\rho^I(D;\omega,e) = \frac{W_\rho^I(D;\omega,e)}{D},$$

so

$$\frac{\partial \bar{m}_\rho^I}{\partial D} = \frac{D\,W_D^I - W^I}{D^2} = \frac{M_\rho(z_\rho^*(D);\kappa,e) - F(\rho,e)/\kappa - \bar{m}_\rho^I(D;\omega,e)}{D}.$$

Because $M_\rho(z;\kappa,e)$ is strictly increasing in z, its average over the staffed upper tail strictly exceeds its value at the lower endpoint of that tail. Hence

$$\frac{\partial \bar{m}_\rho^I(D;\omega,e)}{\partial D} < 0,$$

which proves (21).

Finally, if there exists a unique $\hat{z}_{\rho,e}$ satisfying (22), then the marginal welfare effect changes sign exactly once. Therefore welfare is hump-shaped, with maximizer

$$D_{\rho,e}^I(\omega) = \kappa \int_{\hat{z}_{\rho,e}}^{\bar{z}} h_\rho(z)\,dz.$$

Proof of Proposition 5

Form the Bellman Lagrangian

$$\mathcal{L}_t = \pi(x_t, o_t, \rho_t, e_t) + \beta \mathbb{E}_t[V(x_{t+1})] + \nu_t \left[\bar{S} - [s(\rho_t) + s_0]o_t - e_t N_t\right].$$

At an interior optimum, differentiation with respect to ρ_t yields

$$-\nu_t s'(\rho_t) o_t + \beta \mathbb{E}_t \left[V_q(x_{t+1}) \frac{\partial q_{t+1}}{\partial \rho_t} \right] = 0,$$

which is (34).

Differentiating with respect to e_t gives

$$-\nu_t N_t + \beta \mathbb{E}_t \left[V_g(x_{t+1}) \frac{\partial g_{t+1}}{\partial e_t} + V_q(x_{t+1}) \frac{\partial q_{t+1}}{\partial e_t} \right] = 0.$$

Since $\partial g_{t+1}/\partial e_t = \psi$, this becomes (35).

Finally, differentiation with respect to o_t gives

$$-c - \nu_t[s(\rho_t) + s_0] + \beta \mathbb{E}_t \left[V_N(x_{t+1}) \frac{\partial N_{t+1}}{\partial o_t} + V_q(x_{t+1}) \frac{\partial q_{t+1}}{\partial o_t} + V_B(x_{t+1}) \frac{\partial B_{t+1}}{\partial o_t} \right] = 0.$$

Under the one-period opening lag,

$$\frac{\partial N_{t+1}}{\partial o_t} = 1 \qquad \text{and} \qquad \frac{\partial B_{t+1}}{\partial o_t} = 0,$$

so the condition simplifies to (36).

Proof of Proposition 6

Let

$$\Phi(x, o, \rho, e) = \pi(x, o, \rho, e) + \beta \mathbb{E}\big[V(x'(x, o, \rho, e, \theta')) \mid \theta \big]$$

denote the Bellman objective at state $x = (N, q, g, B, \theta)$. For each feasible opening level o, define the reduced objective

$$T(x, o) = \sup_{(\rho, e):\, [s(\rho)+s_0]o + eN \le \bar{S}} \Phi(x, o, \rho, e).$$

Under the maintained assumptions, higher backlog raises the marginal continuation value of opening because a larger o improves future capacity and future quality composition, and those gains matter more when future waiting losses are larger. The same logic applies when high-demand states are more persistent and when waiting harm or attrition is steeper. Since m^E has increasing differences in (ρ, o) and the continuation value is supermodular in $(N, -B, g, \theta)$, the reduced objective T inherits increasing differences in (o, B), in $(o, \text{demand persistence})$, and in $(o, \text{waiting-loss slope})$.

By Topkis's theorem, the optimal opening correspondence is therefore nondecreasing in backlog, in demand persistence, and in the slope of waiting harm and attrition.

For support fragility, let η denote a parameter that increases either $|\Xi_n|$ or δ_g. A higher η weakens the effect of opening on retained incumbent quality by making overload more damaging or support capital less durable. Since the continuation value is increasing in future quality and future support capital, the reduced objective has decreasing differences in (o, η). Topkis's theorem then implies that the optimal opening rule is weakly decreasing in support fragility.

Proof of Proposition 7

Fix a state x_t and define

$$G(o; x_t) = \sup_{(\rho,e):\, [s(\rho)+s_0]o+eN_t \leq \bar{S}} \Phi(x_t, o, \rho, e),$$

where Φ is the Bellman objective.

The derivative $G_o(o; x_t)$ reflects three forces: the continuation benefit of reducing future backlog, the entrant-quality and future-capacity gains from additional opening, and the marginal diversion cost created by the shared supervisory constraint. Under the assumptions of the proposition, the first two forces are diminishing in o: strict convexity of h and a implies diminishing marginal backlog relief, and concavity of $m^E(\rho, o)$ implies diminishing entrant-quality gains. At the same time, support fragility makes the diversion cost increasingly negative as o rises.

Hence there exists a nonempty region of the state space in which

$$G_{oo}(o; x_t) < 0.$$

To obtain an interior optimum, note that when backlog is positive and waiting losses are sufficiently important, the marginal value of opening at $o = 0$ is positive. By contrast, near the maximal feasible opening $\bar{o}$, the diversion cost dominates, so the marginal value becomes negative. Strict concavity together with the intermediate value theorem then implies a unique interior optimizer

$$0 < o_t^*(x_t) < \bar{o}.$$

The comparative statement about the size of the staged-opening region follows from the same logic. Lower shock persistence reduces the value of an immediate corner response, while a larger marginal value of incumbent support increases the curvature generated by diversion costs. Both effects enlarge the set of states in which the optimum is interior.

Proof of Proposition 8

In the two-state specialization, define the net gain from opening by $\Delta_t(o)$ in (37). For fixed (g_t, N_t, p), the marginal gain from opening is increasing in current backlog because larger B_t makes additional future capacity more valuable in the continuation problem. It is also increasing in shock persistence p, since persistent overload raises the continuation value of

installed capacity. By contrast, greater support fragility or tighter supervisory scarcity lowers the marginal gain from opening because the diversion of supervisor time becomes more costly.

Consider the marginal gain at zero opening, $\Delta_t'(0)$. Under these monotonicity properties, $\Delta_t'(0)$ is increasing in backlog and decreasing in support fragility and supervisor scarcity. Therefore the set of backlog levels for which opening is optimal is an upper interval, which yields a threshold rule

$$B_t \geq B^*(g_t, N_t, p).$$

The stated comparative statics for $B^*(g, N, p)$ follow immediately from the monotonicity of $\Delta_t'(0)$.

For the final claim, suppose shock persistence is sufficiently low and inherited support capital is sufficiently weak. Then the option value of waiting for the shock to dissipate, together with the current cost of diverting supervisory effort away from incumbents, can dominate the future benefit of immediate opening. In that case,

$$\Delta_t'(0) < 0$$

even though the static marginal value of opening is positive. Hence the planner optimally tolerates one more period of overload.

Proof of Proposition 9

By Assumption 3, for every clinician type a and panel size n,

$$|m(a; n; F_z) - m(a; n; F)| \leq L\Delta.$$

Taking posterior expectations conditional on signal z preserves the same bound, so

$$\sup_{z,n} |M_\rho^\Delta(z; n) - M_\rho(z; n)| \leq L\Delta,$$

which proves (39).

The marginal welfare formula in Proposition 4 depends on posterior expected value through the staffed upper tail and on opening cost through a linear term that is unchanged by panel heterogeneity. Therefore perturbing M_ρ by at most $L\Delta$ perturbs the derivative of welfare by at most the same order. This yields (40).